Global Journal of
Computer Science and Technology

discovering thoughts and inventing future

Online ISSN-0975-4172
Volume 9 Issue 5
Print ISSN: 0975-4350

Ver:2.0
January 2010

Global Journal of Computer Science and Technology

Global Journal of Computer Science and Technology

Volume 9 Issue 5 (Ver. 2.0)

Global Academy of Research and Development

Editorial Board Members

Contents of the Volume

From the Chief Author's Desk

The research activities among different disciplines of natural science are backbone of system. The deep and strong affords are the demands of today. Sincere afford must be exposed worldwide. Which, in turns, require international platform for rapid and proper communication among similar and interdisciplinary research groups.

The Global Journal of Computer Science and Technology is to fulfill all such demands and requirements, and functions also as an international platform. Of course, the publication of research work must be reviewed to establish its authenticity. This helps to promote research activity also. We know, great scientific research have been worked out by philosopher seeking to verify quite erroneous theories about the nature of things.

The research activities are increasing exponentially. These great increments require rapid communication, also to link up with others. The balanced communication among same and interdisciplinary research groups is major hurdle to aware with status of any research field.

The Global Journals is proving as milestone of research publication. In view of whole spectrum of Knowledge, the research work of different streams may be considered as branches of big tree. Every branch is of great importance. Thus, we look after the complete spectrum as whole. Global Journals let play all the instruments simultaneously. We hope, affords of global Journals will sincerely help to build the world in new shape.

Dr. R. K. Dixit
Chief Author
cheifauthor@globaljournals.org

θ Scheme (Orthogonal Milstein Scheme), a Better Numerical Approximation for Multi-dimensional SDEs

Klaus Schmitz Abe
Mathematical Institute, University of Oxford, England
klaus.schmitz@gmail.com
November 11, 2009

***Abstract*- Today, better numerical approximations are required for multi- dimensional SDEs to improve on the poor performance of the standard Monte Carlo integration. Usually in finance, it is the weak convergence property of numerical discretizations, which is most important, because with financial applications, one is mostly concerned with the accurate estimation of expected payoffs. However, recent studies for hedging, portfolio optimization, and the valuation of exotic options show that the strong convergence property plays a crucial role.**

When one prices an exotic option or wants to approximate a portfolio, the SDEs used are not important. What really matters is that the SDEs approximate correctly the real distribution of the process. Using this principle, this research suggests that, instead of considering a given no-commutative multi-dimensional SDE that represents our process, we consider another SDE that has the same distribution but with a different strong convergence order. Manipulating the new SDE, which has an extra process θ it becomes commutative and we avoid the simulation of the Lévy Area (extremely expensive with respect to the computational time). The new SDE obtains solutions that in a weak sense, which is in a distributional sense, coincide with those of the original SDE. If certain conditions are satisfied, θ scheme gives a first order strong convergence without the simulation of the Lévy Area. Conversely, for the original nocommutative SDE, the Milstein scheme, neglecting the Lévy Area, has 0.5 order strong convergence. If the conditions are not satisfied, this study confirms experimentally that θ scheme has a better strong approximation than using the standard Milstein scheme in the original SDEs (both schemes neglecting the simulation of the Lévy Area).

AMS subject classifications: 60G20, 65CXX, 65C20, 37H10, 41A25.

Keywords- Discrete time approximation, stochastic schemes, stochastic volatility models, Milstein Scheme, Lévy Area, θ scheme, Orthogonal Milstein Scheme, orthogonal transformation, strong convergence.

I. Introduction

Strong convergence properties of discretization of stochastic differential equations (SDEs) are very important in finance. Usually, it is the weak convergence property of numerical discretization, which is most important, because with financial applications, one is mostly concerned with the accurate estimation of expected payoffs. However, in recent studies for hedging, portfolio optimization, and the valuation of exotic options, the strong convergence property plays a crucial role. One example is the Multilevel Monte Carlo path simulation method (MSL-MC [13],[14]). Using strong convergence properties, the MSL-MC reduces substantially the computational cost for pricing exotic options using stochastic volatility models. For some exotic options, this research shows that the MSLMC is 55 times more efficient than the standard Monte Carlo method using the Euler discretization. It reduces 90% of the computation time. As the MSLMC method, other research in the literature has shown that strong convergence properties are very useful for hedging and in portfolio optimization

For time discrete approximations, the Euler-Maruyama scheme has 0.5 order strong convergence for all multi-dimensional SDEs. The next Taylor approximation, the Milstein scheme, gives first order strong convergence for all 1−Dimensional systems (using one Wiener process). However, for two or more Wiener processes, such as stochastic volatility models and correlated multidimensional SDEs, there is no exact solution for the iterated integrals of second order (Lévy Area), and the Milstein scheme, neglecting the Lévy Area, usually gives the same strong order of convergence as the Euler-Maruyama scheme. The numerical difficulty with the Milstein scheme is how to simulate efficiently the Lévy Area. It is extremely expensive with respect to the computational time.

On the other hand, in some problems, the diffusion coefficients have special properties, which allow the Milstein scheme to be simplified in a way that avoids the use of Lévy Areas. As is well known, if the SDE is commutative (44), the Lévy Areas need not be computed. Unfortunately, for many important practical financial problems (e.g. stochastic volatility models), the diffusion coefficients do not satisfy these conditions. The presented study confirms experimentally the fact that the inclusion of the Lévy Area in a strong scheme cannot be avoided if one wants to achieve one strong order. Only strong order 0.5 already achieved by the Euler scheme, results if one omits the Lévy Area terms in the Milstein scheme. In addition, the difference in the the leading error between Euler and Milstein schemes are rather small.

The purpose of the paper is to show that if certain conditions are satisfied, one can avoid the calculation of the Lévy Area and obtain first convergence order by applying an orthogonal transformation when the multi-dimensional SDE do not satisfy commutativity conditions (44). We introduce a new scheme or discrete time approximation based on an idea of Paul Malliavin where, for certain conditions, a better convergence order is obtained than the standard Milstein scheme without the expensive simulation of the Lévy Area. We demonstrate when the conditions of the 2−Dimensional problem permit this and give an exact solution for the orthogonal transformation (θ scheme).

The convergence analysis in this paper requires the SDE to satisfy global Lipschitz conditions in the drift and diffusion coefficients. This is a standard requirement for this type of analysis. However, most of the SDE models that are mentioned and used in the computational experiments do not satisfy such global Lipschitz conditions. Problems arise at the origin and/or at infinity. The results in this paper give numerical evidence that the conclusions regarding strong order remain true in circumstances where no theory currently exists. However, this is not within the scope of this research.

To simplify the understanding of the orthogonal transformation and theta scheme, the first sections consider only the 2−Dimensional case. In section 2 we give the introduction to θ scheme and how one can obtain it. In section 3, we include four examples using θ scheme applied to stochastic volatility models that are important in financial applications. In section 4, we present the definition of θ scheme for the 2−Dimensional general case. In section 5, we present the definition of θ scheme for the multi-dimensional general case. Finally, we present the conclusions, future work, references, and an appendix.

II. Orthogonal Transformation 2D

To simplify the understanding of the orthogonal transformation, we begin with the 2−Dimensional stochastic case:

$$\begin{aligned} dx &= \mu^{(x)}(x,y,t)\,dt + \sigma(x,y,t)\,d\overline{W}_{1,t}\,, \\ dy &= \mu^{(y)}(x,y,t)\,dt + \xi(x,y,t)\,d\overline{W}_{2,t}\,, \qquad E\left[d\overline{W}_{1,t}, d\overline{W}_{2,t}\right] = \rho\,dt\,. \end{aligned} \tag{1}$$

where $d\overline{W}_{i,t}$ are two Wiener processes and the coefficient functions μ, σ and ξ are assumed to satisfy the linear growth and global Lipschitz conditions ([4], pp. 548) for existence and uniqueness of a strong solution to the SDE (1). Alternatively, (1) can be represented in vector form as:

$$dZ(t) = A_0\,(t,Z)\,dt + \sum_{k=1}^{2} A_k\,(t,Z)\,d\overline{W}_{k,t}\ , \qquad Z \in \mathbb{R}^2\,.$$

This is, in fact, only a symbolic representation for the stochastic integral equation

$$Z(t) = Z(t_0) + \int_{t_0}^{t} A_0(s,Z)\,ds + \sum_{k=1}^{2}\int_{t_0}^{t} A_k(s,Z)\,d\overline{W}_{k,s}\,.$$

The first integral is a deterministic Riemann integral and the second is a stochastic integral. Using the standard definition of constant correlation, one can represent the system (1) in vector form with independent noise as:

$$d\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \mu^{(x)} \\ \mu^{(y)} \end{bmatrix} dt + \begin{bmatrix} \sigma \\ \rho\,\xi \end{bmatrix} dW_{1,t} + \begin{bmatrix} 0 \\ \widehat{\rho}\,\xi \end{bmatrix} dW_{2,t}\,, \tag{2}$$

$$\langle dW_{1,t}, dW_{2,t}\rangle = 0\,, \qquad \widehat{\rho} = \sqrt{1-\rho^2}\,.$$

The 1.0 strong order Milstein scheme for (2) with time step Δt is (Appendix (45)):

$$\begin{aligned} \begin{bmatrix} \widehat{x}_{t+\Delta t} \\ \widehat{y}_{t+\Delta t} \end{bmatrix} &= \begin{bmatrix} \widehat{x}_t \\ \widehat{y}_t \end{bmatrix} + \begin{bmatrix} \mu^{(x)} \\ \mu^{(y)} \end{bmatrix} \Delta t + \begin{bmatrix} \sigma \\ \rho\xi \end{bmatrix} \Delta W_{1,t} + \begin{bmatrix} 0 \\ \widehat{\rho}\xi \end{bmatrix} \Delta W_{2,t} \\ &+ \frac{1}{2}\begin{bmatrix} \sigma\sigma_x + \rho\xi\sigma_y \\ \rho\sigma\xi_x + \rho^2\xi\xi_y \end{bmatrix} (\Delta W_{1,t}^2 - \Delta t) + \frac{1}{2}\begin{bmatrix} 0 \\ \widehat{\rho}^2\xi\xi_y \end{bmatrix} (\Delta W_{2,t}^2 - \Delta t) \\ &+ \frac{1}{2}\begin{bmatrix} \widehat{\rho}\xi\sigma_y \\ \widehat{\rho}\sigma\xi_x + 2\rho\widehat{\rho}\xi\xi_y \end{bmatrix} (\Delta W_{1,t}\Delta W_{2,t}) + \frac{1}{2}[A_1, A_2]\left[\underline{\mathcal{I}}_{(1,2)}\right]_t^{t+\Delta t}\,, \end{aligned} \tag{3}$$

where subscript x and y denote partial derivatives, $\underline{\mathcal{I}}_{(1,2)}$ is the Lévy Area defined by:

$$\left[\underline{\mathcal{I}}_{(1,2)}\right]_t^{t+\Delta t} = \int_t^{t+\Delta t}\int_t^{S} dW_{1,U}\,dW_{2,S} - \int_t^{t+\Delta t}\int_t^{S} dW_{2,U}\,dW_{1,S}\,, \tag{4}$$

and $[A_1, A_2]$ is the Lie bracket defined by (∂_{A_i} is the Jacobian matrix of A_i):

$$[A_1, A_2] = (\partial_{A_2}A_1 - \partial_{A_1}A_2) = \begin{bmatrix} -\widehat{\rho}\xi\sigma_y \\ \widehat{\rho}\sigma\xi_x \end{bmatrix}. \tag{5}$$

As is well-known, the Lévy Areas need not be computed if the SDE is commutative (44). To have this commutativity condition, we need that (1) satisfies:

$$\frac{\partial\sigma}{\partial y} = 0 \ \text{ and } \ \frac{\partial\xi}{\partial x} = 0 \tag{6}$$

If the conditions (6) are satisfied, the coefficients of the Lévy Area (5) are equal to zero and we do not need to simulate (4). Unfortunately, only special cases satisfy the conditions. If we want to use stochastic volatility models, the SDE (1) will never be commutative.

The numerical difficulty with the Milstein scheme is how to simulate efficiently the $\underline{\mathcal{I}}_{(1,2)}$ Lévy Area computationally very expensive). The technique of Gaines and Lyons [3] can be used to sample the distribution for the Lévy Area conditional on $\Delta W_{1,t}$, $\Delta W_{2,t}$.

However there is no generalization of this to higher dimensions apart from the approximation of [16], which has a significant computational cost (hours for a good approximation or small error).

On the other hand, if one makes an orthogonal transformation of the uncorrelated process (2), one does not change the distribution (see Theorem 2 (56)) and gets:

$$\begin{aligned} d\widetilde{x} &= \mu^{(x)}(\widetilde{x},\widetilde{y},t)\,dt + \sigma(\widetilde{x},\widetilde{y},t)\,d\widetilde{W}_{1,t}\,, \\ d\widetilde{y} &= \mu^{(y)}(\widetilde{x},\widetilde{y},t)\,dt + \xi(\widetilde{x},\widetilde{y},t)\,d\widetilde{W}_{2,t}\,, \end{aligned} \tag{7}$$

where:

$$\begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \widehat{\rho} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}. \tag{8}$$

and the angle θ is a function of x and y. If one computes the coefficients of the Lévy Area (Lie bracket) for the new orthogonal process using independent Brownian paths $W_{1,t}, W_{2,t}$:

$$[A_1, A_2] = \begin{bmatrix} -\widehat{\rho}\xi\sigma_y - \sigma^2\theta_x - \rho\sigma\xi\theta_y \\ \widehat{\rho}\sigma\xi_x - \rho\sigma\xi\theta_x - \xi^2\theta_y \end{bmatrix} .$$

To avoid having to simulate the Lévy Area $\underline{I}_{(1,2)}$ to make (7) commutative, one needs the Lie brackets to be identically zero, e.g. impose the following conditions:

$$\begin{aligned} -\widehat{\rho}\xi\sigma_y - \sigma^2\theta_x - \rho\sigma\xi\theta_y &= 0 , \\ +\widehat{\rho}\sigma\xi_x - \rho\sigma\xi\theta_x - \xi^2\theta_y &= 0 . \end{aligned}$$

Simplifying one gets:

$$\begin{aligned} \Phi &\doteq \frac{\partial\theta}{\partial x} = \theta_x = \frac{-1}{\widehat{\rho}}\left(\frac{\xi\sigma_y}{\sigma^2} + \frac{\rho\xi_x}{\xi}\right) , \\ \Psi &\doteq \frac{\partial\theta}{\partial y} = \theta_y = \frac{1}{\widehat{\rho}}\left(\frac{\sigma\xi_x}{\xi^2} + \frac{\rho\sigma_y}{\sigma}\right) . \end{aligned} \tag{9}$$

If one wants to find a solution for $\theta(x, y)$, one must first determine when the system is consistent, or integrable. This requires that:

$$\frac{\partial\Phi}{\partial y} = \frac{\partial^2\theta}{\partial x\,\partial y} = \frac{\partial\Psi}{\partial x} , \tag{10}$$

and the solution for θ is:

$$\theta(x, y) = \int^{(x,y)} (\Phi\, dx + \Psi\, dy) . \tag{11}$$

However, if one applies Itô's lemma, one also obtains the following SDE for $\theta(x, y)$:

$$\begin{aligned} d\theta &= \mu^{(\theta)}dt + \sigma\Phi d\widetilde{\widetilde{W}}_{1,t} + \xi\Psi d\widetilde{\widetilde{W}}_{2,t} , \\ \mu^{(\theta)} &= \mu^{(x)}\Phi + \mu^{(y)}\Psi + \frac{1}{2}\sigma^2\frac{\partial\Phi}{\partial x} + \rho\sigma\xi\frac{\partial^2\theta}{\partial x\partial y} + \frac{1}{2}\xi^2\frac{\partial\Psi}{\partial y} . \end{aligned} \tag{12}$$

If one chooses to define θ in this way, our system becomes a 3–Dimensional Itô process with two Wiener process inputs (θ scheme):

$$\begin{bmatrix} d\widetilde{x} \\ d\widetilde{y} \\ d\theta \end{bmatrix} = \begin{bmatrix} \mu^{(x)} \\ \mu^{(y)} \\ \mu^{(\theta)} \end{bmatrix} dt + \begin{bmatrix} \sigma \\ 0 \\ \sigma\Phi \end{bmatrix} d\widetilde{\widetilde{W}}_{1,t} + \begin{bmatrix} 0 \\ \xi \\ \xi\Psi \end{bmatrix} d\widetilde{\widetilde{W}}_{2,t} . \tag{13}$$

If one computes again the Lie brackets with independent noise, one obtains (see Appendix (55)):

$$[A_1, A_2] = \begin{bmatrix} 0 \\ 0 \\ \widehat{\rho}\sigma\xi\left(\frac{\partial\,\Psi}{\partial x} - \frac{\partial\Phi}{\partial y}\right) \end{bmatrix} . \tag{14}$$

Note that when condition (10) is satisfied, this Lie bracket (14) is identically zero. Because the value of Lie brackets (14) does not depend on the drift for θ it is convenient to set it to zero:

$$\mu^{(\theta)} = 0 .$$

In the remainder of the paper, we shall investigate when particular applications satisfy condition (10), in which case one can discretise either (7) or (13) and when they do not, and in which case one can only discretise (13) or the original untransformed SDE (1). Our objective is to try to achieve higher order strong convergence without the simulation of the Lévy Areas.

When the Lie bracket is not equal to zero, the important question to be considered is how precisely does θ need to be calculated to obtain first strong order convergence in $\widetilde{x}$ and $\widetilde{y}$? For example, does neglecting the Lie bracket affect the accuracy of θ but not in $\widetilde{x}$ and $\widetilde{y}$?

One approach of θ scheme results is given by Ana-Bela Cruzeiro, Paul Malliavin and T. Thalmaier in [2]. Because dW and $d\widetilde{W}$ have the same distribution (see Theorem 2 (56)), they ignore the calculation of θ For example, the 1.0 strong order Milstein scheme for (7) with time step Δt. using (9) is (see Appendix (51)):

$$\begin{bmatrix} \widehat{\widetilde{x}}_{t+\Delta t} \\ \widehat{\widetilde{y}}_{t+\Delta t} \end{bmatrix} = \begin{bmatrix} \widehat{\widetilde{x}}_t \\ \widehat{\widetilde{y}}_t \end{bmatrix} + \begin{bmatrix} \mu^{(x)} \\ \mu^{(y)} \end{bmatrix} \Delta t + \begin{bmatrix} \sigma & 0 \\ \rho\xi & \widehat{\rho}\xi \end{bmatrix} \begin{bmatrix} \Delta\widetilde{W}_{1,t} \\ \Delta\widetilde{W}_{2,t} \end{bmatrix} + \frac{1}{2}R_M , \tag{15}$$

where R_M is equal to:

$$\begin{aligned} R_M = &\begin{bmatrix} \sigma\sigma_x + \rho\xi\sigma_y \\ \rho\sigma\xi_x + \rho^2\xi\xi_y - \frac{\widehat{\rho}^2\xi^2\sigma_y}{\sigma} \end{bmatrix} \left(\Delta\widetilde{W}_{1,t}^2 - \Delta t\right) \\ &+ \begin{bmatrix} -\rho\xi\sigma_y - \frac{\sigma^2\xi_x}{\xi} \\ \widehat{\rho}^2\xi\xi_y - \rho\sigma\xi_x - \frac{\rho^2\xi^2\sigma_y}{\sigma} \end{bmatrix} \left(\Delta\widetilde{W}_{2,t}^2 - \Delta t\right) \\ &+ \begin{bmatrix} 2\widehat{\rho}\xi\sigma_y \\ 2\widehat{\rho}\sigma\xi_x + 2\rho\widehat{\rho}\left(\xi\xi_y + \frac{\xi^2\sigma_y}{\sigma}\right) \end{bmatrix} \Delta\widetilde{W}_{1,t}\Delta\widetilde{W}_{2,t} , \end{aligned}$$

and

$$\begin{bmatrix} \Delta\widetilde{W}_{1,t} \\ \Delta\widetilde{W}_{2,t} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} .$$

Replacing $\Delta\widetilde{W}$ by ΔW in (15) one obtains the Malliavin scheme published in [2] and in book [11]. Note that the advantage of this scheme is that one does not need to simulate the Lévy Area or be concerned about the value of θ every time step. For weak solutions, the Malliavin scheme is a good approach. However, for strong solutions, it has the same or worse strong convergence constant than both the scheme that includes the simulation of θ and the Milstein scheme that does not include the orthogonal transformation (3). For illustration, see the examples in the next section with simulation plots (Figures 1 to 4).

III. Orthogonal Stochastic Volatility Models

In this section, we consider four mean reverting stochastic volatility models (SVM). The aim with a stochastic volatility model is to incorporate the empirical observation that volatility appears not to be constant and indeed varies, at least in part, randomly. The idea is to make the volatility itself a stochastic process. The candidate models have generally been motivated by intuition, convenience and a desire for tractability. In particular, the SVMs presented in this section have all appeared in the literature and have the following generic form:

$$\begin{aligned} dx &= \mu^{(x)}dt + \alpha\, x^{\gamma_1} y^{\lambda_1} d\overline{W}_{1,t}\,, \\ dy &= \mu^{(y)}dt + \beta\, x^{\gamma_2} y^{\lambda_2} d\overline{W}_{2,t}\,, \qquad E\left[d\overline{W}_{1,t}, d\overline{W}_{2,t}\right] = \rho\, dt\,. \end{aligned} \tag{16}$$

If one applies an orthogonal transformation, (16) changes to:

$$\begin{aligned} d\widetilde{x} &= \mu^{(x)}dt + \alpha\, \widetilde{x}^{\gamma_1} \widetilde{y}^{\lambda_1} d\widetilde{W}_{1,t}\,, \\ d\widetilde{y} &= \mu^{(y)}dt + \beta\, \widetilde{x}^{\gamma_2} \widetilde{y}^{\lambda_2} d\widetilde{W}_{2,t}\,, \end{aligned} \tag{17}$$

where $d\widetilde{W}_{i,t}$ are the orthogonal correlated Wiener processes defined in (8). If one would like to obtain an exact solution of θ (11), the integrability condition (10) becomes:

$$\begin{aligned} \frac{\partial \Phi}{\partial \widetilde{y}} &= \frac{\lambda_C \lambda_1 \beta\, \widetilde{y}^{\lambda_C - 1}}{-\widehat{\rho}\alpha\, \widetilde{x}^{\gamma_C+1}} = \frac{\gamma_C \gamma_2 \alpha\, \widetilde{x}^{\gamma_C - 1}}{\widehat{\rho}\beta\, \widetilde{y}^{\lambda_C+1}} = \frac{\partial \Psi}{\partial \widetilde{x}}\,, \\ \gamma_C &= \gamma_1 - \gamma_2 - 1\,, \quad \lambda_C = \lambda_2 - \lambda_1 - 1\,, \end{aligned} \tag{18}$$

so then, for $\alpha, \beta, \lambda_i, \gamma_k \neq 0$, one can conclude that θ is integrable if, and only if, $\lambda_C = \gamma_C = 0$, in which case the solution is:

$$\theta = \left(\frac{\rho\lambda_1\beta + \gamma_2\alpha}{\widehat{\rho}\beta}\right)\log\widetilde{y} - \left(\frac{\rho\gamma_2\alpha + \lambda_1\beta}{\widehat{\rho}\alpha}\right)\log\widetilde{x}\,. \tag{19}$$

Using the θ scheme (13), the 3−Dimensional Itô process for (17) is:

$$\begin{bmatrix} d\widetilde{x} \\ d\widetilde{y} \\ d\theta \end{bmatrix} = \begin{bmatrix} \mu^{(x)} \\ \mu^{(y)} \\ 0 \end{bmatrix} dt + \begin{bmatrix} \alpha\,\widetilde{x}^{\gamma_1}\widetilde{y}^{\lambda_1} \\ 0 \\ \alpha\,\widetilde{x}^{\gamma_1}\widetilde{y}^{\lambda_1}\Phi \end{bmatrix} d\widetilde{W}_{1,t} + \begin{bmatrix} 0 \\ \beta\,\widetilde{x}^{\gamma_2}\widetilde{y}^{\lambda_2} \\ \beta\,\widetilde{x}^{\gamma_2}\widetilde{y}^{\lambda_2}\Psi \end{bmatrix} d\widetilde{W}_{2,t}\,, \tag{20}$$

where:

$$\Phi = \frac{\lambda_1\beta\,\widetilde{y}^{\lambda_C} + \rho\gamma_2\,\alpha\,\widetilde{x}^{\gamma_C}}{-\widehat{\rho}\alpha\,\widetilde{x}^{\gamma_C+1}}\,, \qquad \Psi = \frac{\rho\lambda_1\,\beta\,\widetilde{y}^{\lambda_C} + \gamma_2\alpha\,\widetilde{x}^{\gamma_C}}{\widehat{\rho}\beta\,\widetilde{y}^{\lambda_C+1}}\,.$$

If one computes the Lie brackets:

$$[A_1, A_2] = \begin{bmatrix} 0 \\ 0 \\ \widetilde{x}^{2\gamma_2}\widetilde{y}^{2\lambda_1}\left(\gamma_C\gamma_2\alpha^2\widetilde{x}^{2\gamma_C} + \lambda_C\lambda_1\beta^2\widetilde{y}^{2\lambda_C}\right) \end{bmatrix}\,. \tag{21}$$

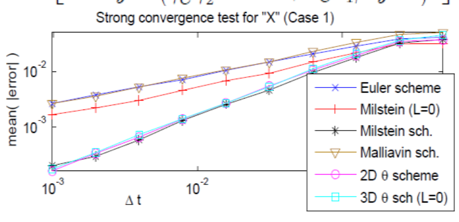


Figure 1: Strong convergence test for x (Case 1).

Even without the condition (18) being satisfied, one can perhaps improve the convergence using the θ scheme of without the simulation of the Lévy Areas. However, this depends on the parameters of our system. In other words, the accuracy is dependent on the value of the Lie bracket of the scheme (21). It give us the bias in the calculation of the value of θ and hence in x and y. Note that when condition (18) is satisfied, this Lie bracket (21) is identically zero.

A. The Quadratic Volatility Model

The first case we consider is the Quadratic Volatility Model:

$$\begin{aligned} dx &= x\,\overline{\mu}\,dt + x\,y\,d\overline{W}_{1,t}\,, \\ dy &= k\,(\varpi_2 - y)\,dt + \beta_2\,y^2 d\overline{W}_{2,t}\,. \end{aligned} \tag{22}$$

The Quadratic Volatility Model is a typical explosive model in financial applications where high or extreme volatility shocks persist through time. In this model, the volatility process itself is an OU process with a mean reversion level ϖ_2. The disadvantages of this model is that the volatility could easily become negative and no closed form solution is available for option pricing.

Because $\lambda_C = 0$, one can use either equation (17) together with (19), or the 3−Dimensional θ scheme (20).

Because of the orthogonal transformation, neither requires the calculation of the Lévy Area. Figure 1 and Table 1 show that, as expected, the Euler scheme and the Milstein scheme with zero Lévy Areas (setting $\underline{L}_{(1,2)} = 0$ in (3)) give 0.5 strong convergence order. On the other hand, the Milstein scheme (3) with a proper value for the distribution of the Lévy Area (by simulating the Lévy Area using N subintervals within each time step) gives 1.0 order strong convergence, as do the two orthogonal θ schemes. We have used the following parameters: $t_o = 0$; $T = 1$; $\rho = -0.50$; $\overline{\mu} = 0.1$; $k = 1.4$; $\varpi_2 = 0.32$; $\beta = 1.22$ and initial conditions $x(t_o) = 1$; $y(t_o) = \varpi_2$.

B. The 3/2 Model (Case 2)

The second case we consider is the following stochastic variance model, usually called the 3/2 Model [10]:

$$\begin{aligned} dx &= x\,\overline{\mu}\,dt + x\sqrt{y}\,d\overline{W}_{1,t}\,, \\ dy &= k\,y(\varpi_{3/2} - y)\,dt + \beta_{3/2}\,y^{3/2} d\overline{W}_{2,t}\,. \end{aligned} \tag{23}$$

The 3/2 Model is an important model in finance, not only because it has a closed form solution for option pricing as simple as the square root model (25), but it also displays a feature of many stochastic volatility models that one does not see in the square root model. That is, even after a change of measure to the riskadjusted process, option prices (relative to the bond price) under the 3/2 model are sometimes not martingales but merely local martingales [10]. When option prices are not martingales, this means that they are not given by the standard expected value formula (e.g. $e^{-rT}E\left[\max\left(S_T - K\right)\right]$ for a call option). So the 3/2 Model, with its closed form solution for European and Digital options, is one of the simplest illustrations of this important phenomenon for financial theory. It was first used by Cox, In ersoll, and Ross ([1], 1985) and further investigated by Heston ([7], 1997) and Lewis ([10], 2000).

Because $\lambda_C = 0$, we obtain almost the same results as Case 1 (Figure 2 and Table 1). The parameters and initial conditions are the same as in Case 1 except for $\varpi_{3/2} = \varpi_2^2$;

$\beta = 2.44$; $y(t_o) = \varpi_2^2$; which are chosen so that x and y will have approximately the same relative volatility.

C. The Garch Diffusion Model (Case 3)

The third case we consider is the following stochastic variance model, usually called GARCH Diffusion Model:

$$\begin{aligned} dx &= x\,\overline{\mu}\,dt + x\sqrt{y}d\overline{W}_{1,t}\ , \\ dy &= k\,(\varpi_1 - y)\,dt + \beta_1\,y d\overline{W}_{2,t}\ . \end{aligned} \tag{24}$$

(24) is described as the diffusion limit of a GARCH-type process. The failure of the usual martingale pricing relation can also occur in this SVM and was first shown by Sin [15] in 1998. These failures are specific examples of the notion that the absence of arbitrage implies that financial claim prices are, in general, only strictly local martingales, not martingales [10]. From a practical point of view, the advantage of this model is that you can estimate its parameters using well-known algorithms that are available as computer software, although no closed form solution is available for option pricing.

In this case, where $\lambda_C = 0.5$, it is not possible to use the scheme $2D-\theta$ since the integrability condition is not satisfied. Figure 3 and Table 1 show that the only schemes that achieve first order convergence are the Milstein and θ ; schemes, which simulate the Lévy Area. However, the simulation results also show that there is a remarkable difference between the original and the orthogonal scheme without the simulation of the Lévy Area, not the improved order of convergence achieved in the first case but a much improved constant of proportionality. The parameters and initial conditions are the same as in Case 2 except for This is chosen $\varpi_1 = \varpi_2^2$; $\beta = 0.78$.

to ensure that x and y will have approximately the same relative volatility as in the first two cases.

D. The Square Root Model (Case 4)

The last case we consider for stochastic variance models is the Heston's Square Root Model:

$$\begin{aligned} dx &= x\,\mu\,dt + x\sqrt{y}d\overline{W}_{1,t}\ , \\ dy &= k\,(\varpi_{1/2} - y)\,dt + \beta_{1/2}\sqrt{y}d\overline{W}_{2,t}\ . \end{aligned} \tag{25}$$

This model was proposed by Heston in 1993 [6]. The volatility is related to a square root process and can be interpreted as the radial distance from the origin of a multi-dimensional OU process. For small dt, this model keeps the volatility positive and is the most popular among all SVM because of its two main features: it has a semi-analytical pricing formula for European and Digital options which is easy to implement, and the solution is typical (it displays the same qualitative properties that one generally expects in time homogenous cases). Furthermore, it can be used to understand how volatility models that do not have analytical solutions behave in many respects.

In this case, $\lambda_c = 1$. Figure 4 and Table 1 show that neither of the Milstein schemes in which the Lévy Areas are set to zero performs very well. Both have order 0.5 strong convergence, and the constant of proportionality is not much better than for the Euler scheme. When the Lévy Areas are simulated correctly, the Milstein and schemes do exhibit the expected first order strong convergence. This demonstrates the importance of the Lévy Areas in this case.

The parameters and initial conditions are the same as in Case 2 except for $\varpi_{1/2} = \varpi_2^2$; $\beta = 0.25$.

Scheme	Description	C-1	C-2	C-3	C-4
Euler scheme	set $\triangle t = dt$, $\Delta W_i = dW_i$ in (2)	0.49	0.50	0.51	0.50
Milstein (L=0)	Milstein (3), set $\underline{L}_{(1,2)} = 0$	0.52	0.54	0.53	0.53
Milstein sch.	Milstein (3), simulate $\underline{L}_{(1,2)}$	**0.94**	**0.95**	**0.96**	**0.96**
Malliavin sch.	Milstein (17), set $\Delta\widetilde{W}_i = dW_i$	0.50	0.52	0.50	0.49
$2D$-θ scheme	Milstein (17) with (19)	**0.96**	**0.95**	n/a[1]	n/a
$3D$-θ sch (L=0)	Milstein (20), set $\underline{L}_{(1,2)} = 0$	**0.96**	**0.95**	0.78	0.63
$3D$-θ scheme	Milstein (20), simulate $\underline{L}_{(1,2)}$	**0.96**	**0.95**	**0.95**	**0.94**

Table 1: Convergence orders γ for SVMs (all cases (22-25))[2].

[1] n/a = not applicable.

[2] Note that the constant proportion factors in the Figures 1 through 4 depend completely on the chosen parameter values in the examples and can be very different for another choice. To calculate the strong order of convergence we have used the theorems in [12] or [13].

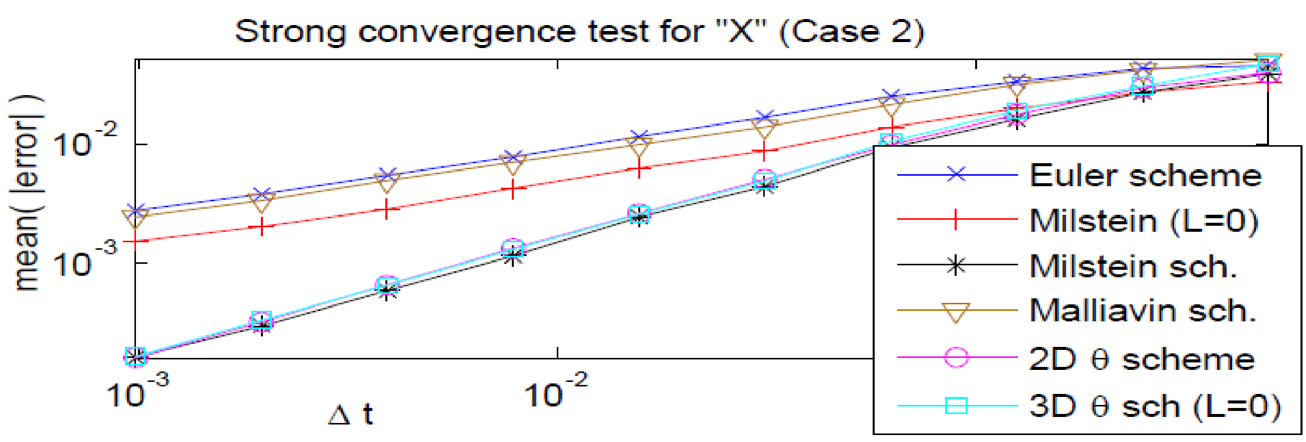


Figure 2: Strong convergence test for x (Case 2).

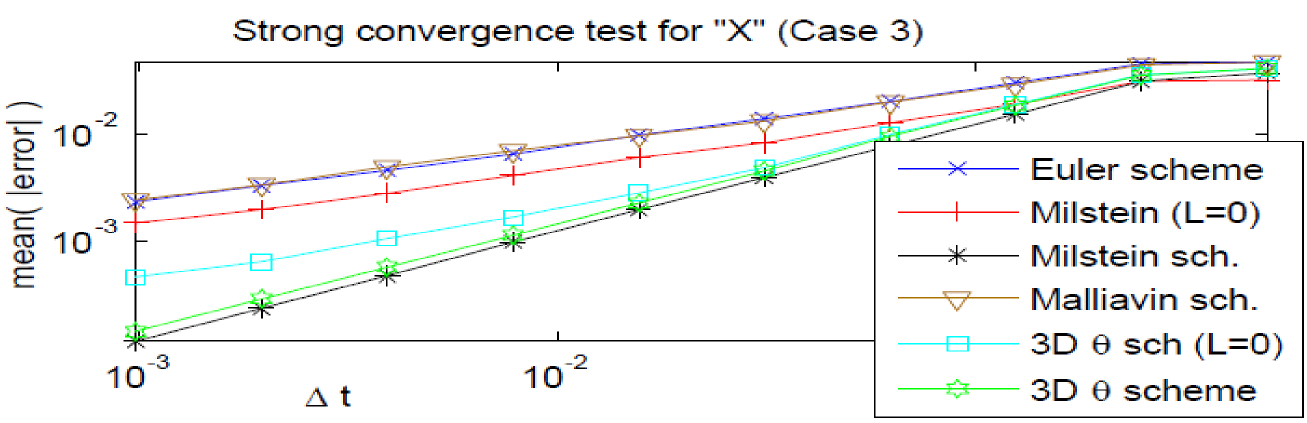


Figure 3: Strong convergence test for x (Case 3).

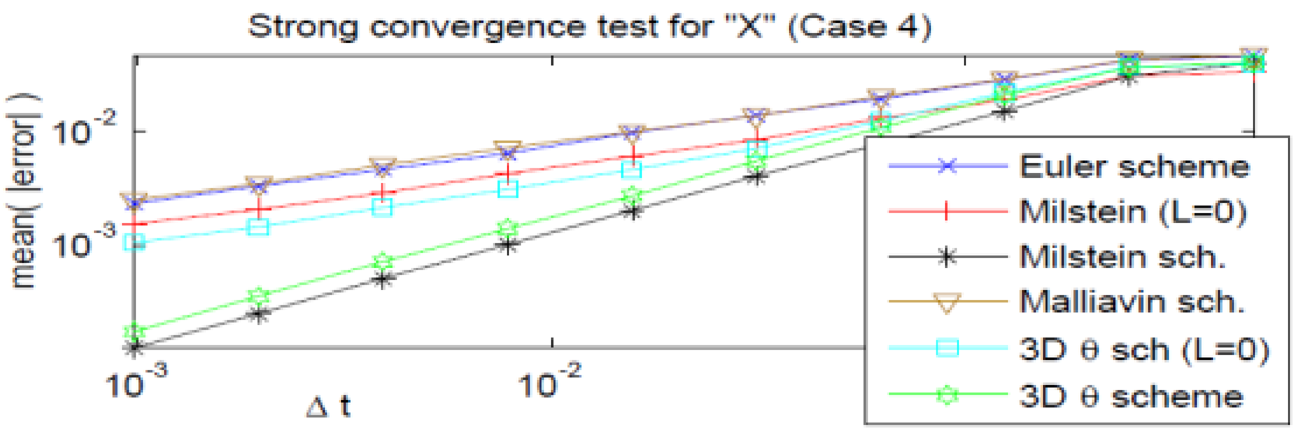


Figure 4: Strong convergence test for x (Case 4).

IV. 2D Orthogonal Milstein Scheme (θ Scheme)

This section presents the definition of the θ scheme for the 2–Dimensional SDE case. At the end of the section, we present two more examples that confirm a better order of convergence than the standard Milstein scheme without the simulation of the Lévy Area when one uses an orthogonal transformation.

A. 2D-θ Scheme

Theorem 1: 2D $-\theta$ Scheme (Exact solution)

If one has a 2–Dimensional Itô stochastic differential equation with two independent Wiener processes:

$$d\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} dt + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}, \tag{26}$$

where a_i, $b_{i,k}$ are smooth functions of t and $X_{i,t}$, satisfying the linear growth and global Lipschitz conditions ([4], pp. 548). If one applies an orthogonal transformation to (26) described by:

$$\begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}, \tag{27}$$

where:

$$\theta_t(X_1, X_2) = \int^{X_1} \Phi dX_1 + \int^{X_2} \Psi dX_2 , \tag{28}$$

and satisfying:

$$\frac{\partial \Psi}{\partial X_1} = \frac{\partial \Phi}{\partial X_2} , \tag{29}$$

then the new orthogonal process has 1.0 strong order convergence using the Milstein scheme neglecting the simulation of the Lévy Area. Conversely, for $H_i^- \neq 0$ (the commutativity condition (44) is not satisfied), the Milstein scheme of (26) with zero Lévy Area has 0.5 strong order convergence. The functions Φ and Ψ are equal to:

$$\Phi \doteq \frac{H_1^-\left(b_{2,1}^2 + b_{2,2}^2\right) - H_2^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} ,$$

$$\Psi \doteq \frac{H_2^-\left(b_{1,1}^2 + b_{1,2}^2\right) - H_1^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} ,$$

where H_i^- are the coefficients of the Lévy Area (Lie bracket) of (26) and are defined by:

$$H_j^- = L_1 b_{j,2} - L_2 b_{j,1} , \qquad L_j := \sum_{k=1}^{d} b_{k,j} \frac{\partial}{\partial X_k} . \tag{30}$$

Proof: The 1.0 strong order Milstein scheme for (26) with time step Δt is (Appendix 42):

$$\begin{bmatrix} \widehat{X}_{1,t+\Delta t} \\ \widehat{X}_{2,t+\Delta t} \end{bmatrix} = \begin{bmatrix} \widehat{X}_{1,t} \\ \widehat{X}_{2,t} \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \Delta t + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} + \frac{1}{2} R_M ,$$

$$\begin{aligned} R_M &= \sum_{j=1}^{2} \begin{bmatrix} L_j b_{1,j} \\ L_j b_{2,j} \end{bmatrix} \left(\Delta W_{j,t}^2 - \Delta t\right) + \begin{bmatrix} H_1^+ \\ H_1^+ \end{bmatrix} \Delta W_{1,t} \Delta W_{2,t} \\ &+ \begin{bmatrix} H_1^- \\ H_2^- \end{bmatrix} \left[\underline{\overline{L}}_{(1,2)}\right]_t^{t+\Delta t} , \qquad H_j^{\pm} = L_1 b_{j,2} \pm L_2 b_{j,1} . \end{aligned}$$

For $H_j^{\pm} \neq 0$, the Milstein scheme is 1.0 strong order convergence when one includes all terms in the equation (see Theorem 10.3.5, page 350 from [9]); otherwise it becomes 0.5 strong order convergence. In general, if X_T is the solution of the SDE (26) and $\widehat{X}_T$ is the numerical approximation using the Milstein scheme, for $H_j^- \neq 0$ and neglecting the simulation of the Lévy Area, one can say:

$$E\left[X_T - \widehat{X}_T\right] \leq \widehat{C}_1 (\Delta t)^{0.5} .$$

On the other hand, if one makes an orthogonal transformation (27) to (26), one obtains:

$$d\begin{bmatrix} \widetilde{X}_{1,t} \\ \widetilde{X}_{2,t} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} dt + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} . \tag{31}$$

The 1.0 strong order Milstein scheme for (31) with time step Δt is (Appendix (48)):

$$\begin{bmatrix} \widehat{\widetilde{X}}_{1,t+\Delta t} \\ \widehat{\widetilde{X}}_{2,t+\Delta t} \end{bmatrix} = \begin{bmatrix} \widehat{\widetilde{X}}_{1,t} \\ \widehat{\widetilde{X}}_{2,t} \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \Delta t + \begin{bmatrix} \widetilde{b}_{1,1} & \widetilde{b}_{1,2} \\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2} \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} + \frac{1}{2} R_M ,$$

$$\begin{aligned} R_M &= \sum_{j=1}^{2} \begin{bmatrix} \widetilde{L}_j \widetilde{b}_{1,j} \\ \widetilde{L}_j \widetilde{b}_{2,j} \end{bmatrix} \left(\Delta W_{j,t}^2 - \Delta t\right) + \begin{bmatrix} \widetilde{H}_1^+ \\ \widetilde{H}_2^+ \end{bmatrix} \Delta W_{1,t} \Delta W_{2,t} \\ &+ \begin{bmatrix} \widetilde{H}_1^- \\ \widetilde{H}_2^- \end{bmatrix} \left[\underline{\overline{L}}_{(1,2)}\right]_t^{t+\Delta t} , \qquad \widetilde{H}_j^{\pm} = \widetilde{L}_1 \widetilde{b}_{j,2} \pm \widetilde{L}_2 \widetilde{b}_{j,1} , \end{aligned}$$

where:

$$\begin{bmatrix} \widetilde{b}_{1,1} & \widetilde{b}_{1,2} \\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} .$$

If one computes the coefficients of the Lévy Area using independent Wiener processes, one gets:

$$\begin{bmatrix} \widetilde{H}_1^- \\ \widetilde{H}_2^- \end{bmatrix} = \begin{bmatrix} H_1^- - \frac{\partial\theta}{\partial X_1}\left(b_{1,1}^2 + b_{1,2}^2\right) - \frac{\partial\theta}{\partial X_2}\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right) \\ H_2^- - \frac{\partial\theta}{\partial X_2}\left(b_{2,1}^2 + b_{2,2}^2\right) - \frac{\partial\theta}{\partial X_1}\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right) \end{bmatrix} . \tag{32}$$

To avoid having to simulate the Lévy Area, one needs (32) to be identically zero, e.g. impose the following conditions:

$$\begin{bmatrix} \widetilde{H}_1^- & \widetilde{H}_2^- \end{bmatrix} = 0 .$$

Simplifying one gets:

$$\Phi \doteq \frac{\partial\theta}{\partial X_1} = \frac{H_1^-\left(b_{2,1}^2 + b_{2,2}^2\right) - H_2^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} ,$$

$$\Psi \doteq \frac{\partial\theta}{\partial X_2} = \frac{H_2^-\left(b_{1,1}^2 + b_{1,2}^2\right) - H_1^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} .$$

To find a solution for θ one must first determine when the system is consistent (integrable); this requires condition (29) and the solution for θ is (28).

B. 3D- θ Scheme

If one has a 2–Dimensional Itô process (26) and applies an orthogonal transformation (27) to it, where the rotation angle θ_t is described using a third SDE:

$$d\begin{bmatrix}\widetilde{X}_{1,t}\\ \widetilde{X}_{2,t}\\ \theta_t\end{bmatrix}=\begin{bmatrix}a_1\\ a_2\\ 0\end{bmatrix}dt+\begin{bmatrix}b_{1,1} & b_{1,2}\\ b_{2,1} & b_{2,2}\\ (\Phi b_{1,1}+\Psi b_{2,1}) & (\Phi b_{1,2}+\Psi b_{2,2})\end{bmatrix}\begin{bmatrix}d\widetilde{W}_{1,t}\\ d\widetilde{W}_{2,t}\end{bmatrix}, \tag{33}$$

then, for sufficiently smooth function $a_i,\ b_{i,k},$ the Milstein scheme for the 3–Dimensional SDE (33) can have better strong convergence than (26) using the Milstein scheme neglecting the simulation of the Lévy Area. The accuracy of θ_t and hence in $X_{i,t}$ depends on the value of the Lie bracket (34) of the process (33):

$$R_{\underline{L}}=\begin{bmatrix}0\\ 0\\ (b_{1,1}b_{2,2}-b_{1,2}b_{2,1})\left(\frac{\partial\Psi}{\partial X_1}-\frac{\partial\Phi}{\partial X_2}\right)\end{bmatrix}. \tag{34}$$

The 1.0 strong order Milstein scheme for (33) with time step Δt is (Appendix (53)):

$$\begin{bmatrix}\widehat{\widetilde{X}}_{1,t+\Delta t}\\ \widehat{\widetilde{X}}_{2,t+\Delta t}\\ \widehat{\theta}_{t+\Delta t}\end{bmatrix}=\begin{bmatrix}\widehat{\widetilde{X}}_{1,t}\\ \widehat{\widetilde{X}}_{2,t}\\ \widehat{\theta}_t\end{bmatrix}+\begin{bmatrix}a_1\\ a_2\\ 0\end{bmatrix}\Delta t+\begin{bmatrix}\widetilde{b}_{1,1} & \widetilde{b}_{1,2}\\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2}\\ b_{\theta,1} & b_{\theta,2}\end{bmatrix}\begin{bmatrix}\Delta W_{1,t}\\ \Delta W_{2,t}\end{bmatrix}+\frac{1}{2}R_M,$$

$$R_M = \sum_{j=1}^{2}\begin{bmatrix}\widetilde{L}_j\widetilde{b}_{1,j}\\ \widetilde{L}_j\widetilde{b}_{2,j}\\ \widetilde{L}_j\widetilde{b}_{3,j}\end{bmatrix}(\Delta W_{j,t}^2-\Delta t)+\begin{bmatrix}\widetilde{H}_1^+\\ \widetilde{H}_2^+\\ \widetilde{H}_3^+\end{bmatrix}\Delta W_{1,t}\Delta W_{2,t}$$
$$+\begin{bmatrix}\widetilde{H}_1^-\\ \widetilde{H}_2^-\\ \widetilde{H}_3^-\end{bmatrix}\left[\underline{L}_{(1,2)}\right]_t^{t+\Delta t}, \qquad \widetilde{H}_j^{\pm}=\widetilde{L}_1\widetilde{b}_{j,2}\pm\widetilde{L}_2\widetilde{b}_{j,1}$$

If one computes the coefficients of the Lévy Area of the last equation (Appendix (55)), one obtains:

$$R_{\underline{L}}=\begin{bmatrix}0 & 0 & \widetilde{H}_3^-\end{bmatrix}^T. \tag{35}$$

If the value of $\widetilde{H}_3^-$ in the Lie bracket $R_{\underline{L}}$ is small enough the accuracy of θ_t is not affected by neglecting this term in the equation and, hence, the 3D Itô process (33) will have better strong convergence than (26) using Milstein scheme neglecting the simulation of the Lévy Area. Note that when condition (29) is satisfied, the Lie bracket (34 or 35) is identically zero $(\widetilde{H}_3^-=0)$.

C. Example Of θ Scheme

Consider the following 2D SDEs:

$$\begin{aligned} dx &= x\,\mu_x\,dt+0.5x^{\gamma}\sqrt{y}d\overline{W}_{1,t}\,, \\ dy &= y\,\mu_y\,dt+0.5\sqrt{x}y^{\lambda}d\overline{W}_{2,t}\,, \qquad E\left[d\overline{W}_{1,t},d\overline{W}_{2,t}\right]=\rho\,dt\,, \end{aligned} \tag{36}$$

where:

$$\mu_x=\mu_y=0.05,\ \ \rho=-0.2,\ \ x(t_o)=1,\ \ y(t_o)=0.3^2\,.$$

If $\gamma=\lambda=1.5,$ then we have the integrability condition 29) or (18) and either Theorem 1 $(2D-\theta \text{ scheme})$ or $3D-\theta$ scheme can be applied. Figures 5 and 6 show that the new orthogonal process of (36) has 1.0 strong order convergence in x and y using the Milstein scheme neglecting the simulation of the Lévy Area. Conversely, Euler, Malliavin and the Milstein schemes with zero Lévy Area have 0.5 strong order convergence in x and y.

If $\gamma=\lambda=1,$ then the integrability condition (29) or (18) is not satisfied and only the 3D– θ scheme can be applied. Figure 7 shows that the only schemes that achieve first order convergence are the Milstein and θ schemes which simulate the Lévy Area. However, Figure 7 shows that there is a remarkable difference between the original and the orthogonal scheme without the simulation of the Lévy Area. The improved order of convergence is not achieved as in the case of $\gamma=\lambda=1.5,$ but there is a much improved constant of proportionality.Note that the constant proportion factors in the Figures 5 through 7 depend completely on the chosen parameter values in the examples and can be very different for another choice.

The numerical results do not simply confirm an outcome that has been rigorously derived in Theorem 1 under global Lipschitz conditions, but are indicative

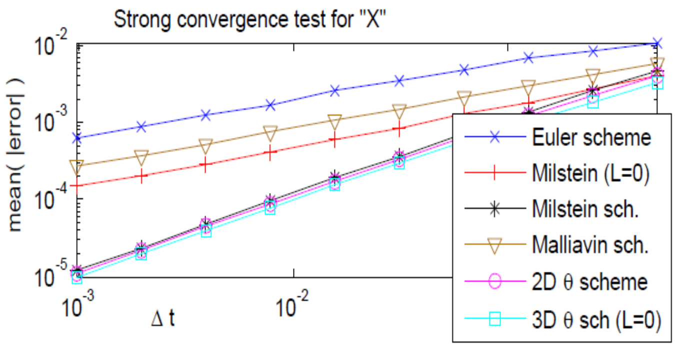


Figure 5: Strong convergence test for

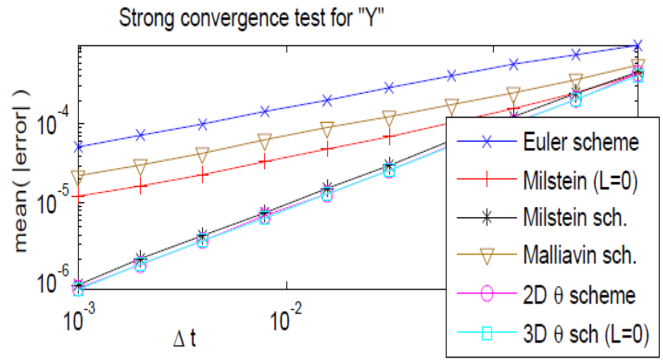


Figure 6: Strong convergence test for

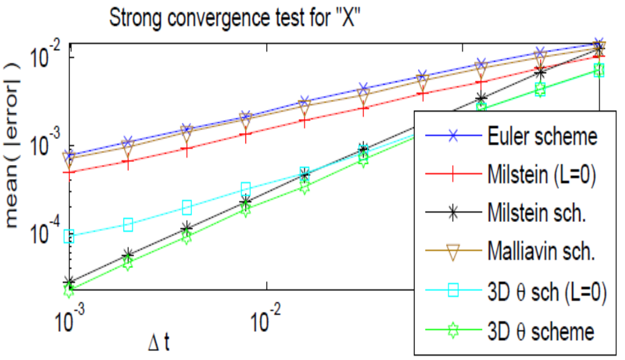


Figure 7: Strong convergence test for

that, under less restrictive assumptions, such results are possible. Though there has been some work completed on convergence analysis under non-global Lipschitz conditions ([17] and [8]), this topic is not covered in this research.

V. θ Scheme (N-Dimensional)

In this section we shall present a summary when one deals with an N-Dimensional SDE and would like to apply an orthogonal transformation to avoid the calculation of the Lévy Area. All models can be described through a SDE of the form:

$$dX_t = \mu\left(X_t, t\right) dt + \sigma\left(X_t, t\right) dW_t \ , \qquad X(t_0) = X_0 \ , \tag{37}$$

where:

$$\begin{aligned} X_t &= X(t) \in \mathbb{R}^d \ , \ W_t \in \mathbb{R}^M \ , \ t \in [t_0, \ldots T] \in \mathbb{R} \ , \\ \sigma\left(X_t, t\right) &= \sigma\left(b_{i,k}\left(X_t, t\right)\right) \in \mathbb{R}^{d \times M} \ , \quad \mu\left(X_t, t\right) = \mu\left(a_i\left(X_t, t\right)\right) \in \mathbb{R}^d \ , \\ & E\left[dW_{j,t} dW_{k,t}\right] = 0 \ , \qquad for \ i \neq k \ , \end{aligned}$$

If one replaces the Wiener process W_t with an orthogonal transform $\widetilde{W}_t$, the probability distribution does not change, and we obtain the set of all orthogonal transformations of our system (37):

$$d\widetilde{X}_t = \mu\left(\widetilde{X}_t, t\right) dt + \sigma\left(\widetilde{X}_t, t\right) d\widetilde{W}_t \ , \tag{38}$$

where:

$$d\widetilde{W}_t = \Gamma\left(\theta_t\right) dW_t \quad \text{and} \quad \Gamma\left(\theta_t\right) = \Gamma\left(\Theta_{i,k}\left(\widetilde{X}_t, t\right)\right) \in \mathbb{R}^{M \times M} \ .$$

Using non-orthogonal Wiener processes, (38) can be represented by:

$$d\widetilde{X}_t = \mu\left(\widetilde{X}_t, t\right) dt + \widetilde{\sigma}\left(\widetilde{X}_t, t, \theta_t\right) dW_t \ , \tag{39}$$

where:

$$\sigma\left(\widetilde{X}_t, t, \theta_t\right) = \sigma\left(\widetilde{b}_{i,k}\left(\widetilde{X}_t, t, \theta_t\right)\right) \in \mathbb{R}^{d \times M} \ , \quad \widetilde{b}_{i,j}\left(\widetilde{X}_t, t, \theta_t\right) = \sum_{s}^{M} b_{i,s} \Theta_{s,j}$$

The 1.0 strong order Milstein scheme for (39) with time step Δt using Itô operators is [9]:

$$Z_{i,t+\Delta_t} = Z_{i,t} + \mu_i \Delta t + \sum_{j=1}^{M} \widetilde{b}_{i,j} \Delta W_{j,t} + \frac{1}{2} R_M \ .$$

Note that the coefficient functions a_i, $b_{i,k}$ are assumed to satisfy the linear growth and global Lipschitz conditions for existence and uniqueness of a strong solution to the SDE (37). If one uses the Lévy Areas, R_M is equal to:

$$R_M = \sum_{j_1, j_2 = 1}^{M} \widetilde{L}_{j_1} \widetilde{b}_{i,j_2} \left(\Delta W_{j_1,t} \Delta W_{j_2,t} - \widetilde{\delta}_{j_1,j_2} \Delta t\right) + \sum_{j_1 < j_2 = 1}^{M} \left(R^{\underline{\overline{L}}}_{(j_1,j_2)}\right)_i \left[\underline{\overline{L}}_{(j_1,j_2)}\right]_t^{t+\Delta t}$$

$$\left(R^{\underline{\overline{L}}}_{(j_1,j_2)}\right)_i = \left(\widetilde{L}_{j_1} \widetilde{b}_{i,j_2} - \widetilde{L}_{j_2} \widetilde{b}_{i,j_1}\right) \ . \tag{40}$$

$\widetilde{\delta}_{j_1,j_2}$ is the Kronecker symbol ($\widetilde{\delta}_{j_1,j_2} = 1$ if $j_1 = j_2$ and zero otherwise) and the Itô operators are defined in (30). Using the definition of the variables, the orthogonal properties and considering the vector fields to be independent of time, the coefficients for the Lévy Area (40) are equal to:

$$\left(R^{\underline{\overline{L}}}_{(j_1,j_2)}\right)_i = \sum_{k=1}^{d} \left((-1)^{k+1} \sum_{s=1}^{M} b_{i,s} b_{k,s} + \theta_k \sum_{s_1 < s_2 = 1}^{M} \left(b_{k,s_1} \frac{\partial b_{i,s_2}}{\partial Z_k} - b_{k,s_2} \frac{\partial b_{i,s_1}}{\partial Z_k} \right) \right)$$

where are the orthogonal functions defined by:

$$\theta_k = (-1)^{k+1} \left(\Theta_{k,k+1_{\widetilde{X}_k}} \Theta_{k,k} - \Theta_{k,k_{\widetilde{X}_k}} \Theta_{k,k+1} \right) \ .$$

To avoid having to simulate the Lévy Areas $\underline{\overline{L}}_{(j_1,j_2)}$, one needs to impose the following conditions:

$$\left(R^{\underline{\overline{L}}}_{(j_1,j_2)}\right)_i = 0 \ .$$

VI. Conclusions

Strong convergence properties of discretizations of stochastic differential equations (SDEs) are very important in financial applications. Numerical examples in the paper demonstrate, as expected, a 0.5 and 1.0 strong order of convergence for Euler and Milstein schemes respectively. To obtain a 1.0 strong order of convergence with the Milstein scheme, one has to apply the scheme to the vector form of the SDE, use independent Wiener processes and compute correctly the double integral or Lévy Area.

We have shown that, under certain conditions, the use of the orthogonal θ scheme in multi-dimensional SDEs can achieve the first order strong convergence properties of the Milstein numerical discretization without the expensive simulation of Lévy Areas and when the commutativity condition is not satisfied. Conversely, the Milstein scheme with zero Lévy Area has a 0.5 strong order convergence.

The bias or error in the computation of the rotation angle θ that makesthe Lie bracket equal to zero in the orthogonal scheme (θ scheme) is crucial to obtain a better convergence order. When the conditions for integrability are satisfied, one can use the formula for θ to obtain the value of the rotation angle and obtain first order strong convergence. Otherwise, one has to use the 3−Dimensional transformation and check the magnitude of the Lie brackets to decide if it is likely to give computational savings in the solution of our system.The numerical results in this research show a better strong order of convergence than the standard Milstein scheme (without the simulation of the Lévy Area) when an orthogonal transformation is applied.

Standard convergence theory for numerical methods for SDEs (e.g. as in Kloeden and Platen [9]) makes a global Lipschitz assumption on the coefficients. However, most of the SDE models that are mentioned and used in the computational experiments, do not satisfy such global Lipschitz conditions (e.g.example (36)). The numerical results are not simply confirming a theory that has been proved; they are giving numerical evidence that the conclusions about strong order remain true in circumstances where no theory currently exists.

When one prices an exotic option or wants to approximate a portfolio, the SDEs used is not important. What really matters is that the SDEs approximate correctly the real distribution of the process. Because of this θ scheme helps to obtain a better strong order of convergence and can be applied for hedging, portfolio optimization and pricing exotic options. In [13] and [14], we have demonstrated that the use of strong convergence of θ scheme reduces substantially the computation cost for pricing exotic options (90% of the computation time).

In summary, this paper proposes a better numerical approximation for multidimensional SDE's. We introduce a new scheme or discrete time approximation where a better convergence order is obtained than that of using the standard Milstein scheme without the simulation of the expensive Lévy Area. We demonstrate when the conditions of the 2−Dimensional problem permit this and give an exact solution for the orthogonal transformation. Our applications are focused on continuous time diffusion models for the volatility and variance with their discrete time approximations (ARV).

For future work, we think that, for multi-dimensional SDE's $(d \geq 3)$, the investigation and test of the multi-dimensional ☐ scheme will be very interesting. For some parameters, it will be obvious that the new orthogonal scheme will provide considerable computational time savings when calculating the strong and weak solutions.

VII. Acknowledgements

We are very grateful to Prof. T. J. Lyons and Prof. M. Giles for several helpful discussions about this research. Their comments were key in obtaining θ scheme (orthogonal Milstein scheme).

VIII. References

1. Cox, J., Ingersoll, J., and Ross, S. (1985): "An Intertemporal General Equilibrium Model of Asset Prices". Econometrica, 53, 363 − 384.
2. Cruzeiro, A. B., Malliavin, P., and Thalmaier, A. (2004): "Geometrization of Monte-Carlo numerical analysis of an elliptic operator: strong approximation". C. R. Acad. Sci. Paris, Ser. I, 338, 481 − 486.
3. Gaines, J. G., and Lyons, T. J. (1994): "Random Generation of Stochastic Area Integrals". SIAM Journal on Applied Mathematics, 54, No. 4, 1132 − 1146.
4. Glasserman, P. (2004): "Monte Carlo Methods in Financial Engineering". Springer.
5. Grimmett, G., and Stirzaker, D. (2004): "Probability and Random Processes". Oxford University Press, Third Edition.
6. Heston, S. L. (1993): "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options". The Review of Financial Studies, Volume 6, Issue 2, 327-343.
7. Heston, S. L. (1997): "A Simple New Formula for Options With Stochastic Volatility". Course notes of Washington University in St. Louis, Missouri.
8. Higham, D. (2004): "An Introduction to Financial Option Valuation". Mathematics, Stochastic and Computation,Cambridge University Press, Cambridge
9. Kloeden, P. E., and Platen, E. (1999): "Numerical Solution of Stochastic Differential Equations". Springer.
10. Lewis, A. L. (2000): "Option Valuation under Stochastic Volatility: with Mathematica Code". Finance Press.
11. Malliavin, P., and Thalmaier, A. (2005): "Stochastic Calculus of Variations in Mathematical Finance". Springer.
12. Schmitz-Abe, K., and Shaw, W. (2005): "Measure Order of Convergence without an Exact Solution, Euler vs Milstein Scheme". International Journal of Pure and Applied Mathematics, Vol. 24, No 3, 365 − 381.
13. Schmitz-Abe, K., and Giles, M. (2006): "Pricing Exotic Options using Strong Convergence Properties". Springer, Progress in Industrial Mathematics at CMI 2006.
14. Schmitz-Abe, K. (2008): "Pricing Exotic Options using MSL-MC". Quantitative Finance, Cambridge, UK, Manuscript RQUF-2008-0196.
15. Sin, C. (1998): "Complications with Stochastic Volatility Models", Advances in Applied Probability, 30, 256 − 268.
16. Wiktorsson, M. (2001): "Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions". The Annals of Applied Probability, Vol. 11, No 2, 470 − 487.
17. Xuerong M., Chenggui Y. and Yin G. (2007): "Approximations of Euler-Maruyama type for

stochastic differential equations with Markovian switching, under non-Lipschitz conditions", Journal of Computational and Applied Mathematics, Volume 205, Issue 2, 15, pp. 936 − 948.

IX. APPENDIX

The Appendix outlines a theorem and mathematical operations required to understand both the Milstein and θ schemes.

A. *Milstein Scheme (Itô Operators)*

We start with the 2D Itô SDE case with a 2D independent Wiener process:

$$d\begin{bmatrix} X_{1,t} \\ X_{2,t} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} dt + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}, \qquad (41)$$

where a_i, $b_{i,k}$ are assumed to satisfy the linear growth and global Lipschitz conditions ([4], pp. 548) for existence and uniqueness of a strong solution to the SDE (41). The 1.0 strong order Milstein scheme for (41) with time step Δt using Itô operators is [9]:

$$\begin{bmatrix} \widehat{X}_{1,t+\Delta t} \\ \widehat{X}_{2,t+\Delta t} \end{bmatrix} = \begin{bmatrix} \widehat{X}_{1,t} \\ \widehat{X}_{2,t} \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \Delta t + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} + \frac{1}{2} R_M, \qquad (42)$$

where using Lévy Areas, R_M is equal to:

$$R_M = \sum_{j=1}^{2} \begin{bmatrix} L_j b_{1,j} \\ L_j b_{2,j} \end{bmatrix} (\Delta W_{j,t}^2 - \Delta t) + \begin{bmatrix} H_1^+ \\ H_2^+ \end{bmatrix} \Delta W_{1,t} \Delta W_{2,t} \qquad (43)$$
$$+ \begin{bmatrix} H_1^- \\ H_2^- \end{bmatrix} \left[\underline{L}_{(1,2)} \right]_t^{t+\Delta t}, \qquad H_j^{\pm} = L_1 b_{j,2} \pm L_2 b_{j,1}.$$

The Itô operators are defined by:

$$L_j := \sum_{k=1}^{d} b_{k,j} \frac{\partial}{\partial X_k}.$$

A more general, but important special case is that of commutative noise in which the diffusion matrix of (41) satisfies the commutativity condition (pp. 348, [9]):

$$L_{j_1} b_{k,j_2} = L_{j_2} b_{k,j_1} \qquad (44)$$

If conditions (44) are satisfied, $H_j^- = 0$ in (43) and we do not need to simulate the Lévy Areas. Doing some computations, (43) is equal to:

$$R_M = \begin{bmatrix} C_{1,X_1} \\ C_{1,X_2} \end{bmatrix} (\Delta W_{1,t}^2 - \Delta t) + \begin{bmatrix} C_{2,X_1} \\ C_{2,X_2} \end{bmatrix} (\Delta W_{2,t}^2 - \Delta t)$$
$$+ \begin{bmatrix} C_{3,X_1} + C_{4,X_1} \\ C_{3,X_2} + C_{4,X_2} \end{bmatrix} \Delta W_{1,t} \Delta W_{2,t} + \begin{bmatrix} C_{3,X_1} - C_{4,X_1} \\ C_{3,X_2} - C_{4,X_2} \end{bmatrix} \left[\underline{L}_{(1,2)} \right]_t^{t+\Delta t}$$

where:

$C_{1,X_1} = b_{1,1} b_{1,1_{X_1}} + b_{2,1} b_{1,1_{X_2}}$	$C_{1,X_2} = b_{1,1} b_{2,1_{X_1}} + b_{2,1} b_{2,1_{X_2}}$
$C_{2,X_1} = b_{1,2} b_{1,2_{X_1}} + b_{2,2} b_{1,2_{X_2}}$	$C_{2,X_2} = b_{1,2} b_{2,2_{X_1}} + b_{2,2} b_{2,2_{X_2}}$
$C_{3,X_1} = b_{1,1} b_{1,2_{X_1}} + b_{2,1} b_{1,2_{X_2}}$	$C_{3,X_2} = b_{1,1} b_{2,2_{X_1}} + b_{2,1} b_{2,2_{X_2}}$
$C_{4,X_1} = b_{1,2} b_{1,1_{X_1}} + b_{2,2} b_{1,1_{X_2}}$	$C_{4,X_2} = b_{1,2} b_{2,1_{X_1}} + b_{2,2} b_{2,1_{X_2}}$

,

Having example (2), we get:

$C_{1,X_1} = \sigma\sigma_x + \rho\xi\sigma_y$	$C_{1,X_2} = \rho\sigma\xi_x + \rho^2\xi\xi_y$
$C_{2,X_1} = 0$	$C_{2,X_2} = \widehat{\rho}^2\xi\xi_y$
$C_{3,X_1} = 0$	$C_{3,X_2} = \widehat{\rho}\sigma\xi_x + \rho\widehat{\rho}\xi\xi_y$
$C_{4,X_1} = \widehat{\rho}\xi\sigma_y$	$C_{4,X_2} = \rho\widehat{\rho}\xi\xi_y$

, (45)

Checking the commutativity conditions (44) in example (2), we have:

$$\widehat{\rho}\,\xi \frac{\partial \sigma}{\partial X_2} = 0, \qquad \widehat{\rho}\,\sigma \frac{\partial \xi}{\partial X_1} = 0.$$

For $\sigma \neq 0$ and $\xi \neq 0$, we need:

$$\frac{\partial \xi}{\partial X_1} = 0 \text{ and } \frac{\partial \sigma}{\partial X_2} = 0$$

B. *2d − θ Scheme (Orthogonal Milstein)*

If one makes an orthogonal transformation to (41), one gets:

$$d\begin{bmatrix} \widetilde{X}_{1,t} \\ \widetilde{X}_{2,t} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} dt + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix}, \qquad (46)$$

where:

$$\begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} = \begin{bmatrix} \cos\theta & \mp\sin\theta \\ \sin\theta & \pm\cos\theta \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix}. \qquad (47)$$

The 1.0 strong order Milstein scheme for (46) with time step Δt using Itô operators is [9]:

$$\begin{bmatrix} \widehat{\widetilde{X}}_{1,t+\Delta t} \\ \widehat{\widetilde{X}}_{2,t+\Delta t} \end{bmatrix} = \begin{bmatrix} \widehat{\widetilde{X}}_{1,t} \\ \widehat{\widetilde{X}}_{2,t} \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \Delta t + \begin{bmatrix} \widetilde{b}_{1,1} & \widetilde{b}_{1,2} \\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2} \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} + \frac{1}{2} R_M, \qquad (48)$$

$$\begin{bmatrix} \widetilde{b}_{1,1} & \widetilde{b}_{1,2} \\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2} \end{bmatrix} = \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} \cos\theta & \mp\sin\theta \\ \sin\theta & \pm\cos\theta \end{bmatrix},$$

where using Lévy Areas, R_M is equal to:

$$R_M = \sum_{j=1}^{2} \begin{bmatrix} \widetilde{L}_j \widetilde{b}_{1,j} \\ \widetilde{L}_j \widetilde{b}_{2,j} \end{bmatrix} (\Delta W_{j,t}^2 - \Delta t) + \begin{bmatrix} \widetilde{H}_1^+ \\ \widetilde{H}_2^+ \end{bmatrix} \Delta W_{1,t} \Delta W_{2,t} \qquad (49)$$
$$+ \begin{bmatrix} \widetilde{H}_1^- \\ \widetilde{H}_2^- \end{bmatrix} \left[\underline{L}_{(1,2)} \right]_t^{t+\Delta t}, \qquad \widetilde{H}_j^{\pm} = \widetilde{L}_1 \widetilde{b}_{j,2} \pm \widetilde{L}_2 \widetilde{b}_{j,1}.$$

Doing some computation, (49) is equal to:

$$R_M = \begin{bmatrix} \widetilde{C}_{1,X_1} \\ \widetilde{C}_{1,X_2} \end{bmatrix} \left(\Delta\widetilde{W}_{1,t}^2 - \Delta t\right) + \begin{bmatrix} \widetilde{C}_{2,X_1} \\ \widetilde{C}_{2,X_2} \end{bmatrix} \left(\Delta\widetilde{W}_{2,t}^2 - \Delta t\right) \qquad (50)$$

$$+ \begin{bmatrix} \widetilde{C}_{3,X_1} + \widetilde{C}_{4,X_1} \\ \widetilde{C}_{3,X_2} + \widetilde{C}_{4,X_2} \end{bmatrix} \Delta\widetilde{W}_{1,t}\Delta\widetilde{W}_{2,t} + \begin{bmatrix} \pm\left(\widetilde{C}_{3,X_1} - \widetilde{C}_{4,X_1}\right) \\ \pm\left(\widetilde{C}_{3,X_2} - \widetilde{C}_{4,X_2}\right) \end{bmatrix} \left[\underline{\widetilde{L}}_{(1,2)}\right]_t^{t+\Delta t} ,$$

where:
To make zero the coefficient of the Lévy Area in (50), one needs:

$$\Phi \doteq \frac{\partial\theta}{\partial X_1} = \theta_{X_1} = \frac{H_1^-\left(b_{2,1}^2 + b_{2,2}^2\right) - H_2^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} ,$$

$$\Psi \doteq \frac{\partial\theta}{\partial X_2} = \theta_{X_2} = \frac{H_2^-\left(b_{1,1}^2 + b_{1,2}^2\right) - H_1^-\left(b_{1,1}b_{2,1} + b_{1,2}b_{2,2}\right)}{\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)^2} .$$

Having example (7), we get:

$\widetilde{C}_{1,X_1} = \sigma\sigma_x + \rho\xi\sigma_y$	$\widetilde{C}_{1,X_2} = \rho\sigma\xi_x + \rho^2\xi\xi_y - \frac{\widehat{\rho}^2\xi^2\sigma_y}{\sigma}$
$\widetilde{C}_{2,X_1} = -\rho\xi\sigma_y - \frac{\sigma^2\xi_x}{\xi}$	$\widetilde{C}_{2,X_2} = -\rho\sigma\xi_x + \widehat{\rho}^2\xi\xi_y - \frac{\rho^2\xi^2\sigma_y}{\sigma}$
$\widetilde{C}_{3,X_1} = \widehat{\rho}\xi\sigma_y$	$\widetilde{C}_{3,X_2} = \widehat{\rho}\sigma\xi_x + \rho\widehat{\rho}\left(\xi\xi_y + \frac{\xi^2\sigma_y}{\sigma}\right)$
$\widetilde{C}_{4,X_1} = \widetilde{C}_{3,X_1}$	$\widetilde{C}_{4,X_2} = \widetilde{C}_{3,X_2}$

. (51)

C. 3D $-$ θ Scheme

We start with the following 3−Dimensional Itô SDE with a 2−Dimensional Wiener process:

$$d\begin{bmatrix} \widetilde{X}_{1,t} \\ \widetilde{X}_{2,t} \\ \theta_t \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_\theta \end{bmatrix} dt + \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{\theta,1} & b_{\theta,2} \end{bmatrix} \begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} , \qquad (52)$$

where:

$$a_\theta = \Phi a_1 + \Psi a_2 ,$$

$$\begin{bmatrix} b_{\theta,1} & b_{\theta,2} \end{bmatrix} = \begin{bmatrix} \Phi b_{1,1} + \Psi b_{2,1} & \Phi b_{1,2} + \Psi b_{2,2} \end{bmatrix}$$

and a_i, $b_{i,k}$ are assumed to satisfy the linear growth and global Lipschitz conditions. The 1.0 strong order Milstein scheme for (52) with time step Δ_t is [9]:

$$\begin{bmatrix} \tilde{\tilde{X}}_{1,t+\Delta t} \\ \tilde{\tilde{X}}_{2,t+\Delta t} \\ \theta_{t+\Delta t} \end{bmatrix} = \begin{bmatrix} \tilde{\tilde{X}}_{1,t} \\ \tilde{\tilde{X}}_{2,t} \\ \theta_t \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ a_\theta \end{bmatrix} \Delta t + \begin{bmatrix} \widetilde{b}_{1,1} & \widetilde{b}_{1,2} \\ \widetilde{b}_{2,1} & \widetilde{b}_{2,2} \\ b_{\theta,1} & b_{\theta,2} \end{bmatrix} \begin{bmatrix} \Delta W_{1,t} \\ \Delta W_{2,t} \end{bmatrix} + \frac{1}{2} R_M , \qquad (53)$$

where using Lévy Areas, R_M is equal to:

$$R_M = \sum_{j=1}^{2} \begin{bmatrix} \widetilde{L}_j\widetilde{b}_{1,j} \\ \widetilde{L}_j\widetilde{b}_{2,j} \\ \widetilde{L}_j b_{3,j} \end{bmatrix} \left(\Delta W_{j,t}^2 - \Delta t\right) + \begin{bmatrix} \widetilde{H}_1^+ \\ \widetilde{H}_2^+ \\ \widetilde{H}_3^+ \end{bmatrix} \Delta W_{1,t}\Delta W_{2,t}$$

$$+ \begin{bmatrix} \widetilde{H}_1^- \\ \widetilde{H}_2^- \\ \widetilde{H}_3^- \end{bmatrix} \left[\underline{\widetilde{L}}_{(1,2)}\right]_t^{t+\Delta t} , \qquad \widetilde{H}_j^{\pm} = \widetilde{L}_1\widetilde{b}_{j,2} \pm \widetilde{L}_2\widetilde{b}_{j,1} .$$

Doing some computation, (53) is equal to:

$$R_M = \begin{bmatrix} \widetilde{C}_{1,X_1} \\ \widetilde{C}_{1,X_2} \\ \widetilde{C}_{1,\theta} \end{bmatrix} \left(\Delta\widetilde{W}_{1,t}^2 - \Delta t\right) + \begin{bmatrix} \widetilde{C}_{2,X_1} \\ \widetilde{C}_{2,X_2} \\ \widetilde{C}_{2,\theta} \end{bmatrix} \left(\Delta\widetilde{W}_{2,t}^2 - \Delta t\right) \qquad (54)$$

$$+ \begin{bmatrix} \widetilde{C}_{3,X_1} + \widetilde{C}_{4,X_1} \\ \widetilde{C}_{3,X_2} + \widetilde{C}_{4,X_2} \\ \widetilde{C}_{3,\theta} + \widetilde{C}_{4,\theta} \end{bmatrix} \Delta\widetilde{W}_{1,t}\Delta\widetilde{W}_{2,t} + \begin{bmatrix} \pm\left(\widetilde{C}_{3,X_1} - \widetilde{C}_{4,X_1}\right) \\ \pm\left(\widetilde{C}_{3,X_2} - \widetilde{C}_{4,X_2}\right) \\ \pm\left(\widetilde{C}_{3,\theta} - \widetilde{C}_{4,\theta}\right) \end{bmatrix} \left[\underline{\widetilde{L}}_{(1,2)}\right]_t^{t+\Delta t} ,$$

where:

$\widetilde{C}_{1,\theta} = \Phi\left(C_{1,X_1} + \Phi b_{1,1}b_{1,2}\right) + \Psi\left(C_{1,X_2} + \Psi b_{2,1}b_{2,2}\right) + \Phi_{X_1}b_{1,1}^2 + \Psi_{X_2}b_{2,1}^2$
$+\Phi\Psi\left(b_{1,1}b_{2,2} + b_{1,2}b_{2,1}\right) + \left(\Phi_{X_2} + \Psi_{X_1}\right)b_{1,1}b_{2,1}$
$\widetilde{C}_{2,\theta} = \Phi\left(C_{2,X_1} - \Phi b_{1,1}b_{1,2}\right) + \Psi\left(C_{2,X_2} - \Psi b_{2,1}b_{2,2}\right) + \Phi_{X_1}b_{1,2}^2 + \Psi_{X_2}b_{2,2}^2$
$-\Phi\Psi\left(b_{1,1}b_{2,2} + b_{1,2}b_{2,1}\right) + \left(\Phi_{X_2} + \Psi_{X_1}\right)b_{1,2}b_{2,2}$
$\widetilde{C}_{3,\theta} = \Phi\left(C_{3,X_1} - \Phi b_{1,1}^2\right) + \Psi\left(C_{3,X_2} - \Psi b_{2,1}^2\right) - 2\Phi\Psi b_{1,1}b_{2,1}$
$+\Phi_{X_1}b_{1,1}b_{1,2} + \Phi_{X_2}b_{1,2}b_{2,1} + \Psi_{X_1}b_{1,1}b_{2,2} + \Psi_{X_2}b_{2,1}b_{2,2}$
$\widetilde{C}_{4,\theta} = \Phi\left(C_{4,X_1} + \Phi b_{1,2}^2\right) + \Psi\left(C_{4,X_2} + \Psi b_{2,2}^2\right) + 2\Phi\Psi b_{1,2}b_{2,2}$
$+\Phi_{X_1}b_{1,1}b_{1,2} + \Psi_{X_1}b_{1,2}b_{2,1} + \Phi_{X_2}b_{1,1}b_{2,2} + \Psi_{X_2}b_{2,1}b_{2,2}$

.

Doing some operations, the Lie bracket is equal to:

$$\begin{bmatrix} \widetilde{H}_1^- \\ \widetilde{H}_2^- \\ \widetilde{H}_3^- \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \pm\left(b_{1,1}b_{2,2} - b_{1,2}b_{2,1}\right)\left(\Psi_{X_1} - \Phi_{X_2}\right) \end{bmatrix} . \qquad (55)$$

Having example (13), we get:

$\widetilde{C}_{1,\theta} = -\widehat{\rho}\left(2\sigma_y\xi_x + \xi\sigma_{xy} + \frac{\xi}{\sigma}\left(-\sigma_x\sigma_y + \rho\left(\sigma_y\xi_y + \xi\sigma_{yy}\right)\right)\right)$
$\widetilde{C}_{2,\theta} = \widehat{\rho}\left(2\sigma_y\xi_x + \sigma\xi_{xy} + \frac{\xi}{\sigma}\rho\left(\sigma_y\xi_y + \xi\sigma_{yy}\right) - \frac{\sigma}{\xi}\xi_x\xi_y\right)$
$\widetilde{C}_{3,\theta} = \rho\left(2\sigma_y\xi_x + \sigma\xi_{xy} + \xi\sigma_{xy} + \frac{\xi}{\sigma}\left(-\sigma_x\sigma_y + \rho\left(\sigma_y\xi_y + \xi\sigma_{yy}\right)\right)\right)$
$+\frac{\sigma}{\xi}\left(\sigma\xi_{xx} + \xi_x\left(\sigma_x - \rho\xi_y\right)\right) + \frac{-\sigma^2\xi_x^2}{\xi^2} + \frac{-\xi^2\sigma_y^2}{\sigma^2}$
$\widetilde{C}_{4,\theta} = \rho\left(2\sigma_y\xi_x + \frac{\xi}{\sigma}\rho\left(\sigma_y\xi_y + \xi\sigma_{yy}\right)\right) + \frac{\xi^2\sigma_y^2}{\sigma^2} + \frac{\sigma^2\xi_x^2}{\xi^2} - \frac{\xi}{\sigma}\left(\sigma_y\xi_y + \xi\sigma_{yy}\right)$

,

and $\widetilde{H}_3^-$ from the Lie bracket (55) is equal to:

$$\widetilde{H}_3^- = \rho\left(\sigma\xi_{xy} + \xi\sigma_{xy} - \frac{\xi}{\sigma}\sigma_x\sigma_y\right) + \frac{\sigma}{\xi}\left(\sigma\xi_{xx} + \xi_x\left(\sigma_x - \rho\xi_y\right)\right)$$

$$+ \frac{-2\xi^2\sigma_y^2}{\sigma^2} + \frac{-2\sigma^2\xi_x^2}{\xi^2} + \frac{\xi}{\sigma}\left(\sigma_y\xi_y + \xi\sigma_{yy}\right) .$$

D. Orthogonal Transformation Theorem

In this section, we shall present a theorem where we prove that if an orthogonal transformation is applied to a standard normal distributed process dW, then the new orthogonal process $d\widetilde{W}$ is independent and has the same distribution as the original process dW.

Theorem 2: Distribution of an Orthogonal Standard Normal Distributed Process

If $dW_{1,t}$, $dW_{2,t}$ are two independent and identically standard normal distributed processes with expectation 0 and variance σ^2, and we apply an orthogonal transformation

$$\begin{bmatrix} d\widetilde{W}_{1,t} \\ d\widetilde{W}_{2,t} \end{bmatrix} = \begin{bmatrix} \cos\theta_t & \sin\theta_t \\ -\sin\theta_t & \cos\theta_t \end{bmatrix} \begin{bmatrix} dW_{1,t} \\ dW_{2,t} \end{bmatrix} , \qquad (56)$$

then:

A. The new orthogonal random process, $d\widetilde{W}_{1,t}$ and $d\widetilde{W}_{2,t}$, are independent.

B. $dW_{i,t}$ and $d\widetilde{W}_{i,t}$ have the same distribution.

Proof:

A. If $dW_{1,t}$ and $dW_{2,t}$ are independent then;

$$E\left[dW_{1,t}dW_{2,t}\right] = 0 \ .$$

Doing the same for the orthogonal Wiener process:

$$E\left[d\widetilde{W}_{1,t}d\widetilde{W}_{2,t}\right] = E\left[\begin{array}{c} dW_{1,t}dW_{2,t}\left(\cos^2\theta - \sin^2\theta\right) \\ -\sin\theta\cos\theta\left(dW_{1,t}^2 - dW_{2,t}^2\right)\end{array}\right] = 0 \ .$$

B. The probability density function (PDF) of an N–Dimensional multivariate normal is [5]:

$$f(x) = f(x_1, x_2, ...x_N) = \frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{\det\left(\sum\right)}} \exp\left(-\frac{1}{2}(x-\overline{\mu})^T \sum^{-1}(x-\overline{\mu})\right) ,$$

where $\overline{\mu} = \left[\mu_1, \mu_2, ..., \mu_N\right]^T$ is the mean and $\sum$ is the covariance matrix (positivedefinite real $N \ x \ N$ matrix)

For dW_t we have:

$$\overline{\mu}_{dW_t} = \left[E\left[dW_{1,t}\right] \ , \ E\left[dW_{2,t}\right]\right] = \left[0 \ , \ 0\right] \ ,$$

$$\begin{aligned}\sum\nolimits_{dW_t} &= \left[\begin{array}{cc} V\left[dW_{1,t}\right] & \rho\sqrt{V\left[dW_{1,t}\right]}\sqrt{V\left[dW_{2,t}\right]} \\ \rho\sqrt{V\left[dW_{1,t}\right]}\sqrt{V\left[dW_{2,t}\right]} & V\left[dW_{2,t}\right]\end{array}\right] \\ &= \left[\begin{array}{cc} \sigma^2 & 0 \\ 0 & \sigma^2 \end{array}\right] .\end{aligned}$$

If dW_t and $d\widetilde{W}_t$ have the same distribution then they have the same mean and covariance matrix:

$$\overline{\mu}_{d\widetilde{W}_t} = \left[E\left[d\widetilde{W}_{1,t}\right] \ , \ E\left[d\widetilde{W}_{2,t}\right]\right] = \overline{\mu}_{dW_t} \ ,$$

$$\begin{aligned}\sum\nolimits_{d\widetilde{W}_t} &= \left[\begin{array}{cc} V\left[d\widetilde{W}_{1,t}\right] & \rho\sqrt{V\left[d\widetilde{W}_{1,t}\right]}\sqrt{V\left[d\widetilde{W}_{2,t}\right]} \\ \rho\sqrt{V\left[d\widetilde{W}_{1,t}\right]}\sqrt{V\left[d\widetilde{W}_{2,t}\right]} & V\left[d\widetilde{W}_{2,t}\right]\end{array}\right] \\ &= \sum\nolimits_{dW_t} \qquad \square\end{aligned}$$

Input Data Processing Techniques in Intrusion Detection Systems – Short Review

GJCST Classifications:
E.m, C.2.0, D.4.6

Suhair H. Amer, and John A. Hamilton, Jr.

***Abstract*—In this paper intrusion detection systems (IDSs) are classified according to the techniques applied to processing input data. This process is complex because IDSs are highly coupled in actual implemented systems. Eleven input data processing techniques associated with intrusion detection systems are identified. They are then grouped into more abstract categories. Some approaches are artificially intelligent such as neural networks, expert systems, and agents. Others are computationally based such as Bayesian networks, and fuzzy logic. Finally, some are based on biological concepts such as immune systems and genetics. Characteristics of and systems employing each technique are also mentioned.**

I. Introduction

When traditionally classifying intrusion detection systems (IDSs) as misuse, anomaly or hybrid, the systems are grouped according to the technique they utilize to detect intrusions. For example, misuse-based IDSs match already stored attack signatures against the audit data gathered while the monitored system is or was running. In anomaly based IDSs, detection utilize models of normal behavior where any deviation from such behavior is identified as an intrusion. Another type of traditional classification is categorizing an IDS according to its setup as network-based, host-based or hybrid. Network based systems monitor network activities whereas a host based system monitor the activities of a single system for intrusion traces [1]. In general, IDSs may apply many techniques to detect intrusions and improve detection such as neural networks, expert systems, agents, Bayesian networks, fuzzy logic, immune systems and genetics. Little attention has been given to classifying the processing techniques applied on the input data provided to the IDS. In this paper we classify input data processing techniques utilized with IDSs that may use and may not use the same processing technique to detect intrusions. In section 2, abstract classification of the different input data processing techniques utilized with IDSs will be presented.
Eleven input data processing techniques associated with IDSs are identified. Then they are grouped into more abstract categories. In section 3, a general description as well as some advantages and disadvantages of each technique and examples of system employing these techniques will be presented.

II. Classification of Input Data Processing Techniques in IDSs

In this paper, we are concerned with the techniques used to process input data that is considered when designing and implementing IDSs. Classifying such techniques are not easy because in the actual implemented system, combination of techniques may be used. However, identifying them individually helps better understand the merits and limitations of each, and how to improve a techniques performance by using another. Eleven techniques are identified [shown at the lower level of diagram 1] that are widely and currently used for processing input data of IDSs. They are then grouped into more abstract categories that are identified at the upper levels of diagram 1. This is important because the characteristics of each technique are highly affected by the category(ies) that it belongs to. In the lower level of Fig. 1, techniques such as Agents and Data Mining belong to the Intelligent Data Analysis category. This is indicated by the dotted relation between Data Analysis and AI categories. The techniques: Expert systems and Fuzzy logic are intelligent model-based-rule-based systems shown by the dotted relation between Rule based and AI categories in Fig. 1. Next is an explanation of each item in Fig. 1, along with some identified characteristics.

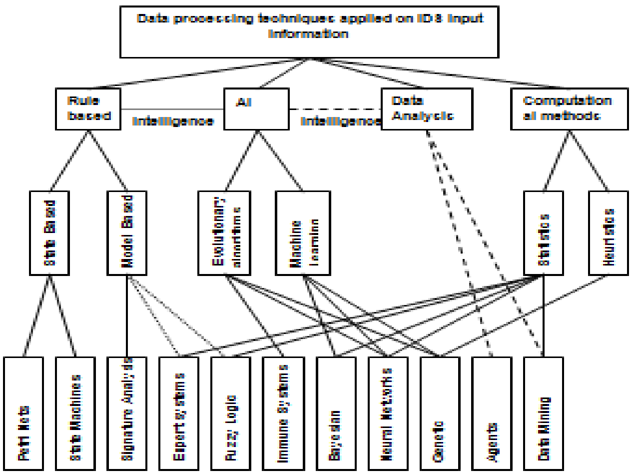


Fig.1. Data processing techniques applied on input data processed by Intrusion Detection Systems

Manuscript received July 31, 2009.
S. H. Amer was with Auburn University, Auburn, AL 36849 USA. She is now with the Department of Computer Science, Southeast Missouri State University, Cape Girardeau, MO 63701 USA (telephone: 573-651-2525, e-mail: samer@semo.edu).
J. A. Hamilton, Jr., is with the Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849 USA (telephone: 334-844-6360, e-mail: hamilton@auburn.edu).

A. Rule Based

If a rule-based IDS is to use input data or audit data, such information will be in a codified rules format of known intrusions. The input data will represent identified intrusive behavior and categorizing intrusion attempts by sequences of user activities that lead to compromised system states. The IDS will take as input the predefined rules as well as the current audit data and check if a rule is fired. In general, using rule bases are affected by system hardware or software changes and require updates by system experts as the system is enhanced or maintained. Such input data technique is very useful in an environment where physical protection of the computer system is not always possible (e.g., a battlefield situation) but require strong protection [http://www.sei.cmu.edu/str/descriptions/rbid.html].

In general, rule based systems can be:

I) State-based: in the audit trails, intrusion attempts are defined as sequences of system states leading from an initial state to a final compromised state represented in a state transition diagram. The two inputs to the IDS will include the audit trail and the state transition diagrams of known penetrations that will be compared against each other using an analysis tool. One advantage of using state based representation of data is that it is independent of the audit trail record and is capable of detecting cooperative attacks and attacks that span across multiple user sessions. However, some attacks cannot be detected because they cannot be modeled with state transitions [http://www.sei.cmu.edu/str/descriptions/rbid.html]

II) Model-based: intrusion attempts in input data can be modeled as sequences of user behavior. This approach allows the processing of more data, provide more intuitive explanations of intrusion attempts and predict intruder's next action. More general representation of penetrations can be generated since intrusions are modeled at a higher level of abstraction. However, if an attack pattern does not occur in the appropriate behavior model it cannot be detected [http://www.sei.cmu.edu/str/descriptions/rbid.html] ..

B. Artificial Intelligence

AI improves algorithms by employing problem solving techniques used by human beings such as learning, training and reasoning. One of the challenges of using AI techniques is that it requires a large amount of audit data in order to compute the profile rule or pattern sets. From the audit trails, information about the system is extracted and patterns describing the system are generated. In general, AI can be employed in two ways: (1) Evolutionary methods (Biologically driven) are mechanisms inspired by biological evolution, such as reproduction, mutation and recombination. (2) Machine learning is concerned with the design and development of algorithms and techniques that allow the learning of computers. The major focus of machine learning research is to extract information from data automatically [2].

C. Data Analysis

With data analysis, data is transformed in order to extract useful information and reach conclusions. It is usually used to approve or disapprove an existing model, or to extract parameters necessary to adapt a theoretical model to an experimental one. Intelligent data analysis indicates that the application is performing some analysis associated with user interaction and then provides some insights that are not obvious. One of the problems faced when applying such an approach is that most application logs (input information) do not conform to a specific standard. Analysis of logs should be performed to find commonalities and different types of logs should be grouped. Another problem is the existence of noise, missing values and inconsistent data in the actual log information. Attackers may take advantage of the fact that logs may not record all information and therefore exploit this point. Finally, real world data sets tend to be too large and multidimensional which requires data cleaning and data reduction [3].

D. Computational Methods

Computational intelligence research aims to use learning, adaptive, or evolutionary algorithms to create programs. These algorithms allow the systems to operate in real time and detect system faults quickly. However, there are costs associated with creating audit trails and maintaining input user profiles as well as some risks. For example, because user profiles are updated periodically, it is possible to accept a new user behavior pattern where an attack can be safely mounted. This is why it is difficult sometimes to define user profiles especially if they have inconsistent work habits. In general, there are two types of IDSs that utilize a computational method: (1) Statistics-based IDS are employed to identify audit data that may potentially indicate intrusive behavior. These systems analyze input audit trail data by comparing them to normal behavior to find security violations. (2) Heuristics-based IDS which can be a function that estimates the cost of the cheapest path from one node to another [http://www.sei.cmu.edu/str/descriptions/sbid.html].

III. Capabilities and Examples of Processing Techniques of Input Data Used by IDSs

Because some IDS data processing techniques are closely interacting and similar, classifying them is complex. However, we believe that the identified eleven categories capture most of the well known types. For example, from Fig. 1, although expert systems and fuzzy logic belong to the categories AI and rule based they have distinguishing characteristics and usages. The output of the expert system is specific; the data that is used to build the system is complete, and the set of rules are well defined. As for fuzzy logic, it is usually used in systems where the output is not well defined and is continuous between 0 and1.

A. Bayesian Networks

Bayesian networks are used when we want to describe the conditional probability of a set of possible causes for a given observed event that are computed from the probability of each cause and the conditional probability of the outcome of each cause. They are suitable for extracting complex patterns from sizable amounts of input information that can also contain significant levels of noise. Several systems have been developed using Bayesian network concepts. In the following system, Scott's [4] IDS is based on stochastic models of user and intruder behavior combined using Bayes' theorem which mitigates the complexity of network transactions that have complicated distributions. Intrusion probabilities can be calculated and dynamic graphics are used to allow investigators to use the evidence to navigate around the system.

B. Neural Networks

Training Neural networks enable them to modify a state of a system by discriminating between classes of inputs. They also learn about the relationship between input and output vectors and generalize them to extract new input and output relationships. They are suitable when identification and classification of network activities are based on incomplete and limited input data sources. They are able to process data from a number of sources, accept nonlinear signals as input and need a large sample size of input information. Finally, neural networks are not suitable when the information is imprecise or vague and it is unable to combine numeric data with linguistic or logical data. In the following system, Bivens et al. [5] employed the time-window method for detection and were able to recognize long multi-packet attacks. They were able to identify aggregate trends in the network traffic in the preprocessing step by looking only at three packet characteristics. Once the system is trained and by using the input data, the neural network was able to perform real-time detection.

C. Data Mining

Data mining refers to a set of techniques that extracts previously unknown but potentially useful data from large stores system logs. One of the fundamental data mining techniques used in intrusion detection is associated with decision trees [6] that detect anomalies in large databases. Another technique uses segmentation where patterns of unknown attacks are extracted from a simple audit and then matched with previously warehoused unknown attacks [7]. Another data mining technique is associated with finding association rules by extracting previously unknown knowledge on new attacks and building normal behavior patterns [8]. Data mining techniques allows finding regularities and irregularities in large input data sets. However, they are memory intensive and require double storage: one for the normal IDS data and another for the data mining. The system of Lee, Solto and Mok's [7] was able to detect anomalies using predefined rules; however, it needed a supervisor to update the system with the appropriate rules of certain attacks. The rule generation methodology developed, first defines an association rule that identifies the relation between rules and specifies the confidence for the rule.

D. Agents

Agents are self contained processes that can perceive their environment through sensors and act on the environment through effectors. Agents trace intruders and collect input information that is related only to the intrusion along the intrusion route and then decide if an intrusion has occurred from target systems across the network. One of the major disadvantages associated with agents is that it needs a highly secure agent execution environment while collecting and processing input information. It is difficult also to propagate agent execution environments onto large numbers of third-party servers. Several systems have been developed utilizing agents. Spafford and Zamboni [9] introduced Autonomous Agents for Intrusion Detection (AAFID) using autonomous agents for performing intrusion detection. Their prototype provides a useful framework for the research and testing of intrusion detection algorithms and mechanisms. Gowadia, Farkas and Valtorta [10] implemented a Probabilistic Agent-Based Intrusion Detection (PAID) system that has cooperative agent architecture. In their model agents are allowed to share their beliefs and perform updates. Agent graphs are used to represent intrusion scenarios. Each agent is associated with a set of input, output, and local variables.

E. Immune Based

Immune based IDS are developed based on human immune system concepts and can perform tasks similar to innate and adaptive immunity. In general, audit data representing the appropriate behavior of services are collected and then a profile of normal behavior is generated. One challenge faced is to differentiate between self and non-self data which when trying to control causes scaling problems and the existence of holes in detector sets.

There have been several attempts to implement immunity-based systems. Some have experimented with innate immunity which is the first line of defense in the immune system and is able to detect known attacks. For example, Twycorss and Aickelin [11] implemented libtissue that uses a client/server architecture acting as an interface for a problem using immune based techniques. Pagnoni and Visconti [12] implemented a native artificial immune system (NAIS) that protects computer networks. Their system was able to discriminate between normal and abnormal processes, detect and protect against new and unknown attacks and accordingly deny access of foreign processes to the server. For adaptive immunity two approaches have been studied: negative selection and danger theory concepts. Kim and Bentley [13] implemented a dynamic clonal selection algorithm that employs negative selection by comparing immature detectors to a given antigen set. Immature detectors that bind to an antigen are deleted and the remaining detectors are added to the accepted population. If a memory detector matches an antigen an alarm is raised. A recent approach to implement adaptive immunity uses the danger theory concept [14]. Danger theory suggests that an immune response reacts to danger signals resulting from damage happening to the cell and not only for being foreign or non-self to the body.

F. Genetic Algorithms

Genetic algorithms are a family of problem-solving techniques based on evolution and natural selection. Potential solutions to the problem to be solved are encoded as sequences of bits, characters, or numbers. The unit of encoding is called a gene, and the encoded sequence is called a chromosome. The genetic algorithm begins with chromosomes population and an evaluation function that measures the fitness of each chromosome. Finally, the algorithm uses reproduction and mutation to create new solutions. In the system of Shon and Moon [15] the Enhanced Support Vector Machine (Enhanced SVM) provides unsupervised learning and low false alarm capabilities. Profile of normal packets is created without preexisting knowledge. After filtering the packets they use a genetic algorithm for extracting optimized information from raw internet packets. The flow of packets that is based on temporal relationships during data preprocessing is used in the SVM learning.

G. Fuzzy Logic

Fuzzy logic is a system of logic that mimics human decision making and deals with the concept of partial truth and in which the rules can be expressed imprecisely. Several systems have been developed using fuzzy logic. Abrahama et al. [16] modeled Distributed Soft Computing-based IDS (D-SCIDS) as a combination of different classifiers to model lightweight and heavy weight IDSs. Their empirical results show that a soft computing approach could play a major role for intrusion detection where the fuzzy classifier gave 100% accuracy for all attack types using all used attributes. Abadeh, Habibi and Lucas [17] describe a fuzzy genetics-based learning algorithm and discuss its usage to detect intrusion in a computer network. They suggested a new fitness function that is capable of producing more effective fuzzy rules that also increased the detection rate as well as false alarms. Finally, they suggested combining two different fitness function methods in a single classifier, to use the advantages of both fitness functions concurrently

H. Expert Systems

Expert systems-based IDSs build statistical profiles of entities such as users, workstations and application programs and use statically unusual behavior to detect intruders. They work on a previously defined set of rules that represent a sequence of actions describing an attack. With expert systems, all security related events that are incorporated in an audit trail are translated in terms of if-then-else rules. The expert system can also hold and maintain significant levels of information. However, the acquisition of rules from the input data is a tedious and is an error-prone process. The system of Ilgun, Kemmerer and Porras [18], is an approach to detect intrusions in real time based on state transition analysis. The model is represented as a series of state changes that lead from an initial secure state to a target compromised state. The authors developed USTAT which is a UNIX specific prototype of a state transition analysis tool (STAT) which is a rule based expert system that is fed with the diagrams. In general, STAT extracts and compares the state transition information recorded within the target system audit trails to a rule based representation of known attacks that is specific to the system.

I. Signature Analysis Or Pattern Matching

In this approach the semantic description of an attack is transformed into the appropriate audit trail format representing an attack signature. An attack scenario can be described, for example, as a sequence of audit events that a given attack generates. Detection is accomplished by using text string matching mechanisms. Human expertise is required to identify and extract non conflicting elements or patterns from input data. The system of Kumar's [19] is based on the complexity of matching. Based on the desired accuracy of detection, he developed a classification to represent intrusion signatures and used different encodings of the same security vulnerability. His pattern specification incorporated several abstract requirements to represent the full range and generality of intrusion scenarios that are: context representation, follows semantics, specification of actions and representation of invariants.

J. State Machines

State machines model behavior as a collection of states, transitions and actions. An attack is described with a set of goals and transitions that must be achieved by an intruder to compromise a system. Several systems have been developed using this technique. Sekar et al. [20] employ state-machine specifications of network protocols that are augmented with information about statistics that need to be maintained to detect anomalies. The protocol specifications simplified the manual feature selection process used in other anomaly detection approaches. The specification language made it easy to apply their approach to other layers such as HTTP and ARP protocols. Peng, Leckie and Ramamohanarao [20] proposed a framework for distributed detection systems. They improved the efficiency of their system by using a heuristic to initialize the broadcast threshold and hierarchical system architecture. They have presented a scheme to detect the abnormal packets caused by the reflector attack by analyzing the inherent features of the reflector attack.

K. Petri Nets

The Colored Petri Nets are used to specify control flow in asynchronous concurrent systems. It graphically depicts the structure of a distributed system as a directed bipartite graph with annotations. It has place nodes, transition nodes and directed arcs connecting places with transitions. In the system of Srinivasan and Vaidehi [22] a general model based on timed colored Petri net is presented that is capable of handling patterns generated to model the attack behavior as sequence of events. This model also allows flagging an attack, when the behavior of one or more processes matches the attack behavior. Their use of a graphical representation of a timed colored Petri net gives a straightforward view of relations between attacks.

IV. Conclusion

Choosing an IDS to be deployed in an environment would seem to be simple, however, with the different components, types and classifications such a decision is quite complex. There have been many attempts to classify IDSs as a mean to facilitate choosing better solutions. In this paper we classified IDSs according to the data processing techniques applied to input information. Careful design of an IDS may allow correct implementation of an IDS. However, the actual merits and limitations of each approach, which is also discussed in this paper, indicate that obtaining complete security and different desirable system characteristics can not be achieved by employing only one type of an implementation approach. The data processing techniques were grouped into general (abstract) categories and were then further expanded into eleven more specialized techniques.

We discussed and summarized the characteristics of each technique followed by examples of developed systems using each technique. Fig. 1, for example, helps us understand that we can use the state machine technique to build an IDS, and that we can add intelligence to it and use the expert system technique with added merits and costs. The merits are the ability to perform and provide intelligent actions and answers. Unrealistic actions or answers can be refuted or ignored. It also borrows from statistics the ability to detect intrusions without prior information about the security flaws of a system. Some of the incurred costs are the conflicting requirement of maintaining high volume of data which affects throughput and selecting the appropriate thresholds that lower false positive and negatives. To conclude, selecting the appropriate technique should be carried out carefully. Each organization should state prior to development the requirements of its agency and the acceptable costs. Accordingly, the selected system should be able to incorporate most of the requirements, as complete security can not be achieved.

V. References

1) H. Debar, M. Dacier, A. Wespi, "Towards A Taxonomy Of Intrusion-Detection Systems," Computer Networks, Vol. 31, pp. 805-822, 1999.
2) S. Peddabachigari, A. Abraham, C. Grosan and J. Thomas, "Modeling Intrusion Detection System Using Hybrid Intelligent Systems," Journal of Network and Computer Applications, Vol. [MAT02], No. 1, pp. 114-132, 2007.
3) Andre' Muscat, "A Log Analysis Based Intrusion Detection System for the Creation of a Specification Based Intrusion Prevention System", CSAW 2003 Proceedings, 2003.
4) S. L. Scott, "A Bayesian Paradigm for Designing Intrusion Detection Systems," Computational Statistics Data Analysis, Vol. 45, No. 1, pp. 69-83, 2004.
5) Bivens, M. Embrechts, C. Palagiri, R. Smith, and B. K. Szymanski, "Network based Intrusion Detection using Neural Networks," Intelligent Engineering Systems through Artificial Neural Networks, Vol. 12, 2002.
6) W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using Artificial Anomalies to Detect Unknown and Known Network Intrusions", In Proceedings of the First IEEE International Conference on Data Mining, San Jose, CA, 2001.
7) W. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach," Artificial Intelligence Review, Vol. 14, No. 6, pp. 533–567, 2000.
8) T. Bass, "Intrusion Detection Systems Multi-sensor Data Fusion: Creating Cyberspace Situational Awareness," Communication of the ACM, Vol. 43, No. 1, pp. 99-105, 2000.
9) H. Spafford and D. Zamboni, "Intrusion Detection Using Autonomous Agents," in Computer Network, Vol. 34, No. 4, pp. 547-570, 2000.
10) Gowadia, C. Farkas, and M. Valtorta, "PAID: A Probabilistic Agent-Based Intrusion Detection System," Computers & Security, 2005.
11) J. Twycross, and U. Aickelin, "Libtissue - Implementing Innate Immunity," Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2006), Vancouver, Canada, 2006.
12) Pagnoni, and A. Visconti, "An Innate Immune System for the Protection of Computer Networks," ACM International Conference Proceeding Series, Vol. 92 archive Proceedings of the 4th international symposium on Information and communication technologies, 2005.

13) J. Kim, and P. J. Bentley, "A Model of Gene Library Evolution in the Dynamic Clonal Selection Algorithm," Proceedings of the First International Conference on Artificial Immune Systems (ICARIS) Canterbury, pp.175-182, 2002.
14) P. Matzinger, "The Danger Model: A Renewed Sense of Self," Science, Vol. 296, pp. [MAT02]1-[MAT02]5, 2002.
15) T. Shon and J. Moon, "A Hybrid Machine Learning Approach to Network Anomaly Detection," Information Sciences: an International Journal, Vol. 177, No. 18, pp. 3799-3821, 2007.
16) Abrahama, R. Jainb, J. Thomasc, and S. Y. Hana, "D-SCIDS: Distributed Soft Computing Intrusion Detection System," Journal of Network and Computer Applications, Vol. 30, pp. 81–98, 2007.
17) S. Abadeh, J. Habibi and C. Lucas, "Intrusion Detection Using a Fuzzy Genetic Based Learning Algorithm," Journal of Network and Computer Applications, Vol. 30, No. 1, pp. 414-428, 2007.
18) Ilgun, R. A. Kemmerer, and P. A. Porras, "State Transition Analysis: A Rule- Based Intrusion Detection Approach," IEEE Transactions on Software Engineering, pp. 181-199, 1995
19) S. Kumar. "Classification and Detection of Computer Intrusions," *Ph.D. dissertation*, Purdue University, 1995.
20) R. Sekar, A. Gupta, J. Frullo, T. Hanbhag, A. Tiwari, H. Yang, and S. Zhou, "Specification-Based Anomaly Detection: a New Approach for Detecting," *International Journal of Network Security*, Vol. 1, No.2, pp. 84–102, 2005.
21) T. Peng, C. Leckie and K. Ramamohanarao, "Information Sharing for Distributed Intrusion Detection Systems," *Journal of Network and Computer Applications*, Vol. [MAT02], No. 3, pp. 877-899, 2007.
22) N. Srinivasan and V. Vaidehi. "Timed Coloured Petri Net Model for Misuse Intrusion Detection." *First International Conference on Industrial and Information Systems*, 8-11 Aug. 2006.

Design of a Dual-band Reconfigurable Antenna

Nazia Hasan[1], Nitin Kumar[1], S.C. Gupta[2], Vipul Sharma[3]
[1]Dehradun Institute of Technology, Dehradun, [2]Indian Institute of Technology, Roorkee , [3]Member IEEE, Gurukul Kangdi Vishwavidyalaya, *Haridwar.*

GJCST Classifications: B.4.1, C.2.0

Abstract: **Reconfigurable antennas have received significant attention for their applications in communications, electronic surveillance and countermeasures, by adapting their properties to achieve selectivity in frequency, bandwidth, polarization and gain. In this paper the design of reconfigurable microstrip antenna operating at two different frequency bands will be presented. The switching between the different frequency bands is achieved by using RF-MEMS switches.**

Keywords- RF-MEMS, reconfigurable, HFSS, microstrip patch.

I. Introduction

With tremendous advancement in technology in the field of communication and the increasing consumer demands, the need for multifunctional wireless communication devices is always felt. Multifunctional systems depend on the co-existence of several antennas and RF components, but as there number of components required in a single system grows, problems such as interference, cost, maintainability, reliability and weight etc may arise. Multifunctional antennas provides a solution to these problems, a multifunctional antenna supports multiple functions in a single antenna unit by supporting more than one frequency or radiate in different patterns etc.

In some designs RF, MEMS (Micro-Electro-Mechanical Systems), solid- state switches or other technologies are used to change the operating frequencies, radiation pattern of the antennas, and are usually named as "Reconfigurable Antennas".[1]

In this paper the design of a reconfigurable antenna operating at two different frequencies is presented. The operating frequency is switched between two values by changing the aperture of the antenna. Operating frequencies are chosen as 2.2 GHz and 3.6 GHz, RF-MEMS switch is used to change the aperture.

II. Microstrip Patch Antenna

The transmission line model (TLM) is used for designing the patch antenna.

The width of the patch is calculated first by,

$$W = \frac{c}{2f}\sqrt{\frac{2}{\varepsilon_r + 1}} \qquad (1)$$

where W is the width of the patch, and ε_r is the substrate dielectric constant. The antenna seems bigger than its physical dimensions due to fringing effect. To take this effect into account a parameter ΔL can be computed from [3]

$$\Delta L = 0.412\,h\,\frac{(\varepsilon_{eff} + 0.3).(W/h + 0.264)}{(\varepsilon_{eff} - 0.258).(W/h + 0.8)} \qquad (2)$$

where h is the height of the substrate and ε_{eff} is the effective dielectric constant given by:

$$\varepsilon_{eff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 10\frac{h}{W}\right]^{1/2} \qquad (3)$$

Since the length has been extended by Δ on each side of the patch, the effective length is given by,

$$L_{eff} = \frac{c}{2f\sqrt{\varepsilon_{eff}}} \qquad (4)$$

Patch resonant length L is given by,

$$L = L_{eff} - 2\Delta \qquad (5)$$

Using the values given by TLM approximation, various parameters for the antenna were calculated for 3.6 GHz. The dielectric substrate chosen here was Rogers RO4032 (ε_r = 3.2) and the height of the substrate h = 1 mm. To feed the patch antenna a microstrip feed line can be attached to the center of one of the radiating edges.

The feed line is a 50-Ohm transmission line and the value of impedance at the edge of the patch is different, so we require a quarter wave transformer to match the impedance of feed line with the patch; this procedure is called as impedance matching. The impedance Z1 of the matching transformer is given by,

$$Z_1 = \sqrt{Z_0 * R_L} \qquad (6)$$

where Z0 is the impedance of microstrip feed and RL is the impedance of the patch.

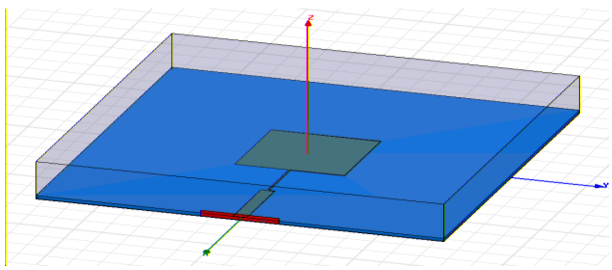


fig.1 Patch at 3.6 GHz

The width of the transformer is calculated by the formula

$$\frac{w}{h} = \frac{8 \exp(A)}{\exp(2A) - 2} \qquad (7)$$

where A is given by

$$A = \frac{Z_0}{60}\left\{\frac{\varepsilon_r + 1}{2}\right\}^{1/2} + \frac{\varepsilon_r - 1}{\varepsilon_r + 1}\left\{0.23 + \frac{0.11}{\varepsilon_r}\right\} \qquad (8)$$

The calculated dimensions are Width W=20.66mm, Length L=20.66mm, Transformer width =0.50mm. These values are based on open loop formulas but the simulator is based on closed loop formulas, so these values need to be readjusted for appropriate results. These were adjusted with the help of HFSS and now the corrected dimensions are Width=19.51mm, Length=19.51mm, Transformer width =0.402mm.

Similarly the dimension of the microstrip patch operating at 2.2 GHz was calculated and adjusted. The adjusted dimensions are Width W=34.2mm, Length L=34.2mm, Transformer width=0.47mm.

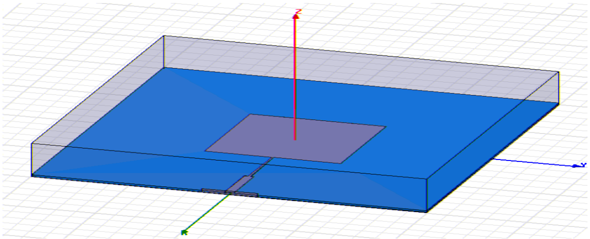


fig. 2 Patch at 2.2 GHz

III. Reconfigurable Microstrip Patch Antenna Design

After designing the microstrip patch at 3.6 GHz, to make this antenna reconfigurable, a square ring is placed around it (as shown in fig.3) with a separation of 420µm between the two. The dimensions of this ring are such as those of the patch operating at 2.2 GHz. The two apertures are connected using 8 RF-MEMS switches at different positions.

When we turn these switches ON the two patches are connected and the whole structure resonates at 2.2 GHz and when the switches are turned OFF only the inner patch resonates at 3.6 GHz. The outer ring acts as a parasitic element in switch OFF condition.

The effect of placing this square ring was that in switch OFF position it and affects the performance of the antenna. Also the isolation provided by the switches in OFF condition affected the performance. As a result the resonant frequency of the antenna was changed and shifted to a higher value than 3.6 GHz, so we need to adjust the effects due to this parasitic ring and we did it by increasing the dimensions of the inner patch maintaining the separation of 420µm.

After adjusting the dimensions in switch OFF position, we turned the switches ON. In ON condition the switches provide insertion loss due to which the parameters of the antenna changes again. And, again we had to readjust the dimensions of the patch.

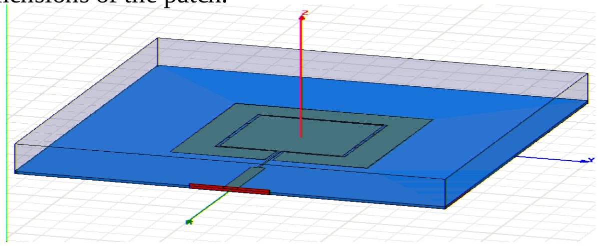


fig. 3 Reconfigurable Antenna

In switch ON position the problems faced were more than in switch OFF position, because we are using the same antenna designed for 3.6 GHz (smaller or inner) along with a square ring, to work at 2.2 GHz (combined). The dimensions of quarter wave transformer used for impedance matching remains the same in switch ON as well as in switch OFF position, which causes the VSWR to increase in earlier case.

The final dimensions of the antenna are:

Switch OFF position: Frequency 3.62 GHz, Length 21.5mm, Width 21.5 mm, Transformer width 0.4506mm.

Switch ON position: Frequency 2.2 GHz, Length 34.2mm, Width 34.2 mm.

Few parameters of the antenna are as shown below

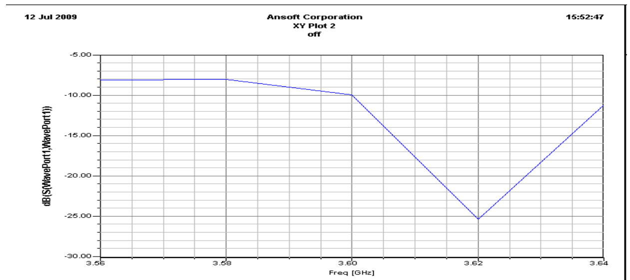


fig. 4 Return loss in switch OFF position (-25 dB approx.)

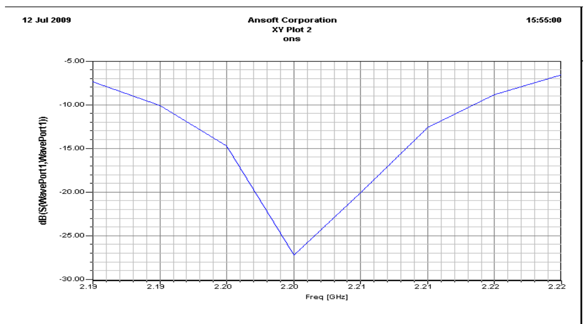


fig. 5 Return loss in switch ON position (-27 dB approx.)

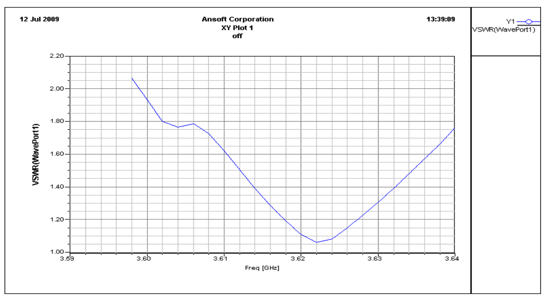


fig. 6 VSWR in OFF state (1.05 approx)

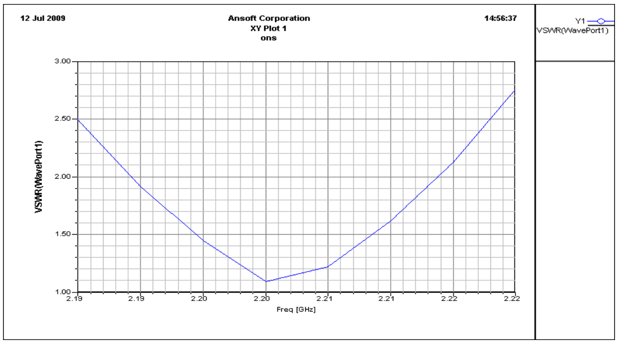


fig. 7 VSWR in ON state (1.05 approx.)

IV. RF-MEMS [4]

The term RF MEMS refers to the design and fabrication of MEMS for RF integrated circuits. By utilizing electromechanical architecture on a miniature- (or micro-) scale, MEMS RF switches combine the advantages of traditional electromechanical switches (low insertion loss, high isolation, extremely high linearity) with those of solid-state switches (low power consumption, low mass, long lifetime). While improvements in insertion loss (< 0.2 dB), isolation (> 40dB), linearity (third order intercept point > 66dBm), and frequency bandwidth (dc-40GHz) are remarkable, the RF MEMS switches are slower and have lower power handling capabilities.

The switch used in our model is based on two; 1 µm thick gold cantilevers with dimensions of 150x80 µm2 with a central conductor (60x40 µm2) joining the two side cantilevers. The cantilever is suspended 101 µm above the substrate and 1µm above the patch. The anchors supporting the structure are 4x4 µm2 of gold and 101µm thick (in switch OFF). A gold hinge (4x30 µm2) joins the anchor with the side cantilever. The metal strip (T-line) joining the two patches is 40 µm wide, there is break of 40 µm in the mid of the T-Line, such that the central conductor of the switch is placed just above it.

To bring the switch in ON state, the gap between the switch and the patch is eliminated by reducing the height of the components of the switch by 1µm, and a voltage of 10V is also applied from the actuation area; defined below the side cantilever. In actual this voltage pulls the cantilever down so that the switch is closed.

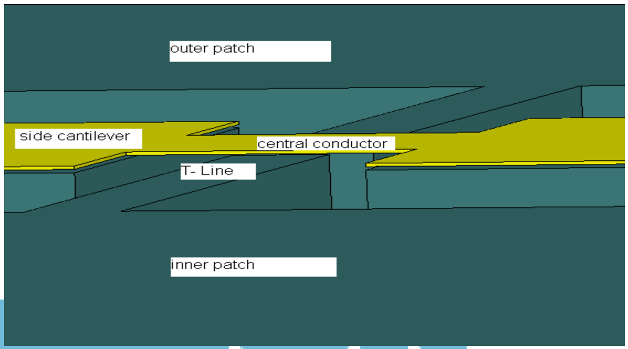


fig. 8 Switch in OFF state

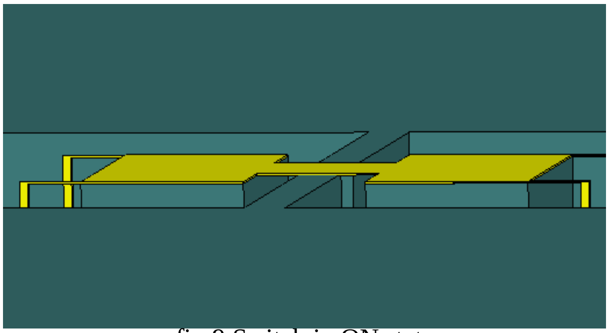


fig.9 Switch in ON state

When the switch is turned ON, the central conductor fills the gap and acts as a bridge for the current to pass through the T-Line. In switch OFF position the break is maintained and there is no path for the current to pass.

The two parameters of the switch: namely insertion loss and isolation are as shown below

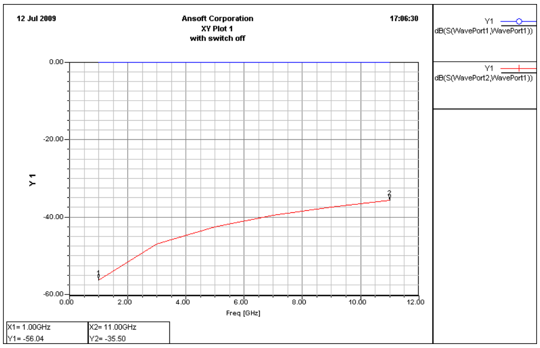


fig.10 S11(blue), S21(red) in OFF state

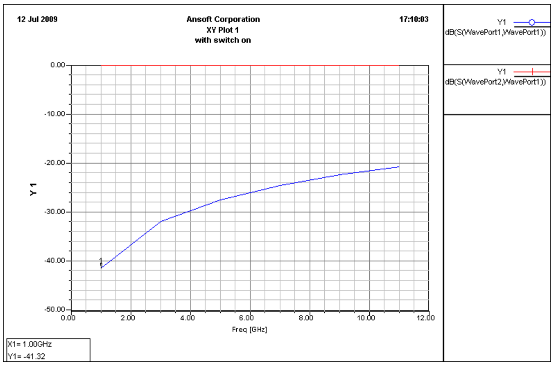


fig.11 S11(blue), S21(red) in ON state

V. Discussion and Conclusions

The dual frequency reconfigurable antenna element is suitable for use in communication applications. However the reliability and life-time problems should also be addressed before commercial use.

The design of a frequency reconfigurable antenna by using the ON/OFF state properties of a low voltage-actuated MEMS switch is presented in this paper. The two parameters of the switch namely Insertion loss and Isolation

affect the performance of the antenna and hence have to be taken care of. Probant experimental verifications have been made by using a 420x40µm2 small piece of copper to model the ON state of the MEMS switch. The results were compared with theoretical ones, for the structure good agreement between simulated and experimental results is achieved.
Use of good quality RF MEMS switches, optimizing their location and performance can improve the working of the antenna.

Here reconfiguration in only one parameter is presented; we can also go for reconfiguration in other parameters such as radiation pattern or polarization. Reconfiguration in more than one parameter using a single antenna is also possible but it greatly increases the complexity of the design.

VI. References

1) Dr. John L. Volakis, “Antenna Engineering Handbook”, 4th Edition,. Chapter 31.
2) Vijay K. Varadan, K.J. Vinoy and K.A. Jose, “RF MEMS and their applications,” John Wiley and Sons, Chichester, UK, 2003.
3) Ramesh Garg, Prakash Bhartia, Inder Bahl, “Microstrip Antenna Design Handbook”, Artech House Inc., Boston, London 2001.
4) Gabriel .M. ReBeiz, “RF MEMS: Theory, design and technology,” John Wiley and sons, New Jersey, 2003.

Updated Congestion Control Algorithm for TCP Throughput improvement in Wired and Wireless Network

GJCST Classifications:
C.2.5, C.2.1

Prof.K.Srinivas
Dept of MCA,KIPGS,KONDAIR,509125 India; srinivas_mnb4u@yahoo.co.in
Dr.A.A.Chari
Dept of SQC,Rayalaseema University, Kurnool 518002, India; directorresearch.ru@gmail.com,
Prof.N.Kasiviswanath
Dept of CSE, GPREC, Kurnool 518002, India; nkv@gmail.co

Abstract- The basic idea proposed in this paper is to determine the Optimal Congestion Window for a TCP Sender in a particular network set-up (that corresponds to the fair share of that connection) and keep this congestion window a constant to a point where the fair share in the network has changed considerably from the instance of the calculation of the size of the last window. At this point, the TCP Congestion Window is recalculated according to the nature of new circumstances. The proposed mechanism is particularly effective over wireless links, which have an inherently loss-prone nature, as Modified TCP's congestion window being independent of packet losses (be it corruption losses or it congestion losses), keeps transmitting at the same rate at before.

I. INTRODUCTION

The well-known challenge in providing TCP congestion control algorithm [1], [2], [12] in wired – cum – wireless environment is that it relies on the packet loss as an indicator of network congestion. In order to ease the congestion scenario and to avoid a congestion collapse, a TCP Reno Sender reduces the congestion window (henceforth referred to as cwnd and expressed in number of segments) and refrains from sending packets. In the wired portion of the network, a congested router is invariably the likely reason of packet loss, while in the wireless portion a noisy, fading radio channel is the more likely cause of loss. This creates problems in TCP Reno since it does not possess the capability to distinguish and isolate congestion loss from wireless loss. Approaches to address this problem have been discussed and compared in the work by Balakrishnan et al. [3]–[4]. Three alternative approaches: end-to-end (E2E), Split Connection, and Localized Link Layer methods were carefully contrasted.

The split-connection approach [13]-[14] violates the semantics of E2E reliability. Secondly, this approach requires a lot of state maintenance at the base station.

In this paper, we propose a TCP Sender side modification of the TCP congestion control algorithm [5]. The crux of idea is that for a given network scenario, the Modified TCP Sender determines its optimal fair share of bandwidth in the link setting its cwnd in a way that it can effectively transmit with a rate that utilizes the fair share of bandwidth. After the cwnd is set to a value optimal for a given network scenario, it is kept constant to the point where the network scenario has changed by a extent significantly altering the connection's fair share. Since the value of cwnd is not decreased at any packet loss indication like retransmission on receipt of a triple DUPKT, or a coarse timeout caused by the expiration of the Retransmission Timer, hence it is not susceptible to performance degradation and cwnd reduction on the occurrences of stray packet losses. This leads to an enhanced performance in the wireless domain, as the losses are never an indication of congestion, rather they are caused due to the inherent loss-prone nature of the radio propagation medium. We provide simulation results in support to our claim that constant cwnd can outperform TCP Reno in static (i.e., certain time interval) network scenarios.

The rest of this paper is organized as follows: section 2 summarizes some related work; section 3 gives the analytical approach; section 4 describes the algorithm used by the sender; section 5 summarizes the results obtained by the simulations; section 6 gives an idea of the challenges faced while implementing such a strategy; and finally, section 7 concludes the paper.

II. RELATED WORK

NCPLD [15] compares the measured rtt with the lowest rtt (or that at the knee of the goodput – load curve). If the former is close to the latter, then the cause of a packet loss is assumed to be wireless errors. TCPW [8]–[11] measures goodput (or reception rate) and uses that rate to set the congestion window whenever a packet is detected lost. If the current goodput is below a certain band around the mean, then the cause of a packet loss is assumed congestion, otherwise the cause of loss is attributed to wireless errors.

This paper uses the TCPW bandwidth estimation scheme and compares the performance of the Modified sender with the TCPW sender. The TCPW [8]–[11] sender monitors ACKs to estimate the bandwidth currently used by, and thus available to the connection. More precisely, the sender uses (1) the ACK reception rate and (2) the information an ACK conveys regarding the amount of data delivered to the destination. The Westwood algorithm is described briefly below.

Let us assume that an ACK is received at the source at time tk, notifying that dk bytes have been received at the TCP receiver. We can measure the following sample bandwidth used by that connection as bk = dk/Δk, where Δk = tk−tk−1 and tk−1 is the time the previous ACK was received

The following discrete-time filter is used which is obtained by discretizing a continuous low-pass filter using the Tustin approximation

$$b'_k = \alpha_k b'_{k-1} + (1 - \alpha_k)(b_k + b_{k-1})/2$$

where b'_k is the filtered estimate of the available bandwidth at time $t = t_k$, $\alpha_k = (2\tau - \Delta_k)/(2\tau + \Delta_k)$, and $1/\tau$ is the cutoff frequency of the filter.

Algorithm after n duplicate ACKs

The pseudo code of the algorithm is the following:

```
if (n DUPKTs are received)
    ssthresh = (BWE * RTTmin)/seg_size;
    if (cwin > ssthresh) /* congestion avoid */
            cwin = ssthresh;
    endif
endif
```

Here, seg_size identifies the length of the payload of a TCP segment in bits.

Algorithm after coarse timeout expiration

The pseudo code of the algorithm is:

```
if (coarse timeout expires)
    ssthresh = (BWE * RTTmin)/seg_size;
    if (ssthresh < 2)
            ssthresh = 2;
    endif;
    cwin = 1;
endif
```

III. Analytical Approach

The logic for using a constant window would be summarized as under:

As in [1] if we measure the network load by average queue length over fixed intervals of some appropriate length, and L_i be the load at instant i, then, for a congested network we have:

$$L_i = N + \gamma L_{i-1} \qquad (1)$$

where N (a constant) accounts for the average arrival rate of the new traffic, and γLi-1 accounts for the traffic left from the last time interval. Evidently, the term γLi-1 arises when the sender is sending at a rate which is greater that its fair share leading to a fraction of packets from the previous round remaining in the network when the packets form the next round arrives in the network. But if the sender is sending at a rate that utilizes its fair share, the γLi-1 vanishes; equation (1) thereby reduces to

$$L_i = N \qquad (2)$$

which is a constant, and this forms the basis for use of a constant congestion window.

IV. TCP Modifications

The key idea here [5] is that we can divide the entire lifetime of a TCP connection into a finite number of slots such that the connection's fair share in the network remains almost same in a particular slot, i.e. we may assume that the network scenario remains *almost static* with such slot. A change in the available share of a network, due to some connections leaving the network or some new connections joining, ends a slot and marks the beginning of the next slot. Our proposal is to use a constant TCP Congestion Window during these slots where the network scenario is assumed to remain unchanged. The beginning of a new slot would trigger a window recalculation and the cwnd would be set according to the connection's available share in that slot.

In the proposed mechanism, we use a bandwidth estimation algorithm similar to that of TCPW to obtain an estimate of the available fair share. The change in the rtt measurements is used as a trigger to move to the recalculation phase from the constant window phase (our model uses the knee region in the rtt curve as in [15] to detect a change in fare share and trigger recalculation). In our model, the Modified TCP Sender moves through three distinct phases during its lifetime: the startup phase followed by mutually interleaved window recalculation phase and constant window phase. The three phases are described with some detail as under.

A. The Startup Phase

At connection setup, the sender has no inkling of the network scenario. In order to impart dynamic nature, the Sender refrains from using typical default values for these essential attributes of the connection. The sender uses a slow start mechanism as in [1]. The sender continues the slow start process for say k rounds, during which it acquires various vital information about the network such as the minimum rtt measurement, a measure of the network bandwidth that the connection etc. After the first *k* rounds, the sender has acquired enough information about the network and hence calculates cwnd for the first time.

B. Window Recalculation Phase

When a change in available fair share is detected by the trigger, the TCP sender enters this phase. This is the most crucial phase of the connection, as in this phase, the cwnd is calculated which is kept a constant during the next phase. Hence the performance of the sender, how well it utilizes its share of the network, depends on the cwnd calculated. Along with the window recalculation process, the current value of the smoothed rtt measurements, obtained by passing the coarse rtt measurements of the individual segment through a low pass filter as suggested by Jacobson [1], is also archived for future reference.

An efficient Bandwidth Estimation Algorithm must be in place to determine the fair share of the connection in the network. The accuracy of this algorithm in determining the network share would determine the performance of the Modified TCP Sender.

C. The Constant Window Phase

During this phase of the connection, the cwnd is kept a constant irrespective of the number of ACKs received or any indications of packet loss like DUPKT or a coarse timeout. The sender keeps track of the rtt estimates from the segments that have been delivered. If the percentage change in the smoothed rtt measurements over the archives rtt measure is greater than a specified threshold, the sender exits the constant window phase and enters the Window Recalculation Phase i.e. if $|rtt_{arc}-rtt_{var}|/rtt_{arc}>\beta$, a window recalculation is made.
The algorithm's pseudo code is as follows

```
if( slow_start_state)
        slow_start(); /* open cwnd by one segment on each ACK arrival */
else
{
        if(|rtt_arc–rtt_var|/rtt_arc>β)    /*fractional increase greater than threshold */
        {
                /* recalculate window and archive the value of rtt_var */
                cwnd_ = (Estimated_Bandwidth* rtt_min) / seg_size_;
                if(cwnd_ < 1) cwnd_=1;
                        rtt_arc = rtt_var ;
        }
}
```

In the pseudo code, seg_size_ identifies the length of the TCP segments in bytes; rtt_{min} is the estimated minimum value of rtt throughout the lifetime of the particular connection, and Estimated Bandwidth is the Bandwidth Estimate obtained by some Bandwidth Estimation Algorithm.

V. PERFORMANCE ANALYSIS

In this section, we report on the basic performance behavior of the modified TCP senders and its fairness among a number of connections sharing a bottleneck link. A performance comparison is made with the TCP Reno and TCP Westwood [8]-[11] sources operating in similar network scenarios. Intermediate node buffer capacity is always set equal to the bandwidth delay product for the bottleneck link based on literature studied. Increasing the buffer capacity further does not have any impact on the performance [15]. The traffic model used is FTP with infinite data to send so that the sender has data to send whenever the network permits, and the packet size is set to 1000 bytes (1040 bytes with headers) in all experiments. The wireless subnet is error prone. In our simulations we have used the conventional TCP Sink which responds with an ACK for every packet received. There is no congestion or error in the ACK path. All simulations have been carried out for a period of 250 seconds with the TCP senders transmitting data for the entire period of simulation. All the simulations have been carried out with 802.11 MAC with a maximum available bandwidth of 1Mbps. A Two Ray Ground propagation model is used with an Omni-directional antenna. The wired subnet is error free while the wireless subnet is prone to varying error rates.

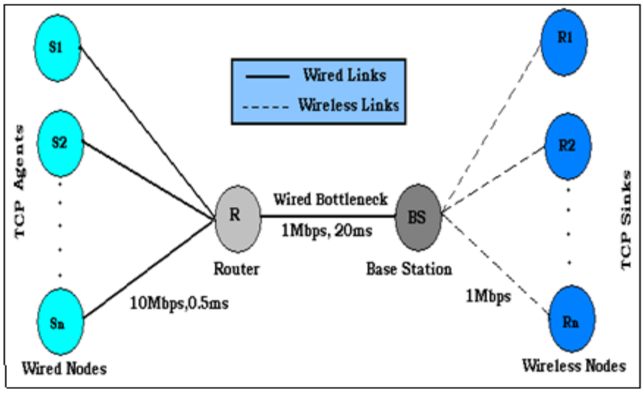


Figure 1: Network Scenario used for simulation

The performance of the Modified TCP Senders, TCP Westwood, and TCP Reno has been compared based on the throughput metric, i.e. the number of data packets received at the sender. We have analyzed the performance of the Constant Congestion Window aspect of the Modified TCP to assert that in the time slots when the share of a connection in the network remains unchanged, a Constant cwnd TCP outperforms the Reno and Westwood sources in situations with wireless errors. Our assumption is that the share of a connection remains unchanged during the entire period of simulation. The optimal cwnd for a given scenario has been evaluated the cwnd of the modified TCP Sender has been set accordingly. One aspect is to be noted that we are not simulating the entire lifetime of a modified TCP sender. Rather, our analysis is concentrated only on the Constant Congestion Window phase of the connection.
All simulations in this paper have been carried out using the LBL network simulator ns2 [6], [7] with appropriate modifications for implementation of the changes in the modified TCP sender. For comparison with TCP Westwood, the corresponding TCPW modules were used [16].
Figure 1 shows a schematic of the scenario used for simulation. A number of TCP connections share a common wired bottleneck that connects the intermediate router to the base station. When there is only one TCP connection in the network, there is no loss due to congestion. As a result, any packet loss is due to wireless errors. Hence, we can evaluate the performance of the Modified congestion control algorithm in scenarios where wireless loss is the only cause for packet loss. As the numbers of source/receiver pairs are increased, gradually the wired link between the router and the base station would become congested. Hence packets will also be lost both due to congestion as well as wireless errors. Hence, the performance of the Modified TCP sender in congested networks can also be evaluated using the same scenario.

A. Constant Bit Error rates

In the scenarios under consideration, the wireless subnet is prone to constant bit error rates. Figure 2 compares the performance of the Constant cwnd senders for different values of cwnd. As is evident, for every scenario, there exists a value of cwnd (in some cases more than one) for which the performance of the TCP Sender is maximum. This is the optimal cwnd for the given network scenario. In figure 2, cwnd is expressed in segments.
As is evident from figure 3, a Constant cwnd TCP outperforms the Reno and Westwood senders operating in similar network conditions. A 10-15% increase in throughput has been obtained as is evident from figure 3. In figure 3, the error rates are expressed as percentage. Figure 4 compares the performance of the TCP variants for multiple connections sharing the wired bottleneck and hence, packet is lost due to congestion as well. The Constant cwnd TCP sender outperforms Reno and Westwood in such scenarios as well.

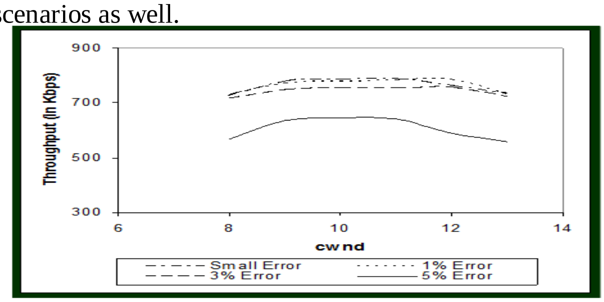


Figure 2: Variation of Throughput with varying cwnd for various bit error rates for single S/R pair.

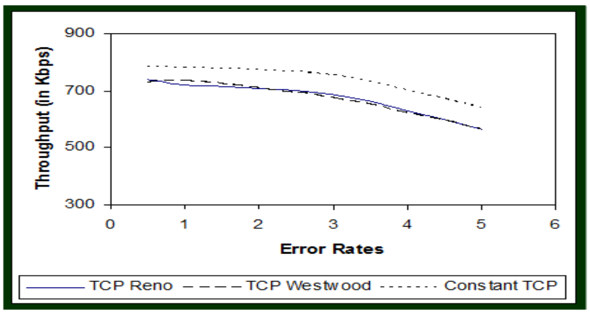


Figure 3: Variation of throughput with varying error rates in scenarios with constant bit error rates for single S/R pair

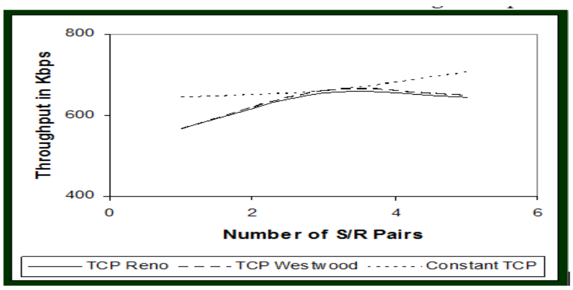


Figure 4: Variation of Throughput with number of TCP connections sharing the link in scenarios with 5% loss in constant bit error

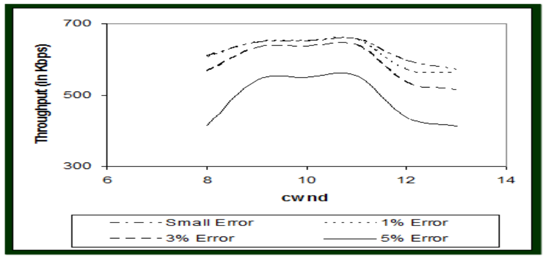


Figure 5: Variation of Throughput with varying cwnd error for various burst error for single S/R pair

B. Burst Error

This subsection compares the performance of Modified TCP, TCP Reno and TCP Westwood based on the throughput metric. In the scenarios under consideration, the wireless subnet is prone to burst error. The burst error is modeled using is a discrete time first order Markov Model. The pattern of errors is described by the transition matrix

$$M=\begin{bmatrix} PBB & PBG \\ PGB & PGG \end{bmatrix}$$

Where is p_{BG} the transition from bad to good, i.e., the conditional probability that successful transmission occurs in a slot given that a failure occurred in the previous slot, and the other entries in the matrix are defined similarly. It is to be noted that represents $1/(1 - p_{BB})$ the average length of a burst of errors, which is described by a geometric random variable.
Figure 5 compares the performance of the Constant cwnd senders for different values of cwnd. As is evident, for every scenario, there exists a value of cwnd (in some cases more than one) for which the performance of the TCP Sender is maximum. This is the optimal cwnd for the given network scenario. In figure 5, cwnd is expressed in segments.
When comparing the performance of the Constant cwnd TCP Sender, the cwnd is set to the optimal value for the given scenario. As is evident from figure 6, a Constant cwnd TCP outperforms the Reno and Westwood senders operating in similar network conditions. A 10-20% increase in throughput has been obtained as is evident from figure 6. In figure 6, the error rates are expressed as percentage.
Figure 7 compares the performance of the TCP variants for multiple connections sharing the wired bottleneck and hence, packet is lost due to congestion as well. The Constant cwnd TCP sender outperforms Reno and Westwood in such scenarios as well.

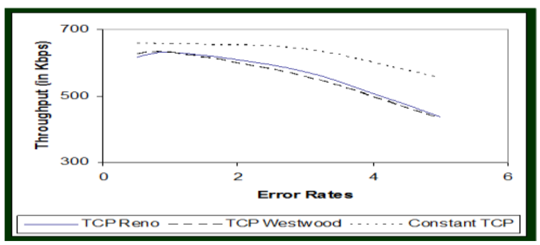


Figure 6: Variation of throughput with varying error rates in scenarios with burst error and for single S/R pair

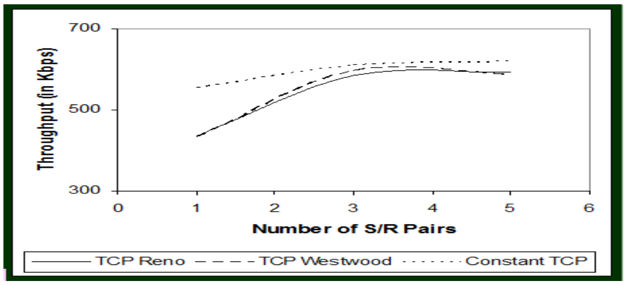


Figure 7: Variation of Throughput with number of TCP connections sharing the link in scenarios with 5% loss in burst error

VI. CHALLENGES IN IMPLEMENTING THE PROPOSED MECHANISM

In the earlier sections of the paper, we have proposed Sender side modification of the TCP congestion control algorithm. There are certain challenges, which need to be met in order for this mechanism to work more efficiently. Firstly, the bandwidth determination algorithm would be precisely able to calculate the available fair share of the connection in the network. An incorrect estimation would negate the performance enhancement, which would be gained by not reducing the window in case of wireless errors. Secondly, the triggering mechanism would be able to efficiently determine a change in the available fair share of the bandwidth in the network. Failure to do so would lead to potential over or under utilization of the available fair share in case the fair share of the connection decreases or increases respectively.

VII. CONCLUSION AND FUTURE WORK

In this paper, we propose a sender side modification of the TCP congestion control algorithm. In addition to this proposal, we have evaluated and compared the performance of the modified TCP sender during a particular phase of its lifetime viz. the Constant Congestion Window Phase. The simulations performed has shown a throughput enhancement of 10-15% as compared to TCP Reno and Westwood in cases with constant bit error rates and about 10-20% in cases with burst error corresponding to a discrete time first-order Markov model.

One important aspect of operation of this modified TCP is that the cwnd should be set to a value optimal for a given connection. For this purpose, an efficient Bandwidth Estimation Algorithm would be designed that would dynamically determine a connection's fair share based on certain observed and measured parameters. We are working on to derive a function that would dynamically determine the cwnd during the window recalculation phase.

VIII. REFERENCES

1) Jacobson and M. J. Karels, Congestion Avoidance and Control, in Proceedings of ACM SIGCOMM, August 1988, vol. 18, pp 314-329.
2) Jacobson, Modified TCP congestion avoidance algorithm. Note sent to end2end-interest mailing list, 1990.
3) H. Balakrishnan et. al., A comparison of mechanisms for improving TCP performance over wireless links, IEEE/ACM Transactions on Networking (December 1997).
4) H. Balakrishnan and R.H. Katz, Explicit loss notification and wireless web performance, in: Proceedings of IEEE GLOBECOM'98 Internet Mini-Conference, Sydney, Australia (November 1998).
5) R. Roy, S. Das, A.K. Ghosh and A. Mukherjee, "Modified TCP Congestion Control Algorithm for Throughput Enhancement in Wired-cum-Wireless Networks", Proceeding of SNCNW 2006, Luela, Sweden.
6) UCB/LBNL/VINT. The ns-2 network simulator, June 2004. http://www.isi.edu/nsnam/ ns/.
7) Fall and K. Varadhan, The ns Manual, December 13, 2003, http://www.isi.edu/nsnam/ns/ns-documentation.html
8) C Casetti, et al, TCP Westwood: End-to-End Bandwidth Estimation for Enhanced Transport over Wireless Links, Kluwer Academic Publishers. Wireless Networks 8, 467–479, 2002
9) Gerla, B.K.F. Ng, M.Y. Sanadidi, M. Valla, R. Wang, TCP Westwood with adaptive bandwidth estimation to improve efficiency/friendliness tradeoffs, Computer Communications xx (2003) 1–18, www.elsevier.com/locate/comcom
10) J. Chen, F. Paganini, R. Wang, M.Y. Sanadidi, M. Gerla, Fluid-flow Analysis of TCP Westwood with RED, IEEE GLOBECOM 2003.
11) R. Wang, K. Yamada, M. Y. Sanadidi, and M. Gerla, TCP With Sender-Side Intelligence to Handle Dynamic, Large, Leaky Pipes, IEEE Journal on Selected Areas in Communications, Vol. 23, No. 2, Feb 2005, pp 235-248.
12) Allman, M., Paxson, V. and Stevens W.R., TCP congestion control, RFC 2581, April 1999.
13) R. Yavatkar and N. Bhagwat, Improving End – to – End Performance of TCP over Mobile Inter networks, in Proceedings of Workshop on Mobile Computing Systems and Applications, December, 1994.
14) A.Bakre and B.R. Badrinath, I-TCP: Indirect TCP for Mobile Hosts, International Conference on Distributed Computing Systems 1995, pages 136–143.
15) N. K. G. Samaraweera, Non- Congestion Packet Loss Detection for TCP error recovery using wireless links, IEE Proc. on Communications, Vol 146, No. 4, Aug 1999, pp 222 – 230.
16) Network Research Lab, Dept. of CS, UCLA, TCP Westwood modules for ns-2, http: //www.cs.ucla.edu/NRL/hpi/tcpw/tcpw_ns2/tcp-westwood-ns2.html

Collaborative Web Recommendation Systems -A Survey Approach

GJCST Classifications: C.2.5, H.3.5, A.1

A. Kumar* and Dr. P. Thambidurai#
*Research Scholar, Department of Computer Science & Engineering, Sathyabama University, Chennai, India. email:akr2020av@yahoo.in
Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry, India. email: ptdurai58@yahoo.com

Abstract*-This paper is a survey of recent work in the field of web recommendation system for the benefit of research on the adaptability of information systems to the needs of the users. This issue is becoming increasingly important on the Web, as non-expert users are overwhelmed by the quantity of information available online, while commercial Web sites strive to add value to their services in order to create loyal relationships with their visitors-customers. This article views to provide a remedy for the negative effects of the traditional one-size-fits-all approach is to enhance the system's ability to adapt its own behavior to the user's characteristics, such as goals, tasks, interests, that are stored in user profiles by implementing a variety of algorithms. The enormous content of information on the World Wide Web makes it obvious candidate for Web Recommendation System research. Web based application facing with large amount of data. In order to produce the portal usage patterns and user behaviors, Web recommendation system consists of three main phases, namely Data Preprocessing, Pattern Discovering and Pattern Analysis. Server log files become a set of raw data where it must go through with all the Web recommendation system phases to produce the final results. Here, Web recommendation system, approach has been combining with the basic Association Rules, Apriori Algorithm to optimize the content of the E-application portal. Finally, this paper will present an overview of results analysis and can use the findings for the suitable valuable actions.

I. INTRODUCTION

An abundant amount of information is created and delivered over electronic media. Users risk becoming overwhelmed by the flow of information, and the users lack adequate tools to help them manage the situation. Information filtering (IF) is one of the methods that are rapidly evolving to manage large information flows. The aim of IF is to expose users to only information that is relevant to them. Many IF systems have been developed in recent years for various application domains. Information filtering systems can help users by eliminating the irrelevant information and by bringing the relevant information to the user's attention. Filters are mediators between the sources of information and their end-users.

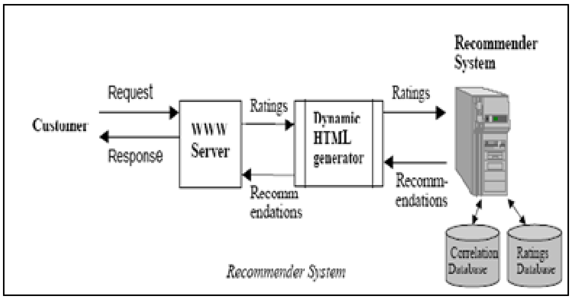
The system is based on a user modeling component [21], designed for building and maintaining long term models of individual Internet users. Presently the system acts as an intelligent interface for the Web search engines. The experimental results we have obtained are encouraging and support the choice of adaptive Information Filtering. Its main goal is the management of the information overload and increment of the semantic signal-to-noise ratio. To do this the user's profile is compared to some reference characteristics. These characteristics may originate from the information item (the content-based approach) or the user's social environment (the collaborative filtering approach).Whereas in information transmission electronic filters are used against syntax-disrupting noise on the bit-level, the methods employed in information filtering act on the semantic level. The range of machine methods employed builds on the same principles as those for information extraction [1]. A notable application can be found in the field of email spam filters. Thus, it is not only the information explosion that necessitates some form of filters, but also inadvertently or maliciously introduced pseudo-information.

The different systems use various methods, concepts, and techniques from diverse research areas like: Information Retrieval, Artificial Intelligence, or Behavioral Science. Various systems cover different scope; have divergent functionality, and various platforms. There are many systems of widely varying philosophies, but all shares the goal of automatically directing the most valuable information to users in accordance with their User Model, and of helping them use their limited reading time most optimally.

When a user interacts with the system for the first time, the user model needs to be made from scratch. In order to quickly build a reliable model an interview is proposed to the user, expressing an interest score for each of the domain categories. The user sets a query to the system that in turn posts it to the external WWW search engine, obtaining documents that are filtered and returned to the user. In the filtering process the systems works using two different levels of refinement, a first, coarse one, and a more elaborate step that takes place only if the first stage succeeds. During the normal usage the system offers a series of panels, being the first the filtering panel [19]. Here at the left is shown the list of documents retrieved by the search engine given the user query.

For an easier usage the system automatically sorts the document lists so to help the user locating the best documents. The user browses the needed documents by double-clicking on them, and then he can express a simple feedback [15] among three different values: very good, good or bad, in order to ease the burden on the user as recommended. In this way the system can modify the user model accordingly to user's preferences [3 and 4]. Furthermore, a system objects browser has been provided in order to allow the user to inspect all the system's data structures with an effective graphical interface to shorten the semantic gap between the user and the system. In the next section the user modeling component is presented.

II. Recommendation Systems



Recommender systems are active information filtering systems that attempt to present to the user information items (movies, music, books, news, web pages) the user is interested in. These systems add information items to the information flowing towards the user, as opposed to removing information items from the information flow towards the user. Typically, a recommender system compares the user's profile to some reference characteristics, and seeks to predict the rating that a user would give to an item they had not yet considered [20]. Recommender systems use collaborative filtering approaches or a combination of the collaborative filtering and content-based filtering approaches, although content-based recommender systems do exist [7].

Web-based Recommender Systems (RS) are recently applied to provide different type of customized information for their users. The Recommender Systems are applied in many areas such as: web-browsing, information filtering, net-news or movie recommender and e-Commerce. The central element of all recommender systems is the user model that contains knowledge about the individual preferences which determine his or her behavior in a complex environment of web-based systems. User modelings as well as RS are characterized by cross-fertilization of various research fields such as: Information Retrieval, Artificial Intelligence, Knowledge Representation, Discovery and Data/Text Mining, Computational Learning and Intelligent and Adaptive Agents. The alternating information environment that is combined of various users, their needs and contexts of use as well as different system platforms necessitates application of recommender systems.

The ever increasing importance of the e-Commerce in the global economy also increases the importance of web-based RS's. RS systems are developed by different domains such as personal agents and adaptive hypermedia. The personalized hypermedia application is defined as a hypermedia system that adapts: the content, structure, and/or presentation of the web objects to each individual user's model. RS's are applied in many different areas from web browsing for purchase recommendation. Montaner et. al in their work present comprehensive taxonomy of the recommender agents. In this taxonomy the following two dimensions are considered: profile generation and maintenance, and profile exploitation. The dimension of profile generation and maintenance considers the following elements: user profile representation, initial profile generation, profile learning technique and relevance feedback [22].

III. Classification of Recommendation System

Many groups have built various types of systems that recommend pages to web users. This section will summarize several of those systems, and discuss how they differ from our approach. The objective of collecting user information is to create a profile that describes user characteristics. The more common techniques are explicit profiling, implicit profiling, and use of legacy data:

Explicit profiling: Each user is asked to fill in a form when visiting the web site. This method has the advantage of letting users specify directly their interests.

Implicit profiling: The user's behavior is tracked automatically by the system. This method is generally transparent to the user. Often, user registration is saved in what is called a cookie that is kept at the browser and updated at each visit. Behavior information is generally stored in a log file.

Legacy data: The Legacy data provides a rich source of profile information for known users.

IV. A Survey on Web Recommendation

Piatetsky-Shapiro et. al., discusses in [5] personalization is a process of gathering and storing information about visitors of a web site, analyzing the stored information, and, based on this analysis, delivering the right information to each visitor at the right time. A personalization component should be capable to recommend documents and/or other web sites, promote products, make appropriate advice, target e-mail, etc. Personalization is increasingly used as a mean to expedite the delivery of information to a visitor, making the site useful and attractive so that the visitor is stimulated to return to it. For this, personalization is one of the e-business web sites.

A personalization component builds and exploits models or profiles of the users interacting with the system. A user profile is a (possibly structured) representation of characteristics of that user, in order to take into accounts his or her needs, goals, and interests.

A. *Recommendation System Using Apriori Algorithm*

R. Agrawal et. al., discusses in [2] that recommendation system using apriori algorithms a classic algorithm for learning association rules [13]. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

As is common in association rule mining, given a set of itemsets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a bottom up approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found.

B. Rule-Based Techniques

Rule-based techniques exploit a set of rules specified in the system in order to drive personalization. Cross-selling is an e-business example of the rule-based technique: a rule could be specified to offer product X to a customer who has just bought product Y. For example, a customer of a book might be interested in current or previous books by the same author or in books on the same subject.

C. Item-Based Collaborative Filtering

Bardul M. Sarwar et. al., projected a different approach in the area of filtering algorithms, that was suggested recently [28] [29], is based on item relations and not on user relations, as in classic Collaborative Filtering. In the Item-based Collaborative Filtering algorithm, we look into the set of items, that the active user, has rated, compute how similar they are to the target item and then select the k most similar items {i1, i2, ..., ik}, based on their corresponding similarities {si1, si2, ..., sik}. The predictions can then be computed by taking a weighted average of the active user's ratings on these similar items. The first step in this new approach is the Representation. Its purpose is the same as with the classic Collaborative Filtering algorithm: represent the data in an organized manner.

The Item Similarity Computation should be calculated. The basic idea in that step is to first isolate the users who have rated two items ij and ik and then apply a similarity computation technique to determine their similarity. Various ways to compute that similarity have been proposed.

D. Content-Boosted Collaborative Filtering

Emmanouil G. Vozalis et. sl., estimated in [31] the basic idea behind Content-Boosted Collaborative Filtering is to use a content-based predictor to enhance existing user data, expressed via the user-item matrix, R, and then provide personalized suggestions through collaborative filtering. The content-based predictor is applied on each row from the initial user-item matrix, corresponding to each separate user, and gradually generates a pseudo user-item matrix, PR. At the end, each row, i, of the pseudo user-item matrix PR consists of the ratings provided by user u_i, when available, and those ratings predicted by the content-based predictor.

Memory-based filtering algorithms include the basic Collaborative Filtering algorithm [30], Item-based Collaborative Filtering [28] and the Algorithm using SVD/LSI for Prediction Generation [31]. Correlation-based vs. Machine Learning based algorithms Billsus and Pazzani attempt [24], through their work described in [32], to transform the formulation of the recommendation problem, as viewed by the classic Collaborative Filtering algorithm, into a Machine Learning problem, where any supervised learning algorithm can be drawn and applied. They are based on the assumption that while correlation based approaches seems to work well in the specific domain.

E. The Weighted Combination of Content-based and Collaborative

Filtering defines two distinct filtering components. The first component implements plain Collaborative Filtering, while the second component implements Content based Filtering. The final rating prediction is calculated as a weighted sum of those components, where the applied weights are decided by how close is the prediction of each component to the actual rating.

V. PRESENT SCENARIO OF RESEARCH IN RECOMMENDATION SYSTEM

Recommender systems have been evaluated in many, often incomparable, ways. In the present scenario the user tasks being evaluated, the types of analysis and datasets being used, the ways in which prediction quality is measured, the evaluation of prediction attributes other than quality, and the user-based evaluation of the system as a whole. In addition to reviewing the evaluation strategies used by prior researchers, we present empirical results from the analysis of various accuracy metrics on one content domain where all the tested metrics collapsed roughly into three equivalence classes. Metrics within each equivalency class were strongly correlated, while metrics from different equivalency classes were uncorrelated.

Dimensions for User Evaluation

Explicit (ask) vs. implicit (observe) A basic distinction is between evaluations that explicitly ask users about their reactions to a system and those that implicitly observe user behavior. The first type of evaluation typically employs survey and interview methods. The second type usually consists of logging user behavior, then subjecting it to various sorts of analyses.

VI. CHALLENGING PROBLEMS IN RECOMMENDATION SYSTEM

The several current challenges of the recommender systems are considered in this section. The first set of challenges concerns issues of bringing people together into communities of interest. A major concern here is respecting people's privacy. The second challenge is to create recommendation algorithms that combine multiple types of information, probably acquired from different sources at different times.

Establishing the user tasks to be supported by a system, and selecting a data set on which performance enables empirical experimentation – scientifically repeatable evaluations of recommender system utility. A majority of the published empirical evaluations of recommender systems to date has focused on the evaluation of a recommender system's accuracy. We assume that if a user could examine all items available, they could place those items in a ordering of preference. Accuracy metric empirically measures how close a recommender system's predicted ranking of items for a user differs from the user's true ranking of preference. Accuracy measures may also measure how well a system can predict an exact rating value for a specific item. Researchers who want to quantitatively compare the accuracy of different recommender systems must first select one or more metrics. In selecting a metric, researchers face a range of questions. Will a given metric measure the effectiveness of a system with respect to the user tasks for which it was designed? Are results with the chosen metric comparable to other published research work in the field? Are the assumptions that a metric is based on true? Will a metric be sensitive enough to detect real differences that exist? How large a difference does there have to be in the value of a metric for a statistically significant difference to exist? Complete answers to these questions have not yet been substantially addressed in the published literature.

The challenge of selecting an appropriate metric is compounded by the large diversity of published metrics that have been used to quantitatively evaluate the accuracy of recommender systems. This lack of standardization is damaging to the progress of knowledge related to collaborative filtering recommender systems. With no standardized metrics within the field, researchers have continued to introduce new metrics when they evaluate their systems. With a large diversity of evaluation metrics in use, it becomes difficult to compare results from one publication to the results in another publication. As a result, it becomes hard to integrate these diverse publications into a coherent body of knowledge regarding the quality of recommender system algorithms.

VII. Future

Recommender systems are a powerful new technology for extracting additional value for a business from its customer databases. The systems help customers find products they want to buy from a business. Recommender systems benefit customers by enabling them to find products they like. Conversely, they help the business by generating more sales. Recommender systems are rapidly becoming a crucial tool in E-commerce on the Web. Recommender systems are being stressed by the huge volume of customer data in existing corporate databases, and will be stressed even more by the increasing volume of customer data available on the Web. New technologies are needed that can dramatically improve the scalability of recommender systems.

Web recommendation system is seen as a fully automated process, powered by operational Knowledge. A number of systems following many approaches have been developed, using methods and techniques from Web recommendation system. In addition to the functions employed by existing systems, many other interesting ones have been neglected so far. The combination of recommendation and customization functionality has been seen as the main solution to the information overload problem and the creation of loyal relations between the Web site and its visitors. However, other functions such as task performance support and user tutoring can certainly improve the experience of a Web site visitor. It should be noted at this point, that Web recommendation is a very active research field and new approaches related to its application appear on a regular basis. As a result, there are a number of unsolved technical problems and open issues. Some of these have been presented in this survey. New techniques and possibly new models for acquiring data are needed. One serious issue concerning data collection is the protection of the user's privacy. A poll by KDnuggets (15/3/2000 to 30/3/2000) revealed that about 70% of the users consider Web recommendation as a compromise of their privacy. Thus, it is imperative that new tools are transparent to the user, by providing access to the data collected and clarifying the use of these data, as well as the potential benefits for the user. At the same time, one should be very careful not to burden the user with long-winded form-filling procedures, as these discourage users from accessing a Web site. Even the simple process of user registration is unacceptable for some Web-based services.

In addition to the various improvements to the Web recommendation system process, there are a number of other issues, which need to be addressed in order to develop effective Web personalization systems. From the open issues that were mentioned in this survey, the treatment of time in the user models can be distinguished as being particularly difficult. The main source of difficulty is that the manner in which the behavior of users changes over time varies significantly with the application and possibly the type of the user. Therefore, any solution to this problem should be sufficiently parametric to cater for the requirements of different applications. It is therefore evident that the integration of Web recommendation system using apriori algorithm has introduced a number of methodological and technical issues, some of which are still open. At the same time the potential of this synergy between the two processes has barely been realized. As a result, a number of interesting directions remain unexplored. This survey has identified promising directions, providing at the same time a vehicle for exploration, in terms of Web recommendation system tools and methods.

VIII. Conclusions

Web using recommendation system is an emerging technology that can help in producing personalized Web-based systems. This article provides a survey of the work in recommendation system, focusing on its application and future. The survey aims to serve as a source of ideas for people working on the recommendation of information systems, particularly those systems that are accessible over the Web. Since the current web is largely unorganized and there is a rapid growth of information volumes, the recommendation system whose major purpose is to reduce irrelevant content and to provide users with more pertinent and tailored information becomes an important research area. A key issue in this area is how to discover user's interest and behavior effectively.

The selection of the Apriori algorithm for performing Web recommendation system is because, Apriori algorithm is a common recommendation technique for association based analysis. By applying this algorithm to the user systems, the relationship between the accessed pages and visitors can be efficiently maintained. The Web usage patterns and user behavior also can analyze by using this algorithm where the descriptive statistic approach cannot perform this analysis. The results and findings for this analysis are more reliable but less of accuracy because of the Apriori algorithm properties where the same selected item sets are always counted. The results or findings from this experimental analysis are surely useful for Web administrator in order to improve Web services and performance through the improvement of Web sites, including their contents, structure, presentation, and delivery.

We hope that the framework and survey presented in this paper will lead to research that is more systematic on recommendation system using apriori algorithm.

IX. References

1) Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. Morgan-Kaufmann Academic Press, San Francisco.
2) R. Agrawal; T. Imielinski; A. Swami: Mining Association Rules Between Sets of Items in Large Databases", SIGMOD Conference 1993: 207-216
3) Peetr Hajek, Tomas Feglar, Jan Rauch, David Coufal. The GUHA method, data preprocessing bvand mining. Database Support for Data Mining Applications, ISBN 978-3-540-22479-2, Springer, 2004
4) Hajek P., Havel I., Chytil M.: The GUHA method of automatic hypotheses determination, Computing 1(1966) 293-308.
5) Piatetsky-Shapiro, G. (1991), Discovery, analysis, and presentation of strong rules, in G. Piatetsky-Shapiro & W. J. Frawley, eds, "Knowledge Discovery in Databases", AAAI/MIT Press, Cambridge, MA.
6) "Web Site Personalisation". IBM High-Volume Web site team, January 2000,http://www-4.ibm.com/software/developer/library/personalization/index.html.
7) Bilgic and R. J. Mooney. Explaining recommendations: Satisfaction vs. promotion. In Beyond Personalization Workshop, IUI, 2005.
8) J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In ACM conference on Computer supported cooperative work, 2000.
9) Mcsherry. Explanation in recommender systems. Artificial Intelligence Review, 24(2):179 – 197, 2005.
10) Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions on the web. In ACM CHI 2001 Conference on Human Factors in Computing Systems, pages 490-497, Seattle WA, 2001.
11) Billsus and M. Pazzani. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling (UM '99), Banff, Canada, 1999.
12) Jennings and H. Higuchi. A user model neural network for a personal news service. User Modeling and User-Adapted Interaction, 3(1):1-25, 1993.
13) R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th Int'l Conference on Very Large Databases (VLDB'94), Santiago, Chile, Sep 1994.
14) John and P. Langley. Estimating continuous distributions in bayesian classifiers. In Eleventh Annual Conference on Uncertainty in Artificial Intelligence, pages 338-345, San Francisco, 1995. Morgan Kaufmann Publishers. .
15) McCarthy, J. Reilly, L. McGinty, and B. Smyth. Thinking positively - explanatory feedback for conversational recommender systems. In Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop,, 2004.
16) R. Rafter and B. Smyth. Conversational collaborative recommendation-an experimental analysis. *Artif. Intel. Rev*, 24(3-4):301–318, 2005.
17) Badrul M. Sarwar et al., on "Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering" Department of Computer Science and Engineering University of Minnesota, Minneapolis, MN 55455, USA
18) Resnick, P., N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In Proceedings of the 1994 Computer Supported Cooperative Work Conference, 1994.
19) Shardanand, U. and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In Proc. of the Conf. on Human Factors in Computing Systems, 1995.
20) Hill, W., L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In Proceedings of CHI'95.
21) Hanani U, Shapira B and Shoval P. Source: User Modeling and User-Adapted Interaction, Volume 11, Number 3, August 2001 , pp. 203-259(57)

22) Oard, D., Kim, J. Implicit Feedback for Recommender Systems. In Proceedings of AAAI Workshops on Recommender Systems. July 1998.
23) Claypool, M., Le, P., Waseda, M., Brown, D. Implicit Interest Indicator. Computer Science Technical Report Series, WPI Computer Science Department, Worcester, Massachussets, July 2000.
24) Billsus, D., and Pazzani, M. J. (1998). Learning Collaborative Information Filters. In Proceedings of ICML '98. pp. 46-53.
25) Swearingen, k., sinha, R. 2001. Beyond Algorithms: An HCI Perspective on Recommender Systems. In *SIGIR 2001* Workshop on Recommender Systems.
26) Cosley, d., lam, s. K., Albert, i., konstan, j. A. and riedl, j. 2003. Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. *CHI Letters 5*. TURPIN, A., HERSH, W. 2001. Why Batch and User Evaluations Do Not Give the Same Results. In Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, 17-24
27) NEWMAN, W., "Better or just different? On the benefits of designing interactive systems in terms of critical parameters. In Designing Interactive Systems" (DIS97*),* ACM Press, New York, 2001 NY. 239-246.
28) Bardul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl, "Item-based collaborative filtering recommendation algorithms," in 10th International World Wide Web Conference (WWW10), Hong Kong, 2001
29) George Karypis, "Evaluation of item-based top-n recommendation algorithms," in CIKM 2001, 2001.
30) Bardul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl, "Analysis of recommendation algorithms for ecommerce," in Electronic Commerce, 2000.
31) Bardul M. Sarwar, "Sparsity, Scalability, and Distribution in Recommender Systems", Ph.D. thesis, University of Minnesota, 2001.
32) Emmanouil G. Vozalis, Konstantinos G. Margaritis, "Recommender Systems: An Experimental Comparison of two Filtering Algorithms**",** 2003, In Proceedings of the Ninth Panhellenic Conference in Informatics .

Fast Association Rule Mining Algorithm for Spatial Gene Expression Data

M.Anandhavalli [#1], M.K.Ghose[#2], K.Gauthaman[*3]

GJCST Classifications: H.2.8, B.6.m, G.1.3

Abstract- **One of the important problems in data mining is discovering association rules from spatial gene expression data where each transaction consists of a set of genes and probe patterns. The most time consuming operation in this association rule discovery process is the computation of the frequency of the occurrences of interesting subset of genes (called candidates) in the database of spatial gene expression data. A fast algorithm has been proposed for generating frequent itemsets without generating candidate itemsets along with strong association rules. The proposed algorithm uses Boolean vector with relational AND operation to discover frequent itemsets. Experimental results shows that combining Boolean Vector and relational AND operation results in quickly discovering of frequent itemsets and association rules as compared to general Apriori algorithm .**

Keywords-Spatial Gene expression data, Association Rule, Frequent itemsets, Boolean vector, relational AND operation, Similarity Matrix.

I. INTRODUCTION

There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as spatial gene expression database. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data might have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Analysis of these massive genomic data has two important goals:

I) To try to determine how the expression of any particular gene might affect the expression of other genes

II) To try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells?

The most popular pattern discovery method in data mining is association rule mining. Association rule mining was introduced by [4]. It aims to extract interesting correlations, frequent patterns, associations, or casual structures among sets of items in transaction databases or other data repositories. The relationships are not based on inherent properties of the data themselves but rather based on the co-occurrence of the items within the database.

The associations between items are commonly expressed in the form of association rules. In this setting, attributes which represents items are assumed to have only two attributes and thus referred as Boolean attributes. If an item is contained in a transaction, the corresponding attribute value will be 1; otherwise the value will be 0. Many interesting and efficient algorithms have been proposed for mining association rules for these Boolean attributes, for examples, Apriori [3], DHP [6], and partition algorithms [7]. Currently most association mining algorithms are dedicated to frequent itemsets mining. These algorithms are defined in such a way that they only find rules with high support and high confidence. A characteristic of frequent itemsets mining is that it relies on there being a meaningful minimum support level that is sufficiently high to reduce the number of frequent itemsets to a manageable level. A huge calculation and a complicated transaction process are required during the frequent itemsets generation procedure. Therefore, the mining efficiency of the Apriori-like algorithms is very unsatisfactory when transaction database is very large particularly spatial gene expression database.

In this paper, an attempt has been made to propose a novel algorithm for mining association rule from spatial gene expression data.

II. MATERIALS AND METHODS

A. *Spatial Gene Expression Data*

The Edinburgh Mouse Atlas gene expression database (EMAGE) is being developed as part of the Mouse Gene Expression Information Resource (MGEIR) [1] in collaboration with the Jackson Laboratory, USA. EMAGE (http://genex.hgu. mrc.ac.uk/Emage/database) is a freely available, curated database of gene expression patterns generated by in situ techniques in the developing mouse embryo. The spatial gene expression data are presented as N×N similarity matrix. Each element in the matrix is a measure of similarity between the corresponding probe pattern and gene-expression region. The similarity is calculated as a fraction of overlap between the two and the total of both areas of the images. This measurement is intuitive, and commonly referred to as the Jaccard index [2]. When a pattern is compared to itself, the Jaccard value is 1 because the two input spatial regions are identical. When it is compared to another pattern, the Jaccard Index will be less than one. If the Jaccard Index is 0, the two patterns do not intersect. If a Jaccard Index value is close to 1, then the two patterns are more similar.

However, biologists are more interested in how gene expression changes under different probe patterns. Thus, these similarity values are discretized such that similarity

Manuscript received "21Sep.2009"

#M.Anandhavalli, Dr. M.K.Ghose, Department of Computer Science and Engineering, Sikkim Manipal Institute of Technology, Majitar, East Sikkim, India - 737136. e-mail: anandhigautham@gmail.com.

*Dr. K.Gauthaman Department of Pharmacognosy, Himalayan Pharmacy Institute, Majitar, East Sikkim-737136, India.

measure greater than some predetermined thresholds and converted into Boolean matrix.

B. Data Preprocessing

Preprocessing is often required before applying any data mining algorithms to improve performance of the results. The preprocessing procedures are used to scale the data value either 0 or 1. The values contained in the spatial gene expression matrix had to be transformed into Boolean values by a so-called discretization phase. In our context, each quantitative value has given rise to the effect of four different discretization procedures [2]: Max minus x% method, Mid-range-based cutoff method, x% cut off and x% of highest value method.

Max minus x% procedure consists of identifying the highest expression value (HV) in the data matrix, and defining a value of 1 for the expression of the gene in the given data when the expression value was above HV – x% of HV where x is an integer value. Otherwise, the expression of the gene was assigned a value of 0 (Figure 1a).

Mid-range-based cutoff (Figure 1b) identifies the highest and lowest expression values in the data matrix and the mid-range value is defined as being equidistant from these two numbers (their arithmetic mean). Then, all expression values below or equal to the mid-range were set to 0, and all values strictly above the mid-range were set to 1.

x% of highest value approach (Figure 1c) identifies data in which its level of expression is in the 5% of highest values. These are assigned the value 1, and the rest were set to 0.

Value greater than x% approach (Figure 1d) identifies the level of expression and assigns the value 1 when it is greater than given percentage and the rest are set to 0.

From these four different procedures resulted in different matrix densities, the first and last procedure resulted in the same number of Boolean 1 results for all gene expressions, whereas the second and fourth procedure generated same densities of 1, depending on the gene expression pattern throughout the various data matrix. From the similarity matrix, two different sets of transactions are constructed, which in turn lead to two different types of association rules.

I) The items I are genes from the data set, where a transaction T ⊆ I consists of genes that all have an expression pattern intersecting with the same probe pattern.

II) The items I are the probe patterns, where a transaction T ⊆ I consists of probe patterns all intersecting with the expression patterns in the same image.

To create the first type of transactions, we take for each probe pattern r, every gene g from which its associated gene expression pattern ge satisfies the minimum similarity β, i.e., similarity(r, ge) > β, to form the itemsets.

The second type of transactions is created in a similar way. For each gene expression pattern g in the database we create an itemsets that consists of a set of probe patterns that intersect with the gene expression pattern ge. Each probe pattern r must satisfy the minimum similarity β, i.e., similarity(r, ge) > β, to get included in the itemsets.

	α (Input)	α (after discretization)
a	0.096595	0
b	0.123447	0
c	0.291310	1
d	0.126024	0
e	0.155819	0
f	0.288394	1
g	0.000000	0
h	0.215049	1

Fig.1a Results of Max minus 25% method

	α	α (after discretization)
a	0.096595	0
b	0.123447	0
c	0.291310	1
d	0.126024	0
e	0.155819	1
f	0.288394	1
g	0.000000	0
h	0.215049	1

Fig.1b. Results of Mid-range-based cutoff

	α	α (after discretization)
a	0.096595	0
b	0.123447	0
c	0.291310	1
d	0.126024	0
e	0.155819	1
f	0.288394	1
g	0.000000	0
h	0.215049	1

Fig.1c. Results of x% of highest value approach

	α	α (after discretization)
a	0.096595	0
b	0.123447	0
c	0.291310	1
d	0.126024	0
e	0.155819	0
f	0.288394	1
g	0.000000	0
h	0.215049	1

Fig.1c. Results of x% of highest value approach.

Fig.1. Schematic description of the discretization protocols used.

C. Association Rule Mining

The Apriori-like algorithms adopt an iterative method to discover frequent itemsets. The process of discovering frequent itemsets need multiple passes over the data. The algorithm starts from frequent 1-itemsets until all maximum frequent itemsets are discovered. The Apriori-like algorithms consist of two major procedures: the join procedure and the prune procedure. The join procedure combines two frequent k-itemsets, which have the same (k-1)-prefix, to generate a (k+1)-itemset as a new preliminary candidate. Following the join procedure, the prune procedure is used to remove from the preliminary candidate set all itemsets whose k-subset is not a frequent itemsets [3]. From every frequent itemset of k>=2, two subsets A and C, are constructed in such a way that one subset C, contains exactly one item in it and remaining k-1 items will go to the other subset A. By the downward closure properties of the frequent itemsets these two subsets are also frequent and their support is already calculated. Now these two subsets may generate a rule A →C, if the confidence of the rule is greater than or equal to the specified minimum confidence.

D. *Algorithm Details*

I) Let I={i1, i2, …, in} be a set of items, where each item ij corresponds to a value of an attribute and is a member of some attribute domain Dh={d1, d2, …, ds}, i.e. ij Є Dh. If I is a binary attribute, then the Dom (I)={0,1}. A transaction database is a database containing transactions in the form of (d, E), where d ∈ Dom(D) and E ⊆ I.

II) Let D be a transaction database, n be the number of transactions in D, and minsup be the minimum support of D. The new_support is defined as new_support = minsup × n.

III) Proposition 1: By Boolean vector with AND operation, if the sum of „1‟ in a row vector Bi is smaller than k, it is not necessary for Bi to involve in the calculation of the k- supports.

IV) Proposition 2: According to [5], Suppose Itemsets X is a k-itemsets; |F_{K-1}(j)| presents the number of items „j‟ in the frequent set F_{K-1}. There is an item j in X. If | F_{K-1}(j)| is smaller than k-1, itemset X is not a frequent itemsets.

V) Proposition 3: |F_K| presents the number of k-itemsets in the frequent set F_K. If |F_K| is smaller than k+1, the maximum length frequent itemsets is k.

The proposed algorithm for finding the association rules in terms of spatial gene expression data in the form of similarity matrix consists of five phases as follows.

1. Transforming the similarity matrix into the Boolean matrix
2. Generating the set of frequent 1-itemsets F1
3. Pruning the Boolean matrix
4. Generating the set of frequent k-itemsets Fk(k>1)
5. Generating association rules from the generated frequent itemsets with confidence value greater than a predefined threshold (minconfidence).

A detailed description of the proposed algorithm is described as follows:

Input: Spatial Gene Expression data in similarity matrix (M), the minimum support, and minimum confidence.

Output: Set of frequent itemsets F and Association rules.

1. Normalize the data matrix M and transformed into Boolean Matrix B;
   // Frequent 1-itemset generation
2. For each column Ci of B
3. If sum(Ci) >= new_support
4. F1 = { Ii};
5. Else delete Ci from B;
   // By Proposition 1
6. For each row Rj of B
7. If sum(Rj) < 2
8. Delete Rj from B;
   // By Proposition 2 and 3
9. For (k=2; | Fk-1| > k-1; k++)
10. {
    // Join procedure
11. Produce k-vectors combination for all columns of B;
12. For each k-vectors combination { Bi1, Bi2,…Bik}
13. { E= Bi1 ∩ Bi2 ∩.…∩Bik
14. If sum(E) >= new_support
15. Fk = { Ii1, Ii2,…Iik}
16. }
    // Prune procedure
17. For each item Ii in Fk
18. If |Fk(Ii)| < k
19. Delete the column Bi according to item Ii from B;
20. For each row Rj of B
21. If sum(Bj) < k+1
22. Delete Bj from B;
23. k=k+1
24. }
25. Return F = F1UF2….UFk
26. For all Fk k >= 2 do
27. For all i <= k do
28. c=Fk[i]
29. a = Fk – c
30. if((new_support(Fk)/ new_support(a) >= minconfidence
31. declare a → c is a rule
32. enddo
33. enddo

III. RESULTS AND DISCUSSION

The proposed algorithm was implemented in Java and tested on Linux platform. Comprehensive experiments on spatial gene expression data has been conducted to study the impact of normalization and to compare the effect of proposed algorithm with Apriori algorithm. Figure 2 and 3 gives the experimental results for execution time (generating frequent itemsets and finding rules) vs. user specified minimum supports and shows that response time of the proposed algorithm is much better than that of the Apriori algorithm. In this case, confidence value is set 100% for the rule generation, which means that all the rules generated are true in 100% of the cases.

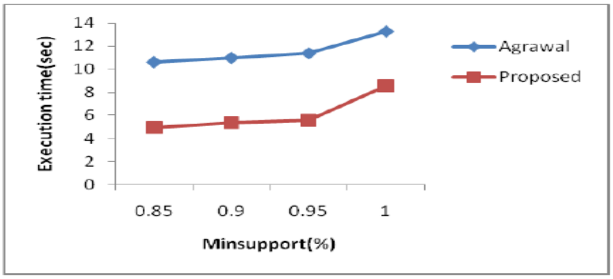


Fig.2. Performance on Stage 14 of EMAGE Spatial Gene expression data (Minsupport vs. Execution time)

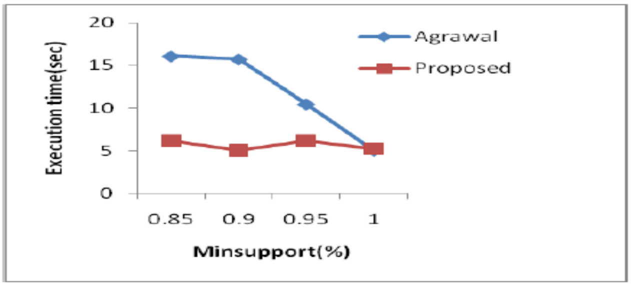


Fig.3. Performance on Stage 17 o EMAGE Spatial Gene expression data (Minsupport vs. Execution time)

Figure 4 and 5 gives the experimental results for memory usage vs. user specified minimum supports and results show that proposed algorithm uses less memory than that of Apriori algorithm because of the Boolean and relational AND bit operations.

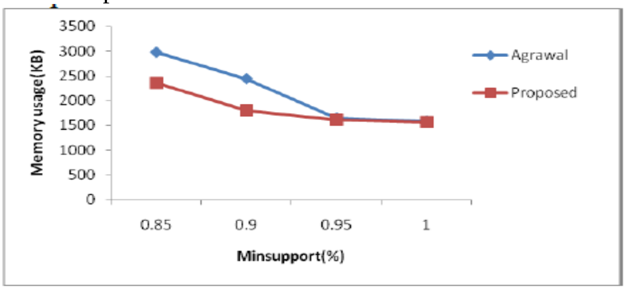


Fig.4. Performance on Stage 14 of EMAGE Spatial Gene expression data (Minsupport vs. Memory usage)

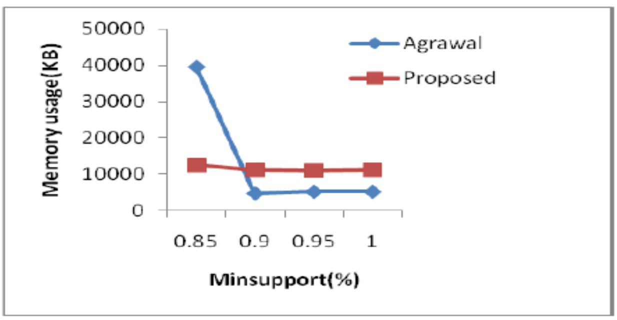


Fig.5. Performance on Stage 17 of EMAGE Spatial Gene expression data (Minsupport vs. Memory usage)

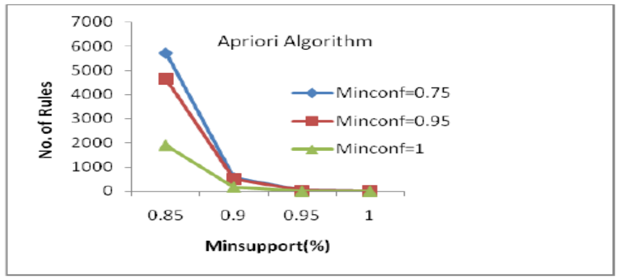


Fig.6. Association rules and Minimum support in Apriori algorithm

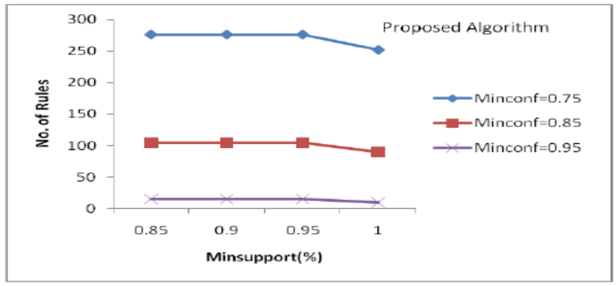


Fig.7. Association rules and Minimum suppport in Proposed algorithm

The number of association rules decreases along with an increase in minimum support (or minimum confidence) under a given specific minimum confidence, which shows an appropriate Minsupport (or Minconf) can constraint the number of association rules and avoid the occurrence of some association rules so that it cannot yield a decision. These results have shown in Figures 6-7. The results are as expected and quite consistent with our intuition.

IV. Conclusion

In this paper, a novel method of mining frequent itemsets and strong association rules from the spatial gene expression data is proposed to generate frequently occur genes very quickly. The proposed algorithm does not produce candidate itemsets, it spends less time for calculating k-supports of the itemsets with the Boolean matrix pruned, and it scans the database only once and needs less memory space when compared with Apriori algorithm. Finally, the large and rapidly increasing compendium of data demands data mining approaches, particularly association rule mining ensures that genomic data mining will continue to be a necessary and highly productive field for the foreseeable future.

V. Acknowledgment

This study has been carried out as part of Research romotion Scheme (RPS) Project under AICTE, Govt. of India.

VI. References

1) Baldock,R.A., Bard,J.B., Burger,A., Burton,N., hristiansen,J., Feng,G.,Hill,B., Houghton,D., Kaufman,M., Rao,J. et al. (2003) EMAP and EMAGE: a framework for understanding spatially organized data, Neuroinformatics, 1, 309–325.
2) Pang-Ning Tan, Micahel Steinbach, Vipin Kumare,.: Intoduction to Data Mining Pearson Education , second edition, pp.74, 2008.
3) Agrawal, R. & Srikant, R. (1994) Fast Algorithms for Mining Association Rules in large databases. In Proceedings of the 20th International Conference on Very Large Databases pp. 487-499. Santiago, Chile.
4) Agrawal, R., Imielinski, T., & Swami, A. (1993) Mining association rules between sets of items in large databases. Proceedings of the ACM SICMOD conference on management of data pp. 207-216. Washington, D.C.
5) Xu, Z. & Zhang, S. (2003) An Optimization Algorithm Base on Apriori for Association Rules. Computer Engineering 29(19), 83-84.
6) J S. Park and M -S. Chen and PS. Yu, “An effective hash-based algorithm for mining association rules”, in proceedings of the ACM SIGMOD International Conference on Management of Data", San Jose, CA, May 1995, pp 175-186.
7) A Savasere, E. Ommcinskl and S Navathe, “An efficient algorithm for mining association rules in large databases”, In Proceedings Of the 21st International Conference on Very Large Databases, Zurich, Switzerland, September 1995, pp 432-444.

Fair Load Balancing in Wireless Networks

Mr.S.Venkatesan[1] Prof.Dr.C.Manoharan[2]
[1]Research Scholar, Anna University, Coimbatore,Tamil Nadu, India
venkats_be@yahoo.co.in
[2] Principal, VSA Educational Institutions, Salem,Tamil Nadu, India
c_m_66@yahoo.co.in

GJCST Classifications: C.2.4, C.2.1

***Abstract*-The traffic load of wireless LANs is often unevenly distributed among the access points (APs), which results in unfair bandwidth allocation among Mobile Users. We argue that the load imbalance and consequent unfair bandwidth allocation can be greatly reduced by intelligent association control. In this paper, we present an efficient solution to determine the user-AP associations for bandwidth allocation. We show the strong correlation between fairness and load balancing, which enables us to use load-balancing techniques for obtaining optimal fair bandwidth allocation. As this problem is NP-hard, we devise algorithms that achieve constant factor approximation. In our algorithms, we first compute a distributed association solution, in which users can be associated with multiple APs simultaneously with variable bandwidth. This solution guarantees the fairest bandwidth allocation in terms of Max-min fairness; we obtain the integral solution from the fractional solution by distributed association algorithm. We also consider time fairness and present a polynomial-time algorithm for optimal integral solution and it is ensure that zero percent data loss.**

Keywords- Distributed Association algorithms, IEEE 802.11 WLANs, load balancing.

I. INTRODUCTION

Recent studies on operational Wireless LANS (WLANs) have shown that the traffic load is often distributed unevenly among the access points (APs) for Mobile Users (MU). In WLANs, by default, each user scans all available channels to detect its nearby APs and associate itself with the AP that has the strongest received signal strength indicator (RSSI), while ignoring its load condition. As users are, typically, not uniformly distributed, some APs tend to suffer from heavy load while adjacent APs may carry only light load or be idle. Such load imbalance among APs is undesirable as it hampers the network from providing fair services to its users. As suggested in existing studies the load imbalance problem can be alleviated by balancing the load among the APs via intelligently selecting the user-AP association, termed association control. Association control can be used to achieve different objectives. For instance, it can be used to maximize the overall system throughput by shifting users to idle or lightly loaded APs and allowing each AP to serve only the users with maximal data rate. Clearly, this objective is not a desired system behavior from the fairness viewpoint. A more desirable goal is to provide network-wide fair bandwidth allocation, while maximizing the minimal fair share of each user. This type of fairness is known as maxmin fairness. Informally, a bandwidth allocation is max-min fair if there is no way to give more bandwidth to any user without decreasing the allocation of a user with less or equal bandwidth. In this paper, we present efficient user-AP association control algorithms that ensure maxmin fair bandwidth allocation and we show that this goal can be obtained by balancing the load on the APs.

II. REVIEW OF LITERATURE

Load balancing in WLANs has been intensely studied. In [1], association algorithm has been proposed for efficient bandwidth allocation with constant bandwidth. [3]- [4] on operational Wireless LANS (WLANs) have shown that the traffic load is often distributed unevenly among the access points (APs). In WLANs, by default, each user scans all available channels to detect its nearby APs and associate itself with the AP that has the strongest received signal strength indicator (RSSI), while ignoring its load condition. As users are, typically, not uniformly distributed, some APs tend to suffer from heavy load while adjacent APs may carry only light load or be idle. Such load imbalance among APs is undesirable as it hampers the network from providing fair services to its users. As suggested in existing studies [6]-[7] the load imbalance problem can be alleviated by balancing the load among the APs via intelligently selecting the user- AP association, termed association control. Association control can be used to achieve different objectives. In [7]-[9], different association criteria are proposed. These metrics typically take into account factors such as the number of users currently associated with an AP, the mean RSSI, the RSSI of the new user and the bandwidth a new user can get if it is associated with an AP in [8]. Various WLAN vendors have incorporated proprietary features in the device driver's firmware [10], [11]. In these proprietary solutions, the APs broadcast their load conditions to the users via the Beacon messages and each user chooses the least loaded AP. Propose to associate new users with the AP that can provide a minimal bandwidth required by the user. If there is more than one such AP, the one with the strongest signal is selected. Most of these heuristics only determine the association of newly arrived users. Tsai and Lien [8] propose to reassociate users when some conditions are violated. Load balancing in cellular networks is usually achieved via dynamic channel allocation (DCA) [12].

III. WIRELESS AND WIRED BOTTLENECKS

However, the wireless link is generally considered as the bottleneck. This assumption is not always valid. For instance, consider a WLAN where the APs are connected to the infrastructure

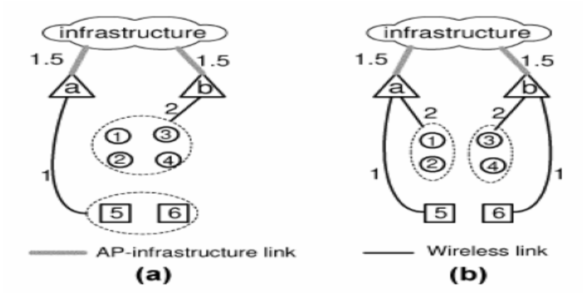


Fig. 1. Examples of bottlenecks both over the wireless and the wired links. (a) An unfair association. (b) The optimal association.

T1 lines, whose capacity is around 1.5 Mb/s, as illustrated in Example 2. Example 2 demonstrates the need to consider both the wireless and the wired links for load balancing.

Example 2: Consider a wireless system with 2 APs, a and b , and 6 users, enumerated from 1 to 6, as depicted in Fig. 1. Users 1,2,3 and 4 experience a bit rate of 2 Mb/s from both APs,[A] while users 5 and 6 have a bit rate of 1 Mb/s from both APs. The[B] APs are connected to a fixed network with T1 lines with capacity[C] of 1.5 Mb/s. In the following, we consider two possible associations and we analyze the average bandwidth that they provide to the users.

Case I: A fair user association only from the wireless perspective- Consider the association depicted in Fig. 1(a). Here, the system can allocate a bandwidth of 0.5 Mb/s to each user over the wireless links. However, while AP a can allocate a bandwidth of 0.5 Mb/s to users 5 and 6 on its T1 line, AP b can only provide 3/8 Mb/s to its associated users over its line. In this case, the wireless link of AP is the bottleneck that affects the bandwidth allocation. Meanwhile, the wired link is the bottleneck of AP.

Case II: A fair user association- Consider the association shown in Fig. 1(b). This association provides a bandwidth of 0.5 Mb/s to each user over the wired and wireless channels. Observe that in this case different users may gain different service time on the wireless links and wired backhauls. For instance, user 5 captures 1/3 of the service time of the T1 link of AP, while, it is served 1/2 of the time by its wireless channel. This ensures that user 5, indeed, receives a bandwidth of 0.5 Mb/s.

IV. Fairness and Load Balancing

In this section, we provide formal definitions of fair bandwidth allocation and load balancing. Additionally, they describe some useful properties that we need for constructing our algorithmic tools. In the following, we consider two association models from this. The first is a single-association model, so-called an integral- association, where each user is associated with a single AP at any given time. This is the association mode used in IEEE 802.11 networks. The second is a multiple-association model, also termed a fractional-association that allows each user to be associated with several APs and to get communication services from them simultaneously. Accordingly, a user may receive several different traffic flows from different APs, and its bandwidth allocation is the aggregated bandwidth of all of them. This model is used to develop our algorithmic tools for the integral- association case. For both association models, we denote by Ua all the users that are associated with AP a ЄA and denotes the set of APs that user u Є U is associated with.

V. Distributed Association Algorithm

In this section, after exploring the details of distributed AP selection algorithm for APs and MUs, we also analyze the stability and overhead of the proposed algorithm.

A. *Association Algorithm for APs and MUs*

By exchanging information among MUs and APs, the proposed association scheme can be summarized as Algo.1 as shown in Fig. 2. In legacy IEEE 802.11 standard, the management packets from the AP do not contain any field indicating the AP load information. To realize the proposed scheme, it is required to add one additional field to the beacon and probing packets. Moreover, due to the dynamic nature of the wireless network and the mobility of MUs, the APs should keep updating the AP load by iterative moving average as

$$y_a(t+T_\Omega) = \alpha y_a(t) + (1-\alpha)\sum_{u\in U_a(t)} d_{ua}(t)$$

where TΩ is the fixed updating interval and 0≤α≤1 is the weighting parameter to tradeoff previously estimated AP load and current value. If a MU is not associated with any AP in the network, it immediately scans all channels by sending probe request messages and receives response packets from the available APs. By detecting the respective RSSI levels to the APs, each MU can determine the most suitable physical data rate for transmitting packets. The proposed AP selection strategy is to let each MU choose the AP with least estimated load by supposing that it will be associated with all available APs. That is, if the newly joining MU u can be served by a subset of APs Au ∈ A, the estimated AP load on a ∈ Au supposing the association of MU u with AP a will be updated as

Algorithm 1 Association algorithm for each AP and MU.

Periodical operation on each AP a **with interval** T_Ω**.**
1. Periodically update its AP load by Eq. (2).

Periodical operation on each MU u **with interval** T_Δ**.**
1. Exchange the probing packets with AP.
2. Calculate the estimated AP load by Eq. (3).
3. **if** u is a newly MU joining the WLAN **then**
4. The MU u selects the AP as $argmin_{a\in A_u}\tilde{y}_a(t)$.
5. **else** /* *u is already associated with AP a* */
6. **if** switching to a' lead to $y_a(t) - y_{a'}(t) > \delta$ **then**
7. MU u switches the association to a'.
8. **end if**
9. **end if**

Fig. 2. The distributed algorithm for load balancing in WLANs.

$$\tilde{y}_a(t) = y_a(t) + d_{ua}(t) \quad (\forall a \in A_u).$$

Then the MU will select an AP as argmina∈Auỹa(t). After the MU joins the WLAN, it will keep periodically (with period TΔ) detecting the load information from the neighboring APs and change its association if the AP loads can be further decreased. This operation is not only necessary to reduce the effect introduced by the joining order of MUs but also required for the MU to be adaptive to the dynamic wireless environment and topology changes. The period TΔ, configured to be more than 10 seconds, is much longer than the load-updating period TΩ on the AP.

B. Association Algorithm for APs and MUs

In dynamic WLANs, the association of MUs should vary with the network conditions. However, it is not intuitively obvious that the proposed distributed algorithm is self-stabilizing for static networks. That is, MUs continually looking to balance the AP loads will eventually converge to a stable result in static topology. Here we can show that indeed this process does stabilize.

Theorem 1: For a fixed population WLAN with APs and static MUs that implement the above distributed association algorithm with δ = 0, the switching operations of the MUs in Algo. 1 reaches a stable state where MUs cease changing associated APs2.

Proof: The core part of the proof is that a monotonic property of global lexicographic ordering [15] decrement holds whenever one MU switches its association. Lexicographic order, a concept borrowed from economics, can be used to compare the extent of fairness between two vectors. Given two vectors A and B; the method to determine the lexicographic order is to compare the corresponding values index by index after sorting the original vectors. According to Algorithm 1, assuming one MU switch from AP a to AP b, the AP loads of them are denoted as ya, yb, y' a, and y' b , respectively. Straightforwardly, we will have yb < ya, y' b < ya, and y' a < ya, where the lexicographic order has been decreased. Since the lexicographical order cannot be infinitely decreased, we can conclude that the Algo. 1 will stop after finite number of operations.

The introduced overhead by the proposed algorithm on the AP is straightforwardly low. On each MU, the most time consuming operation is the periodically probing process in every TΔ seconds. However, this probing process only takes around 300ms according to measurements. Comparing with the interval TΔ, the overhead is almost negligible.

VI. Performance Evaluation

In this section, we first introduce the numerical evaluation based on the developed simulation program. The program is able to simulate dynamic and large-scale topology to clearly show the achievable benefits of the proposed scheme. We then provide NS2 [16] simulation results for a medium-size topology with suddenly roaming clients. Finally, we also explain our prototype implementation on a testbed built with normal computers. To measure the performance, we use total throughput Σ u Є U Ɵu as the metric to measure the overall efficiency and Jain's fairness index [17] to denote the degree of load balancing in the network.

$$\frac{(\sum_{u \in U} \theta_u)^2}{|U| \sum_{u \in U} \theta_u^{\,2}}$$

[2]δ = 0 is the loosest condition to activate the switching operation.

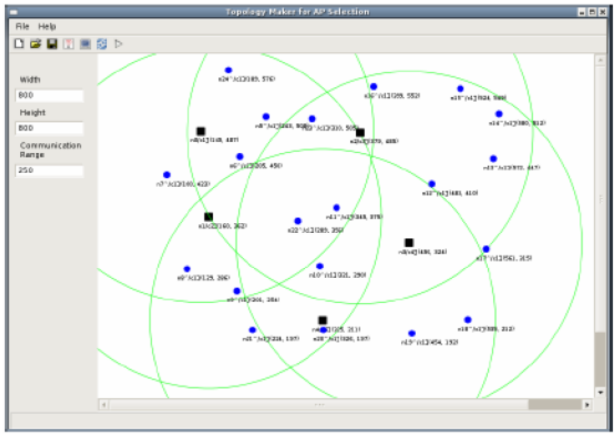


Fig. 3. The snapshot of developed numerical simulator.

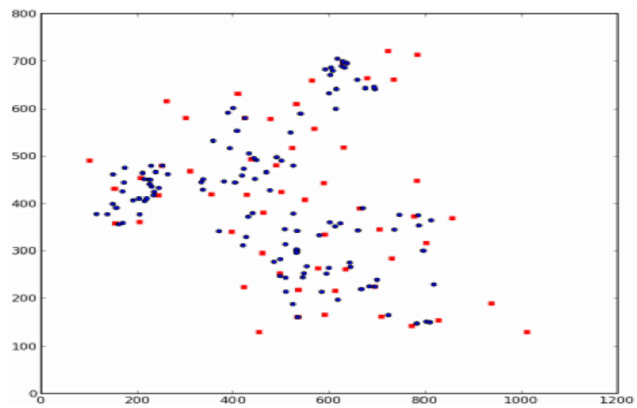


Fig. 4. A realistic scenario with measured mobility for numerical simulation. The red squares denote the APs and the blue circles denote the MUs at the beginning of simulation.

VII. Numerical Simulation for Realistic Scenario

In order to evaluate the proposed scheme for large-scale topologies, we have developed a discrete-event simulator based on SimPy [18], which is a Python framework for discrete-event simulation applications. Users can manually place the APs and MUs in the GUI (Graphic User Interface). The generated scenario can also be saved and loaded for future use. The snapshot of the program interface is captured and shown in Fig. 3. To accelerate the simulation, the complex behavior of IEEE 802.11 MAC is simplified and the throughput is calculated by the throughput model given in [12]. We use a set of measured trace files provide by [19], which collected the 20 minutes measurement data by capturing the realistic mobility patterns of the MUs in the campus of Dartmouth University. From the measurement results, we pick up 56 APs and 126 MUs with their mobility placed in a rectangle topology of size 1100×1000m2 as shown in Fig. 4.

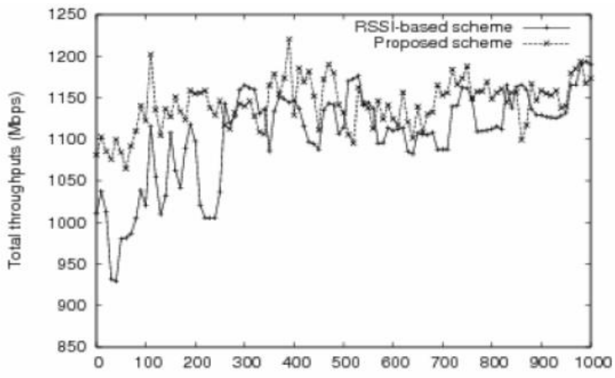


Fig. 5. The throughput difference between RSSI-based scheme and proposed scheme w.r.t simulation time for the realistic topology shown in Fig. 5.

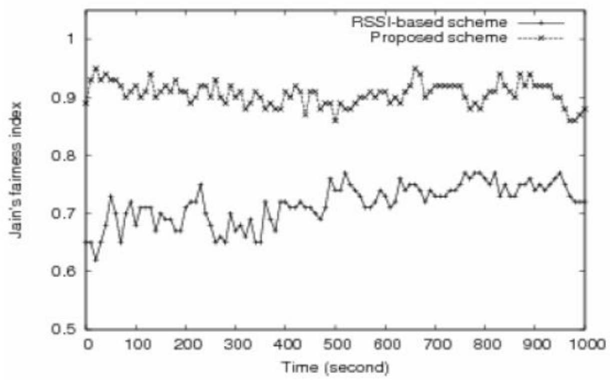


Fig. 6. The Jain's fairness value difference between RSSI-based scheme and proposed scheme w.r.t simulation time for the realistic topology shown in Fig. 5.

According to Fig. 5 and Fig. 6, we can observe that the total throughput achieved by the proposed scheme is generally the same or sometimes higher than that of the default RSSI based scheme. However, the value of fairness metric has been apparently (between 20%-30%) improved after applying the proposed scheme. On the other hand, we also find that it mostly takes only one probing and reassociation operation for the MUs to reach a steady state when they move around in the topology.

VIII. CONCLUSION

In this paper, we have explored the load balancing scheme to guarantee the throughput fairness among the MUs. To achieve this, we have proposed a distributed and self-stabilized association scheme for the MUs in the multi-rate WLANs. The proposed scheme gradually balances the AP loads in a distributed manner. With extensive simulations, we can observe that it can significantly improve, or sometimes nearly double, the extent of throughput fairness among the MUs with low overhead. To show the feasibility of the proposed scheme, we have implemented a prototype on normal computers by modifying open source wireless driver software packaged Our research is oriented for practical WiFi products and can be implemented with small additional modification to achieve apparent load balancing in deployed WLANs.

IX. ACKNOWLEDGEMENT

I express my deep gratitude to almighty who guided me in the right to make this research a success. I would like to express my sincere and deep sense of gratitude to Management and Principal of VMKV Engineering College, Salem, Tamilnadu, India. for giving the chance to do my research work. I wish to express my sincere thanks to and VSA Group of Institutions, Salem, Tamilnadu, India for providing excellent research centre for doing my research work. I owe a high debt of gratitude to my guide Prof.Dr.C.Manoharan Principal VSA Group of Institutions for his valuable guidance during my research work. A special word of thanks to the entire department staffs members for their constant help support and valuable guidance while doing my research. And last, but not least, I would also like to thank me beloved PARENTS and my friends for their external source of encouragement, inspiration, moral support which it would have been almost impossible to complete this research.

X. REFERENCES

1) Y. Bejerano, S.-J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," in Proc. ACM MobiCom'04, 2007,volume 15, pp. 560–573.
2) Y. Bejerano, S.-J. Han, and L. Li, "Fairness and load balancing in wireless LANs using association control," in Proc. ACM MobiCom'04, 2004, pp. 315–329.
3) Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Char- acterizing user behavior and network performance in a public wireless LAN," in Proc. ACM SIGMETRICS, 2002, pp. 195–205.
4) Kotz and K. Essien, "Analysis of a campus-wide wireless network," in Proc. ACM MobiCom'02, 2002, pp. 107–118.
5) Balazinska and P. Castro, "Characterizing mobility and network usage in a corporate wireless local-area network," in Proc. USENIX MobiSys, 2003.
6) Papanikos and M. Logothetis, "A study on dynamic load balance for IEEE 802.11b wireless LAN," in Proc. COMCON, 2001.
7) Balachandran, P. Bahl, and G. M. Voelker, "Hot-spot congestion relief and service guarantees in public-area wireless networks," ACM IGCOMM Comput. Commun. Rev., vol. 32, no. 1, pp. 59–59, 2002.

8) T.-C. Tsai and C.-F. Lien, "IEEE 802.11 hot spot load balance and QoSmaintained seamless roaming," in Proc. National Computer Symp.(NCS), 2003.
9) ORINOCO AP-600 Data Sheet, Proxim Wireless Networks, 2004.
10) Data Sheet for Cisco Aironet 1200 Series, Cisco Systems Inc., 2004.
11) Katzela and M. Nagshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," IEEE Pers. Commun., vol. 3, no. 3, pp. 10–31, Jun. 1996.
12) S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in Proc. IEEE INFOCOM, 2003.
13) SimPy. http://simpy.sourceforge.net/.
14) CRAWDAD. http://crawdad.cs.dartmouth.edu/index.php.
15) Madwifi-ng. http://madwifi.org/.

Maximized Optimization Algorithm For Distributed Traffic Control Laws by Combining Traffic Engineering and Quality of Service

[1]Mr.R.SENTHIL KUMAR, [2]Mr.K.THANGADURAI
[1]Asst.Professor, [2]Lecturer,
Department Of Computer Science and Engineering,
Vinayaka Mission's University, V.M.K.V.Engineering College,
Tamil Nadu, India.
{E_Mail:trsk_kavi2042@yahoo.co.in, gold_rosen@sify.com}

GJCST Classifications:
C.2.4, C.2.1, C.2.5, K.6.4, J.m

***Abstract-* With the Internet evolved into a global commercial infrastructure, there has been a great demand for new applications of global reach, for which today's Internet protocols cannot adequately support. The real-time applications have stringent delay and delay jitter requirements, which cannot be adequately supported by today's Internet protocols.**

As a result, in recent years, a large number of new Internet protocols were developed in an attempt to meet this demand. Multi-Protocol Label Switching (MPLS) has been envisioned as an ideal platform upon which guaranteed services could be developed. Service guarantee is achieved by setting up and managing a set of primary and Backup class-of-service (CoS) aware label switched paths across an IP domain. In addition to MPLS, this approach requires a suite of protocols be implemented, e.g., DiffServ for quality of service (QoS), path protection / fast rerouting for link failure recovery (FR), and constraint-based routing for traffic engineering (TE).

The proposed thesis develop a family of distributed traffic control laws (DCLs), which allows optimal, multiple CoSs, multipath based rate adaptation and load balancing. The DCLs drive the network to an operation point where a user defined global utility function is maximized. The proposed family of DCLs has, the capability to enable optimal, scalable QoS, and Traffic Engineering, simultaneously.

I. INTRODUCTION

The transport control protocol (TCP) window-based congestion control algorithms use minimum information from the network as input to allow fully distributed traffic control. In other words, the only needed feedback information for the TCP window-based congestion control is whether the forwarding path is congested or not.
This allows the TCP source node to infer path congestion by counting the number of repetitive acknowledgments of the same packet or measuring end-to-end round-trip delay, making TCP a truly end-to-end protocol without the assistance of the underlying internetworking layer infrastructure. This has made the proliferation of the Internet applications at global scale possible.
An excellent example is the fast, ubiquitous adoption of World Wide Web due to its use of TCP as its underlying transport. However, as the Internet has evolved into a global commercial infrastructure, there has been a great demand for new applications of global reach, for which today's Internet protocols cannot adequately support.
For example, real-time applications, such as voice over IP (VoIP) and videophone, have stringent delay and delay jitter requirements, which cannot be adequately supported by today's Internet protocols. As a result, in recent years, a large number of new Internet protocols were developed in an attempt to meet this demand.
For example, multiprotocol label switching (MPLS) has been envisioned as an ideal platform upon which guaranteed services could be developed. Service guarantee is achieved by setting up and managing a set of primary and backup class-of-service (CoS) aware label switched paths across an IP domain.
In addition to MPLS, this approach requires a suite of protocols be implemented, e.g., DiffServ for quality of service (QoS), path protection/fast rerouting for link failure recovery (FR), and constraint-based routing for traffic engineering (TE). This, however, means that, to adequately support real-time applications, a whole suite of protocols with significant involvement of the IP core nodes need to be developed.
This raises serious concerns about the scalability and complexity of using these protocols to support real-time applications at a global scale.
Hence, a key question to be answered is whether it is possible to enable the above service, quality features, including QoS,

II. LITERATURE REVIEW

The existing algorithms focus on TCP types of traffic including both empirical algorithms and algorithms based on control theory [10]. These algorithms assume a single path, and the approaches taken are not optimization based.
In the existing scheme flows with different ingress–egress, node pairs share the same network resources. Degree of interaction between different flows due to the resource constraints was very poor in the existing Distributed traffic control laws [1].

Since flows with different ingress–egress node pairs share the same network resources, the key challenge in the design of DCLs is the fact that there is a high degree of interaction between different flows due to the resource constraints. One existing approach to get around this is to incorporate a link congestion cost into the overall utility function, which replaces the link resource constraints. Then, the problem is solved using a gradient type algorithm, resulting in families of DCLs that support point-to-point multipath load balancing for rate adaptive traffic [6, 7].

Some of the existing methods developed a family of DCLs based on nonlinear control theory. This family of DCLs can be applied not only to usual rate adaptive traffic with point-to-point multipath, but also to rate adaptive traffic with minimum service requirements and/or maximum allowed sending rate and to services with targeted rate guarantee, all allowing for point-to-point multipath.

The only needed feedback from the network is the number of congested links along the forwarding paths [2, 5]. Moreover, the technique applies to any utility function that can be expressed as a sum of concave terms.

Due to the needed use of the number of congested links in a forwarding path as the input to a DCL, the existing family of DCLs requires explicit congestion feedback from the network. The existing scheme can only be applied to a connection-oriented network, such as an MPLS enabled IP network [9].

In the proposed system, the DCLs control the traffic independently at different traffic source nodes, e.g., edge nodes or end-hosts. A salient feature of this family of DCLs is that the needed information feedback from the network is minimum, i.e., whether a forwarding path is congested or not, which can be inferred at the source node itself, the same way as TCP congestion notification. This makes it possible to allow this family of DCLs to be operated end-to-end.

III. System Model

The traffic flows can be described by a fluid flow model, where the only resource taken into account is link bandwidth. For simplicity, first restrict ourselves to the point-to-point multipath only and address the point-to-multipoint and multicast cases later.

The system model, consider a computer network where calls of different types are present. Types denote an aggregate of calls with the same ingress and egress node, as well as service requirements i.e., calls that share a given set of paths connecting the same ingress–egress node pair and whose service requirements are to be satisfied by the aggregate, not by individual calls. When the edge nodes coincide with the end-hosts, the control laws developed in this paper become end-to-end control laws working at the transport layer servicing individual application flows.

A. Discretization, Delays and Quantization

The issues handled in implementing the control laws implement a discrete time version of the control algorithms, uses finite word length which leads to a quantization of the possible data rate values and there is delay in the propagation of the congestion information. All of these lead to a well-known phenomenon called oscillation. Even in this case, the discretization of the control laws is approximately optimal.

B. Congestion Detection and Notification

To maintain the transport or higher layers abstraction, a source inferred congestion detection and notification mechanism is desirable for the implementation of this family of DCLs in a connectionless IP network. However, unless the transport or higher layer protocol that implements this family of DCLs is defined, the exact source inferred congestion detection and notification mechanism cannot be decided.

For example, if a DCL in this family is used in association with a TCP-like reliable transport protocol, a source inferred congestion detection and notification mechanism based on, for example, ACK counts can then be adopted. On the other hand, if the DCL is used in association with an UDP-like unreliable transport protocol, the forwarding path congestion may be detected and notified by periodically sending an echo packet to the destination node and measuring the round-trip time of the echoed packet.

The source inferred congestion detection and notification approaches can also be used in the context of a connection-oriented network, such as an MPLS one. In addition, other mechanisms can be employed, e.g., mechanisms using a signaling protocol for congestion detection and notification.

C. Failure Detection and Notification

The node/link failure detection and notification may or may not be integrated with the congestion detection and notification mechanism. Again, they are dependent on the actual protocol that implements a DCL in this family. For example, a source inferred congestion detection and notification using echo packets to infer path congestion may also be used to infer possible node/link failures. On the other hand, in an MPLS network, the path protection mechanism under development can be leveraged to allow failure detection and notification, separate from the congestion detection and notification mechanisms.

D. Design Parameters

The behavior of the algorithm under different choices of the Design parameters are

i. Oscillation Reduction Functions

The adaptive oscillation reduction has a big impact on performance. Considering the behavior for a constant, the maximum value allowed for the time variation. The observed oscillation is clearly larger in magnitude. Moreover, due to the larger oscillations, convergence to a larger neighborhood of the optimal is obtained, and departures from the average target rates for AF are also larger (providing a worse service to these users). On the other hand, the transient response is faster due to larger data rate derivatives.

ii. Discretization Step

Another parameter that has a bearing in the performance of the algorithm is the discretization step. In order to show its influence, it was chosen as 10 ms. clearly, oscillations are also larger in this case. However, the response is still acceptable, and a smaller could be used to limit the magnitude of the spikes.

iii. Scaling of the Utility Function

The scaling of the utility function does not alter the solution of the optimization problem at hand. It does, however, change the bounds on the quantities. Due to the exponential dependence on the gradient, it is advisable to choose a small value of such that the resulting value of is in the order of 1. Simulations have shown that the algorithm is very sensitive to with the amplitude of the oscillations increasing substantially when one increases this parameter.
However, convergence to a neighborhood of the optimal is still achieved as one can expect. In addition, the AF constraints are satisfied in the average but large departures from the imposed average rate can happen for high values.

IV. EXPERIMENTAL EVALUATION

The new family of DCLs provides the much-needed mathematical foundation that allows the use of source inferred congestion detection and notification to maintain layer abstraction. The new family of DCLs allows the rate control to be decoupled from the congestion detection mechanisms in use. This means that any queue management algorithm and queue scheduling discipline used in the core nodes can coexist with the family of DCLs running at the edge nodes or end-hosts.
The implementation of any DCL in this family, only needs to consider the two end nodes, provided that a source inferred congestion detection and notification is available. However, having said that, different queue management algorithms and queue scheduling disciplines do have an impact on the overall performance for any end-to-end traffic control mechanism.
As a result, there are two key components in the implementation of the family of DCLs, i.e., the implementation of the DCL in the edge nodes or end-hosts and the design of source inferred congestion detection and notification mechanisms. The system model focus on the issues related to the design of source-inferred congestion detection and notification mechanisms.

V. CONCLUSION

The proposed family of DCLs can be applied to a connectionless IP network to enable sophisticated service quality features, solely based on a set of shortest paths from any given ingress node to a set of egress nodes. The distributed traffic control laws (DCLs) allows optimal, multiple CoSs, multi-path based rate adaptation and load balancing.
The DCLs drive the network to an operation point where a user defined global utility function is maximized. The mathematical formulation allows both point-to-point multi-path and point-to-multipoint multi-path, the family of DCLs can be applied to a connectionless IP network to enable sophisticated service quality features, solely based on a set of shortest paths from any given ingress node to a set of egress nodes.
A core node may be CoS and multipath agnostic and may employ any queue management / scheduling algorithms, e.g., simple FIFO queues, at its output ports. This family of DCLs allows fast time scale TE through multi-path load balancing. The proposed DCLs can automatically repartition the traffic in an optimal way among the rest of the multipath when path failures occur. The proposed scheme can be applied for both connection oriented and connection less networks.

VI. REFERENCES

1) Bernardo A. Movsichoff, Constantino M. Lagoa, and Hao Che “ End-to-End Optimal Algorithms for Integrated QoS, Traffic Engineering, and Failure Recovery “ IEEE /ACM Transactions on networking, Vol. 15, No. 4, August 2007
2) T. Alpcan and T. Basar, “Global stability analysis of an end-to-end congestion control scheme for general topology networks with delay,” in Proc. 42nd IEEE Conf. Decision Control, Dec. 2003, vol. 1, pp. 1092–1097.
3) Bonomi, D. Mitra, and J. B. Seery, “Adaptive algorithms for feedback based flow control in high-speed, wide-area networks,” IEEE J. Sel. Areas Commun., vol. 13, pp. 1267–1283, Sep. 1995.
4) Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang, “Endpoint admission control: Architectural issues and performance,” in Proc. ACM SIGCOMM 2000, Stockholm, Sweden, Aug./Sep.2000, pp.57-69.
5) Chiu and R. Jain, “Analysis of the increase/decrease algorithms for congestion avoidance in computer networks,” J. Comput. Netw. ISDN Syst., vol. 17, no. 1, pp. 1–14, Jun. 1989.
6) Elwalid, C. Jin, S. Low, and I. Widjaja, “Mate: MPLS adaptive traffic engineering,” in Proc. IEEE INFOCOM’01, Anchorage, AK, Apr. 2001, vol. 3, pp. 1300–1309.
7) A. Movsichoff, C. M. Lagoa, and H. Che, “Decentralized optimal traffic engineering in connectionless networks,” IEEE J. Sel. Areas Commun., vol. 23, pp. 293–303, Feb. 2005.
8) Paganini, Z. Wang, J. C. Doyle, and S. H. Low, “Congestion control for high performance, stability, and fairness in general networks,”

IEEE/ACM Trans. Netw., vol. 13, no. 1, pp. 43–56, 2005.

9) Sharma and F. Hellstrand, "Framework for multi-protocol label switching (MPLS)-based recovery," IETF RFC 3469, 2003.
10) Utkin, Sliding Modes in Control and Optimization. Berlin, Germany: Springer-Verlag, 1992, vol. 66, Communications and Control Engineering Series.
11) L.Ying G.E. Dullerud and R.Srikant ,"Global Stability Of Internet congestion controllers with heterogeneous delays ," in proc.2004 Amer. Control conf.,2004, vol. 4, pp. 2948-2953.

Security Algorithm for Cryptosystems chaotic map

R.Gnanajeyaraman[1], K.Prasadh[2],Vignesh[3]
[1] Research scholar, Vinayaka Missions University, Salem, Tamilnadu, India r.gnanajeyaraman@gmail.com
[2] Research scholar, Vinayaka Missions University, Salem, Tamilnadu, India. ksprasaadh@gmail.com
Vinayaka Missions University, Salem, Tamilnadu, India.

GJCST Classifications:
E.3, K.6.5, D.4.6

Abstract*- In Chaotic the properties of sensitivity to initial conditions, control parameters and pseudo-randomness chaotic maps have been widely used in data encryption recently. The chaotic based cryptosystems are suitable for large-scale data encryption such as images, videos, or audio data. This paper propose a novel higher dimensional chaotic system for audio encryption, in which variables are treated as encryption keys in order to achieve secure transmission of audio signals. Since the highly sensitive to the initial condition of a system and to the variation of a parameter, and chaotic trajectory is so unpredictable. As a result, we obtain much higher security. The higher dimensional of the algorithm is used to enhance the key space and security of the algorithm. The security analysis of the algorithm is given. The experiments show that the algorithm has the characteristic of sensitive to initial condition, high key space; digital audio signal distribution uniformity and the algorithm will not break in chosen/known-plaintext attacks.
Keywords- Audio encryption; Chaos; Security; Higher dimensional chaotic maps;

I. INTRODUCTION

The techniques of secure communication by which one can transmit confidential messages secretively are of practical interest in several areas, including databases, internet banking, production of communication channels etc. Based on the structure of the encryption algorithm we can classify cryptosystems into two categories, namely, stream cipher and block cipher. In a stream cipher algorithm, the message is encrypted bit-by-bit with the application of a secret key generator and the decryption can be done by using the same algorithm as in encryption, and with the same secret-key generator. In practice shift register, non-linear combination generator clock controlled generators ect. are used as key generators. Unlike the stream cipher, in block cipher a group of bits of fixed length is encrypted block by block. In another classification, which is based on method of distribution of secret key, one classifies the cryptosystems into classes-private (symmetric) key and public (asymmetric) key cryptosystems. In the private key cryptosystems, sender and receiver use the same key for the encryption and decryption, respectively. These systems are efficient and their security depends on the statistical properties of the random bit sequence and the length of the secret key.

Most of the existing cryptosystems, except a few, utilize number theory, algebra, combinatory and computer arithmetic, est. as mathematical tools for constructing the algorithms for the encryption and decryption.

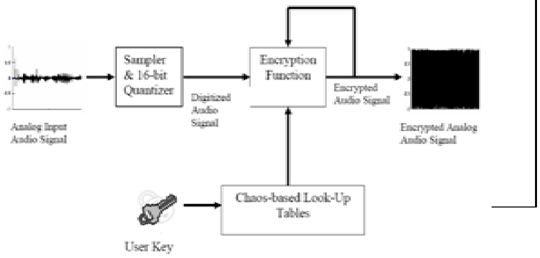
Now a day it has been proved that in many aspects chaotic maps have analogous but different characteristics as compared with conventional encryption algorithms such as DES, IDEA and RSA. These are not suitable for practical audio encryption. At the start of last decade only, the use of the continuous as well as discrete chaotic dynamical systems has been made to develop cryptosystems. The chaotic systems are characterized by the sensitivity on initial conditions, system parameters, and unpredictability of evolution of its orbits. These thinks are used for encryption [1-6].

The chaotic systems for encrypting audio files are highly unpredictable and random-look nature this attractive feature may lead to novel chaotic applications. In most of the previous chaotic algorithm uses, XOR or XNOR is the basic technique for encryption. These techniques are very weak against statistical, chosen/known-plain-text attacks and moreover its security to brut-force attack is also questionable for avoiding the above risk and to increase the security level of the algorithm we propose a higher dimensional chaotic map [10]. Here we use multiple lookup table for encrypt the audio files all the values will be encrypted using cipher block chaining method. Arnold's cat map is a good candidate for permutation, thus it is extended to eight dimensional versions, called 8D cat map and then used for encryption. Taking advantage of the exceptionally good properties of mixing and sensitivity to initial conditions and parameters of the chaotic 8D cat map, the proposed scheme incorporates newer architecture of chaos-based chipper block chain mode of encryption is proposed. It is used to interlink all the previous values of audio signal with the current value of the audio signal so without knowing the previous value of the audio signal value we cannot decrypt the current audio signal value this will increase the security one more level. Higher the dimension and higher the number of keys increases the complexity and keyspace of the algorithm.

II. AUDIO ENCRYPTION SCHEME BASED ON HIGHER DIMENSIONAL CHAOTIC MAP

In this paper, a chaos-based audio encryption system, in the framework of cipher block chaining architecture is proposed. The block diagram of the system is shown in fig. . An analog audio input

Fig.1. Chaos based Audio Encryption Scheme.



A. Chaos-Based Look-Up Tables.

Taking advantage of the exceptionally good properties of mixing and sensitive to initial conditions Chaos-based Look-Up tables are used for encrypting audio files. In chaos there are many maps it is found that Arnold's Cat map [7] is a good candidate for permutation, thus it is extended to a higher dimensional version, called NthD cat map, and then used for this purpose. Here the relation between nth and n+1th data are de-correlated much higher than 2D and 3D maps so the authentication of our algorithm is increased much higher than the lower dimensional chaos maps. A higher dimensional Cat map is formed as follows:

Table.1. Look-Up Table generated using higher Dimensional cat map.

Sequence A			Sequence B				Sequence H		
Iterated values in Ascending order	Iteration number	Index Number	Iterated values in Ascending order	Iteration number	Index Number	…	Iterated values in Ascending order	Iteration number	Index Number
0.0045568	5	0	0.0030982	1633	0	…	0.003287	60206	0
0.010197	15	1	0.0092543	19	1		0.004382	7101	1
0.012603	111	2	0.012718	165	2		0.004449	214	2
0.026510	2488	3	0.013579	1455	3		0.0053193	45	3
0.037699	206	4	0.014931	124	4		0.0077386	1183	4
0.039211	18800	5	0.018185	10321	5		0.010336	248	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.99911	573	65536	0.9919	1380	65536		0.019168	7	65536

Is sampled at a frequency well above the Nyquist frequency of the signal. Then an 16-bit quantization is used to convert the analog signals into its equivalent decimal value. By masking these data with a random key stream generated by a chaos-based pseudo-random key stream generator, the corresponding encrypted audio is formed. The details of each component are to be discussed in the following sections. As demonstrated in our simulation, this approach is more secure in all the ways.

$$\begin{bmatrix} A_{n+1} \\ B_{n+1} \\ \vdots \\ H_{n+1} \end{bmatrix} = x \begin{bmatrix} A_n \\ B_n \\ \vdots \\ H_n \end{bmatrix} \bmod 1. \qquad (1)$$

Where

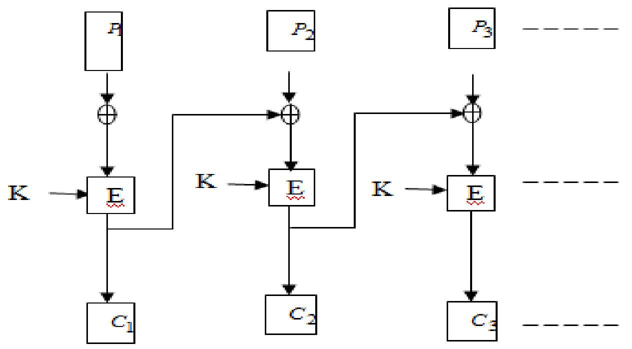
$$x = \begin{bmatrix} 1 & a_{12} & 0 & \cdots & 0 \\ b_{12} & 1+a_{12}b_{12} & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & a_{13} & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ b_{13} & 0 & 1+a_{13}b_{13} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 & 0 & \cdots & a_{1m} \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & 0 \\ b_{1m} & 0 & 0 & 0 & 1+a_{1m}b_{1m} \end{bmatrix}$$
$$\times \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & a_{23} & \cdots & 0 \\ 0 & b_{23} & 1+a_{23}b_{23} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdots \begin{bmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & 1 & a_{m-1,m} \\ 0 & \cdots & 0 & b_{m-1,m} & 1+a_{m-1,m}b_{m-1,m} \end{bmatrix} \qquad (2)$$

Where a_{ij}, b_{ij} are integers in $[0, 2^L - 1]$.

This higher-dimensional Cat map is used as our pre-processing unit for generating Look-Up tables. The initial key stream arranged in tables {T_1, T_2, T_3…}. Here 'm' is the number of tables that is used for encrypting the audio signal L is number of bits used by the key stream in this algorithm L=16 and m=8.

B. Encryption Function

On the encryption block, one uniform distributed random number is generated that sequence will select the tables for encryption. After selecting the table, the digital value of the audio signal is mapped to the iteration number of the chaotic sequence. For encrypting the nth digit, we add the n-1st cipher digit value with nth plain value that resultant value will be mapped with the table value. This type of encryption will increase the security in one more level. Because here all the pixels are interlinked if we want to decrypt the nth digit, first, we will know the n-1st plain digit then only we can decrypt the nth digit. This type of encryption is called cipher block chain mode of encryption the block diagram of encryption function is given in fig. 2.



P_i = Plain Digitized Audio value
E_k = Encryption Algorithm
C_i = Digital Cipher Audio value
K_e = Key

Fig.2. The construction of the Cipher Block Chaining.

In this block diagram, the previous value of cipher digit is added with current plain digit of the audio signal. At the time of addition, the resultant value will be more than 65,536 in sum time. If the value is more than 65,536, we require more than 16 bit for avoiding this some modifications down on the result for that the following formula is used

$$MP_n = C_{n-1}+P_n \bmod 65{,}536. \quad (3)$$

Where MP is the modified plain digit that will be encrypted with key Ke and the resultant cipher digit is C. After encryption, the cipher value of the audio is brought into original range. In this way, all the audio files will be encrypted. For decryption do the reverse process of the encryption.

III. SECURITY DESCRIPTION

In the proposed cryptosystem, 8D chaotic cat map is used for encrypt audio files. The cryptosystems security is determined by the used chaotic maps Key space, Key Sensitivity and Plaintext Sensitivity. Here, chaotic map's properties are in close relation with the cryptosystem's security. At first, its parameter is used as confusion key. Thus, parameter sensitivity is in close relation with key sensitivity. The higher the parameter sensitivity is, the higher the key sensitivity is, and the stronger the cryptosystem is. Thus, the chaotic map with higher parameter sensitivity is preferred in this cryptosystem. Secondly higher the initial-value sensitivity is, the smaller the correlation between adjacent pixels is, and the more random the encrypted audio file is. Therefore, the chaotic map with higher initial-value sensitivity is preferred in this crypto system. In encryption iteration, time is in close relation with cryptosystem's security. The more the iteration time is, the larger the cryptosystem's key space is if different keys are used in different iteration.

IV. SECURITY ANALYSIS

A. *Key Space*

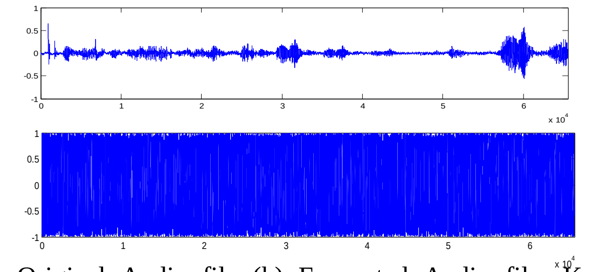
In the proposed cryptosystem, 8 keys are used for encryption. If n is the iteration time, and different keys are used in different iterations, then the key space is

$$S=(K_1\cdot K_2\cdot K_3\cdots K_8)^n. \quad (4)$$

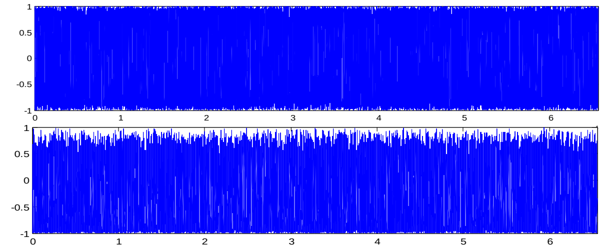
According to Eq.(1)-(3), the key space is determined by the parameter space of chaotic map. As can be seen, the cryptosystem's key space S increases with rise of parameter space $K_1\cdot K_2\cdot K_3\cdot\cdots K_8$ or iteration time n. Parameter space can be increased with number of keys here it is 8. The iteration time can be chosen according to security and complexity requirements in this algorithm 16bit Quantization is used so n=65,536. Taking $1\times N$ sized Audio file with L level of Quantization with different Keys the complexity of the Cat map is $N^n\cdot L^n$.

B. *Key Sensitivity*

High key sensitivity is required by secure cryptosystems, Which means that the cipher text cannot be decrypted correctly although there is only a slight difference between encryption or decryption keys? This guarantees the security of a cryptosystem against brute-force attacks to some extent.

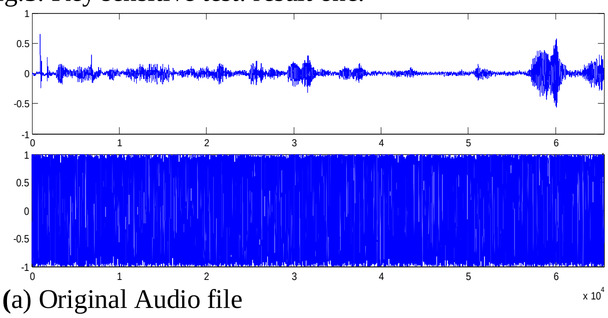


(a) Original Audio file (b) Encrypted Audio file: Key= 1234353463465

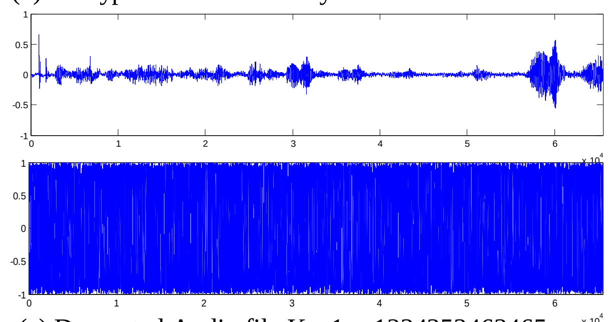


(c) Encrypted Audio file Key1 = 1234353463464
(d) Difference Audio file

Fig.3. Key sensitive test: result one.



(a) Original Audio file
(b) Encrypted Audio file: Key1= 1234353463465



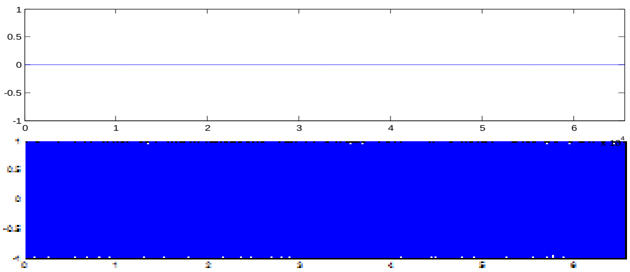
(c) Decrypted Audio file Key1 = 1234353463465
(d) Decrypted Audio file Key1 = 1234353463464

Fig.4. Key sensitive test: result 2.

1) First, an audio file is encrypted by using the test keys key1 value is "1234353463465."
2) Then, the least significant bit of the key1 is change to "1234353463464" all the remaining seven keys are same which is used to encrypt the same audio file.

3) Finally the above two encrypted audio files with two slightly different keys, are compared. The result is: the image encrypted by the key "1234353463465" has 9999% of difference from the image encrypted by the key "1234353463464" in terms of Quantized digital values, although there is only one digit difference in one of the key, Fig.3 Shows the test result.

C. Encryption On Uniform Audio File



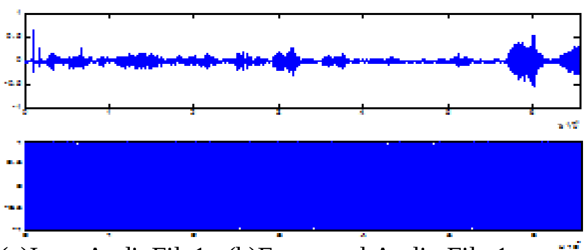
(a) Uniform Audio file (b) Encrypted Uniform Audio file

Fig.5. Uniform Audio Encryption

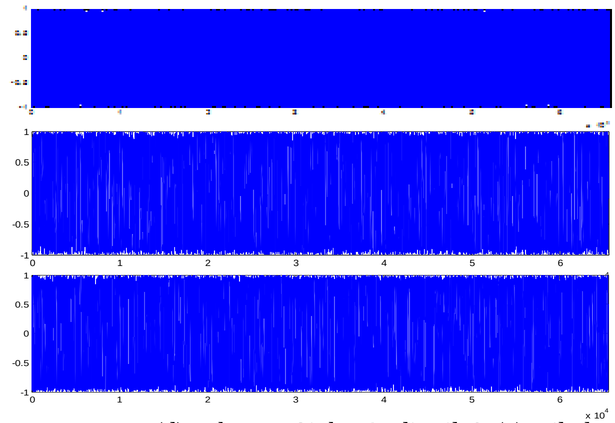
An Audio file with uniform sound is taken that file is encrypted using this algorithm. After encryption, there are no patterns or uniformity on the encrypted audio file. This will ensure even in uniform audio our algorithm make more randomness so the security level of the algorithm is higher.

D. Chosen/Known-Plaintext Attack

Chosen/Known-plain text attacks are such attacks in which one can access/choose a set of plain texts and observe the corresponding cipher texts. In today's networked world, such attacks occur more and more frequently. For a cipher with a higher level of security, the security against both known-plaintext and chosen-plaintext attacks are required. Most of the XOR-ing based techniques are not robust against this attack [10]. Apparently, even when the secret key is changed for each plaintext, these methods are insecure against chosen/known-plaintext attacks. The mask audio I_m is obtained by simply XOR-ing the plain audio I with its corresponding cipher audio I'. XOR-ing the I_m mask with unknown cipher audio J', if we get the unknown plain audio J then the algorithm fails in Chosen/Known-plaintext attack, otherwise the algorithm safe against Chosen/Known-plaintext attack. Fig.7 demonstrates an unsuccessful chosen/known-plain text attack in the proposed algorithm.



(a)InputAudioFile1 (b)Encrypted Audio File 1



(c) XOR - Mask (d)UnknownCipherAudioFile2 (e)Failed to crack the Audio File2

Fig.7. Unsuccessful chosen/Known-plaintext attack on proposed algorithm

V. Conclusions

Telemedicine can provide access to health care in previously unserved or underserved areas. These areas include both rural and inner city or barrier locations, All of them use internet is the basic medium for transferring information. But internet is public medium anyone can get information's from any ware, so the privacy is very less for improving the privacy over the public network we go for encryption. Most of the previous algorithms are return for encrypting only text messages. If we use the same method for encrypting media files like medical audio files, it will not be an efficient one. Because correlation of audio file is higher than the text file, so we go for higher dynamic system.

The dynamic response of chaotic system is highly sensitive to initial values and the variation of a parameter and chaotic trajectory is very unpredictable. Therefore, this paper propose a higher dimensional chaotic system for encrypting medical audio files in telemedicine which is transferring medical files on unsecured internet and telephone network. After calculating correlation coefficient γ and conducting FIPS PUB 140-1 to test on the higher dimensional chaotic system of this paper, it is found that the chaotic system proposed in this paper is obtain optical scatter characteristics this has ensure our algorithm is more secure. In this algorithm the number of keys are increased so the key space and the complexity of the algorithm although increases accordingly. Even in public channel, when the encrypted audio files are stolen, an intruder cannot decrypt and recover the original audio file. From the above security analysis and various attacks, this algorithm gives better results. Even the audio file is uniform; the algorithm will give better results. Therefore, this algorithm is useful for mission critical applications of medical transaction over the unsecured public networks.

VI. References

1) Habutsu T., Nishio Y., Sasase I., and Morio S., "A secret key cryptosystem by iterating chaotic map," Lect. Notes comput. Sci, Advances in Cryptology-EuroCrypt'91, 1991, vol. 547, page(s): 127-140.
2) Pichler F. and Scharinger J., "Finite dimensional generalized baker dynamical systems for cryptographic applications," Lect. Notes in Comput. Sci, 1996, vol. 1030, pp. 465-476.
3) Masuda N. and Aihara K., "cryptosystems with discretized chaotic maps," IEEE Trans. On Circuts and Systems – I: Fundamental Theory and Applications, 202, vol. 49, no. 1, page(s): 28-40.
4) Yen J. C. and Guo J. I., "A new chaotic image encryption algorithm," Proc. 1998 National symposium on Telecommunications, Dec. 1998, page(s): 358-362.
5) Yen J. C. and Guo J. I., "Efficient hierarchical image encryption algorithm and its VLSI realization," IEE Proceeding Vis. Image Signal Process, April, 2000, vol. 147, no.2, page(s): 430-437.
6) Chen G., Mao Y. B. and Chui C. K., "A symmetric image encryption scheme based on 3D chaotic cat maps," Int. J. chaos, Solitons and Fractals, 2004, 21, page(s): 749-761.
7) Chen G, Dong X. From chaos to order: methodologies, perspectives and applications. Singapure: World Scientific; 1998.
8) Bluman AG. Elementary statistics. 3rd ed. New york: McGraw-Hill; 1998.
9) IPS PUB 140-1. Security requirement for cryptographic modules. Federal Information Processing standards Publication; 1994.
10) Daniel Socek, shujun Li, Spyros S. Magliveras and Borko Furko. "Enhanced 1D Chaotic Key-Based Algorithm for Image Encryption" IEEE Int. Conf. On Security and Privacy for emerging areas in communication. Sep. 2005, Page(s) 406-417.

Reducing Packet Delay and Loss in Heterogeneous Mobile Wireless Networks

[1]Kirubakaran.S, [2]Manoharan C,

[1]Research scholar, Anna University, Coimbatore, India.
Email: kiruba.me07@gmail.com

[2] Principal, VSA School of Engineering, Salem, India.
Email: c_m_66@yahoo.co.in

GJCST Classifications:
C.2.1, H.1.1, E.4

Abstract- **Efficient communications are crucial for disaster response and recovery. However, most current public safety land mobile radio (LMR) networks only provide narrowband voice service with limited support of low-speed data services due to the packet delay and loss and less resource utilization. So high resource utilization techniques are needed. The Session Initiation Protocol (SIP) and a joint radio resource management framework is a current technology to support the interoperability between cellular and LMR networks.SIP is seamless handoff scheme. In this paper, we study to enhance the interoperability of LMR with commercial wireless cellular networks, by which a wide variety of benefits can be offered to disaster responders, including new multimedia services, increased data rates, and low cost devices. Our approach is based on FDS and IDS to reduce the packet delay and loss in the interoperable between cellular and LMR networks. In addition, an optimal bandwidth utilization scheme is proposed to maximize the overall radio resource utilization. In our first experiment, we are applying the proposed approach to the interoperable cellular and LMR networks. The two dynamic scheduler FDS and IDS is reduced the packet delay and loss. The proposed approach is a novel one in interoperable cellular and LMR networks besides being the first approach to tackle the issue of packet delay and loss in the heterogeneous mobile wireless network. The proposed scheduler is good to guarantee service availability and continuity quality of service (QoS) for disaster responders .Our approach also used for high resource utilization in the same network.**

Keywords **-** Interoperability, public safety land mobile radio, Packet delay, and loss.

I. Introduction

Disaster response and recovery require timely interaction and coordination of disaster responders in order to save lives and property. Efficient communications are crucial during disasters. With recent advances of wireless technologies, mobile wireless networks play an increasingly important role in disaster response. Currently, public safety land mobile radio (LMR) is used by public safety agencies for coordinating teams and providing rapid emergency response.

During disasters, efficient communications are crucial for disaster responders in disaster response and recovery. For example, it is desirable for the disaster responders to have the access to the Internet to share real-time multimedia information with off-site commanders and specialists providing expert assistance. However, these communication services are not available in the current public safety LMR. Whereas in commercial cellular networks, less service availability means less revenue; in public safety arena, less service availability may affect lives. Therefore, it is attractive to enhance the interoperability of these two wireless networks, by which a wide variety of benefits can be offered to disaster responders, including new multimedia services (e.g., video), increased user data rates and low cost devices.

A. *Interoperable Cellular and Public Safety LMR*

Natural disasters or terrorist attacks often occur in a localized region, we assume that the coverage of the LMR is under the coverage of the cellular network. The mobile devices used by disaster responders are equipped with multiple radio interfaces that enable them access both the LMR and the cellular network within the coverage of the LMR. However, for commercial users, only the cellular network can be accessed. IP-based multimedia services (e.g., video streaming) are available to disaster responders via the cellular network, and mission-critical services (e.g., tactical group voice) are provided to them via the LMR. Since disaster responders are free to move in the interoperable LMR/cellular systems, the support of handoff between these two networks, which provides ongoing service continuity, is needed in this integration. In this interoperable system, disaster responders are efficiently communicated with state-of-the-art applications In the interoperable cellular and public safety LMR networks, disaster responders can access the services in cellular networks that are not available in public safety LMR networks to increase the service availability. Furthermore, when a disaster responder moves out of the coverage of public safety LMR networks with an ongoing communication session, the session should be handoffed to cellular networks instead of being dropped to provide the communication continuity.

There are some schemes proposed for the interoperability of heterogeneous wireless networks. Authors of [6] propose a location management scheme, including location update and paging, in heterogeneous systems. The optimal conditions under which vertical handoff should be performed is studied in [7]. Authors of [8] and [9] study several admission control schemes in cellular/WLAN integrated networks to improve the performance of voice and data services. Scalable routing techniques are proposed for heterogeneous mobile networks. The commercial wireless cellular user community is two orders of magnitude larger than the public safety LMR base. Consequently, the R&D investments in commercial wireless cellular networks dwarf those made in public safety LMR networks.

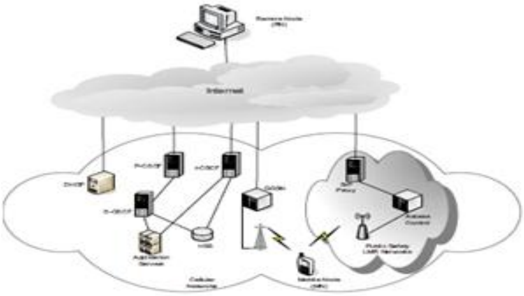


Fig. 1: Interoperable wireless cellular and public safety LMR networks

B. *State of the Art*

Many methods have been proposed to enhance the interoperability between cellular and LMR networks based on Session Initial Protocol (SIP) [11] and a joint radio resource management framework, which are different from the schemes in previous work [5]-[10]. SIP is designed by the Internet Engineering Task Force (IETF) to provide application-layer signaling for voice and multimedia session management, which can achieve true end-to-end mobility management. In addition, SIP has excellent extensibility and scalability due to its operation at the highest layer and use of text-based control messages. Several wireless technical fora (e.g., 3GPP, 3GPP2 and MWIF) have agreed to use SIP to provide session management. However, traditional SIP-based handoff scheme may have considerable handoff delays due to the exchange of application layer messages, which may be unacceptable for real-time multimedia services [12].
All of these are discussed with hand off delay, location management scheme, several admission control schemes and scalable routing techniques, while the need is reducing the packet delay and loss. There is critical need for an approach that is able to optimize the packet delay and loss to improve the multimedia services in interoperable heterogeneous mobile wireless network for achieving QoS.

II. RESEARCH APPROACH

This research work detects the problem of providing disaster response and recovery in interoperable Heterogeneous Mobile Wireless Networks. This goal is achieved by enhancing the radio resource management, reducing the packet delay and loss. We rigorously formulate this enhancing the radio resource management, reducing the hand off delay and the packet loss problem in the context of heterogeneous mobile wireless networks and present feedback based dynamic scheduler based on discrete time control theoretic approach that provide guarantees on the quality of service and the service availability.

A. *Feedback Dynamic Scheduler*

FDS and IDS algorithms will be designed using feedback control theory. We will assume that both algorithms, running at the HC, allocate the WLAN channel bandwidth to wireless stations hosting real-time applications, using HCCA functionalities. This allows the HC to assign TXOPs to ACs by taking into account their specific time constraints and transmission queue levels [13]. We will refer to a WLAN system made of an access point and a set of quality of service enabled mobile stations (QSTAs). Each QSTA has up to four queues. Let T_{CA} be the time interval between the starting of two successive CAPs.Every time intervals $T_{CA,}$ which is assumed to be constant, the HC must allocate the bandwidth that will drain each queue during the next CAP. We assume that at the beginning of each CAP, the HC is aware of all the queue levels.$q_{i,}$i=1..M, at the beginning of the previous CAP, where M is the total number of traffic queues in the WLAN.
The following discrete time linear model describes the dynamics of the queue:

$$q_i(n+1) = q_i(n) + d_i(n)T_{CA} + u_i(n)T_{CA}, \quad i = 1,\ldots,M \qquad (1)$$

where qi≥0 is the queue level at the beginning of the nth CAP; ui≤0 is the average depletion rate (i.e., its absolute value represents the bandwidth assigned to drain the queue);di(n)=dis-dicp(n) is the difference between dis(n)≥0, which is the average input rate at the queue during the nth TCA interval, and diCP(n)≥0, which is the average output rate at the queue during the nth TCAinterval. The input di(n) is unpredictable since it depends on the behavior of the source that feeds the ith queue and on the number of packets transmitted Thus, from a control theoretic perspective, di(n) can be modeled as a disturbance [25]. Without loss of generality, the following piece-wise constant model for the disturbance di(n) can be assumed:

$$d_i(n) = \sum_{j=0}^{+\infty} d_{0j} \cdot 1(n - t_j) \qquad (2)$$

Where 1(n)is the unitary step function. Due to the assumption (2), the linearity of the system described by (1), and the superposition principle that holds for linear systems, we will design the feedback control law by considering only a step disturbance: di(n)=d0 .1(n) [25].

B. *Closed-Loop Control Scheme*

Our goal is to design a control law that drives the queuing delay Ti, experienced by each frame going through the nth queue, to a desired target value TiT that represents the QoS requirement of the AC associated to the queue. We will consider the closed loop control system shown in Fig. 2, where the set point qiT has been set equal to zero, which means that we would ideally target empty queues

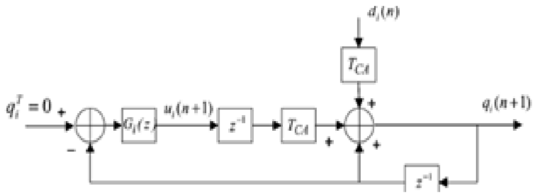


Fig 2: Closed loop control scheme

Regarding the transfer function $G_i(z)$of the controller, we will focus on two very simple controllers: a proportional (P) controller $G_i(z)=k_{pi}$, and a proportional-integral (PI) controller $G_i(z)=k_{pi.}(1+z/z-1,1/T_{Ii})$.

C. *Computational Complexity of the Bandwidth Allocation Algorithms*

Herein, we estimate the computational complexity of the proposed allocation algorithms.

Proposition 1: In a WLAN system with M active traffic streams, the computational complexity of the FDS algorithm is O(2M).

Proof: Every time interval TCA, the HC computes the bandwidth assignment for each one of the M active traffic streams. With FDS, from Fig. 2, the control law is

$$u_i(n+1) = -k_{pi} \cdot q(n). \qquad (3)$$

Thus, (3) becomes

$$\mathrm{TXOP}_i(n) = \beta_i q(n) + H(n) \qquad (4)$$

Where $\beta_i = k_{pi} T_{CA} / C_i$. As a consequence a single bandwidth assignment consists of two multiplications and one sum. The first multiplication takes into account the term $\beta_i q(n)$. the second one estimates the protocol overhead, which is proportional to $\beta_i q(n)$. Thus, we need *2M* multiplication plus *M* sums for each T_{CA} interval, i.e., the *O(2M).*

Proposition 2: In a WLAN system with M active traffic streams, the computational complexity of the IDS algorithm is O(4M).

Proof: For each active stream, the HC computes the bandwidth . When IDS is used, the control law is

$$u_i(n+1) = -k_{pi} \cdot q_i(n) - \frac{k_{pi}}{T_{Ii}} \sum_{h=0}^{n} q_i(h) \qquad (5)$$

which can be also written as

$$u_i(n+1) = u_i(n) + k_{pi} q_i(n-1) - k_{pi}(1 + 1/T_{Ii}) q_i(n). \qquad (6)$$

Thus, (6) becomes

$$\mathrm{TXOP}_i(n) = \frac{T_{CA}}{C_i}[u_i(n) + k_{pi} q_i(n-1) - \chi q_i(n)] + H(n) \qquad (7)$$

Where $\chi = k_{pi}(1 + 1/T_{Ii})$.

Now, considering that the overhead is estimated using 1 multiplication, a single bandwidth assignment consists of 4 multiplications and 3 sums. Consequently, we need *4M* multiplication plus a *3M* sums for each interval. Thus, the computational complexity is *O(4M)*.

From the above propositions, we can conclude that the computational complexities of both FDS and IDS scale linearly with the number of active streams. Thus, such schemes can be easily implemented in real wireless network interface cards.

III. EVALUATION AND PRELIMINARY RESULTS

Table I reports the average and peak superframe utilization in HCCA mode, which is defined as the sum of TXOPs allocated during CAPs over the superframe duration. It shows that the Simple scheduler requires the highest average quota of WLAN resources. The reason is that the simple scheduler does not adapt the quota of allocated resources to the actual load because it provides a CBR service.

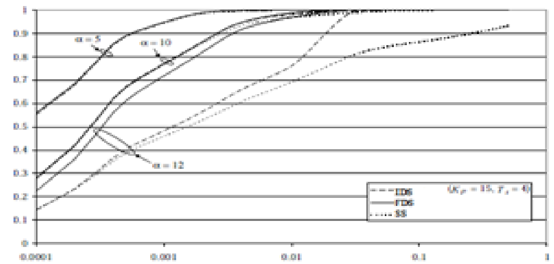


Fig 3: Packet delay

For the same reason, the peak superframe utilizations achieved by the Simple scheduler for α = 10 and α = 12, i.e., at high traffic load, are smaller than those provided by IDS and FDS. This result clearly highlights that the proposed control schemes enable more proper usage of the bandwidth and allows the bandwidth requirements of the real-time flows to be tracked.

α	IDS ($k_{pi} = 15s^{-1}, T_{I_i} = 4$)			FDS			SS		
	Superframe Utilization		Overall Goodput	Superframe Utilization		Overall Goodput	Superframe Utilization		Overall Goodput
	Average [%]	Peak [%]	FTP flows [Mbps]	Average [%]	Peak [%]	FTP flows [Mbps]	Average [%]	Peak [%]	FTP flows [Mbps]
5	6.8	29.5	4.57	6.95	31	4.58	25.62	32.42	4.57
10	15.2	80.2	4.33	15.22	82	4.38	52.4	62.58	4.27
12	29.3	80.9	2.52	17.66	80	4.09	63.75	74.98	2.32

Table I Results

IV. CONCLUSION

We have studied the interoperability problem in public safety LMR networks and commercial cellular networks for disaster response. The interoperability can be enhanced by using S-SIP and Optimal radio resource management. We have presented Feedback dynamic scheduler mechanism to maximize the overall radio resource utilization while guaranteeing the QoS such as reduced packet loss.

Further study is in progress to consider other QoS requirement in the interoperable cellular /LMR wireless networks.

V. REFERENCES

1) TIA,“Apco project 25 system and standards definition," TIA/EIA Telecomm. Syst. Bull., TSB102-A .

2) H. Holma and A. Toskala, WCDMA for UMTS: Radio Access for Third Generation Mobile Communications. New York: Wiley, 2004.
3) J. Gozalvez, “HSDPA goes commercial," IEEE Veh. Technol. Mag., vol. 1, pp. 43-53, Mar. 2006.
4) S. Frattasi, H. Fathi, F. Fitzek, R. Prasad, and M. D. Katz, “Defining 4G technology from the user’s perspective," IEEE Network, Jan. 2006.
5) Balachandran, K. C. Budka, T. P. Chu, T. L. Doumi, and J. H. Kang, “Mobile responder communication networks for public safety," IEEE Commun. Mag., pp. 56-64, 2006.
6) Assouma, R. Beaubrun, and S. Pierre, “Mobility management in heterogeneous wireless networks," IEEE J. Select. Areas Commun., vol. 24, pp. 638-648, Mar. 2006.
7) Stevens-Navarro, Y. Lin, and V. W. S. Wong, “An MDP-based vertical handoff decision algorithm for heterogeneous wireless networks," vol. 57, pp. 1243-1254, Mar. 2008.
8) W. Song, H. Jiang, and W. Zhuang, “Performance analysis of the WLAN-first scheme in cellular/WLAN internetworking," vol. 6, pp. 1932-1952, May 2007.
9) W. Song, Y. Cheng, and W. Zhuang, “Improving voice and data services in cellular/WLAN integrated networks by admission control," vol. 6, pp. 4025-4037, Nov. 2007.
10) T. Yagyu, M. Jibiki, and K. Yoshida, “SMART: scalable mobility adaptive routing techniques for heterogeneous mobile networks," in *Proc. IEEE WCNC’07*, 2007.
11) J. Rosenberg and *et al.*, “SIP: session initiation protocol," IETF RFC 3261, June 2002.
12) W.Wu, N. Banerjee, K. Basu, and S. K. Das, “SIP-based vertical handoff between WWANs and WLANs," *IEEE Wireless Commun.*, vol. 12 pp. 66-72, June 2005.

VI. Acknowledgements

The authors would like to express profound gratitude and heartfelt thanks to the Chairman, VSA Group of Educational Institutions, for his valuable support in this work. The authors would like to extend their sincere thanks to the Chairman, Vinayaka Missions University, for his encouragement to complete this piece of work.

Congestion Analysis of IEEE 802.11 Wireless Infrastructure Local Area Networks

GJCST Classifications: C.2.1, C.2.5

Amardeep Kaur; Lecturer, ECE department, Dehradun Institute of Technology, Dehradun, Uttrakhand, India

***Abstract*- In this paper, the possible effects of the problem of congestion in Wireless infrastructure LAN are discussed. It presents the discrete-event simulation, provides detailed, accurate network simulation results and it observed a wide variety of network statistics for Congestion Control in Wireless LAN. The software simulation package, OPNET (Optimized Network Engineering Tool), can be best described as a set of decision support tools, providing a comprehensive development environment for the specification, simulation and performance analysis of communication networks, computer systems and applications and distributed systems. Discrete event simulations are used as the means of analyzing the system performance and behavior. OPNET simulations were carried out to estimate the effect of congestion situation on the global performance of the network model. A tradeoff of various congestion parameters such as dropped data, load, throughput, retransmission attempts and received data traffic has been observed by creating different scenarios. Simulations are carried out at 11Mbps data rate is 900 simulations-seconds. There are six sections in this paper, section I deal with the introduction to the wireless LAN and the causes behind the occurrence of Congestion in general, section II deal with the problem of Congestion in Wireless environment, section III deal with the suggestions regarding the possible solution to the problem of congestion, section IV and V are to deal with the simulation and results and at last the paper is concluded by summarizing the important results.**

Keywords- Wireless LAN, IEEE 802.11, OPNET

I. INTRODUCTION

In Wireless LAN, Congestion is much more critical problem as compared to the Wired LAN because the error rate is much higher in Wireless LANs and it does not permit to allow a single collision to occur in the network, which will lead to the drastic reduction in throughput. Also unlike the wired networks, congestion measurement and analysis in IEEE 802.11 wireless networks is more difficult due to factors such as time-variant channel capacity, contention among neighboring nodes, interference variable quality of radio signals, transmitted power etc. Also detecting collision in wireless medium is not always possible. Congestion occurs when the amount of data sent to the network exceeds the available capacity, the routers are no longer able to cope up the demand and they begin losing packets. At very high traffic rate, the performance collapses completely, and almost no packets are delivered. Congestion can be brought about by several factors viz shortage of buffer space, slow links and slow processors [1-2]

A. *Shortage of Buffer Space*

If large capacity buffers are used in order to compensate for Shortage of buffer space, many short-term congestion problems will be solved but this will cause undesirably long delays

B. *Slow Links*

Though the problem of congestion caused due to slow links will be solved if high-speed links become available but this is not always the case, sometimes increases in link bandwidth can aggravate the congestion problem because higher speed links may make the network more unbalanced. Higher speed load can make the congestion condition in the switch worse [3-5].

C. *Slow Processors*

On improving the processor speed, faster processors will transmit more data in unit time. If several nodes begin to transmit to one destination simultaneously at their peak rate, the target will be overwhelmed soon.

II. CONGESTION IN WIRELESS ENVIRONMENT

Traditional problems of wireless communications and wireless networking are

1) The channel is unprotected from outside signals
2) The wireless medium is significantly less reliable than wired media
3) The channel has time-varying and asymmetric propagation properties
4) hidden-terminal and exposed-terminal phenomena may occur

In the event of packet loss, appropriate action is not easily taken, as identifying the cause of the loss is difficult.

There have been various mechanisms proposed to help classify the reason for packet loss, but all add extra complexity, may not be compatible with existing protocols and none seem to cover all possible causes [5-7].

III. POSSIBLE SOLUTIONS

There are two general solutions to the problem of congestion

1) Congestion avoidance
2) Congestion control

Congestion avoidance attempts to predict when congestion is about to occur and reduces the transmission rate at this time. The algorithm should operate in such a manner to keep response time v/s load and throughput v/s load operating to the left of the location of the knee in Fig 1.1.

Congestion control attempts to take fuller advantage of the network resources by transferring data at a rate close to the capacity of the network. The capacity of the network can be viewed as the point at which any increase in traffic will increase the delay but not the throughput. Congestion control algorithms, like that of TCP, attempt to increase traffic until the capacity of the network is reached, and then slow the transmission rate. Thus these algorithms attempt to operate to the left of the cliff in Fig. 1.1.

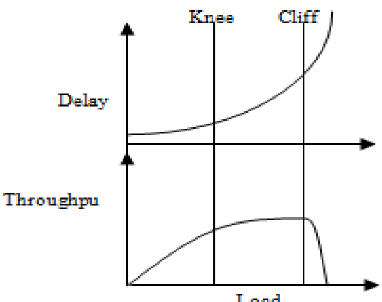


Fig 1.1 Load versus Delay and Throughput

IV. Simulation

The occurrence of a high density of nodes within a single collision domain of an IEEE 802.11 wireless network can result in congestion, thereby causing a significant performance bottleneck. Effects of congestion include drastic drops in network throughput, unacceptable packet delays, packet drops, retransmissions, and session disruptions. OPNET simulation was carried out to estimate the effect of congestion situation on the performance of the network model. Simulation was run at 11 Mbps data rate except congested node and total simulation time used was 900 simulations-seconds.

To observe the congestion in IEEE 802.11 networks, WLAN node model was created on OPNET IT guru academic edition 9.1 (Figure 1.2)

Table 1.1 a Parameter Setting of WLAN Network

WLAN environment	Campus
Workspace area	100m x 100m
Node model	wlan_station_adv

Table 1.1 b Parameter Setting of WLAN Network

Number of nodes	6 (PCF_WKSTN)
Access Point	1 (AP_0)

compared with periphery nodes in order to study congestion.

The packet size distribution is exponential with a mean of 92 bytes. The inter arrival time is exp (0.02) for all the nodes unless otherwise specified. Since the packet size is exponentially distributed with mean of 92 bytes, RTS/CTS exchange is required for most of the packets. All the wireless station nodes and the access point use Frequency Hopping Spread Spectrum at the physical layer. All the nodes employ the PCF basic CSMA/CA access mechanism. The nodes transmit at a maximum data rate of 2 Mbps. Packets received at a node with power less than 7.33 E-14 Watts will find receiver to be busy. The packet transmissions with a power higher than this threshold are considered as valid. Unless the default transmission power is changed, all the WLAN packets should reach their destinations with sufficient power to be valid packets if the propagation distance between the source and destination is less than 300 meters as required by the IEEE 802.11 WLAN standard.

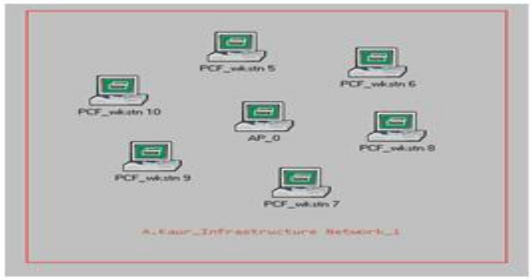


Figure 1.2 Node Arrangement

The distance between any two-periphery nodes is about 50 meters. In the simulation model considered here, all the nodes are static. The simulations were carried out for 900 simulation seconds and repeated many times in order to ascertain validity.

V. Results

The buffer size, bandwidth, and data rates of AP_0 have been reduced as compared to other nodes in order to study its impact on the performance of the network. Various global parameters and individual node parameters were observed. The global parameters were chosen as data dropped, load, throughput, and their variations against simulation time are shown in Figs 1.3, 1.4 and 1.5.

Individual node parameters were chosen as retransmissions attempts and data traffic received and are plotted in Figures 1.5 and 1.6.

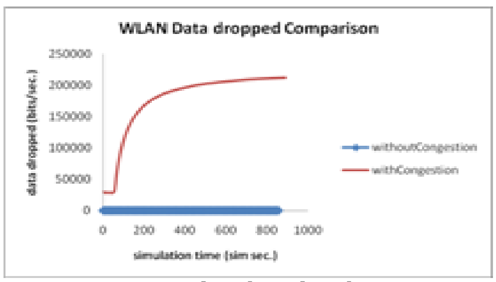


Fig-1.3: Data Dropped with and without Congestion

Observing the global data dropped (Fig. 1.3), it was observed that data dropped in the network is very high as compared to the situation when all nodes were having exactly similar buffer size, bandwidth, and data rates. Initially, it was estimated 22269 times the without congestion situation. Thereafter, it increases until 206562 for 900 simulation seconds.

Another global parameter chosen is load (Fig-1.4) of the network. It was observed that as compared to the situation when all nodes were having exactly similar attributes the load reduces to 54.7% initially which further dips to 14.2% for 900 simulation seconds duration

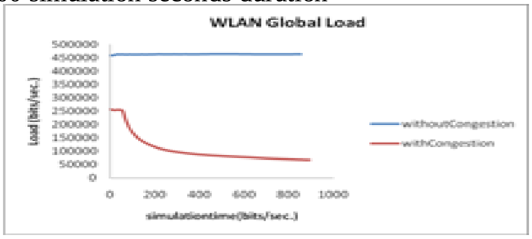


Fig-1.4: Global Load with and without Congestion

Another global parameter chosen is Throughput of the network (Fig-1.5). The throughput is constant throughout the simulation time of 900 simulation seconds. In that case it has been observed that in Congestion situation, the throughput is constant at 0, making a reduction of approximately 100% as compare to the situation of no congestion. That means in Congestion the throughput is totally zero

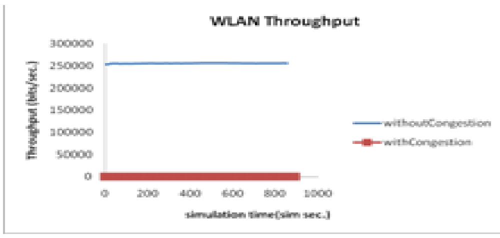


Fig-1.5: Global Throughput with and without Congestion

The impact of congestion situation on various individual nodes was also studied. The node parameters chosen for this study were Retransmission attempts and data traffic received. Retransmission is another important node parameter affected by congestion situation. The comparison of retransmissions of various nodes against simulation time is shown in Fig.1.6. The retransmission in PCF_WKSTN7 varies between (329-313). Likewise, the retransmissions values corresponding to other nodes-namely PCF_WKSTN10,PCF_WKSTN5, PCF_WKSTN6, PCF_WKSTN8 and PCF_WKSTN9 are varying between (294-305), (280-310), (287-308) (280-312) and (280-312), respectively. While that of AP_0 is maximum that varies between (524-617). That means the Congested node will have to do maximum number of retransmissions in order to receive any data packet.

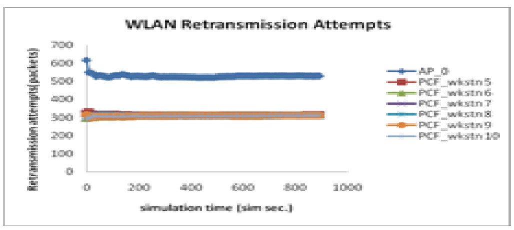


Fig-1.6:Node Comparison of Retransmission Attempts

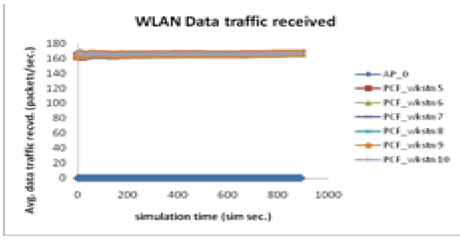


Fig-1.7: Node Comparison of Data Traffic Received

Variation of traffic received by various nodes against simulation time is shown in Fig.1.7. Traffic received in congested node (AP_0) is minimum as compared to other nodes. Likewise, the traffic-received values corresponding to other nodes- namely, PCF _WKSTN5,PCF _WKSTN6 ,PCF_WKSTN7, PCF_WKSTN8 PCF_WKSTN9 and PCF_WKSTN10 are varying between (161-166), (163-167), (163-166), (162-166), (163-166) and (163-168), respectively. While that of AP_0 the value is just 0. That means the Congested node does not receive any data traffic.

VI. Conclusion

In this paper, the performance of wireless infrastructure networks in terms of congestion has been studied through OPNET simulator and the results are presented for global as well as for the individual parameters. It was observed that global parameter viz. data dropped increases with congestion situation. However, the load and throughput reduces in similar situation. Individual node parameters such as the number of retransmission attempts is maximum in congested node as compared to the other nodes and received data traffic in congested node were found to be zero throughout the simulation duration of 900 simulation seconds.

VII. References

1) Habib, I.W., and Saadawi, T.N.," Controlling flow and avoiding congestion in broadband networks", IEEE comm.. Magazine, vol.30 no.10 pp. 46-53, oct.1991
2) Jain, R., "Congestion Control in Computer Networks: issues and trends", IEEE Network Magazine, Vol. 4, no.5 pp. 24-30, May 1990.
3) Tanenbaum, A.S.,"Computer Networks" , IIIrd Edition,Prentice hallof India Pvt.ltd.,New Delhi
4) "Performance of Infrastructure Mode Wireless LAN Access Network based on OPNET Simulator" Saeed A. Bawazir, Saleh H. Al-Sharaeh, Department of Computer Science,Normal, AL 35762,USA
5) Stallings, W.," Wireless Communication and Networks", Fifth Indian Reprint2004
6) www.opnet.com
7) IEEE 802.16-2004, IEEE standard for local and metropolitan area networks part 16: air interface for fixed broadband wireless access systems, IEEE, Oct. 1,2004
8) Gerla,M.,"High-speed local area networks" IEEE Spectrum 1991:26-31.

Histogram based Image Spam Detection using Back propagation Neural Networks

M. Soranamageswari and Dr. C. Meena

GJCST Classifications:
I.2.10, K.6.5, F.1.1, I.4.0, I.5.4

***Abstract*-In general words, image spam is a type of e-mail in which the text message is presented as a picture in an image file. This prevents the text based spam filters from detecting and blocking such spam messages. In our study we have considered the valid message as "ham" and the invalid message as "spam". Though there are several techniques available for detecting the image spam (DNSBL, Greylisting, Spamtraps, etc.,) each one has its own advantages and disadvantages. On behalf of their weakness, they become controversial to one another. This paper includes a general study on image spam detection using some of the well-liked methods. The methods comprise, image spam filtering based on File type, RGB Histogram, and HSV histogram, which are explained in the following sections. The finest method for detecting the image spam from the above-mentioned methods can be determined from the above study.**

Keywords- File Type, HSV Histogram, Image Spam, RGB Histogram

I. Introduction

Spam can be uttered as Unsolicited Bulk E-mail (UNBE). The most extensively predictable category of spam is e-mail spam. Most UNBE is designed for elicitation, phishing, or advertisement. E-mail spam is the practice of sending unwanted e-mail message through junk mail, frequently with commercial content, in large quantities to an indiscriminate set of recipients. A Spam message also holds its hand with Instant Messaging System. This Instant Messaging spam, which is also known as "Spim", makes use of instant messaging system. Mobile Phone Spam is directed at the text messaging service of a mobile phone. In the similar fashion spam targets on video sharing sites, real time search engines, online game messaging and so on.

In the mid 1990s when the internet was opened up for the general public Spam in e-mail started to become a problem. In the following years the growth of spam was exponential and today it comprises some 80-85 percent of all the e-mail in the world, by conservative estimate [3]. Spam messages have its wings stretch into all kinds of applications in recent years. More techniques are adopted by several service providers to eliminate spam messages and not all are noteworthy. All these techniques are losing their potency as spammers become more agile. The spammers have turned their towards image spam in the recent years.

The embedded image carries the target message and most email clients display the message in their entirety. Most of them emails also have similar properties as image-based emails; existing spam filters are no longer capable of detecting between image-based spam and image ham [1]. This provides a way for the spammers to easily foil the spam filters. The text messages embedded in all image spam will convey the intent of the spammer and this text is usually an advertisement and often contains text, which has been blacklisted by spam filters (drug store, stock tip, etc).

M. Soranamageswari is with LRG Govt Arts College for Women, Tirupur. (Email: swarnakannappan@rediff.com).
Dr. C. Meena is with the Avinashilingam University for Women in India. (E-mail: meena_cc@avinuty.ac.in;

II. Neural Networks

This image spam can be identified using various methods. By using supervised and semi, supervised learning algorithms in neural networks image spam can be detected. A neural network is a computational model or mathematical model that tries to simulate the structure and/or functional aspects of biological neural networks. An artificial neural network is an adaptive system that change its structure based on external or internal information that flows through the network during the learning phase. Thus, neural networks are non-linear statistical data modeling tool. Different types of artificial neural networks that can be trained to perform image processing are feed forward neural networks, Self-Organizing Feature Maps, Learning Vector Quantizer network. All these networks contain at least one hidden layer, with fewer units than the input and the output layers. In particular, the Back propagation neural network algorithm performs gradient-descent in the parameter space minimizing an appropriate error function. The Parameters like mode of learning, information content, activation function, target values, input normalization, initialization, learning rate, and momentum decide the performance of the back propagation neural networks. The back propagation neural network can be used for compression of various types of images like natural scenes, satellite images, and standard test images. In this paper, back propagation neural network is implemented to detect the image spam. Back propagation algorithm is a widely implemented learning algorithm in ANN. This algorithm implemented is based on error correction learning rule. The error propagation contains two passes through the different layers of the network, a forward pass and a backward pass. In the first case, the synaptic weights of the networks are fixed, whereas in the latter case, the synaptic weights of the network are adjusted in accordance with an error-correction tool. A neural network has wide range of applications. Their application areas include data processing including filtering, clustering, blind source separation and compression, classification including pattern and sequence recognition, novelty detection and sequential decision making and function approximation or regression analysis, including time series prediction, fitness approximation and modeling.

III. RELATED WORK

Many discussions have been carried out previously on image spam detection. Under this section, we have an overview of the literature.

Marco Barreno et al., in [2], explains the different types of attacks on the machine learning algorithms and the systems, a variety of defenses against those attacks, and the ideas that are important to secure the machine learning. This approach illustrates the methods that spammers handle to attack a system to design an image spam. The issue of machine learning security goes beyond intrusion detection systems and spam e-mail filters. The different measures of defenses involved in their discussion are robustness, detecting the attacks, disinformation, randomization for targeted attack, and cost of countermeasures.

A modification of Latent Dirichlet Allocation (LDA), Known as multi-corpus LDA technique was introduced by Istvan Biro et al., in [4]. In their proposal, they created a bag-of-words document for every web site and run LDA both on the corpus of sites labeled as spam and as non-spam. This assisted them to collect spam and non-spam topics during their training phase. They implemented these collections on an unseen test site to detect the spam messages. This method in combination with web spam challenge 2008 public features, and the connectivity sonar features is used to test images. Using logistic regression to aggregate these classifiers, the multi-corpus LDA yields an improvement of around 11 percent in F-measures and 1.5 percent in ROC.

Spam web page detection through content analysis is put forth by Alexandros Ntoulas et al., in [11], which projected some earlier undefined techniques for automatic spam message detection. They also discussed the effectiveness of those techniques in isolation and when aggregated using some classification algorithms, which proved to be truth worthy in detecting the image spam..

Bhaskar Mehta et al., in their paper [2], describe two solutions for detecting image-based spam after considering the characteristics of image spam. The first utilized the visual features for classification, and offers an accuracy of about 98 percent, i.e. an improvement of at least 6 percent on comparison with the existing solutions. Second, they used SVMs to train classifiers using judiciously decided color, texture, and shape features. This approach helped them in dtecting near duplication in images. The strategies for Image spam detection discussed in their work are near-duplicate detection in images, visual features for classification, an algorithm for classification of visual features, optical character recognition (OCR).

Clustering based spam detection is put forth by Chun Wei et al., in [23], which propagates a fuzzy-matching algorithm to group subjects found spam emails, which are generated by malware. The subjects similar to each other are found out using a dynamic programming. The main proposal is that the recursive seed selection strategy allows the algorithm to detect similar patterns even the spammer creates a variation of the original pattern. This proved to be an effective approach in detecting and grouping spam emails using templates. Clustering algorithm is utilized to find the similarity of strings, similarity of spam subjects and for clustering spam subjects.

Seongwook Youn and Dennis McLeod in [24], describes the method of filtering gray e-mail using personalized ontologies. Their work in [24], explains a personalized ontology spam filter to make decisions for gray e-mail. Gray e-mail is a message that could reasonably be estimated as either spam or ham. A user profile has been created for each user or a class of users to handle gray e-mail. This profile ontology creates a blacklist of contacts and topic words.

A. Image Spam Classification Based Of Text Properties

Image Spam classification based on text properties includes finding the location of texts in images or videos. Texts are usually designed to attract attention and to reveal information. The Connected Component based (CC) and the texture-based approaches are the two leading approaches used in the past to extract the characteristics for the text detection task. These characteristics include coherence in space, geometry and color [5]. In CC-based methods [6], the image is segmented into a set of CCs and is grouped into potential text regions based on their geometric relation. These potential regions are then examined using some rule-based heuristics, which makes use of the characteristics like size, the aspect ratio and the orientation of the region. The efficiency of this method becomes questionable when the text is multi-colored, textured, with a small font size, or overlapping with other graphical objects.

In texture-based methods [7], it is assumed that the texts have distinct textural properties, and this can be used to distinguish then from the background. Even though this method perform well for images with noisy, degraded, or complex texts and/or background it seems to be time-consuming as texture analysis is essentially computational intensive.

An increase in the use of internet in the recent years, had led to tremendous growth in volumes of text documents available on internet. Accordingly, the management and organization of text has become an important task. So a set of predefined categories of these text documents known as Text Categorization is maintained. A number of machine learning algorithms such as K-nearest Neighbor, Centroid classifier, Naive Bayes (NB), Winnow and Support Vector Machines (SVM) are extensively used to deal with Text Classification. OCR technique is used to isolate text from image. Carrying out semantic analysis of text embedded into images attached to e-mails first requires text extraction by OCR techniques.

Naïve Bayes is a simple classifier most commonly used in pattern recognition, while it has the assumption that the feature attributes are independent, the accuracy of the Naïve Bayes classification is typically high [8]. Support Vector Machine (SVM) has been widely applied to most practical applications because of its superiority in handling high dimensional data. The parameter tuning [9, 10], and the Thresholding [11, 12], are the two common techniques that are applied prior and posterior to the SVM algorithms respectively.

B. Image Spam Classification Based On Content

One popular practice when creating spam pages is "Keyword Stuffing". [13] In Content based image spam detection we investigate whether an excessive number of words within a web page (excluding markup) is a good indicator spam. In the next step, have to determine whether there is excessive appearance of keywords in the title of a page. Uncommon practice that was observed in manually tagged data set is the use of "composite" words in spam pages.

Content-based Naive Bayes (PGRAM) is another technique for the classification of Image spam. In [14], the task of spam detection has floated the idea of a partial Naïve Bayes approach, biased towards low false positive rates. It also uses word tokens, but filters out predefined common tokens.

The content and the header of the incoming e-mail is mostly analyzed by the available anti-spam techniques [15]. They try to infer something about the kind of the material contained in the message by looking for specific pattern typical of a spam message. For these reasons, these filters are known as "content based." There are many anti-spam techniques available that falls under this category.

Blacklist and White list filters check whether the incoming message is from a known and trusted email address. Rule based filters correlate a score to every incoming email calculated according to a set of rules based on typical features of spam messages (fake SMTP components, Keywords, HTML formatting, etc) [16]. In case the score exceeds the given threshold value it is recognized to be a spam message. Major problem in this method is that, since its semantics are not well defined, it is difficult to aggregate rules and ascertains a threshold that limits the number of false positives. Spam Assassin [17], results from the successful implementation of the above-mentioned technique.

IV. PROPOSED METHODOLOGY

In this work, we proposed two new image spam classifiers based on file properties and histogram of an image. The proposed techniques can be explained in the following sections.,

A. Image Spam Detection Based On Their File Type

One method of detecting the image spam is based on their file type. Image spam e-mails will mostly contain images in JPEG or GIF file types. The basic features (see table1.) that can be derived from an image at an extremely low computational cost are the width and the height denoted in the header of the image file, the image file type and the file size. In this study, we focus on the all file formats that are commonly seen in e-mail, which are the Graphics Interchange Format (GIF), and the Joint Photographic Experts Group (JPEG) format, Bitmap (BMP) and Portable Network Graphic (PNG). By parsing the image headers of the image files using a minimal parse a general idea of the image dimensions (width and the height), can be obtained; as this does not involve any decompression or decoding on any actual image data the dimensions can be obtained rather faster.

In the case of GIF files there will be presence of virtual frames [18], which may be either larger or smaller than the actual image width. And this issue can be detected by decoding the image data. The problem imposed in the case of the corrupted images is that the lines near to the bottom of the image will not decode properly. Any further decoding of the image data from that point of corruption will be decisive. In order to obtain the amount of information that we gain from above features, the signal to noise ratio is defined. It is calculated as the distance of the arithmetic means of the spam and ham classes divided by the sum of corresponding standard deviation.

$$S2N = \left| \frac{\mu_{spam} - \mu_{ham}}{\sigma_{spam} + \sigma_{ham}} \right|$$

Where, μ_{spam} is the Mean value of the spam,

μ_{ham} is the Mean value of ham,

σ_{spam} is the standard deviation of spam,

σ_{ham} is the standard deviation of ham.

Table 1. Image Features

Features	Description
f1	Image width denoted in header
f2	Image height denoted in header
f3	Aspect Ratio: f1/f2
f4	File Size
f5	File Area: f1.f2
f6	Compression: f5/f4

This feature analysis reveals a fact that the binary features reflect the percentage of images in the respective formats. The feature f6 is the most informative feature beyond the binary image format feature. Most legitimate images in e-mails ("ham") are JPEG images. The f3 is the aspect ratio of the image (i.e.) f1/f2. The feature f6 captures the amount of compression achieved by calculating the ratio of pixels in an image to actual image size. The compression is better if more number of pixels is stored per byte. This stimulates us to classify image spam with a similar supervised learning idea like Data Modeling.

B. Image Spam Detection Rgb Histogram

Image spammers implement different randomization techniques to introduce noise into spam images. This makes the single feature not able to detect all the variations introduced into an image. Hence, color histogram filter can be used to detect image spam in this case [19]. Spam image formation algorithms are intended to defeat well-known vision algorithms like OCR (Optical Character

Recognition). The color histogram is a trouble-free feature and can be calculated very effectively by one simple pass of the whole image.

A 64-dimensional color histogram can be used in a RGB color space. The values in each of the color space (R, G, and B) can be divided into 4 bins of equal size resulting in 4X4X 4 = 64 bins in total. For each bin, the amount of color pixel that falls into that particular bin is calculated. Finally, it is normalized so that the sum equals to one. The distance between the two color histogram features can be determined using L1 distance. For D-dimensional, real-valued feature vectors of an image, the L1 distance of the pair of points $X = (X_1,....,X_D)$ and $Y = (Y_1,....,Y_D)$ has to form [19]:

$$d(X,Y) = \sum_{i=1}^{D} |X_i - Y_i|$$

Frankel et al in [20], quantifies color saturation as the fraction of total number of pixels in the image for which difference max(R, G, B) – min(R, G, B) is greater than some threshold value T. The threshold value can be set by the evaluator. This fraction is evaluated for both text and non-text regions on the image. This leads to two color saturation features.

C. *Image Spam Detection Using Hsv Histogram*

The HSV color space is fundamentally different from the widely known RGB color space since it separates out the intensity from the color information. The HSV stands for the Hue, Saturation, and Value based color model. The Value represents intensity of a color, which is decoupled from the color information in the represented image.

A three dimensional representation of the HSV color space is a hexacone, in which the central vertical axis represents the intensity [21]. Hue defines the angle relative to the red axis. Similarly, saturation is the depth or purity of the color and is measured from the radical distance from the central axis with value between 0 at the center to 1 at the outer surface.

The intention of the spammers is that the spam messages designed by them should be easily noticeable by the users. Hence, it is obvious that the spammers use highly contrasting colors to design their spam messages. This property is defined as "Conspicuousness" meaning "Obvious to the eye". This histogram is converted into three bins and passed into neural networks and their epoch value is set to 300 and the goal in BPNN is set to 0.001.

V. Experimental Results

In order to design and evaluate our spam detection algorithms, we used a collection of 5000 random images from spam archive dataset. Accuracy (A), Precision (P), and Recall (R), are some of the well-known performance measures. If the value of precision is high, it obviously indicates that the false negative is high. In other words, the detector has misclassified many spam messages as ham message. On the other hand, a high recall indicates that the false positive is high, i.e. many legitimate messages (ham) are misjudged as spam. We concern about the trade-off that exists between the spam and ham when we consider precision and recall values.

These measures are defined below and used in this study.

$$Accuracy(A) = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

TP is the number of e-mail that is spam and correctly predicted as spam; FP is the number of e-mail that is legitimate but predicted as spam; TN is the number of e-mail that is legitimate and is truly predicted as legitimate (ham); and FN the number of e-mail that is spam but predicted as legitimate.

Table 2. Confusion Matrix

Prediction	Observer	
	Legitimate	Spam
Legitimate	TN	FN
Spam	FP	TP

Table 3. Shows the comparison of the Accuracy (A), Precision (P), and Recall (R) for different approaches of spam detection. The approaches being spam detection based on file properties, RGB histogram, and HSV histogram.

Approach	Accuracy (A)		Precision (P)		Recall (R)	
	Ham	Spam	Ham	Spam	Ham	Spam
File properties	90.5 %	86.6 %	84.5 %	80.6 %	88.3 %	85.7 %
RGB histogram	94.6 %	92.1 %	88.7 %	84.1 %	90.5 %	89.6 %
HSV histogram	96.5 %	95.4 %	90.5 %	88.7 %	92.0 %	91.4 %
Combination of RGB and HSV histogram	99.3 %	99.1 %	98.3 %	95.5 %	96.8 %	95.9 %

Table 3. Comparison of Accuracy, Precision, Recall

Based on spam detection with the help of file properties, the signal to noise ratio of the GIF Images and the JPEG images are tabulated below. The Table. 4 [18] evidently illustrate the Signal to Noise ratio for calculated for spam and ham messages that were of GIF format.

In the similar way Table. 5 [18], explains the calculation of Signal to Noise ratio for detecting the image spam for JPEG format only. This demonstrates the signal to noise ratio for different features that were mentioned in Table. 1. Based on the Signal to Noise ratio obtained for different features of an image it is possible to isolate spam message from the ham message. This minimizes the rate of false positive obtained.

Feature	S2N	μ_{spam}	μ_{ham}	σ_{spam}	σ_{ham}
f1	0.188	519.2	257.0	176.4	1216.5
f2	0.143	356.3	165.1	128.7	1208.7
f3	0.043	1.76	53.8	4.58	1206.1
f4	0.100	15269.6	29347.1	13459.1	127587.5
f5	0.767	195339.6	42098.9	107180.16	92658.9
f6	0.524	16.97	5.00	10.4	12.50

Table 4. Feature Quality (GIF Only)

Feature	S2N	μ_{spam}	μ_{ham}	σ_{spam}	σ_{ham}
f1	0.289	422.08	618.40	133.16	546.64
f2	0.308	305.50	496.66	129.20	491.59
f3	0.040	2.05	2.12	2.005	14.98
f4	0.272	21601.06	203686.40	12787.30	655880.90
f5	0.323	127524.60	539062.50	71339.82	1202866.95
f6	0.265	6.70	4.82	3.90	3.15

Table 5. Feature Quality (JPEG Only)

Fig. 2 shows the comparison chart of different histogram based approaches in determining the image. The comparison of precision and recall value is shown below.

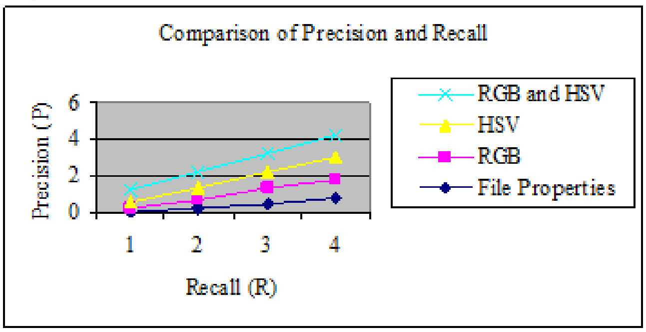


Fig. 2 Comparison of precision and recall values for different types of HSV and RGB histogram generation.

VI. Conclusion

This paper reveals a general study on Image spam, classification of image spam on the basis of text properties and content properties, and some of the methodologies in detecting the image spam. The detection of image spam using their file properties seems to be an effective method in detecting the spam. This method eliminates only 80 percent of the spam messages and this makes the method not suitable for most of the cases. The spam messages need further filtering after the file type detector to completely eliminate the spam e-mails. The second approach of image spam detection using histogram seems to be advanced method of the first described. Since, this method implements the distance measurement it seems to be more convenient in detecting spam than the former approach. The latter discussed HSV histogram method of image spam detection is the most advanced method in eliminating the spam messages. This method utilizes the color moments to determine the saturation level of the contrasted colors. This seems to be effective in spam detection. This method minimizes the low false positive rate to minimum. HSV histogram approach provides improved performance in detecting the image spam than the methods of spam detection using their file type and color histogram. HSV based histogram provides better performance at varying brightness and contrast settings.

VII. References

1) Bhaskar Mehta, Saurabh Nangia, Manish Gupta, and Wolfgang Nejdl, “Detecting Image Spam using Visual features and near duplicate detection,” WWW 2008.
2) Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph and J. D. Tygar, “Can Machine Learning Be Secure?” ASIACCS 2006.
3) http://www.maawg.org/about/MAAWG20072Q_Metrics_Report.pdf
4) Istvan Biro, Jacint Szabo, and Andras A. Benczur, “Latent Dirichlet Allocation in Web Spam Filtering,” AIR web 2008, April 22, 2008 Beijing, China.
5) Qiang Zhu, Mei-Chen Yeh, and Kwang-Ting Cheng, “Multimodal fusion using learned text concepts for image categorization”.
6) S. Loncaric, “A survey of shape analysis techniques,” Pattern Recognition, vol. 31, pp. 983-1001, 1998.
7) I. Kim, K. Jung, and J. H. Kim, “Texture-based approach for text detection in images using Support vector machine and continuously adaptive mean shift algorithm,” IEEE Transaction on Pattern Analysis and Machine Intelligence, 2003.
8) P. Domingos and M. Pazzani, “On the Optimality of simple Bayesian classifier under zero-one Loss”. Machine Learning, 1997.
9) C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
10) Kolcz and J. Alspector, “SVM-based filtering of e-mail spam with content-specific misclassification costs,” In proc. of textDM, 2001.

11) H. Drucker, D. Wu, and V. Vapnik, "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, 10(5), 1999.
12) Sculley and G. Wachman, "Relaxed online support vector machines for spam filtering" In proc. of SIGIR, 2007.
13) Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly, "Detecting spam web pages through Content Analysis."
14) P. Graham, "A plan for spam," 2003.
15) Giovanni Donelli, "Email Interferometry."
16) W. W. Cohen, "Learning Rules that Classify E-Mail," AT & T Laboratories, 1996.
17) Apache Group, SpamAssassin, http://spamassassin.apache.org, last visited May, 2005.
18) Sven Krasser, Yuchun Tang, Jeremy Gould, Dmitri Alperovitch, and Paul Judge, "Identifying Image Spam based on Header and file properties using C4.5 decision trees and support vector machine," IEEE Workshop on Information Assurance 2007.
19) Zhe Wang, William Josephson, Qin Lv, Moses Charikar, and Kai Li, "Filtering image spam with near-duplicate detection."
20) Frankel, M. Swain, and V. Athitsos, "Webseer: An Image search engine for the World Wide Web," Univ. of Chicago Technical Report TR96-14, 1996.
21) Stockman and L. Shaprio, "Computer vision," Prentice Hall, 2001.
22) Y. Rui, T. S. Huang, and S. –F. Chang, (1999), "Image retrieval: Current techniques, Promising Directions, and Open issues," Journal of Visual Communications and Image Representations, vol. 10, no. 4, pp.39-62, 1999.
23) Chun Wei, Alan Sprague, and Gary Warner, "Clustering Malware-generated Spam Emails with a novel Fuzzy String Matching Algorithm," SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.
24) Seongwook Youn and Dennis McLeod, "Spam Decisions on Gray E-mail using Personalized Ontologies," SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A.

Fuzzy Congestion Control Scheme in ATM Networks

GJCST Classifications: C.2.1, C.2.5, I.2.3

Parag Jain Doctoral Candidate, Bhagwant Univ. Prof. & Head, RIT Roorkee
Sandip Vijay MIEEE, Fellow ACEEE Prof. & Head, DCMTE
S. C. Gupta Professor Emeritus I.I.T. Roorkee

Abstract- **Fuzzy approach to control congestion in ATM networks is inevitable in research areas. . A control scheme that dynamically regulates traffic flow according to changing network conditions however requires the understanding of network dynamics. To minimize congestion, for a gradual change we proposed fuzzy approach. In our scheme, burst length as well as buffer occupancy are represented by triangular membership functions of fuzzy sets. However, these improvements are achieved at the cost of higher time complexity.**

I. INTRODUCTION

A Major development in high-speed networking is the emergence of B-ISDN's and ATM. ATM has been designed to support various classes of multimedia traffic with different bit rates and QoS requirements. Due to the unpredictable fluctuations and burstiness of traffic flow within multimedia networks, congestion can occur frequently. Therefore, it is necessary to design appropriate congestion control mechanisms to ensure the promised QoS is met. Shift in the network's performance bottleneck from channel transmission speed to propagation delay of the channel and the processing speed at the network switching nodes [1].

Consequently, congestion prevention can be interpreted as the problem of matching the admitted traffic to the network resources. This, in turn, could be viewed as a classical problem of feedback control i.e. matching the output to the input of dynamical systems [2]. In feedback controls, when possible traffic congestion is detected at any network element, feedback signals are sent back to all sources. ATM layer congestion control refers to the set of actions taken by the network to minimize the intensity, spread, and duration of congestion. Feedback flow control is one of the solutions which has been reported in the literature [3],[4],[5].

The growing success of fuzzy logic in various fields of applications, such as control, decision support, knowledge base systems, data base information retrieval and pattern recognition, is due to its inherent capacity to formalize control algorithms that can tolerate imprecision and uncertainty, emulating the cognitive processes that human beings use every day[6],[7],[8]. Fuzzy logic system have been successfully applied to deal with congestion control related problems in ATM networks and have provided a robust mathematical frame work for dealing with real world imprecision [9],[10]. The fuzzy approach exhibits a soft behavior, which means a greater ability to adapt itself to dynamic, imprecise, and bursty environments. Comparative studies [11] have shown that the fuzzy approaches significantly improve system performance compared with conventional approaches.

In conventional schemes, a binary threshold divides the buffer space in two parts: below or equal to the threshold level, for every arriving cell is given entry to the network and above the threshold every cell is rejected. In fixed threshold case as described by Bonde et. al. [11], two states of buffer - block and admit can be replaced by fuzzy sets. We have proposed the use of fuzzy logic for dynamic feed-back threshold scheme. In applied fuzzy scheme, burst length as well as buffer occupancy are represented by triangular functions.

II. FUZZY EXPERT SYSTEM

Fuzzy logic provides a general concept for description and measurement. Unlike traditional Aristotelian two-valued logic, in fuzzy logic, fuzzy set membership occurs by degree over the range [0,1], which is represented by a membership function. The function can be linear or non-linear.

A. From fuzzy Set to Fuzzy Events

Fuzzy set theory, compared to other mathematical theories, is perhaps the most easily adaptable theory to practice. The main reason is that a fuzzy set has the property of relativity, variability, and inexactness in the definition of its elements. Instead of defining an entity in calculus by assuming that its role is exactly known, we can use fuzzy sets to define the same entity by allowing possible deviations and inexactness in its role. This representation suits well the uncertainties encountered in practical life, which make fuzzy sets a valuable mathematical tool.

III. MODEL OF FUZZY CONTROLLER

Fuzzy systems are defined with a strong mathematical basis, which are rule-based systems. A fuzzy system is made of a fuzzifier, a defuzzifier, an inference engine, and a rule base as shown in Fig. 1. The role of the fuzzifier is to map the crisp input data value to fuzzy sets defined by their membership functions depending on the degree of "possibility" of the input data. The goal of the defuzzifier is to map the output fuzzy sets to a crisp output value. It combines the different fuzzy sets with different degrees of possibility to produce a single numerical value.

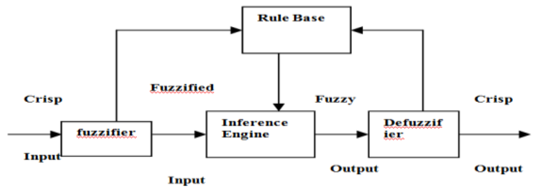


Fig.1: Model of Fuzzy Controller

The fuzzy inference engine defines how the system should infer through the rules in the rules base to determine the output fuzzy sets. Ray-Guang Cheng et. al. developed a model of a fuzzy traffic controller in which inputs linguistic variables are chosen so that the controller is a closed-loop system with the stable and robust operation.

The heart of a Fuzzy system is a rule base, which consists of a set of If-Then rules. The rules are statements in which some words are characterized by continuous membership functions. For example, IF the link is close to congestion THEN reduce the input rate, the words close to congestion are characterized by a membership function as shown in the Figs.2 (a) and 2(b), where congestion is considered happening when the link utilization is above 0.8.

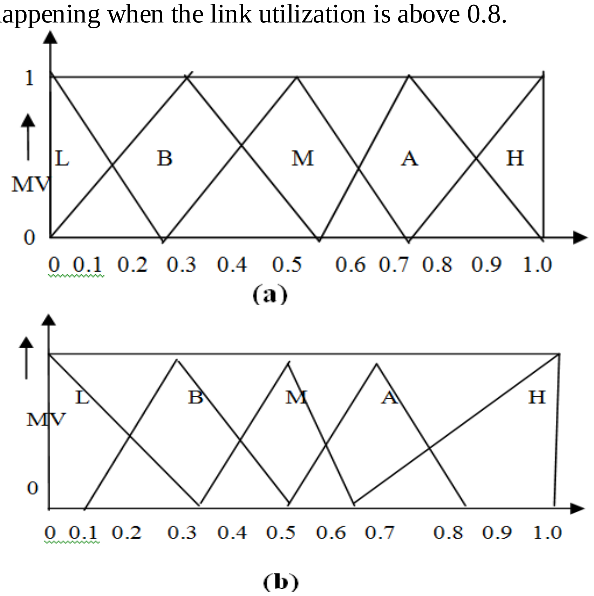


Fig. 2: A Typical Representation of Buffer Occupancy as well as Burst Length by Fuzzy Sets

L, B, M, A and H represent Low, below medium, Medium, above medium and High membership sets respectively. M.V. represents membership values.

The fuzzy system encodes expert knowledge about the system to be implemented rather than modeling the actual system; therefore it resembles a rule based expert system. However, unlike expert system fuzzy system does not fail when faced with a control situation in which no rule is defined. Instead, controls are inferred using the membership function to generate approximate control actions.

IV. Applied Fuzzy Approach

For applied scheme out-put buffer divided into various number of equal parts viz. two, three, and four for this purpose, then the feedback had applied after 50%, 33%, and 25% completion of the buffer space i.e. when N=2, 3, and 4 respectively. Depending upon which threshold has been crossed, the network gets a *mild warning*, ---------or *an ultimatum*. A gradual change is more intuitive here; this has been incorporated with fuzzy logic. In applied fuzzy scheme, burst length as well as buffer occupancy are represented by triangular functions as shown in Fig. 2. The degree of membership of a particular set, associated with each valid buffer occupancy can be read from this figure. This quantification of membership is called fuzzification. From these membership values and corresponding sets, blocking to be offered, again in fuzzy terms can be find out. This process is called rule-based inference. As an example, a typical rule is when buffer occupancy is high and burst length is high, number of blocked cells is also high as shown in Lookup Tables 3.1(a), and 3.1(b).

Blocking

BL	L	B	M	A	H
L	L	L	L	B	B
B	B	B	M	M	M
M	B	M	M	A	H
A	M	M	A	H	H
H	H	H	H	H	H

Table 3.1 (a): Lookup Table

BL	L	B		A	H
L	L	L		M	M
B	L	B	B	M	A
M	B	B	M	A	A
A	M	M	A	H	H
H	M	A	H	H	H

Blocking

Table 3.1(b) : Lookup Table

Then, by applying suitable defuzzification method, the percentage blocking to offered at that particular buffer occupancy level and at given burst length can be determined. For defuzzification, with the set such as shown in the Table 3.2, weighted Average is used.

SET	L	B	M	A	H
I	0.05	0.25	0.50	0.75	0.95
II	0.05	0.20	0.40	0.60	0.80

Table 3.2: Defuzzification Table

L- Low Set, B: Below Medium Set, M: Medium Set, A: Above Medium Set, H: High Set

A typical example is explained as follows: Let us assume that buffer occupancy as well as burst length both are characterized by the fuzzy set described in Fig. 2(a). Also, maximum buffer size is kept at 8 and maximum burst length is assumed to be 8. Suppose, at the time of the new arriving cell burst, buffer occupancy = 5 and arriving burst length = 6. When normalized with respect to maximum value of 8, these variables are mapped as buffer occupancy =0.625 and burst 0.75. Using fuzzy set of Fig. 2(a) for fuzzification it is seen that, buffer occupancy is a member of set M with associated value 0.5 and a member of set A with associated value 0.5. Burst length is a member of set A with associated value 1.0 and a member of set H with associated value 0.0. Using Table 3.1(a) and min-max method of evaluation, we get:

Buffer occupancy M(0.5) and burst length A(1.0)=>Blocking of A(0.5)

Buffer occupancy M(0.5) and burst length H(0.0)=>Blocking of H(0.0)

Buffer occupancy A(0.5) and burst length A(1.0)=>Blocking of H(0.5)

Buffer occupancy A (0.5) and burst length H(0.0)=>Blocking of H(0.0)

Thus, taking maximum of the four values associated with H, blocking has membership of set A with value (0.5) and membership of set H with value (0.5). Using these sets with weighted average of membership values, percentage blocking offered can be found. For defuzzification, set 1 of Table 3.2 is used.

$$\text{Blocking}=\frac{[(0.5\times 0.75)+(0.5\times 0.95)]}{(0.5+0.5)} = 0.85$$

Thus the percentage blocking to be offered, as per the proposed scheme is 85%. Based on this method of determining percentage blocking for the incoming cells, an ATM node is simulated, and performance of the scheme has been compared with static and dynamic feed-back schemes.

V. Results

The simulation results are shown in the Tables 1-6 and Graphs 1-6 indicated that the over all performance of the ATM switch improved when we applied fuzzy logic to Dynamic Feed-back Threshold scheme. In this work, link bandwidth is taken as 155.5 Mbps. So minimum delay suffered by a cell is 2.827 µs. Each input VBR source i, i=1, 2,……., N is modeled by two state ON-OFF Interrupted Bernoulli Process (IBP). We first considered switch of size 10×10, with input length = 4 and output length = 8. We had applied a constant threshold (C.Th.) = 4, the size of the output buffer (Bop) is kept 10.

Out-put buffer had divided into equal number of parts viz. two, three, and four. The feedback had applied after 50%, 33%, and 25% of the buffer space gets filled i.e. when N =1, and 2 respectively under Dynamic Threshold Feed-back (D.Th.Fb.) scheme. Simulation results are taken for these values of N, after applying different load conditions. A gradual change is more intuitive here, this has been done with fuzzy logic in our proposed Fuzzy Feed-back (F.Fb.) scheme.

The results are obtained for the three important performance indices i.e. throughput, average cell delay and cell loss probability Vs load. The performance of the new proposed scheme has been compared with Constant Threshold and Dynamic feed-back Threshold based schemes. From the results we observed that all the QoS parameters as described above are the function of offered load and number of buffer parts(N), but for Constant Threshold Scheme these parameters don't depend on the value of N.

For low loads ($L \leq 0.5$) all the schemes provide about the same throughput (100% - 99%). Which shows that all the incoming cells are served by the switch, so we will limit our discussion to higher loads. For moderate loads ($0.5 \leq L \leq 0.7$), due to rigidness of the Constant Threshold the throughput decreases from 99% to 97%, but the remaining schemes again have same results. The reason is that since after the completion of every 50%, 33% and 25% of the buffer space the network gets a proper signal to control the incoming burst of cells. At higher loads ($0.7 \leq L \leq 0.9$), the throughput decreases up-to 96% for Constant Threshold Scheme. At these load conditions the value of throughput increases gradually for Dynamic Feed-back Scheme with respect to N, while remains constant for proposed Fuzzy Scheme.

The value of average cell delay for Constant Threshold Scheme increases rapidly as offered load changes from lower to higher. This parameter is again doesn't depend upon value of N. But the value of average cell delay is very low for Dynamic Feedback Threshold Scheme. For proposed Fuzzy Scheme it is minimum.

Like average cell delay, the results show that Proposed Fuzzy Scheme has minimum value of average cell loss probability too. The reason can be explained as follows:

The threshold function determines, for each cell-burst, how many of the arriving cells to admit into the buffer. This function bears significant influence on the performance of the network including the fraction of cells lost due to dropping or excessive delays and the delay distribution of the cells. The traditional 'fixed' scheme utilizes a binary threshold: admit or no-admit, depending on the occupancy of the buffer. In the proposed Fuzzy Scheme, blocking decision is based on triangular membership function.

Table-1: Ave.Throughput Vs Load,When N=1

Load	*C-Th*	*D-Th-FB*	*F-FB*
0.1	1	1	1
0.3	1	1	1
0.5	1	0.993	0.9931
0.7	0.9733	0.9729	0.973
0.9	0.9607	0.9606	0.9729
	C.Th.= Constant Threshold		
	D.Th.-Fb.=Dynamic Threshold Feed-back		
	F.Fb.=Fuzzy Feed-back		

Graph-1 :Ave. Throughput Vs Load for Switch size 10 10, N=1

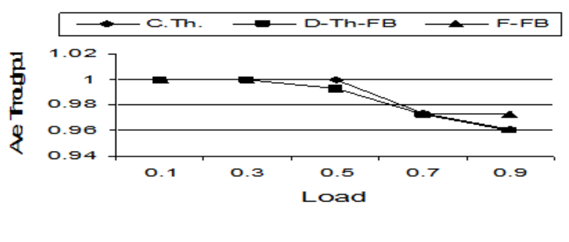


Table-2: Ave. Cell Delay Vs Load, When N=1

Load	*C-Th*	*D-Th-FB*	*F-FB*
0.1	0.0062	0.0062	0.0062
0.3	0.5469	0.01696	0.05469
0.5	0.5673	0.3113	0.3113
0.7	0.7422	0.3783	0.3113
0.9	0.6494	0.3169	0.3113

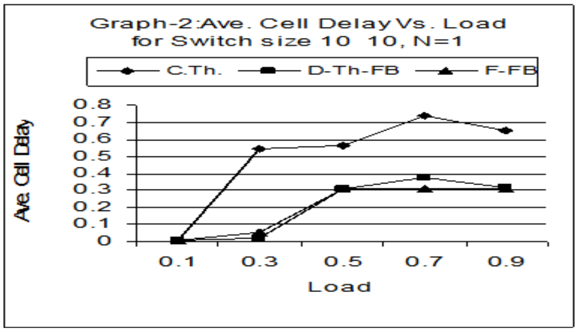


Table-3:Ave.Cell Loss Prob.Vs Load,WhenN=1			
Load	*C-Th*	*D-Th-FB*	*F-FB*
0.1	0	0	0
0.3	0	0	0
0.5	0.0061	0.00695	0.00695
0.7	0.0268	0.02702	0.02702
0.9	0.0393	0.03937	0.03937

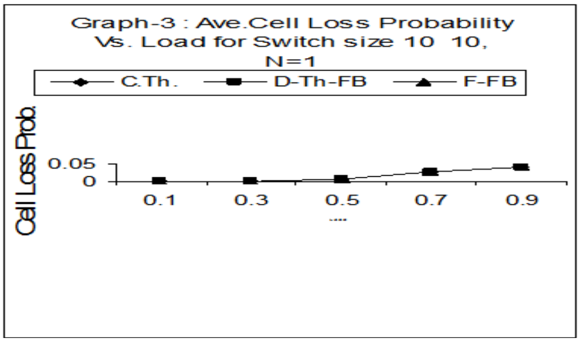


Table 4: Ave.Throughput Vs Load,When N=2			
Load	*C-Th*	*D-Th-Fb.*	*F-Fb.*
0.1	1	1	1
0.3	1	1	1
0.5	0.9939	0.9992	0.9992
0.7	0.9733	0.9969	0.9985
0.9	0.9608	0.9766	0.9962
	C.Th. = Constant Threshold		
	D.Th.Fb.=Dynamic Threshold Feed-back		
	F.Fb.=Fuzzy Feed -back		

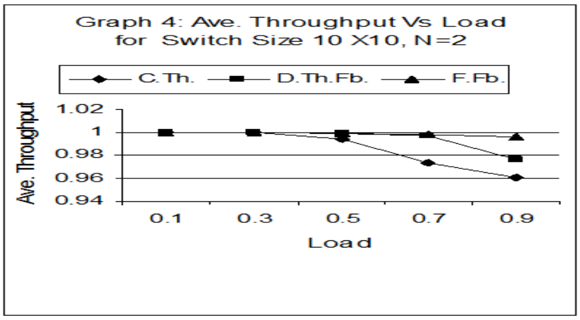


Table 5: Ave. Cell Delay Vs Load, When N=2			
Load	*C-Th*	*D-Th-FB*	*F-FB*
0.1	0.0062	0.0062	0.0062
0.3	0.5469	0.0195	0.01628
0.5	0.5673	0.0401	0.0433
0.7	0.7422	0.1223	0.0878
0.9	0.6494	0.3419	0.1269

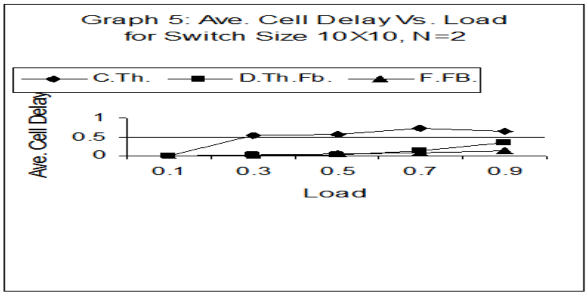
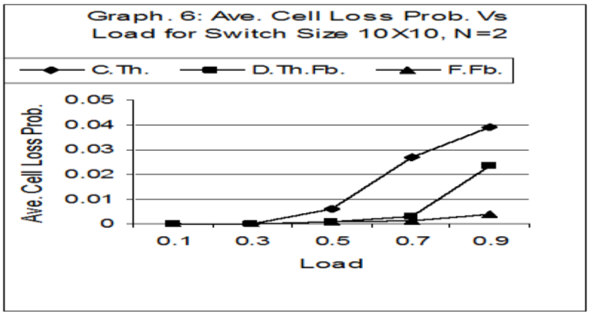


Table 6:AveCell Loss ProbVs Load,WhenN=2			
Load	*C-Th*	*D-Th-FB*	*F-FB*
0.1	0	0	0
0.3	0	0	0
0.5	0.0061	0.0008	0.0007
0.7	0.0268	0.0031	0.0015
0.9	0.0393	0.0234	0.0038



VI. DISCUSSION & CONCLUSION

As shown in the previous section 2.4, the new fuzzy logic based scheme performs very well in comparison to the fixed threshold scheme. However, the price paid, is in terms of increased time complexity. The fixed threshold case has a time complexity of the order of 0(l) because it has to be decided only once if incoming burst can be accommodated or not. In this Section an analysis is carried out to determine worst case time complexity of the suggested scheme.

Step A - Fuzzification

If simple space of buffer occupancy is represented by n sets and if burst length sample space is represented by in Sets, in the worst case, this is bounded by time complexity of 0(n) and 0(m) respectively

Step B - Look-up Table

At the and of fuzzification process, member functions along with membership values obtained. The set of membership functions can be at the most *m* in case of burst length and at most *n* in the case of buffer occupancy. Now, for each membership function of buffer occupancy, every membership function of burst length is taken and the look-up table is referred and corresponding entry is noted down. Whole of this process takes constant time. As this process is referred for a total *m.n* of times, the time complexity of this step is 0(*m.n).*

Step C - Inference Engine

Assuming that, these m.n data elements are clustered in m groups, each group having *n* data elements, it can be easily seen that finding out the minimum element in *n* data elements is of time complexity 0(n). As this procedure is repeated for *m* times, the total time complexity of finding out minimum is 0(m.n). It can be safely assumed that de-fuzzification sample space is represented by not more than I element where m.n. This assumption is valid as at the most *m.n* look-up table entries are referred only. So finding out maximum for these entries takes time of the order of 0(m.n) again. Total time complexity of this step is thus 0(m.n) +O (m.n).

Step D - Defuzzification

Here, the fuzzy sets passed on along with the membership values are defuzzified to determine the crisp output value. Without losing generality, it can be assumed n$\geq$m. In this case. we can say that at the most *m.n* membership functions along with the values are passed on from step 3. For each function, respective de-fuzzification value is retried and multiplying with associated weight tubes a constant time. So worst case time complexity is 0(*m.n*). Addition of all the elements after this has time complexity of 0(*m.n*) and divided by addition of respective weight (complexity of 0(*m.n*) again). Division takes constant time out of time.

So the total time complexity of this step (O(m.n) +O(m.n) +O(m.n)).

Total Time Complexity

Total time complexity of fuzzy logic is obtained by adding the individual time complexities as follows:

O(m) +O(n) +O(m.n) +O(m.n) +O(m.n) +O(m.n) +O(m.n) +O(m.n)

Which by using the result from , turns out to be of time complexity of 0(m.n). For n>=**m** the time complexity of the new scheme can be expressed as 0(n').

Thus new scheme can be easily implemented for small value of n. For large n time complexity may become a liability for the new scheme.

In this paper, we have introduced fuzzy approach to control congestion in ATM networks. When a number of bursty traffic sources add cells, the network is inevitably subject to congestion. Various traditional approaches to congestion management reported in the literature, utilize ‘fixed’ threshold, i.e., either binary or a limited number of predetermined values based on the cell priorities, to determine when to permit or refuse entry of cells into the buffer. The aim is to achieve a desired tradeoff between the number of cells carried through the network, propagation delay of the cells, and the number of discarded cells. Conventional thresholds suffer from some fundamental limitations. One of the limitations is the difficulty of obtaining complete statistics on input traffic to a network. As a result, it is not easy to accurately determine the equivalent capacity or effective thresholds for multimedia high-speed networks in various bursty traffic flow conditions. Besides, these approaches/schemes provide optimal solutions only under a steady state. From these membership values and corresponding sets, blocking to be offered, again in fuzzy terms can be find out. Then, by applying suitable defuzzification method, the percentage blocking to offered at that particular buffer occupancy level and at given burst length can be determined. A comparative study has revealed the proposed scheme is able to achieve lower average delay and higher throughput than the constant as well as dynamic feed-back threshold schemes and that too with lower cell loss probability.

VII. References

1) Bae, J.J., and Suda, T.,: “Survey of traffic control schemes and protocols in ATM networks,” Proc. of the IEEE, vol. 79, no. 2, pp. 170 - 189, Feb. 1991.
2) Benmohamed, L., and Meerkov, S.M., : “Feedback control of congestion in packet switching networks : The case of a single congested node,” IEEE / ACM Trans. on Networking, vol. 1, no. 6, pp. 693-708, Dec. 1993.
3) Atai, A., and Hui, J., : “A rate-based feedback traffic controller for ATM networks,” Proc. IEEE ICC 94, New Orleans, pp. 1605-1615, 1994.
4) Liu, Y.C., and Douligeris, C., : “Rate regulation with feedback controller in ATM networks-a neural network approach,” IEEE JSAC, vol. 15, no. 2, pp. 200-210, Feb.1997.
5) Liu,Y., and Douligeris, C., : “Adaptive vs. static feedback congestion controller,” Proc. IEEE GLOBECOM '95, Singapore, pp. 291-295, Nov. 1995.
6) Berkan, R. C., and Trubatch, S.L., : “Fuzzy system design principles : Building fuzzy If-Than rule Bases,” Published by IEEE Press New Yark, pp. 7, 41-44, 227-99, 1997.
7) Munakata, T., and Jani, Y., : “Fuzzy systems : An overview,” Commun. ACM., vol. 37, no. 3, pp. 69-76, Mar. 1994.
8) Shi, Y., et al., : “Implementation of evolutionary fuzzy systems,” IEEE / ACM Trans. on Networking, vol. 7, no. 2, pp. 109-119, April 1999.
9) Bensaou, B., et. al., : “Estimation of the cell loss ratio in ATM networks with a fuzzy system and application to measurement–based call admission control,” IEEE/ACM Trans. on Networking, vol. 5, no. 4, pp. 572-583, August1997.
10) Cheng, R.G., and Chang, C.J., : “Design of a fuzzy traffic controller for ATM networks,” IEEE / ACM Trans. on Networking, vol. 4, no. 3, pp. 460-469 June 1996.
11) Bonde, A.R., and Ghose, S., : “A comparative study of fuzzy versus ‘fixed’ threshold for robust queue management in cell switching networks,” IEEE/ACM Trans. Networking, vol. 2, no. 4, pp. 337-344, Aug. 1994.

A Survey on Shortest Path Routing Algorithms for Public Transport Travel

GJCST Classifications: C.2.2, I.2.0 , I.2.9, K.4.m

S.Meena Kumari & Dr.N.Geethanjali

Abstract- **Routing is the act of moving information across a network from a source to a destination. Along the way, at least one intermediate node typically is encountered. Routing is often contrasted with bridging, which might seem to accomplish precisely the same thing to the casual observer. Routing involves two basic activities: determining optimal routing paths and transporting information groups through a network. Routing also refers to path finding between source and destination. This literature review investigates some of the gateways to path finding in different networks that are listed in present research literature. A selected set of different approaches are highlighted and set in a broader context, illustrating the various aspects of path finding in different networks. Because path finding is applicable to many kinds of networks, such as roads, utilities, water, electricity, telecommunications and computer networks alike, the total number of algorithms that have been developed over the years is immense. The aim of this survey is to compromise a selected cross-section of approaches towards path finding and the related fields of research, such as transportation GIS, network analysis, operations research, artificial intelligence and robotics, to mention just a few examples where path finding theories are employed.**

Keywords- Routing, Shortest path algorithms, ITS, KSP.

I. INTRODUCTION

This paper projects about various shortest path algorithms of routing in transportation networks. Routing algorithm is the key element in any networks performance, and thus it can be seen as the brain of the network. “Was it possible to find a path through the city crossing each of its seven bridges once and only once and then returning to the origin?” – This was Euler’s famous “Königsberg bridge” question, dating back as far as 1736. It is often seen as the starting point of modern path finding. The basis of what is now known as graph theory was formed by Euler’s methods and this theory in turn paved the way for path finding algorithms. In long–distance road travelling, where successful route planning, prior to travelling and en-route is essential to finding the optimal path from origin to destination. “Optimal” refers to shortest time, shortest distance, or least total cost, the latter being of major concern in some parts of the country, where travelling by car may mean many costly ferry crossings and expensive to all roads roads in order to get from one’s departure to one’s arrival. Path finding in a fixed static network, set costs for traversing the network, and path finding in a dynamic network, the cost of traversing the network varies over the time of traversing. Because path finding is applicable to many kinds of networks, such as roads, utilities, water, electricity, telecommunications and computer networks alike, the total number of algorithms that have been developed over the years is immense. The aim of this survey is to compromise a selected cross-section of approaches towards path finding and the related fields of research, such as transportation GIS, network analysis, operations research, artificial intelligence, and robotics, to mention just a few examples where path finding theories are employed. Road networks are the backbone of modern society. Consequently, the reliability of this road network is thus a decisive factor not only in terms of market outreach and competition, but also in terms of continuity, to ensure a 24/7 operation of the community we live in. Any threat to the reliability of the road network constitutes a vulnerable spot, a weakness, that need to addressed in order for the network not to fail, given the right (in fact: “wrong”) circumstances. This is of particular concern when considering sparse, rural networks, because what by urban standards is a minor degradation (i.e. car accident, resulting in queuing, delays and diversions) may have severe consequences if occurring in a rural setting (i.e. blocking the only access road for hours, even days or weeks).One hazard to transportation networks that has emerged recently and what may become an increasing concern in the near future are the effects global climate change, with extreme weather and precipitation patterns not seen before, and thus closing or degrading links that were thought invulnerable to such threats (Askildsen, 2004).

S.Meena Kumari, Assistant Professor, Department of M.C.A., G.Pulla Reddy Coillege, Kurnool, A.P., India.
(telephone: 9160333625 email: sreemeena@rediffmail.com)
Dr.N.Geethanjali, Associate Professor, Department of C.S.T., Sri KrishnaDevaraya University, Anantapur, A.P., India.
(telephone: 9440896370 email: anjalicsd@yahoo.com)

II. INTELLIGENT TRANSPORT SYSTEMS

In paper, the recent decade’s road transportation systems have undergone considerable increase in complexity and congestion proclivity. From a user point of view, what matters most in relation to a road network is the following: Can I, at the desired time of departure, get from A to B by using the intended route and means of transport, and arrive at a desired time, which would be the best case. Or, does there exist no route or means of travel at all that can take me from A to B at the desired time of departure, let alone within arriving at the desired time, which to the user would be the worst case. This gave rise to the field of ITS, Intelligent Transport Systems, with the goal to apply and merge advanced technology to make transportation more safe and efficient, with less congestion, pollution and environmental impact. In working towards this goal, ITS can take many different forms. Vehicle location and navigation systems are one of these forms and have come along with the emerging field of transport telematics. Transport telemetric implies the large-scale integration and implementation of telecommunication and informatics technology in the field of transportation, penetrating all areas and modes of transport, the vehicles, the infrastructure, the organization, and management of transport.

Zhao (1997) distinguishes between route planning and route guidance as two key elements in vehicle location and navigation systems as part of ITS. Route planning is the process that helps vehicle drivers plan a route prior to driving a specific part of his or her journey. Route guidance is the real-time process of guiding the driver along the route generated by a route planner.

Huang et al. (1995) discriminates route guidance even further, distinguishing between centralized and decentralized route guidance. In the former, vehicles conduct their own path finding using on-board computers and static road maps in CD-ROMs, and applying heuristic search algorithms. Centralized route guidance relies on traffic management centers (TMC) to answer path queries submitted by vehicles linked to it. In this case, Huang et al. (1995) describe a central server holding a materialized view of all shortest paths at that given time, accessed by lookup requests from the vehicles equipped with this system. Although not explicitly stated, it can be assumed that this also is the case in the Advanced Traveler Information System (ATIS) detailed by Shekar and Fetterer (1996) or the ADVANCE project portrayed by both Revels (1998) and Zhao (1997). Boyce et al. (1997) provide a detailed evaluation study of the ADVANCE project for further reference.

III. SHORTEST PATH ALGORITHMS

Efficient management of networks requires that the shortest route from one point (node) to another is known; this is termed as the shortest path. It is often necessary to be able to determine alternative routes through the network, in case any part of the shortest path is damaged or busy. The analysis of transportation networks is one of many application areas in which the computation of shortest paths is one of the most fundamental problems. These have been the subject of extensive research for many years. The shortest path problem was one of the first network problems studied in terms of operations research. Fixed two specific nodes s and t in the network, the goal is to find a minimum cost way to go from s to t. Several algorithms for computing the shortest path between two nodes of a graph are known. This one is due to Dijkstra (1959). Each node is labeled with its distance from the source node along the best-known path. Initially, no paths are known, so all nodes are labeled with infinity. As the algorithm proceeds and paths are found, the labels may change, reflecting better paths. A label may be tentative or permanent. Initially, all labels are tentative. When it is discovered that a label represents the shortest possible path from the source to node, it is made permanent and never changed thereafter. A network consists of arcs, or links, and nodes. The fastest path is calculated as a function associated with the cost of travelling the link. Even though the different research literature tends to group the types of shortest paths problems slightly different, one can discern, in general, between paths that are calculated as one-to-one, one-to-some, one-to-all, all-to-one, or all-to-all shortest paths. In software packages solving static network shortest path problems the software usually aggregates a once-off all-to-all calculation for all nodes, from which subsequent routes then are derived. Clearly, this approach is not feasible for dynamic networks, where the travel cost is time-dependent or randomly varying. However, the majority of published research on shortest paths algorithms has dealt with static networks that have fixed topology and fixed costs. A few early attempts on dynamic approaches, referenced by Chabini (1997), are Cooke and Halsey (1966) and Dreyfus (1979). Not more than a decade ago, Van Eck (1990) reports several hours as an average time for a computer to churn through an all-to-all calculation on a 250-nodes small-scale static network, and several days on a 16.000-nodes large-scale network.

One way of dealing with dynamic networks is splitting continuous time into discrete time intervals with fixed travel costs, as noted by Chabini (1997). Thus, understanding shortest path algorithms in static networks becomes fundamental to working with dynamic networks.

A. Shortest Path In Static Networks

Several algorithms and data structures for algorithms have been put forward since the classic shortest path algorithm by Dijkstra (1959). In its modified version, this algorithm computes a one-to-all path in all directions from the origin node and terminates when the destination has been reached. Deonardo and Fox (1979) introduce a new data structure of reaching, pruning, and buckets. The original Dijkstra algorithm explores an unnecessary large search area, which led to the development of heuristic searches, among them the A* algorithm, that searches in the direction of the destination node. This avoids considering directions with non- favorable results and reduces computation time.

A significant improvement is seen in the bi-directional search, computing a path from both origin and destination, and ideally meeting at the middle. In relation to this search technique, it should be remarked that Jacob et al. (1998) discard bi-directional algorithms as impractical in their computational study of routing algorithms for realistic transportation networks.

Zhan and Noon (1996) had a comprehensive study of shortest path algorithms on 21 real road networks from 10 different states in the U.S., with networks ranging from 1600/500 to 93000/264000 nodes/arcs. In this study, Dijkstra-based algorithms, however differing in data structure, outperform other algorithms in one-to-one or one-to-all fastest path problems.

In summary, the A* algorithm, along with Dijkstra-based algorithms, are preferred in most of the literature researched by the author the author. It is in fact noteworthy that the

Dijkstra algorithm has prevailed to the present date, proving its universal validity.

B. K-Shortest Path In Dynamic Networks

It paper is a result of the recent advances in computer and communications technology, together with the developments in ITS, that have flared a renewed interest in dynamic networks. This interest in the concept of dynamic management of transportation has also brought forward a set of algorithms that are particularly aimed at optimizing the run-time of computations on large-scale networks.

Chabini (1998) lists the following types of dynamic shortest path problems depending on (a) fastest versus minimum cost (or shortest) path problems; (b) discrete versus continuous representation of time; (c) first-in-first-out (FIFO) networks versus non-FIFO networks, in which a vehicle departing at a later time than a previous vehicle can arrive at the destination before the pervious vehicle; (d) waiting is allowed at nodes versus waiting is not allowed; (e) questions asked: one-to-all for a given departure time or all departure times, and all-to-one for all departure times; and (f) integer versus real valued link travel costs.

Fu and Rilett (1996) investigate what they call the dynamic and stochastic shortest path problem by modeling link travel times as a continuous-time stochastic process. The aim of their research was to estimate travel time for a particular path over a given time period. They deviate from the mainstream appraisal of the A* algorithm and advocate the k-shortest path. The reason for this is that standard shortest path algorithms may fail to find the minimum expected paths, particularly when dealing with non-linear optimization, as is the case in developing travel time estimation models. However, in lieu of real data, their research is based on a hypothetical change pattern in travel time.

Based on the research of path finding algorithms, in static networks, Chabini (1997) remarks that a time-space expansion representation can be used in dynamic networks, applying discrete time intervals with fixed costs. Hence, depending on how time is treated, dynamic shortest path problems can be subdivided into two types: discrete and continuous. In the discrete case, if using 15-second time intervals, a full 24-hour implementation would involve calculations on 5760 time discretization, multiplied with the number of nodes and links. Chabini (1997) makes a distinct separation between fastest time paths, in which the cost of a link is the travel time of that link, and minimum cost paths, in which link costs can be of a general form. The difference between these two is nonetheless not explored until Chabini (1998).

Chabini (1997) identifies two key questions in dynamic path finding: (1) what are the fastest paths from one origin to all destinations departing at a given time, and (2) what are the fastest paths from all nodes to one destination for all departure times. He sees the latter as the most significant in relation to ITS, which is true, if one assumes that ITS aims at finding the best path for multiple vehicles with the same destination. In Chabini (1998) the focus extends slightly. Now three questions are put forward: (1) one-to-all fastest path at a given time interval, (2) all-to-one fastest path for all departure times and (3) all-to-one minimum cost path for all departure time intervals.

Chabini (1997, 1998) places emphasis on the all-to-one minimum cost path as the key algorithm with relation to ITS, the reason being that only a limited set of all network nodes are destination nodes in realistic road networks, while there is a considerably larger number of nodes that will be origin nodes. (Moving vehicles tend more to converge to the same goal than to spread in all directions)

Horn (1999) continues along the research trails of Chabini (1997) and Fu and Rilett (1996), but uses a less detailed articulation of travel dynamics, reflecting as he puts it, the recognition that information about network conditions in most parts of the world are most likely to be sparse and that merely estimates of average speed on individual network links are available in most cases. With the presumption that these estimates allow for variation in speed, congestion and delays at nodes, he studied a number of Dijkstra variant algorithms that address these particular conditions. Most important, he propounds an algorithm that calculates an approximation of shortest time path travel duration (path travel time), independent of the particular navigation between nodes. For an experienced vehicle driver, estimated travel time may be more important than the exact route that is to follow. This is a noteworthy addition to the fastest path algorithms in dynamic networks.

C. K-Shortest Path

The shortest path through a network is the least cost route from a given node to another given node and this path will usually be the preferred route between those two nodes. When the shortest path between two nodes is not available for some reason, it is necessary to determine the second shortest path. If this too is not available, a third path may be needed. The series of paths thus derived are known collectively as the k-shortest paths(KSP) and represent the first, second, third,…, k^{th} paths typically of least length from one node to another. The k-shortest path problem is a variant of the shortest path problem, where one intends to determine k paths $p_1,\ldots, p_k$ (in order), between two fixed nodes. The k-shortest paths represent an ordered list of the alternative routes available.

In obtaining the KSPs, it is normally necessary to determine independently the shortest path (k=1) between the two given nodes before computation of the remaining k-1 shortest paths can be carried out. The term shortest does not just apply to the distance between two nodes, but can involve any single component made up of one or more factors, including cost, safety or time, that put a weighting on the route. KSP algorithms are thus widely used in the fields of telecommunications, operations research, computer science and transportation science.

A. W. Brander and M. C. Sinclair made a comparative study of k-shortest path algorithms. Four algorithms were selected for more detailed study from over seventy papers written on this subject since the 1950's. The network was represented as a graph G = (V, E) where V is a finite set of n nodes or

vertices V(G) and E is a finite set of m edges (i.e. links or arcs) E(G) that connect the nodes. The work presented was driven by the desire to find a faster algorithm to calculate the KSPs between nodes in a network. The two original algorithms: Yen and Lawler were implemented to provide a reference to the expected speed and improvement available. Katoh was included as it represented a comparatively recent update and modification to Yen. The fourth algorithm Hoffman was implemented after further study as it was felt that it had the potential to outperform the other algorithms.
Based on solving the k-shortest path problem, Jose L.Santos focused on three codes of Removing path algorithm, Deviation path algorithm-first version and Deviation path algorithm-second version were described and compared on rand and grid networks using random generators. Codes were also tested on the USA road networks. One million paths were ranked in less than 3 seconds on random instances with 10,000 seconds for real-world instances. Dreyfus and Yen cite several additional papers on this subject going back as far as 1957.
Shi-Wei LEE and Cheng-Shong WU proposed an algorithm for finding the k-best paths connecting a pair of nodes in a graph G. Graph extension is used to transfer the k-best paths problem to a problem which deploys well-known maximum flow (MaxFlow) and minimum cost network flow (MCNF) algorithms. Two kinds of path finding procedures are often needed in the design of reliable communication networks. The first one is to find k shortest paths between a pair of nodes. Those paths may be simple or allow loops.
For the k-shortest simple paths problem, Lawler proposed the best known algorithm in computation order O(k(m+nlogn) in undirected graphs, where n and m are the no of nodes and links of the input network. For the directed counterpart, Katoh et al. gave the best known bound in O(k n(m+nlogn)). Recently Eppstein developed an efficient KBP algorithm for finding the k shortest paths allowing loops in O(m+nlogn+k), for highly reliable communication network. The solution output by KBP is a real optimal solution for k disjoint paths and it is very useful for planning highly reliable communication networks.
Francesca Guerriero, Roberto Musmanno, Valerio Lacagnina and Antonio Pecorella dealt with the problem of finding the k shortest paths from a single origin node to all other nodes of a directed graph. The data structure used is characterized by a set of k lists of candidate nodes, and the proposed methods differ in the strategy used to select the node to be extracted at each iteration.

IV. Conclusion

Evaluation of any heuristic method is subject to the comparison of a number of criteria that relate to various aspects of algorithm performance. Examples of such criteria are running time, quality of solution, ease of implementation, robustness, and flexibility (Barr et al., 1995; Cordeau et al., 2002). Since heuristic methods are ultimately designed to solve real world problems, flexibility is an important consideration. An algorithm should be able to easily handle changes in the model, the constraints and the objective function. As for robustness, should not overly be sensitive to differences in problem characteristics: a robust heuristic should not perform poorly on any instance. Moreover, an algorithm should be able to produce good solutions every time it is applied to a given instance. This is to be highlighted since any heuristics are non-deterministic, and contain some random components such as randomly chosen parameter values. The output of separate executions of these non-deterministic methods on the same problem is in practice never the same. This makes it difficult to analyze and compare results. Using only the best results of a non-deterministic heuristic, as is often done in the literature, may create a false picture of its real performance. So based on the heuristics we would like to do further research work on public transport travel using K-Shortest path algorithm (based on Dijkstra's algorithm), considering user preferences.

V. References

1) Eppstein, D., Finding the K Shortest Paths, SIAM COMPUT 28(2), 652-673.
2) Peytchev E., Coggan J. (2002). See before you go, Traffic Technology International.
3) Patriksson, M. (1994). The Traffic Assignment Problem - Models and Methods. Utrecht: VSP.
4) Ruihong H. & Zhong-Ren P. (2002). An Object-Oriented GIS Data Model for Transit Trip Planning Systems, Journal of the Transportation Research Board: Transportation Research Records 1804 , 205-211.
5) Qiujin Wu & Joanna Hartely., Using K-Shortest Paths Algorithms To Accommodate User Preferences In The Optimization of Public Transport Travel.
6) Benjamin Zhan, F. (February 1998). Shortest Path Algorithms: An Evaluation using Real Road Networks. Transportation Science 32(1).
7) Cherkassky, B. V., Goldberg, A.V. & Radzik, T. (1993). Shortest Paths Algorithms: Theory and Experimental Evaluation. Technical report (93-1480), Computer Science Department, Stanford University.
8) Bellman R. E. (1958) On a Routing Problem. In Journal Quarterly of Applied Mathematics 16, 87-90
9) Dreyfus, S.E. (1969). An Appraisal of Some Shortest-path Algorithms. In Journal Operations Research 17, 395-412.
10) Qiujin Wu, Joanna Hartely, & David AL-Dabass. Time-Dependent Stochastic Shortest Paths Algorithms For A Scheduled Transportation Network.

HierarchyMap: A Novel Approach to Treemap Visualization of Hierarchical Data

Aborisade D. O. And Oyelade O. J

GJCST Classifications:
D.2.12, I.2.8, E.1, G.4, G.2.2

Abstract- **The HierarchyMap describes a novel approach for Treemap Visualization method for representing large volume of hierarchical information on a 2-dimensional space. HierarchyMap algorithm is a new ordered treemap algorithm. Results of the implementation of HierarchyMap treemap algorithm show that it is capable of representing several thousands of hierarchical data on 2-dimensional space on a computer and Portable Device Application (PDA) screens while still maintaining the qualities found in existing treemap algorithms such as readability, low aspect ratio, reduced run time, and reduced number of thin rectangles. The HierarchyMap treemap algorithm is implemented in Java programming language and tested with dataset of Departmental and Faculty systems of Universities, Family trees, Plant and Animal taxonomy structures.**

Keywords-Treemaps, Aspect ratio, HierarchyMap, Hierarchical data, Tree-like structure, Node.

I. INTRODUCTION

Large volume of data we use today are represented in hierarchical structures, such structures in their natural forms includes information about Corporate Organizations, University/Departmental Structures, Family trees, Manuals Directory, Internet Addressing, Library Cataloging, Computer Programs, Animal and Plant Taxonomy, e.t.c. The contents and organization of these structures are easily understood if they are small, but very difficult to understand when the structures become large (Mark Bruls, et al.,2000). These problems lead to the concept of Treemaps (Shneiderman and Johnson, 1991). Treemap describes the notion of turning a tree into a planar space-filling map. It is described as space-filling visualization method capable of representing large hierarchical collections of quantitative data. A treemap works by dividing the display area into a nested sequence of rectangles whose areas correspond to an attribute of the dataset, effectively combining aspects of a Venn diagram and a pie chart (Shneiderman et al., 2002). With Treemaps, large hierarchical structures can be viewed without any difficulty because the Treemap visualization method maps hierarchical information into a rectangular 2-dimensional display in a space-filling manner such that 100% of the designated display space is utilized. Interactive control allows users to specify the presentation of both structural (depth bounds, etc.) and content (display properties) information (Shneiderman, 1992). This is in contrast to traditional static methods of displaying hierarchically structured information, which generally makes either poor use of display space or hide vast quantities of information from users. With the Treemap method, sections of the hierarchy containing more important Information can be allocated more display space while ortions of the hierarchy, which are less important to the specific task, can be allocated more space. Although treemaps are originally designed to visualize files on a hard drive (Shneiderman, 1992), it has been applied to a wide variety of areas ranging from financial analysis, business intelligence, money market, stock portfolio to sports reporting (Wattenberg, 1999). A key ingredient of a treemap is the algorithm used to create the nested rectangles that make up the map. These set of rectangles are referred to as the layout of the treemap.

In this work, we developed and implemented a novel HierarchyMap Algorithm. The idea behind this algorithm is to layout information from an hierarchy structures on nested rectangles which we called HierarchyMap Treemap. With this algorithm, every attribute in a hierarchical structure is represented by a rectangular node on the treemap. Each rectangle on the treemap corresponds to an attribute of the dataset. Each of these nodes representing the main attributes of tree-like structures is made to generate the information of sub-nodes of a lower level of the hierarchical structures. This process would continue until all the information in the different levels of the tree hierarchy are displayed one after the other on the same 2-dimensional screen.

II. RELATED WORKS

There are various methods that have been applied to display structure of information, and one of these techniques is the traditional tree diagram where elements are shown as nodes and relations are shown as links from parent to child nodes. More improved techniques have been presented to enhance the efficiency and qualities of such diagram both in 2-dimensional and 3-dimensional space (Furnas , 1986), Knuth, 1973), (Bruggemenn, 1989), and (Card et al.,1991). These techniques have been found to be effective for small trees, but generally ineffective when more than hundreds elements have to be visualized simultaneously. The major reason for this limitation is that node and link diagrams use

---

Manuscript received "19th December 2009"

1st author: Department of Computer and Mathematical Sciences, College of Natural and Applied Sciences, Crawford University, Faith City, Igbesa, Nigeria.
(Telephone: +234-8056535109
Email: adoj_olan@hotmail.com)
2nd author: Department of Computer and Information Sciences, School of Natural and Applied Sciences, Covenant University, Ota, Nigeria.
(Telephone: +234-8035755778
Email: ola2000faith@yahoo.co.uk)

the display space inefficiently as depicted in the Figure 1 below:

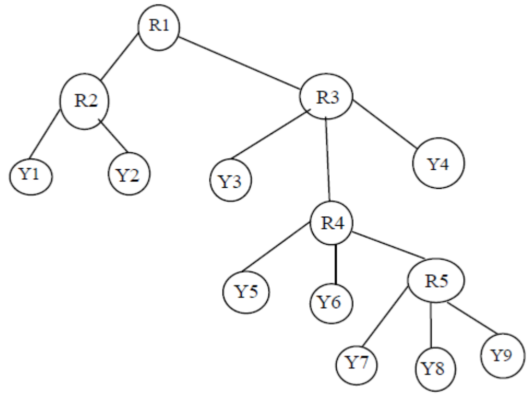


Fig. 1: Tree diagram for representing Hierarchical Data Structure (Mark Bruls et al., 2000)

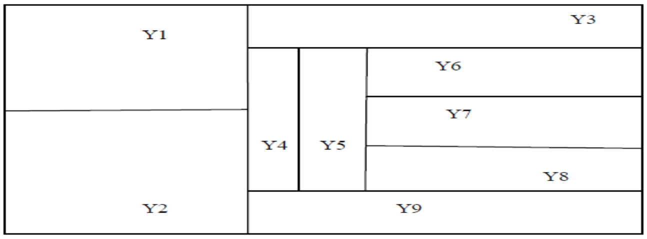


Fig. 2: TreeMap representing the Hierarchical Data Structure in fig. 1 (Mark Bruls et al., 2000)

A treemap as shown in Figure 2 above was developed and introduced to solve the problem of this space usage by using the full display space to visualize the contents of the tree (Johnson and Shneidermann, 1991), (B. Shnerdermann, 1992). As illustrated in Figure 2 above, Slice and Dice treemap algorithm splits the display rectangles along horizontal and vertical lines while recursively traversing a hierarchically structured dataset in top–down direction (Shneiderman,1992). Slice- and –Dice treemap are very effective when size is the most important feature to be displayed. However, this method also has the problem of creating layouts that contain many rectangles with a high aspect ratio. Therefore, many other treemap layout algorithms have been proposed. In order to overcome this limitations. These include Cluster and Squarified treemap algorithms,

Cluster treemap uses a simple recursive algorithm that reduces overall aspect ratios (Wattenberg, 1999), while Squarified treemap algorithm presented the layout of the children in one rectangle as a recursive procedure squarify (Bruls et al., 2000). This procedure lays-out the rectangles in horizontal and vertical rows. When a rectangle is processed, a decision is made between two alternatives, either the rectangle is added to the current row, or the current row is fixed and a new row is started in the remaining sub-rectangle. This decision depends only on whether adding a rectangle to the row will improve the layout of the current row or not.

These methods also have their drawbacks; changes in the data set can cause dramatic discontinuous changes in the layout produced by both cluster treemaps and squarified treemaps. This rapid layout changes also cause an unattractive flickering that draws attention away from other aspects of the visualization and makes it hard to find items on the treemap. Another problem with Cluster and Squarified treemap is that, its layouts fail to preserve order of information as it is done with slice and dice treemap. Many ordered treemap algorithms were introduced to address the limitations in slice-and-dice, Cluster, and Squarified treemap algorithms. The motivating factor here is to seek for the creation of layout in which items that are next to each other in a given order are adjacent in the treemap. Ordered treemaps include Pivot by Split Size, Pivot by Middle, Split and Strip treemap algorithm. These ordered treemaps generally change relatively smoothly under dynamic updates and roughly preserve order, produce rectangles with low aspect ratios compared to that of cluster and squarified treemap (Shneiderman et al. 2002).

Pivot- by- middle algorithm selects the pivot to the middle item of the list so as to create a balanced layout. With this idea, this algorithm is not sensitive to changes as Pivot –by-Split Size. The pivot is taken to be the item (rectangle) with the largest area. Pivot –by-Split-size selects the pivot that will split the list into approximately equal total areas. These two algorithms create layouts that roughly preserve order and are relatively efficient, but fail to produce layouts with relatively low aspect ratio.

Strip algorithm is a modification of the Squarified treemap algorithm. It works by processing input rectangles in order, laying them out in horizontal or vertical strips of varying thickness. It is efficient in that it produces a layout with better readability than the basic ordered treemap algorithm, and reasonable aspect ratios and stability (Shneiderman et al. 2002).

III. METHODS

A. Development of HierarchyMap Algorithm

The algorithm for the HierarchyMap treemap is as follows: Infotree(treedata nodes) T={t1,t2 ,t3, ……….., tn} and a 2-D space divided into four equal rectangles.

i. If the number of hierarchical items to be displayed is zero (i.e. T=0) , then no display.
ii. If the number of hierarchical items to be displayed is 1 (i.e T=1), then Set 2-D space to the item.
iii. If the number of items is greater than 1, split the rectangular 2-D space into four equal sizes and recursively divides each of the resultant item into fours until all items in the list are exhausted such that ∀ ti ∈ T1, ∀ tj ∈ T2, ∀ tK ∈T3,……………………∀ tn ∈ Tn : ti ≤ ti+1 ≤ tj ≤ tj+1 ≤ tk ≤ tk+1 ≤ . . . tn ≤ tn+1.
iv. An attribute of each hierarchical item corresponds to an area of each of the nested rectangles is defined as area(R) in such a manner that their areas correspond to the size of the elements of T1, T2 T3, and T4 where area (R1) ≈ area (R2) ≈ area (R3) ≈ . . . . . area (Rn).

The algorithm accepts inputs data in hierarchical form. These input items in their hierarchical order are stored, read and lay-out on nested rectangles which make up a treemap on the computer screen. The entire 2-dimensional computer screen is divided first into four equal parts, each of the successive parts is then repeatedly divided into four parts in such a way that the resultant rectangles are grouped according to the nodes level to be represented in the entire hierarchical data. This is to ensure that the order of the items to be displayed is maintained. These items are then linked to each of the resultant rectangles that make up the treemap. Each rectangle that represents the node level of tree data can then clicked repeatedly to display the sub-node elements. Every other nodal rectangle on the treemap could be clicked to display their own sub-node elements in a similar manner. In this process, several thousands of items of information could be displayed and viewed in a single space of 2-dimensional treemap.

IV. RESULTS AND DISCUSSION

HierarchyMap algorithm is tested with a several number of sample data of the information structures such as University system, Family system, and Animal Taxonomy. The results of this implementation are represented in Figures 3,4 and 5 respectively. Figure 3 shows the treemap appearance with no information, Figure 4 shows the treemap representation of ten different families Structure and the adjustment of each of the rectangles to reduce their aspect ratio, improve their readability, reduction of thin rectangles . Finally, Figure 5 shows the HierarchyMap for the combination of several tree structures capable of displaying thousands of information. It also shows the adjustment change of the rectangles to demonstrate its optimum measures of the three treemap metrics (i.e. aspect ratio, readability, ordering and capability for change) as data is updated.

The results of this implementation also shows that this HierarchyMap algorithm is similar to other existing treemaps in that, it lays out hierarchical information on nested rectangles, and added further advantage by making it possible to display very large volume of hierarchical information by continuous clicking of node level rectangle, which we have demonstrated in the implementation.

Figure 3: HierarchyMap showing nested rectangles without information

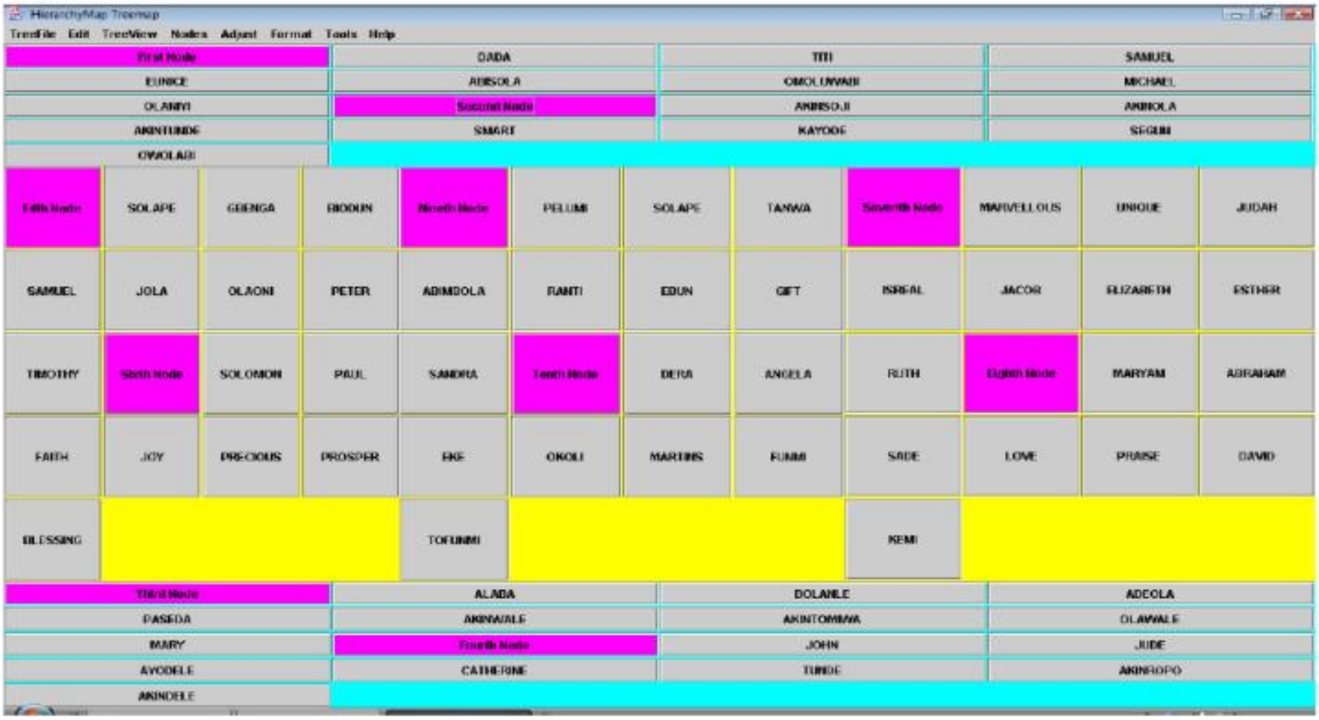


Figure 4: HierarchyMap representing ten different family Structures

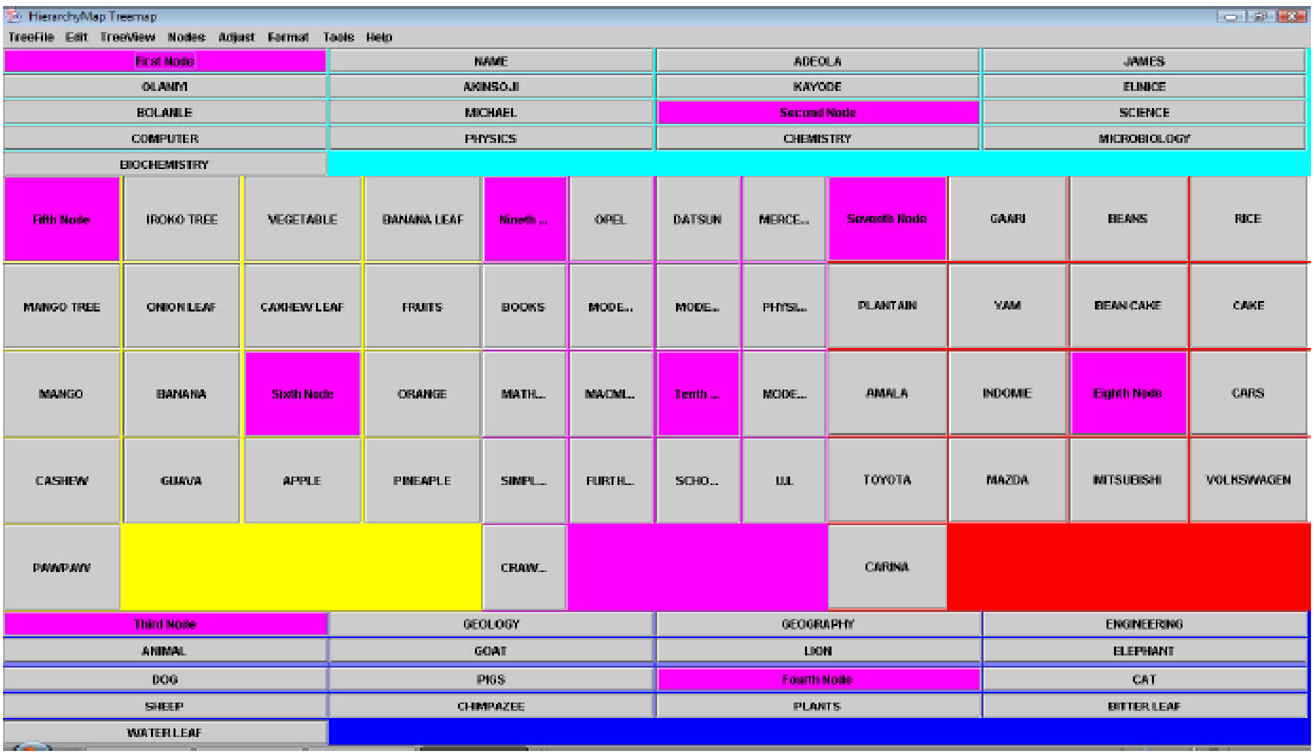


Figure 5: HierarchyMap representing a combination of several hierarchical Structures.

V. Conclusions

In this work, we developed and implemented a novel treemap called HierarchyMap algorithm, which improved on the limitations of the existing treemap algorithms such as Slice-and-dice, Cluster, Squarified, Strip, etc. and added a new feature, which enable viewing of several thousands of hierarchical information by clicking on any of the nodal rectangles. The result showed that the HierarchyMap treemap algorithm has the capability for adjustment change whenever data are updated; it also improved on readability, preservation of order, low aspect ratio, and reduced number of thin rectangles. The combination of these treemap metrics makes HierarchyMap a promising treemap algorithm for the future.

VI. References

1) Bruggemann-Klein and D. Wood. Drawing trees nicely with tex. Electronic Publishing, 2(2):101–115, 1989.
2) Johnson and B. Shneiderman. Treemaps: A space-filling approach to the Visualization of Hierarchical Information Structures. In Proc. of the 2nd International IEEE Visualization Conference, pages 284–291, October 1991.
3) Shneiderman. Tree visualization with treemaps:A 2-D space-filling approach. ACM Transactions on Graphics, 11(1):92–99, September 1992.
4) Bruls S., M., Huizing, K., and Van Wijk, J., 2000. Squarified treemaps. In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization (VisSym), 33–42.
5) Bederson, B., Shneiderman, B., and Wattenberg, M. 2002. Ordered and quantum treemaps: Making effective use of 2D space to display hierarchies. ACM Transactions on Graphics 21, 4, 833–854.
6) D.E. Knuth. Fundamental algorithms. Art of computer programming. Volume 1. Addison-Wesley, Reading, MA, 1973.
7) G.W. Furnas. Generalized fisheye views. In Proc. of ACM CHI'86, Conference on Human Factors in computing systems, pages 16–23, 1986.
8) Herman H, Maurer. Data Structures and Programming Techniques. Prentice- All Incorporation. 1977.
9) J. Bingham and S.Sudarsanum. Visualising large hierarchical clusters in Hyperbolic space. Bioinfomatics Chapter 16:pg. 660-661, 2000. Malin Koksal, Visualization of threaded discussions forums on hand-held devices, Masters Thesis at NADA, 2005.
10) Russel Winder and Graham Roberts, Developing Java Software, John Wiley & Sons. 1998.
11) S.K. Card, G.G. Robertson, and J.D. Mackinlay. The information visualizer, an information workspace. In Proc. of ACM CHI'91, Conference on Human Factors in Computing Systems, pages 181–188, 1991.
12) Wattenberg, M. 1999. Visualizing the stock market. In Extended Abstracts onBHuman Factors in Computing Systems (CHI), ACM Press, 188–189.

A Survey on Congestion Control

GJCST Classifications: C.2.1, C.2.5, A.1

Dr E. Chandra Head, Dept of Computer Application, DJ Academy for Managerial Excellence, Coimbatore
B. Subramani Head, Dept of Computer Technology, Dr. N.G.P Arts and Science College, Coimbatore

***Abstract*- Modern Telecommunication, Computer Networks and both wired and wireless communications including the Internet, are being designed for fast transmission of large amounts of data, for which Congestion Control is very important. Without proper Congestion control mechanism the congestion collapse of such networks would become highly complex. Congestion control for streamed media traffic over network is a challenge due to the sensitivity of such traffic towards. This challenge has motivated the researchers over the last decade to develop a number of congestion control protocols and mechanisms that suit the traffic and provides fair maintenance for both unicast and multicast communications. This paper gives out a brief survey of major congestion control mechanisms, categorization characteristics, elaborates the TCP-friendliness concept and then a state-of-the-art for the congestion control mechanisms designed for network. The paper points the pros and cons of the congestion control mechanism, and evaluates their characteristics.**

Keywords- TCP-Friendliness, Goals, and Metrics of Congestion Control and UDP Traffic

I. Introduction

Congestion control over network, for all types of media traffic, has been an active area of research in the last decade [1]. This is due to the flourishing increase in the audiovisual traffic of digital convergence. There exists a variety of network applications built on its capability of streaming media either in real-time or on demand such as video streaming and conferencing, voice over IP (VoIP), and video on demand (VoD). The number of users for these network applications is continuously growing hence resulting in congestion.

All the networks applications do not use TCP and therefore do not allow fair allocation with the available bandwidth. Thus, the result of the unfairness of the non-TCP applications did not have much impact because most of the traffic in the network uses TCP-based protocols. However, the quantity of audio/video streaming applications such as Internet audio and video players, video conferencing and analogous types of real-time applications is frequently increasing and it is soon expected that there will be an increase in the proportion of non-TCP traffic. In view of the fact that these applications commonly do not amalgamate TCP-compatible congestion control mechanisms, network applications treat challenging TCP-flows in an unreasonable manner. All TCP-flows reduce their data rates in an attempt to break up the congestion, where the non-TCP flows maintains to send at their original rate. This highly unfair condition will lead to starvation of TCP-traffic i.e.., congestion collapse [2], [3], which describes the disagreeable situation where the accessible bandwidth in a network is almost entirely occupied by packets which are discarded because of the congestion before they reach their destination.

For this reason, it is desirable to define suitable congestion control mechanisms for non-TCP traffic that are compatible with the rate-adaptation mechanism of TCP. These mechanisms should make non-TCP applications TCP-friendly, and thus lead to a fair distribution of bandwidth. Unicast is a one-to-one form of communication in networks where multicast is one-to-many. Multicast is advantageous over unicast particularly in bandwidth reduction, but unicast is until the extensively widen communication form network.

II. Theory of Congestion Control System

Congestion control concerns in controlling the network traffic in a telecommunications network, to prevent the congestive collapse by trying to avoid the unfair allocation of any of the processing or capabilities of the networks and making the proper resource reducing steps by reducing the rate of packets sent.

A. Goals and Metrics of Congestion Control

Goals that are taken for the evaluation process of a congestion control algorithm are:

i. To accomplish a high bandwidth utilization.
ii. To congregate to fairness quickly and efficiently.
iii. To reduce the amplitude of oscillations.
iv. To sustain a high responsiveness.
v. To coexist fairly and be compatible with long established widely used protocols.

The Metrics [24] that have been set for Congestion control are:

i. Convergence Speed - The Convergence speed estimates time passed to reach the equilibrium state.
ii. Smoothness - The Smoothness reflects the magnitude of the oscillations through multiplicative reduction and it depends on the oscillations size.
iii. Responsiveness - The Responsiveness is measured by the number of steps or the round trip times (RTTs) to attain equilibrium.

The discrepancy between Responsiveness and Convergence Speed is that the responsiveness is related to a single flow and the convergence is related to the System.

I. Efficiency - The Efficiency is the standard flow throughput per step or round trip time (per RTT), when the system is in equilibrium.
II. Fairness: The Fairness characterizes the fair allocation of resources between the flows in a shared bottleneck link.

III. CLASSIFICATION OF CONGESTION CONTROL ALGORITHMS

The Congestion Control Algorithms are classified mainly based on the below criterion:

i. Can be classified by the type and size of the feedback received from the network
ii. Can be classified by increasing the deploy ability on the network. Only the sender needs for the modification (or) sender and receiver need modification (or) only the router needs for the modification (or) tall the three: sender, receiver and routers needs for the modification.
iii. Can be classified by the aspect of performance. To make improvements in performance: high bandwidth networks, lossy links, fairness, advantage to short flows, variable-rate links
iv. Can be classified by the fairness criterion it uses: max-min, proportional, "minimum potential delay"

A. Classification of Congestion Control by Network

Congestion control algorithms can be categorized using network awareness as a criterion. The following are the three categories for the congestion control mechanisms.

The Black box consists of a collection of algorithms based on the concept that reflects on the network as a black box, pretentious of no knowledge of its state much other than the binary feedback upon congestion.

The Grey box is grey group approaches that use the measurements to estimate accessible bandwidth and the level of contention or even the provisional characteristics of congestion. Because of the opportunity of wrong estimations and measurement dimensions, the network is considered as a grey box.

The Green box contains the bimodal congestion control through which it can calculate explicitly the fairs hare, also the network-assisted control, where as the network communicates through its transport layer. Hence, this is considered as green box.

i. The Black Box

The black box classified congestion control is also called the Blind Congestion Control method and this methodology uses the Additive Increase Multiplicative Decrease (AIMD) algorithm. The AIMD implements the TCP window adjustments. Stability is achieved with these algorithms in situations where the demand of competing flows exceeds the available bandwidths of the channel. The congestion control mechanism in the conventional TCP is based on the fundamental idea of AIMD. In TCP-Tahoe, TCP-NewReno and TCP-Sack, the preservative increase phase is adopted exactly as in AIMD, where the protocols mechanisms are in the congestion control phase. In case of a packet drop, instead of the multiplicative reduction, a more conservative method is used in TCP-Tahoe. The congestion window resets and the protocol mechanisms enter again the slow-start phase. On the other hand, in TCP-NewReno and TCP-Sack, when the sender receives 3 DACKs, a multiplicative reduction is used for the both windows and slow-start threshold phase is applied. In such case, the protocol mechanism remains at the Congestion control phase. When the retransmission timeout expires, they enter the slow-start phase as in TCP-Tahoe.

Highspeed-TCP - Highspeed-TCP modifies the response function in environments with high delay-bandwidth product, increases the congestion window more belligerently upon getting an acknowledgment, and reduces the window more gently upon a loss event.

BIC-TCP - Binary Increase Congestion Control Protocol uses a hollow raise of the sources rate following each congestion event until the window is equivalent to that before the event, to maximize the utilization time of the network.

CUBIC TCP - It is a less aggressive and more systematic derivative of BIC, where the window is a cubic function of time because of the final congestion event, with the modulation point set to the window former to the event.

AIMD-FC - A current advancement of AIMD is Additive Increase Multiplicative Decrease with Fast Convergence is not based on a new algorithm, but on an optimization of AIMD and the convergence procedure that enables the algorithm to congregate faster and attain higher efficiency.

Binomial Mechanisms - Binomial Mechanisms form is a new class for the nonlinear congestion control algorithms named Binomial Congestion Control Algorithms. These algorithms are called binomial because of the control mechanism that is based on the contribution of two additional algebraic terms with different exponents.

SIMD Protocol - SIMD is a TCP-friendly nonlinear congestion control algorithm that that controls the congestion by utilizing history information.

GAIMD - General AIMD Congestion Control generalizes congestion control mechanism of AIMD by parameter zing the additive increase value α and multiplicative decrease ratio β.

ii. The Grey Box

The Grey Box is also called as Measurement-based Congestion Control. Standard TCP relies on packet losses as an implicit congestion signal from congested links. There are a number of reasons for indicating the congestion one of the common reasons is the packet loss:

Random bit corruption is the main cause for the packet loss and is caused when bandwidth is still available.

Acknowledgement-based loss detection at the sender side can be affected by the cross-traffic on the reverse path.

Packet loss, as a binary feedback, cannot indicate the level of contention before the occurrence of congestion.

Therefore, an efficient window adjustment tactic should reflect various network conditions, which cannot all be captured simply by packet drops. Several measurement-based transport protocols gather information on current network conditions.

TCP Vegas -- The queuing delay is estimated by TCP Vegas. To make a constant number of packets per flow the window is linearly increased and decreased in the network.

FAST TCP -- FAST achieves the same equilibrium as Vegas, but uses proportional control instead of linear increase, and intentionally scales the gain down as the bandwidth increases with the aim of ensuring stability.

TCP-Westwood -- A loss causes the window to be reset to the sender's estimation of the bandwidth-delay product in TCP-Westwood which is the minimum measured round trip times the experimental rate of getting acknowledgement.

TFRC -- TFRC is based on the rate-based congestion control mechanism, which intends to efficiently compete for bandwidth with flows in the network.

TCP-Real -- TCP-Real mechanism is based on a receiver-oriented and measurement-based congestion control mechanism that improves the overall performance of TCP over heterogeneous both wired or wireless networks and over asymmetric paths.

TCP-Jersey -- TCP-Jersey is also based on the TCP scheme that focuses on the competence of the transport mechanism in the network.

iii. The Green Box

The Green box contains the bimodal congestion control mechanism by which it can calculate explicitly the fairs hare of the system flow in the network. Bimodal Mechanism -- Bimodal Congestion Avoidance and Control mechanism for each flow the fair-share of the total bandwidth that should be allocated is measured at any point during the execution of the system flow.

Random Early Detection – In Random Early Detection (RED) packets are randomly dropped in proportion to the router's queue size, triggering multiplicative reducing in some flows.

Explicit Congestion Notification – In Explicit Congestion Notification (ECN) routers are enabled to probabilistically mark a bit in the IP header instead of dropping the packets, to intimate the end-hosts of imminent congestion when the length of the queue exceeds a threshold [23].

VCP -- The variable-structure congestion control protocol (VCP) uses two ECN (Explicit Congestion Notification) bits to explicitly get the feedback of the network state of congestion.

IV. Congestion Control Algorithms

A. Drop Tail Algorithm

F. Postiglione et al., discussed that the drop Tail (DT) algorithm [15] has a great accuracy, simplest and most commonly used algorithm in the current networks, which drops packets from the tail of the full queue buffer. The main advantages of this algorithm are simplicity, suitability to heterogeneity and its decentralized nature. However, this algorithm also has some serious disadvantages, such as lack of fairness, no protection against the misbehaving or non-responsive flows (i.e., flows where the sending rate is not reduced after receiving the congestion signals from gateway routers) and no relative Quality of Service (QoS). QoS is of particular concern for the continuous transmission of high-bandwidth video and multimedia information [15]. This type of transmitting the content is difficult in the present Internet and network with DT.

B. Random Early Detection Algorithm

B. Braden et al., discussed that the Random Early Detection Algorithm (RED) had been proposed to be mainly used in the implementation of AQM (Active Queue Management) [4]. On the arrival of each packet, the average queue size is calculated by using the Exponential Weighted Moving Average (EWMA) [5]. The computation of the average queue size is compared with the minimum and the maximum threshold to establish the next action.

C. Choke Algorithm

Konstantinos Psounis et al., proposed CHOKe algorithm [6 and 7], whenever the arrival of a new packet takes place at the congested gateway router, a packet is drawn at random from the FIFO buffer, and the drawn packet is then compared with the arriving packet. If both belong to the same flow in the network then both are dropped, else the packet that was chosen randomly is kept integral and the new incoming packet is admitted into the buffer with a probability depending on the level of congestion. This computation of the probability is the same as in RED. It is a simple and stateless algorithm where no special data structure is required. However, this algorithm is not present well when the number of flows is huge when compared to the buffer space.

D. Blue Algorithms

Rong Pan et al., discussed the basic idea behind the RED queue management system is to make early detection of the incipient congestion and to feed back this congestion notification and allowing them to decrease their sending rates accordingly. The RED queue length gives very less information about the number of contending connections in a shared link of the network.

BLUE and Stochastic Fair Blue Algorithms (SFB) were designed to overcome the drawbacks of the problems caused by the RED techniques, the TCP flows are protected by using packet loss and link idle events against non-responsive flows. SFB is highly scalable and enforces fairness using an enormously miniature amount of state information and a small amount of buffer space. The FIFO queuing algorithm identifies and limits the non-responsive flows based on secretarial similar to BLUE [7].

E. Random Exponential Marking Algorithm

According to Debanjan Saha the Random Exponential Marking Algorithm (REM) [8] is a new technique for congestion control, which aims to achieve a high utilization of link capacity, scalability, negligible loss and delay. The main limitations of this algorithm are: it does not give incentive to cooperative sources and a properly calculated and fixed value of φ must be known globally.

F. Fair Queuing Algorithms

Alan Demers et al., proposed the Fair Queuing Algorithms [9] and Stochastic Fair Queuing Algorithms [10] are mainly used in the multimedia integrated services networks for their fairness and delay bounding in the flow. The frame-based class of FQ is called Weighted Round Robin [11], where a router queue scheduling method is used in which queues are serviced in round robin fashion in fraction to a weight assigned for each flow or queue.

G. Virtual Queue Algorithm

The Virtual Queue Algorithm (VQ) is a radical technique proposed by Gibben and Kelly [12]. In this scheme, a virtual queue is maintained in link with the same arrival rate as the real queue. However, the capacity of the virtual queue is smaller than the capacity of a real queue. When the packets are dropped virtual, then all packets already enqueued in the real queue and all new incoming packets are marked until the virtual queue becomes empty again.

H. Adaptive Virtual Queue Algorithm

R.J. Gibben et al., discussed in the Adaptive Virtual Queue algorithm [13] the capacity of the link and the desired utilization maintains a virtual queue at the link. The capacity and buffer size of the virtual queue is the same as that of the real queue. At the arrival of each packet, the virtual queue capacity is updated. The adaptation of virtual queue algorithm does not suitably follow the varying traffic pattern at flow in the network, and it is also FIFO based methodology.

V. TCP-FRIENDLINESS

TCP is a connection-oriented unicast protocol provides reliable data transfer with flow and congestion control. TCP maintains a congestion window, which controls the number of exceptional unacknowledged data packets in the network. The sender can send packets only as long as free slots are available because the data send will consume slots of the window. When an acknowledgment for exceptional packets is received, the window is shifted so that the acknowledged packets can leave the window and the same number of free slots becomes available for the upcoming data. TCP performs slow start, and the rate roughly doubles each round-trip time (RTT) to quickly increase its fair share of bandwidth. In its steady state, TCP uses an additive increase, multiplicative decrease mechanism to react to congestion by the detection of additional bandwidth. TCP increases the congestion window by one slot per round-trip time when there is no sign of loss. In case of packet loss is indicated by a timeout, the congestion window is reduced to one slot, and TCP reenters the slowstart phase.

TCP-friendliness can be measured through the consequence of a non-TCP flow on the competing TCP flows under the same conditions regarding throughput and other parameters. A non-TCP unicast flow can be TCP-friendly if it does not influence the long-term throughput for any of the synchronized TCP flows by a factor that is more than that done by a TCP flow under the same conditions. A multicast flow is said to be TCP-friendly if it separately views for each sender-receiver pair of the multicast flow TCP-friendly.

A. TCP-Friendliness Vs UDP Traffic

One of the grave drawbacks of FIFO-based queue management is that there is no way to homogenize the connections which send more than their bandwidth share and are non-responsive or very slow in response [18] to congestion collapse indication. In order to present, a fair share of accessible bandwidth to all TCP-friendly connections that is amenable to the congestion collapse indication and the misbehaving in connections should be successfully synchronized by a queue management algorithm. One possible methodology is to solve the above consequences is to use per-flow queuing to discriminate against the non-TCP-friendly connections and to present fair bandwidth share to connections. It is also possible to provide an inducement to TCP-friendly connection in terms of financial benefits. Another possible method is to append a new concept of service i.e., differentiated services to connections. Thus, the differentiated services are being studied by the Differentiated Services Working Group in the IETF [17].

VI. CLASSIFICATION OF CONGESTION CONTROL PROTOCOLS

Congestion control protocols are classified into four major categories according to a number of features in their mechanism of work [22]. The following shows the valid categories of classification.

A. Window-Based Congestion Control

Window-Based protocols are built based on the technique of congestion window-based mechanism, and the congestion window is used at the sender or receiver side [25]. A slot in that window is reserved for each packet, when the sent packet is acknowledged to be received the slot becomes free and allows transmission only when free slots are valid. In absence of congestion the size of window increases and decreases when congestion occurs in the network [14].

B. Rate-Based Congestion Control

Rate-Based protocols are built based on the adaptation of their rate of transmission according to some incorporated feedback algorithm that intimates about congestion when it exists. Rate-based algorithms can be subdivided into simple mechanisms and Congestion control. The results of sawtooth throughput shape are used and this type of schemes usually is not fully compatible with the streaming media applications on which the Simple schemes are based. The current research tends to make the adjustment rate mechanisms ensuring the fairest antagonism between TCP and non-TCP flows equally in the network.

C. Single-rate Congestion Control

Single-rate congestion control mechanisms are usually adopted by all the unicast congestion control protocols. Transmission in unicast has only one recipient, so sending rate is adapted in accordance to the recipient's status. Multicast transmission can adopt the single-rate approach also, where the sender streams the data with same rate to all recipients of the multicast group in the network.

D. Multi-rate Congestion Control

Multi-rate congestion control uses the layered multicast approach, because multi-layering enables to divide data of the sender into different layers to be sent to different multicast groups. Every receiver joins the largest possible number of groups permitted by the bottleneck in the way to sender. The quality of data to be sent to this receiver becomes high when joining more multicast groups. This feature is most evident in multicast video sessions where more the groups that the recipient subscribes in, is more layers that the recipient receives, and also more better the quality of video is. Meanwhile, for other mass data, the transfer time is decreased by additional layers [21]. By the usage of this mechanism, congestion control is achieved absolutely through the group management and routing mechanisms of the primary multicast protocol.

VII. AREAS OF FUTURE RESEARCH

As in the case with an evolving research area, several unsolved issues remain. One particular problem is the lack of comparison congestion control protocols standard methods. A test background that investigates different important aspects such as fairness and scalability of the flow, combined with measures to directly compare the protocol performance [20] would be very handy which also provides standardized suite of test scenarios. While such a test background is not sufficient to walk around all details of a precise protocol, it would provide a sensible basis for more objective comparisons of the protocols.

In many cases, the imitation scenarios presented for a protocol concentrate on a few broad-spectrum scenarios and are frequently too simple to capture behavior and various characteristics of protocol in non-standard situations. Traffic conditions in the network are getting too complex to be modeled in all the aspects by a network simulator, making it significant to estimate the protocols also under real-time applications. We already discussed the various characteristics and behavior of single-rate and multi rate congestion control. It may well be possible that different forms of congestion control are practical maybe with router support that do not show signs of the disadvantages of these methods. While TCP-friendliness is a practical fairness measure in today's network, it is also possible that future network architectures will agree to or necessitate different definitions of fairness. Also the fairness definitions for multicast and many methodologies are still subject to research.

We presented one possible factors and methods to overcome and also briefly addressed a dissimilar form where multicast flows are allowable to use a higher percentage of bandwidth than the unicast flows are, but these can be by no means the only promising fairness definitions. A further area of research is the enhancement of the models for TCP network traffic that are used for some of the rate based congestion control mechanisms. Existing TCP formulae are based on several assumptions that are often not met in real-time conditions. One feature of congestion control mechanism is, that is not openly related to the traffic discussed in this paper (i.e., streaming media traffic) but highly relevant to congestion control in common is how to treat the short-lived flows that consists of only a few data packets. The TCP congestion control, as well as the congestion control schemes presented in this paper, requires that flows persistence for a certain quantity of time period. If not those forms of congestion control are insignificant.

VIII. CONCLUSION

In this paper, we presented a survey on current trends and advancements in the area of TCP-friendly congestion control. We discussed the necessity for TCP-friendly congestion control for both non-TCP based unicast traffic and multicast communication and thus provided an overview of the design space for such congestion control mechanisms. This paper briefly surveys of various congestion control algorithms. It seems that at present there is no single algorithm that can resolve all of the problems of congestion control on computer networks and the Internet. More research work is needed in this direction. It is also to note that not almost all of the surveyed papers have employed any statistical techniques to verify their simulation results. The above discussed are the theory of congestion its goals and merits and the most common factors for the occurrence of congestion and the methods to overcome the congestion collapse. This paper in brief discusses the congestion control algorithms based on the network awareness and various common congestion control algorithm used and its protocols. The paper also discusses the TCP- friendliness and the characteristics of the TCP and non-TCP flows and also the discussed issues that remain to be solved.

IX. REFERENCES

1) W. Chung, "Congestion control for streaming media," Ph. D. dissertation, Polytechnic Inst., Worcester, 2005.

2) Sally Floyd and Kevin Fall, “Promoting the use of end-to-end congestion control in the Internet,” IEEE/ACM Transactions on Networking, vol. 7, no. 4, pp. 458–472, Aug. 1999.
3) Braden, D. Clark, J. Crowcroft, B. Davie, S. Deering, D. Estrin, S. Floyd, V. Jacobson, G. Minshall, C. Partridge, L. Peterson, K. Ramakrishnan, S. Shenker, J.Wroclawski, and L. Zhang, “RFC 2309: Recommendations on queue management and congestion avoidance in the Internet,” Apr. 1998, Status: INFORMATIONAL.
4) Braden, D. Clark, et al. Recommendations on Queue Management and Congestion
5) Avoidance in the Internet. Network Working Group, RFC2309, Apr 1998.
6) Young P. Recursive Estimation and Time-Series Analysis. Sringer-Verlag, 2000.
7) Rong Pan, Balaji Prabhakar, and Konstantinos Psounis. CHOKe, A Stateless Active Queue Management Scheme for Approximating Fair Bandwidth Allocation. IEEE INFOCOM, Mar 2000.
8) Debanjan Saha Wu-chang Feng, Dilip D. Kandlur and Kang G. Shin. BLUE: A New Class of Active Queue Management Algorithms. Technical Report CSETR- 387-99, University of Michigan, April 1999.
9) Steven H. Low Sanjeewa Athuraliya, Victor H. Li and Qinghe Yin. REM: Active Queue Management. IEEE Network, 2001.
10) Alan Demers, Srinivasan Keshav, and Scott Shenker. Analysis and simulation of a fair queueing algorithm. SIGCOMM Symposium on Communications Architectures and Protocols, pages 1–12, Sep 1989. Austin,Texas.
11) P. E. McKenney. Stochastic Fairness Queueing. Proc. IEEE INFOCOM, 2:733– 740, June 1990.
12) N. Pekergin. Stochastic Bounds on Delays of Fair Queueing Algorithms. Technical Report PRISM, UVSQ 10, Universit’e de Versailles-St-Quentin, 1998.
13) R.J. Gibben and F.P. Kelly. Resource pricing and evolution of congestion control. Automatica, 1999.
14) Srisankar Kunniyur. Analysis and Design of an Adaptive Virtual Queue Algorithm for Active Queue Management. ACM SIGCOMM, 2001.
15) De Marco, F. Postiglione, M. Longo, “Run-time adjusted congestion control for multimedia: experimental results”, Journal of Interconnection Networks (JOIN), vol. 5, no. 3, pp. 249-266, 2004.
16) Greg Ewing, Krys Pawlikowski, and Don McNickle. Akaroa 2 User’s Manual. University Of Canterbury, Christchurch, New Zealand, Aug 2001.
17) Park and W. Willinger, Self-Similar Network Traffic and Performance valuation, John Wiley and Sons, 2000.
18) The Internet Engineering Task Force, IETF home page, http://www.ietf.org/
19) S. Floyd and K. Fall, “Promoting the Use of the End-to-End Congestion Control in the Internet,” IEEE/ACM Trans. Net., vol. 7, no. 4, Aug. 1999, pp. 458–72.
20) Szilveszter Nadas, Zoltan Nagy and Sandor Racz Ericsson Research, Traffic Analysis and Network Performance laboratory, 2009.
21) Warrier, S. Janakiraman, S. Ha and I. Rhee, DiffQ: Differential Backlog Congestion Control for Multi-hop Wireless Networks, INFOCOM 2009.
22) Sally Floyd, ICSI Center for Internet Research and Eddie Kohler, UCLA Computer Science Department, “Datagram Congestion Control Protocol”, USA, 2009
23) "Adding Explicit Congestion Notification (ECN) Capability to TCP's SYN/ACK Packets" A. Kuzmanovic, A. Mondal, S. Floyd and K. Ramakrishnan, AT&T Labs Research, June 2009.
24) "Metrics for the Evaluation of Congestion Control Mechanisms", S. Floyd, Ed., March 2008.
25) "Congestion Control in the RFC Series", M. Welzl, W. Eddy, University of Innsbruck, October 30, 2008.

Document Clustering using Linear Partitioning and Reallocation using EM Algorithm

GJCST Classifications: H.3.3, I.4.6, I.1.2

[1] MS. P.J.Gayathri [2] MRS. S.C. Punitha [3] Dr.M. Punithavalli
[1]M.Phil scholar, P.S.G.R. Krishnammal College for Women, Coimbatore, India.
[2] HOD, Department of Computer science and Information Technology, P.S.G.R. Krishnammal College for Women, Coimbatore, India.
[3] Director, Department of Computer science and Information Technology, Sri Ramakrishna college of Arts and Science for Women, Coimbatore, India.
[1]gaya3jayaram79@gmail.com, [2] saipunith @yahoo.co.in

Abstract*- Document clustering is a subset of the larger field of data clustering, which borrows concepts from the fields of information retrieval (IR), natural language processing (NLP), and machine learning (ML), there exist a wide variety of unsupervised clustering algorithms. In this paper presents a novel algorithm for document clustering based with an enhancement on the features of the existing algorithms. This paper illustrates the Principal Direction Divisive Partitioning (PDDP) algorithm and describes its drawbacks and introduces a combinatorial framework of the PDDP algorithm and then describes the simplified version of the EM algorithm called the spherical Gaussian EM (sGEM) algorithm. The PDDP algorithm recursively splits the data samples into two sub - clusters using the hyper plane normal to the principal direction derived from the covariance matrix, which is the central logic of the algorithm. However, the PDDP algorithm can yield poor results, especially when clusters are not well separated from one another. To improve the quality of the clustering results problem, it is resolved by reallocating new cluster membership using the sGEM algorithm with different settings. Furthermore, based on the theoretical background of the sGEM algorithm, it can be obvious to extend the framework to cover the problem of estimating the number of clusters using the Bayesian Information Criterion. Experimental results are given to show the effectiveness of the proposed algorithm with comparison to the existing algorithm.

Keywords- Introduction, Document clustering via linear partitioning hyper planes, The proposed Spherical Gaussian EM algorithm, Results and Discussions conclusion and future work.

I INTRODUCTION

Clustering has been applied to various tasks in the field of Information Retrieval. The Document clustering has become one of the most active area of research and the development. One of the challenging problems is document clustering that attempts to discover the set of meaningful groups of documents where those within each group are more closely related to one another than documents assigned to different groups. The resultant document clusters can provide a structure for organizing large bodies of text for efficient browsing [15].

Document clustering referred to as Text clustering is closely related to concept of data clustering. It is a more specific Technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. The process of clustering aims to discover natural groupings, and thus present an overview of the classes in a collection of documents. Clustering can either produce disjoint or overlapping partitions. In an overlapping partition, it is possible for a document to appear in multiple clusters. The first challenge in a clustering problem is to determine which features of a document are to be considered discriminatory. A majority of existing clustering approaches choose to represent each document as a vector, therefore reducing a document to a representation suitable for traditional data clustering approaches [18].

A wide variety of unsupervised clustering algorithms has been intensively studied in the document-clustering problem. Among the algorithms that remain the most common and effectual, the iterative optimization clustering algorithms have been demonstrated reasonable performance for document clustering, e.g. the Expectation Maximization (EM) algorithm and its variants, and the well-known K--means algorithm. The K-means algorithm can be considered as a special case of the EM algorithm, which has vast vicinity [3] by assuming that each cluster is modeled by a spherical Gaussian, each sample is assigned to a single cluster, and all mixing parameters are equal. The competitive advantage of the EM algorithm is that it is fast, scalable, and easy to implement. Hence, it has been chosen to enhance the algorithm, Expectation Maximization is proposed, Spherical Gaussian EM algorithm.

Principal Direction Divisive partitioning algorithm was developed by Boley [1], which is a hierarchal clustering algorithm that performs by recursively splitting the data samples into two sub clusters. It applies the concept of the Principal Component Analysis for the requirement of the principal eigenvector, which is not computationally expensive. It can also generate a hierarchal binary tree that inherently produces a simple taxonomic ontology. The clustering results produced by the PDDP algorithm compare favorably to other document clustering approaches, such as the agglomerative hierarchal algorithm and associative rule hyper graph clustering. In some cases, the clusters are not well separated from one another, it can yield poor results.

The proposed methodology overcomes the disadvantages of the PDDP algorithm that uses the PCA for analyzing the data and combines it with the EM algorithm as the proposed work. In PDDP splits the data samples into two sub clusters based on the hyper plane normal to the principal direction derived from the covariance matrix of the data. When the principal direction is not representative, the corresponding hyper plane tends to produce individual clusters with wrongly partitioned contents. One practical way to deal with this problem is to run the EM algorithm on the partitioning results. A simplified version of the EM algorithm called the spherical Gaussian EM algorithm is presented for performing such task. Furthermore, based on the theoretical background of the spherical Gaussian EM algorithm, naturally extending this framework to cover the problem of estimating the number of clusters using the Bayesian Information Criterion [9].

The paper is organized as follows. Section 2 briefly reviews some important backgrounds of the PDDP algorithm, and addresses the problem causing the incorrect partitioning. Section 3 presents the proposed algorithm, spherical Gaussian EM algorithm. Section 4 discusses the idea of applying the BIC to our algorithm. Section 5 explains the Artificial Intelligence in EM algorithm. Section 6 explains the data sets and the evaluation method, and shows experimental results. Finally, this paper concludes in Section 7 with some directions of future work.

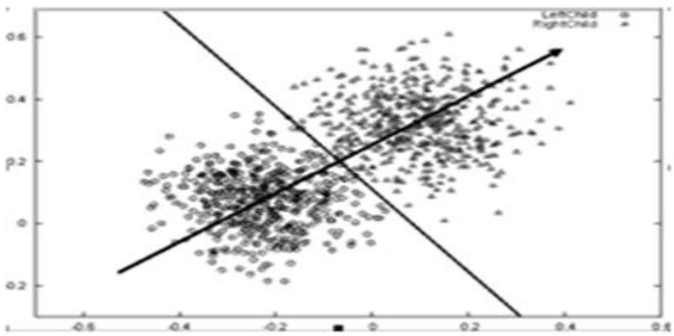


Figure1 The Principal direction and the linear partitioning Hyper plane on the 2d2k dataset.

II DOCUMENT CLUSTERING VIA LINEAR PARTITIONING HYPER PLANES

Considering a one-dimensional data set, e.g. real numbers on a line, the question is how to split this data set into two groups. One simple solution may be the following procedures. The mean value of the data set is first found and then it is compared to each point with the mean value. If the point value is less the mean value, it is assigned to the first group. Otherwise, it is assigned to the second group. The problem arises when it has a dimensional data set. Based on the idea of the PDDP algorithm, this problem can be dealt by projecting all the data points onto the principal direction the principal eigenvector of the covariance matrix of the data set, and then the splitting process can be performed based on this principal direction. In geometric terms, the data points are partitioned into two sub clusters using the hyper plane normal to the principal direction passing through the mean vector [1]. This hyper plane is referred as the linear partitioning hyper plane. Figure 1 illustrates the principal direction and the linear partitioning hyper plane on the 2d2k data set, containing 1000 points distributed in 2 Gaussians.

The PDDP algorithm begins with all the document vectors in a large single cluster. This procedure continues by recursively splitting the cluster into two sub clusters using the linear partitioning hyperactive plane according to the discriminant functions of the algorithm. This procedure terminates by splitting based on some heuristic, e.g. a pre defined number of clusters. Finally, a binary tree is yielded out as the output, whose leaf nodes form the resulting clusters. To keep this binary tree balanced, it selects an unsplit cluster to split by using the scatter value, measuring the average distance from the data points in the cluster to their centroid.

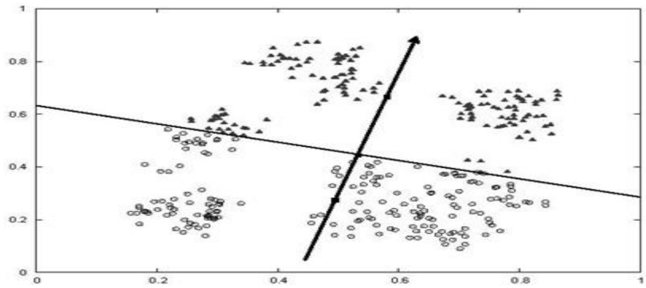


Figure2 Two partitions after the first iteration.

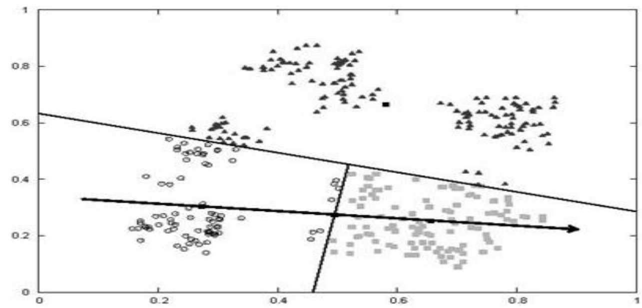


Figure 3 Three partitions after the second iteration

The severe problem of the PDDP algorithm is that it cannot achieve good results when clusters are not well separated from one another. This figure 2 and 3 illustrates this drawback. Figure 2 shows two partitions produced by performing the first iteration of the PDDP algorithm on a dimensional data set. The data set consists of 334 points. The actual class labels are not given, but one can observe that it is composed of five compact clusters [8]. Based on the principal direction and the corresponding linear partitioning hyper plane, it can be seen that the PDDP algorithm starts with significantly wrong partitioning on the middle left hand cluster. Figure 3 shows three partitions after the second iteration. If the partitioning is further performed without making some adjustments, the resulting clusters become worse. This indicates that the basic PDDP algorithm can produce poor solutions in some distributions of the data, which cannot be known in advance. In addition, it may require some information to suggest whether to split the particular cluster or whether to not split on further.

III THE PROPOSED SPHERICAL GAUSSIAN EM ALGORITHM

It is possible to refine the partitioning results by reallocating new cluster membership. The basic idea of the reallocation method [12] is to start from some initial partitioning of the data set, and then proceed by moving objects from one cluster to another cluster to obtain an improved partitioning. Thus, any iterative optimization-clustering algorithm can be applied to do such operation. The problem is formulated as a finite mixture model, and applies a variant of the EM algorithm for learning the model.

The most critical problem is how to estimate the model parameters. The data samples are assumed to be drawn from the multivariate normal density in Rd also assume that features are statistically independent, and a component cj generates its members from the spherical Gaussian with the same covariance matrix [5]. Figure 4 gives an outline of a simplified version of the EM algorithm. The algorithm tries to maximize log Lc at very step, and iterates until convergence. For example, the algorithm terminates when Δ log Lc < δ, where δ is a pre defined threshold.

begin

Initialization: Set $(\mathbf{z}_i)_j^{\langle 0\rangle}$ from a partitioning of the data, and $t \leftarrow 0$.

repeat

E-step: For each $\mathbf{d}_i, 1 \le i \le n$ and $c_j, 1 \le j \le k$, find its new component index as:

$$(\mathbf{z}_i)_j^{\langle t+1\rangle} = \begin{cases} 1, & \text{if } j^* = \operatorname{argmax}_j \log(P^{\langle t\rangle}(c_j|\mathbf{d}_i;\theta_j)) \\ 0, & \text{otherwise.} \end{cases}$$

M-step: Re-estimate the model parameters:

$$P(c_j)^{\langle t+1\rangle} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{z}_i)_j^{\langle t+1\rangle}$$

$$\mathbf{m}_j^{\langle t+1\rangle} = \sum_{i=1}^{n}\mathbf{d}_i(\mathbf{z}_i)_j^{\langle t+1\rangle} \Big/ \sum_{i=1}^{n}(\mathbf{z}_i)_j^{\langle t+1\rangle}$$

$$\sigma^{2\langle t+1\rangle} = \frac{1}{n\cdot d}\sum_{i=1}^{n}\sum_{j=1}^{k}\|\mathbf{d}_i - \mathbf{m}_j\|^2(\mathbf{z}_i)_j^{\langle t+1\rangle}.$$

until $\Delta \log L_c(\Theta) < \delta$;

end

Figure 4 A brief SGEM Algorithm.

A. Estimating Number Of Document Clusters

The clustering algorithm is applied to a new data set having little knowledge about its contents, fixing a predefined number of clusters is too strict and inefficient to discover the latent cluster structures. The finite mixture model of EM algorithm covers the problem of estimating the number of clusters in the data set. A model selection technique is applied called the Bayesian Information Criterion (BIC) [9]. Generally, the problem of model selection is to choose the best one among a set of candidate models.

The BIC contains two components, where the first term measures how well the parameterized model predicts the data, and the second term penalizes the complexity of the model [4]. Thus, the model selected has the largest value of the BIC,

$$M^* = \operatorname{argmax}_i \mathrm{BIC}(M_i).$$

As a result, the value is directly obtained of the first term of the BIC from running the sGEM algorithm. However, it can also be compute it from the data according to the partitioning. The number of parameters is the sum of k − 1 component probabilities, k · d centroid coordinates, and 1 variance.

Boley's subsequent work [2] also suggests a dynamic threshold called the centroid scatter value (CSV) for estimating the number of clusters. This criterion is based on the distribution of the data. Since the PDDP algorithm is a kind of the divisive hierarchical clustering algorithm, it gradually produces a new cluster by splitting the existing clusters. As the PDDP algorithm proceeds, the clusters get smaller. Thus, the maximum scatter value in any individual cluster also gets smaller. The idea of the CSV is to compute the overall scatter value of the data by treating the collection of centroids as individual data vectors. This stopping test terminates the algorithm when the CSV exceeds the maximum cluster scatter value at any particular point.

The CSV is a value that captures the overall improvement, whereas the BIC can be used to measure the improvement in both the local and global structure. As mentioned earlier, in the splitting process, some information is needed to make the decision whether to split a cluster into two sub clusters or keep its current structure. The BIC is first calculated locally when the algorithm performs the splitting test in the cluster. The BIC is calculated globally to measure the overall structure improvement. If both the local and global BIC scores improve, it is then split the cluster into two children clusters.

IV RESULTS AND DISCUSSIONS

- *Data Sets And Setup Information*

The 20 Newsgroups data set consists of 20000 articles evenly divided among 20 different discussion groups [10]. This data set is collected from UseNet postings over a period of several months. Many categories fall into confusable clusters. For example, five of them are computer discussion groups, and three of them discuss religion. The Bow toolkit [11] is used to construct the term document matrix (sparse format). The UseNet headers are used, and also eliminated the stop words and low frequency words (occurring less than 2 times). Finally 59965×19950 term document matrix is obtained for this data set.

The well-known tf·idf term weighting technique is also applied. Let di = (w_{i1}, wi2, wim)T, where m is the total number of the unique terms. The tf·idf score of each wik can be computed by the following formula:

$$w_{ik} = tf_{ik} \cdot \log(n/\, df_k)$$

Where tf_{ik} is the term frequency of w_k in d_i, n is the total number of documents in the corpus, and df_k is the number of documents that w_k occurs. Finally, each document vector is normalized using the L_2 norm. For the purpose of comparison, the basic PDDP algorithm is chosen as the baseline. The number of clusters k is varied in the range [2, 2k], and no stopping criterion was used. Then we applied both the CSV and the BIC to the above settings in order to test the estimation of the number of clusters.

- *Evaluation Method*

Since all the documents are already categorized, comparing clustering results with the true class labels can perform evaluation. In our experiments, the normalized mutual information (NMI) is been used [16]. In the context of document clustering, mutual information can be used as a symmetric measure for quantifying the degree of relatedness between the generated clusters and the actual categories. Particularly, when the number of clusters differs from the actual number of categories, mutual information is very useful without a bias towards smaller clusters, by

Data set	Criterion	Algorithm	k found	NMI	Time (sec.)
20 Newsgroups	CSV	PDDP	34	0.443	15.838
		sGEM	34	0.482	105.39
	BIC	PDDP	25	0.426	14.70
		sGEM	25	0.463	78.45

Table 1: Clustering results by varying stopping criteria on 20 Newsgroups data Sets.

Normalizing this criterion to take values between 0 and 1, the NMI can be calculated as follows

Where nh is the number of documents in the category h, nl is the number of documents in the cluster l, and nh,t is the number of documents in the category h as well as in the cluster l. The NMI value is 1 when clustering results exactly match the true class labels, and close to 0 for a random partitioning [17].

- *ExperimentaL Results*

Figure 5 shows the clustering results on the 20 Newsgroups data set. In this data set, it can be seen that the proposed algorithm perform relatively better than the basic PDDP algorithm. However, performing the global refinement after the local refinement as in EM degrades the quality of the clustering results. The global refinement with the sGEM algorithm leads to more decisions to move each document from its cluster to other candidate clusters.

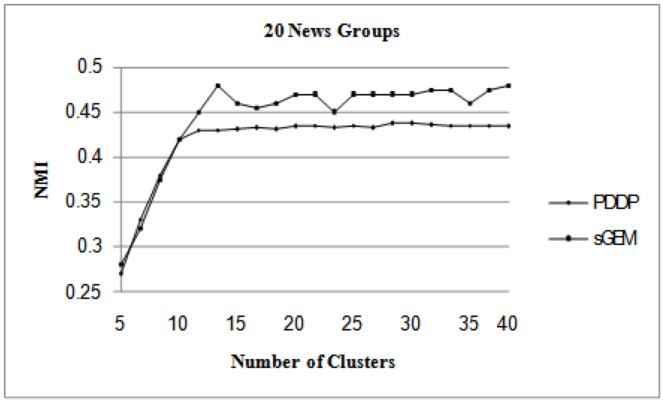


Figure 5: NMI results on the 20 Newsgroups data set.

$$NMI = \frac{\sum_{h,l} n_{h,l} \log(n \cdot n_{h,l}/n_h n_l)}{\sqrt{(\sum_h n_h \log(n_h/n))(\sum_l n_l \log(n_l/n))}}$$

Cj	Purity	Entropy	H	Ep	Ec	Em	Ei	Ef	Emu	Et	Ev	Ea	Er	Eo	Emm	Ecu	Es	E	S	P	T	B
9	1.000	0.000	.	.	.	.	.	.	.	.	.	.	.	.	.	25	.	.	.	.	.	.
10	1.000	0.000	.	30	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
2	0.998	0.005	488	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
1	0.978	0.036	.	.	.	.	3	.	.	.	.	.	.	.	.	.	.	.	132	.	.	.
3	0.900	0.137	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	54	.	4
5	0.878	0.166	.	.	.	.	5	.	.		.	.	.	.	.	.	.	.	.	2	5	86
7	0.865	0.184	.	4	.	.	45	.	.	1	1	1	.	.	.	.	.	.	.	.	.	.
8	0.719	0.363	.	82	.	4	3	.	12	5	.	1	.	1	.	5	.	1	.	.	.	.
11	0.718	0.308	.	1	1	.	.	.	.	28	1	.	1	7	.	.	.	.	.	.	.	.
6	0.680	0.351	.	3	6	.	8	1	.	85	21	1	.	.	.	.	.	.	.	.	.	.
4	0.425	0.372	1	.	.	.	.	.	1	.	.	.	.	48	.	.	.	.	.	.	44	19
0	0.216	0.837	4	128	37	17	54	229	112	67	31	21	157	9	14	44	18	8	9	58	11	32

Table 2 Confusion matrix generated by using sGEM and the BIC

Cj	Purity	Entropy	H	Ep	Ec	Em	Ei	Ef	Emu	Et	Ev	Ea	Er	Eo	Emm	Ecu	Es	E	S	P	T	B
4	1.000	0.000	122	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	0.995	0.010	212	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1
8	0.994	0.013	155	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	0.992	0.015	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	132	.	.	.
10	0.898	0.139	1	.	.	.	.	.	.	1	.	.	.	.	.	.	.	.	.	53	.	4
1	0.564	0.458	.	79	.	3	.	2	8	10	`	1	30	5	1	.	.	.	.	.	.	.
0	0.517	0.281	.	30	1	.	.	.	.	26	1	.	.	.	.	.	.	.	.	.	.	.
5	0.517	0.587	1	23	3	1	1	.	104	7	2	4	25	2	4	12	2	3	5	1	.	.
9	0.507	0.377	.	.	.	.	3	.	.	.	.	.	.	43	.	.	.	.	.	2	53	104
12	0.485	0.383	.	8	1	.	1	79	.	1	2	.	66	.	.	3	1	1	.	.	.	.
2	0.480	0.312	.	4	.	.	.	45	.	47	1	1	.	.	.	.	.	.	.	.	.	.
11	0.474	0.536	.	7	14	.	9	100	3	36	35	1	.	1	1	2	2	.	.	.	.	.
6	0.387	0.492	.	36	1	.	3	47	.	8	.	.	65	.	.	2	6	.	.	.	.	.
14	0.309	0.695	.	4	21	7	42	2	4	25	13	.	.	1	8	2	.	2	.	.	2	3
13	0.209	0.796	3	57	3	9	11	3	6	25	.	17	2	17	1	22	2	2	4	58	5	30

Table 3 Confusion matrix generated by using sGEM and the CSV

V CONCLUSION AND FUTURE WORK

This paper presents several strategies for improving the basic PDDP algorithm. When the principal direction is not representative, the corresponding hyper plane tends to produce individual clusters with wrongly partitioned contents. By formulating the problem with the finite mixture model. This paper describes the sGEM algorithm has tremendous improvement when compared to the PDDP algorithm in several ways for refining the partitioning results. Preliminarily experimental results on two different document sets are very encouraging.

In future work, intends to investigate other model selection techniques for approximating the number of underlying clusters. Recently, work by [7] has demonstrated that estimating the number of clusters in the kmeans algorithm using the Anderson Darling test yields very promising results, and seems to outperform the BIC. The statistical measure can also be applied for this algorithm in further enhancement.

VI REFERENCES

1) Boley, D. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 2(4):325–344, 1998.
2) Boley, D., and Borst, V. Unsupervised clustering: A fast scalable method for large datasets. CSE Report TR99029, University of Minnesota, 1999.
3) Bradley, P. S., and Fayyad, U. M. Refining initial points for kmeans clustering. In Proceedings of the Fifteenth International Conference on Machine Learning, pages 91–99, 1998.
4) Chickering, D., Heckerman, D., and Meek, C. A bayesian approach to learning bayesian networks with local structure. In Proceedings of the thirteenth Conference on Uncertainty in Artificial Intelligence, pages 80–89. Morgan Kaufmann, 1997.
5) Dasgupta, S., and Schulman, L. J.A tworound variant of em for gaussian mixtures. Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI), 2000.
6) Golub, G., and Loan, C. V. Matrix Computations. The Johns Hopkins University Press, Baltimore, 1989.
7) Hamerly, G., and Elkan, C. Learning the k in k - means.In Proceedings of the seventeenth annual conference on neural information processing systems (NIPS), December 2003.
8) He, J., Tan, A.H., Tan, C.L., and Sung, S.Y. On Quantitative Evaluation of Clustering Systems. In W.Wu and H. Xiong, editors, Information Retrieval and Clustering. Kluwer Academic Publishers, 2003.
9) Kass, R. E., and Raftery, A. E. Bayes factors. Journal of the American Statistical Association, 90:773–795, 1995.
10) Lang, K. Newsweeder: Learning to filter netnews. In Proceedings of the Twelfth International Conference on Machine Learning, pages 331–339, 1995.
11) McCallum, A. K. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/~mccallum/bow.
12) Rasmussen, E. Clustering algorithms. In W. Frakes and R. BaezaYates, editors, Information retrieval: data structures and algorithms. Prentice Hall, 1992.
13) Salton, G., and Buckley, C. Termweighting approaches in automatic text retrieval. Information Processing and Management: an International Journal, 24(5):513–523, 1988.

14) Steinbach, M., Karypis, G., and Kumar, V.A comparison of document clustering techniques. KDD Workshop on Text Mining, 1999.
15) Strehl, A., Ghosh, J., and Mooney, R. J. Impact of similarity measures on webpage clustering. In Proceedings of AAAI Workshop on AI for Web Search, pages 58–64, 2000.
16) Strehl, A., and Ghosh, J. Cluster ensembles a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 3:583–617, 2002.
17) Zhong, S., and Ghosh, J.A comparative study of generative models for document clustering.SDM Workshop on Clustering High Dimensional Data and Its Applications, 2003.
18) Nicholas.O.Andrews and Edward.A. Fox. Recent Developments in Document Clustering, Virginia Tech, Blackburg,VA 24060, 2007.

Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus

GJCST Classifications:
F.1.1, J.3, I.2.6

T.Jayalakshmi and Dr.A.Santhakumaran

***Abstract-* The proposed method is a classification problem to diagnose Type II diabetes mellitus using improved Gradient Descent back propagation algorithm. The objective of this research is to increase the performance of the network in terms of accuracy. The accuracy was increased by using three key concepts: missing data replacement, data preprocessing and introducing the Performance Vector (PV) in the search direction. The results of the network have been tested using Pima Indian Diabetes Dataset. This experimental system improves the performance more than 7% than the standard Gradient Descent method.**

Keywords **-** Artificial Neural Networks, Data Preprocessing, Diabetes Mellitus, Gradient Descent, and Missing Data Replacement.

I INTRODUCTION

Diabetes mellitus is now a big growing health problem as it is fourth biggest cause of death worldwide particularly in the industrial and developing countries [Rajeeb Dey and Vaibhav Bajpai, Gagan Gandhi and Barnali Dey]. It is one of the most common chronic diseases, which can lead to serious long-term complications and death. There are two major types of diabetes; Type I and Type II. Type I diabetes is usually diagnosed in children and young adults and was previously known as Juvenile diabetes [Siti Farhanah, Bt Jaafar and Darmawaty Mohd Ali]. Type II diabetes is the most common form of diabetes.

The design and implementation of intelligent system with human capabilities is the starting point to design Artificial Neural Networks (ANN). Artificial neural networks are computational systems whose architecture and operation are inspired from the knowledge about biological neural cells (neurons) in the brain [Madiha J.Jafri, Vince D.Calhoun]. ANNs is a network of many simple processors called units, linked to certain neighbors with varying coefficients of connectivity called weights that represent the strength of these connections. The basic unit of ANNs called an artificial neuron, simulates the basic functions of natural neurons. It receives inputs process them by simple connections and threshold operations and outputs a result.

T.Jayalakshmi is with CMS College of Science and Commerce, Coimbatore. (Email: jayas20@rediffmail.com)
Dr. A.Santhakumaran is with Salem Sowdeswari College, Salem

ANN have been successfully used to solve classification problems in several domains, specifically the back propagation algorithm is very often the favorite to train feed forward neural networks [T.Jayalakshmi, A.Santhakumaran]. Figure.1 shows the schematic representation of a multilayer perceptron with eight input neurons, two hidden layers with eight hidden neurons and one output layer with single neuron. Each of the input neuron connects to each of the hidden neurons, and each of the hidden neurons connects to the output neurons.

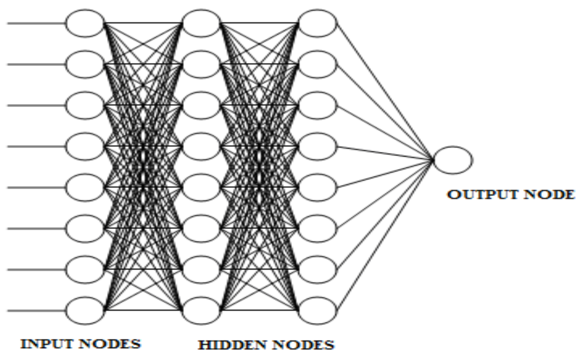


Figure.1 Schematic Representation of a Multi Layer Perceptron

Gradient-based methods are one of the most widely used error minimization methods used to train back propagation networks. Back propagation algorithm is a classical domain dependent technique for supervised training. It works by measuring the output error calculating the gradient of this error, and adjusting the ANN weights and biases in the descending gradient direction. Back propagation is the most commonly used and the simplest feed forward algorithm used for classification.

This paper suggests a simple modification in the search direction to improve the training efficiency, by modifying the search direction vector. The proposed method improves the accuracy at the maximum to classify the Type II diabetes. The paper is organized as follows: Section 2 discusses the improved gradient descent method. Section 3 describes the experimental results and Section 4 concludes the paper.

II BACKGROUND STUDY

Chee-peng Lim, Jenn-Hwai Leong and Mei-Ming Kuan proposed a hybrid neural network comprising Fuzzy ARTMAP and Fuzzy C-Means clustering for pattern classification with incomplete training and test data. To handle missing data in the training samples a three-phase procedure is to be proposed. FAM first trained with complete training samples. Training samples with missing features can be presented, and the missing values can be estimated and replaced using two FCM-based algorithms. Then network training is conducted using all complete and estimated samples. To handle test samples with missing features, a non-substitution FCM-based approach is employed to yield a predicted output quickly. Marisol Giardina, Yongyang Huo, Francisco Azuaje, Paul McCullagh, and Roy Harper makes the investigation about the data acquired from diabetic patients at the Ulster Hospital in Northern Ireland in terms of statistical descriptive indicators and missing values. They made a comparative study of several missing value estimation techniques. This paper reported an exploratory statistical analysis on Type II diabetes databases. It included a comparison of missing value estimation methods, which is a problem that has received relatively little attention from the medical information community. This study is part of the preprocessing phase in the development of supervised and unsupervised machine learning systems for assessing coronary heart disease risk in diabetic patients. HT Nguyen, M Butler, A Roychoudhry, AG Shannon, J Flack and P Mitchell proposes and develops an appropriate integrated for the classification of diabetic retinopathy using a multilayer feed forward neural network. The principal advantages of automated grading are quantitative accuracy and repeatability. Md Monirul Isalm, Md Faijul Amin, Suman Ahmmed and Kazuyuki Murase describes an adaptive merging and pruning algorithm for designing ANNs. This new algorithm prunes hidden neurons by merging and adds hidden neurons by splitting repeatedly or alternatively. The decision when to merge or add hidden neurons is completely dependent on the improvement of hidden neurons learning ability or the training progress of ANNs respectively. Aurangzeb Khan, and Kenneth Revertt describes a rough set theory can be utilized as tool for analyzing relatively complex decision tables like the Pima Indian Diabetes Database. They conclude that in future the missing values filled with 0's can be corrected to improve the accuracy figure. Rajeeb Dey and Vaibhav Bajapi, Gagan Gandhi and Barnali Dey present a work for a classification problem applied to diagnosis of diabetes mellitus using back propagation algorithm of artificial neural network. The database used for training and testing the ANNs have been collected from Sikkim Manipal Institute of Medical Sciences Hospital. They propose that the effectiveness of data normalization in terms of network performance is reflected clearly in the results. Xingbo Sun, Pingxian Yang proposed a novel variant activation sigmoid function with four parameters. The improved BP algorithm based on this is educed and discussed. The efficiency and advantage of the method proved the classification results for the Chinese wines micrographs based on the improved and traditional BPNN.

The activation function can adjust the step, position and mapping scope simultaneously, so it has stronger non-linear mapping capabilities. Michael Rimer, Tony Martinez presents a classification-based objective functions, an approach to training artificial neural networks on classification problems. It directly minimizes classification error by back propagating error only on misclassified patterns from culprit output nodes. Mehmet Onder Efe presents a comparison of neuronal activation functions utilized mostly in neural network applications. This paper dwells on the widely used neuronal activation functions as well as two new ones composed of sines and cosines and a sync function characterizing the firing of a neuron. Pasi Luuka study the suitability of similarity derived from Yu's norms used in a similarity classifier. Usually a similarity classifier uses similarity based on Lukasiwich structure with a generalized mean. A similarity classifier has proved to be a good method in classifying medical data sets. He also tested two different preprocessing methods, PCA and entropy minimization, and their effects..

III METHODOLOGY

The performance of the proposed method is demonstrated by employing the improved gradient descent method. The performance criteria used in this research focus the accuracy of classification. This paper compares the results of normal gradient descent method and the improved gradient method in terms of accuracy.

A. Pima Indian Diabetes Dataset

The Pima Indian Diabetes dataset contains 768 samples with two-class problem. The problem posed here is to diagnose whether a patient would test positive or negative for diabetes. The diagnosis can be carried out based on personal data (age, number of times pregnant) and results of medical examination (blood pressure, body mass index, result of glucose tolerance test etc.) There are 500 samples of class 1 and 268 of class 2. There are eight attributes for each sample. The data set is difficult to classify [Suman Ahmmed, Khondaker Abdullah.A Mamum and Monirul Islam].

B. Network Architecture

The proposed method was implemented with four layer feed forward back propagation neural networks. i.e. one input layer, two hidden layers and one output layer. The architecture has eight input neurons, eight hidden neurons and one output neuron. The output neuron classifies the presence or absence of the diabetes. The network used for training is back propagation neural network. Gradient descent training was used to train the network, which will minimize the mean square error between network output and the actual output. During the training, the tan sigmoid activation function is used for hidden and output layers. The learning rate initialized for the network is 0.01, performance goal is 1e-08 and the number epoch is 500. Weights and biases are initialized to random values in the range of -1 to +1. The reason to initialize weights with small values is to prevent saturation.

C. Missing Data Replacement

Neural network training could be made more efficient by performing certain preprocessing steps on the network inputs and targets. Network input processing functions transforms inputs into better form for the network use. The first key concept used in the research is the missing data analysis. The problem of missing data poses difficulty in the analysis and decision-making processes. Decision-making is highly depending on these data, requiring methods of estimation that are accurate and efficient. The data set used in this research contains missing values, which is a common one in the medical environment. The proposed method interprets the incomplete data into appropriate data set using K-nearest neighbor method. The technique K-nearest neighbor method replaces missing values in data with the corresponding value from the nearest-neighbor column. The nearest-neighbor column is the closest column in Euclidean distance. If the corresponding value from the nearest-neighbor column is also contains missing value the next nearest column is used.

D. Data Preprocessing

The second key concept is data preprocessing. The preprocessing process for the raw inputs has great effect on preparing the data to be suitable for the training. Without this preprocessing, training the neural networks would have been very slow. It can be used to scale the data in the same range of values for each input feature in order to minimize bias within the neural network for one feature to another. Data preprocessing can also speed up training time by starting the training process for each feature within the same scale. It is especially useful for modeling application where the inputs are generally on widely different scales. The proposed method preprocesses the data using Principle Component Analysis (PCA) method. PCA is a very popular preprocessing method. Principal Component's normalization is based on the premise that the salient information in a given set of features lies in those features that have the largest variance. This means that for a given set of data, the features that exhibit the most variance are the most descriptive for determining differences between sets of data. This is accomplished by using eigenvector analysis on either the covariance matrix or correlation matrix for a set of data.

E. Improved Gradient Descent Algorithm

Gradient descent is the most widely used class of algorithm for supervised learning of neural networks. The most popular training algorithm of this category is batch back propagation. It is the first order method that minimizes the error function by updating the weights using the steepest descent method.

$$w(t+1) = w(t) - \eta \, \Delta E \, (w(t))$$

E is the batch error measure; ΔE is the gradient vector which is computed by applying the chain rule of the layers of feed forward neural networks. The parameter η is the heuristic called learning rate. The optimal value of η depends on the shape of the error function. The improved gradient descent algorithm can train any network as long as its weight, net input, and transfer functions have derivative functions. Back propagation is used to calculate derivatives of performance with respect to the weight, bias, and Performance Vector (PV). Each variable is adjusted according to gradient descent (dX). It can be calculated as

$$dX = \eta * \Delta E * PV$$

Where X is the weight and bias values and dX is the search direction vector. PV is the Performance Vector which takes the values in the range of $10<PV<100$ which improves the performance accuracy in a better manner.

Algorithm

i. Create an architecture consists of eight input nodes in the input layer, eight hidden nodes in two hidden layers, one output node in the output layer. Assign the nodes to each layer
ii. Replace the missing data with K-nearest neighbor method
iii. Preprocess the input data using PCA method
iv. Initialize the weights and bias to random values
v. Initialize the network parameters.
vi. Calculate the gradient using
   $dX = \eta * \Delta E * PV$
vii. Train the network with initialized parameters, and with sigmoid activation function.
viii. Calculate the error using MSE method
ix. Repeat the process until the maximum epochs are reached or the desired output is identified or the minimum gradient is reached.

IV EXPERIMENTAL RESULTS

A computer simulation has been developed to study the improved gradient descent method with reconstruction of missing values, preprocessing of data and the effectiveness of performance vector. The simulations have been carried out using MATLAB. Various networks were developed and tested with random initial weights. The network is trained five times, the performance goal is achieved at different epochs, and the classification accuracy is measured. The results of standard gradient descent and improved gradient descent are shown in the performance table (Table 1). The Gradient Descent Neural Network investigation uses a Pima Indian Dataset.

To evaluate the performance of the network the entire sample was randomly divided into training and test sample. The model is tested using the standard rule of 80/20, where 80% of the samples are used for training and 20% is used for testing. In this classification method, training process is considered successful when the MSE reaches the value 1e-08. On the other hand the training process fails to converge when it reaches the maximum training time before reaching the desired MSE. The training time of an algorithm is defined as the number of epochs required to meet the stopping criterion

No. Of Runs	Standard GD (Accuracy)	Improved GD (Accuracy)
1	98.0392	100
2	95.4248	100
3	93.4641	99.3464
4	87.5817	100
5	92.1569	99.3464
Average (Accuracy)	93.33	99.73

Table.1 Performance table

V CONCLUSION

This paper demonstrates the new improved gradient descent approach to classify the diabetic data. The improved gradient descent algorithm includes three key aspects such as replacement of missing value technique, data pre-processing and introducing the performance vector. The computational model used in this paper is to classify a type II diabetes using Pima Indian Dataset. This algorithm proves better average classification than the standard gradient descent method.

VI REFERENCES

1) Aurangzeb Khan, Kenneth Revert (2004), Data Mining the PIMA Dataset Using Rough Set Theory with a Special Emphasis on Rule Reduction, 8th International Multitopic Conference, IEEE Proceedings of INMI.
2) Chee-Peng Lim, Jenn-Hwal Long and Mei-Ming Kuan (2005) , A hybrid Neural Network System for Pattern Classification Tasks with Missing Features, IEEE transactions on pattern analysis and machine Intelligence, Vol 27, No 4, pp.648-653.
3) HT Nguyenl, M Butler, A Roychoudhry, AG Shannon, J Flack, P Mitchell (1996), Classification of Diabetic Retinopathy using Neural networks, 18th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp.1548-1549.
4) Jayalakshmi. T, A.Santhakumaran (2009), Diagnose the Type II diabetes using Feed Forward Back Propagation Neural Networks, CiiT International journal of intelligent system and machine learning, Vol. 2, No. 1.
5) Jayalakshmi. T (2009), Improving the convergence of Back Propagation Neural Network by Tuning Learning Rate Parameter, International journal of emerging technologies and applications in engineering, technology and sciences, Vol.2. Issue 2, pp.298-301.
6) Madiha J.Jafri, Vince D.Calhoun (2006), Functional Classification of Schizophrenia Using Feed Forward Neural Networks, Proceedings of the 28th IEEE EMBS Annual International Conference, pp.6631-6634.
7) Marisol Giardina, Yongyang Huo, Francisco Auaje, Paul McCullagh, Roy Harper (2005), A missing Data Estimation Analysis in Type II Diabetes Databases, Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems.
8) Md. Monirul Islam, Md. Abdus Sattar, Md. Faijul Amin, Xin Yao and Kazuyuki Murase (2009), A New Adaptive Merging and Growing Algorithmfor Designing Artificial Neural Networks, IEEE transactions on systems, man, and cybernetics, VOL. 39, NO. 3. pp.2003-2008.
9) Mehmet Onder Efe (2008), Noval neuronal activation functions for feed forward neural networks, Springer, pp 63-79.
10) Rajeeb Dey and Vaibhav Bajpai, Gagan Gandhi and Barnali Dey (2008), Application of Artificial Neural Network technique for Diagnosing Diabetes Mellitus, IEEE Region 10 Colloquium and the Third ICIIS, PID number 155, pp.1-4.
11) Siti Farhanah Bt Jaafar and Darmawaty Mohd Ali(2005), Diabetes Mellitus Forecast Using Artificial Neural Network, Asian Conference on Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research, pp.135 – 139.
12) Suman Ahmmed, Khondaker Abdullah.A Mamum and Monirul Islam(2007),A Novel Algorithm for Designing Three Layered Artificial Neural Networks, Int.J.Soft.comput.,2(3), pp.450-458.
13) Xingbo Sun, Pingxian Yang(2008),Back propagation Neural Networks With Improved Activation Function and Its Application in the Micrographs Classification, Fourth International Conference on Natural Computation, 2008. ICNC '08. VOL 2, pp.319 – 324.

A Comparative Study on Fingerprint Protection Using Watermarking Techniques

GJCST Classifications:
K.6.5, K.5.1

[1]Ms. D.Mathivadhani, [2]Dr. C. Meena.

***Abstract*- Protection of biometric data is gaining interest and digital watermarking techniques are used to protect the biometric data from either accidental or intentional attacks. Among the various biometrics, fingerprints are more famous in the authentication area, as they are unique to each person and are mainly used for the establishment of instant personal identity. However, they are susceptible to accidental and intentional attacks, when transmitted over network. Thus, a protective scheme is needed which will preserve fidelity and prevent alterations. This paper considers two techniques that protects of fingerprint biometric data using digital watermarking techniques. Both the techniques discussed are based on Discrete Wavelet Transformation (DWT). From the experimental results, it can be concluded that both techniques provide adequate security to the fingerprint data without degrading visual quality. Further, the verification performance after dewatermarking is also analyzed.**

Keywords-watermarking, fingerprint, discrete wavelet transformations, copyright control.

I INTRODUCTION

In recent years, tremendous growth has been witnessed in the development of modern technologies like Internet, P2P, MMS, etc., which in turn, has made an important evolution towards digital distribution of data via network. The digital transmission has introduced flexible, cost-effective communication models that are beneficial in commerce transactions. At the same time, they also possess some serious drawbacks. It allows individuals, other than the owner, to manipulate, duplicate, or access media information without owner's permission (Lin, 2000). This has forced academicians, industrials, and researchers to focus on copyright protection of the intellectual contents. Watermarking is a technique that is used to solve this problem. 'Digital watermarking' is a technique which allows an individual to add hidden copyright notices or other verification messages to digital audio, video, or image signals and documents without affecting the overall quality of the original content. 'Digital watermarks' are information added to digital data which can be detected or extracted later to make a claim on the original content.

The type of information added can be textual data or an image. It is a method that is widely used by industrialists, educationalists, and researchers to protect their digital content along with other data protection methods like encryption, digital signature, etc.

Apart from working as copyright protection scheme, watermarking can also be used as an authentication mechanism to prove the validity of data or owner. Common Applications of digital watermarking include media's ownership protection and proof of ownership, authentication In addition, tampering detection, hiding information and hiding copyright or access control information, etc. Since the interest in protecting Intellectual Property Rights (IPR) is high, the most studied applications of watermarking are proving ownership, authenticating and fingerprinting. However, as the embedded information can be generically be anything that can be represented by bits, interest in other kinds of applications that take advantage of the watermarking techniques is also rising. Two types of watermarking techniques are available for different applications, namely, "Robust Watermarking" and the second is called "Fragile Watermarking". Both are used for different application purposes, such as copyright protection and multimedia authentication, respectively. A third type of watermarking technique that is slowly becoming popular is the "Biometric Watermark." Biometric watermark is a technique that creates a link between a human subject and the digital media by embedding biometric information into the digital object.

Watermarking biometric data is growing importance and is under research for authentication systems. According to Low et al. (2009), biometric watermarking was introduced as the synergistic integration of biometrics and digital watermarking technology. In the battle of copyright piracy, several technological approaches and solutions have been suggested and implemented (Schaathun, 2006). The watermark is nowadays used in conjuction with several biometrics including fingerprint (Jain, 2000), signature (Maiorana et al., 2007), face (Tzouveli et al., 2005), hand [Jain et al., 2008], voice (Lee et al., 2005), iris (Ryoung et al., 2007), retina (Coatrieux et al., 2006).

Fingerprints are unique biometrics that is mainly used for the establishment of instant personal identity. However, they are susceptible to accidental and intentional attacks, when transmitted over network. Thus, a protective scheme is needed which will preserve fidelity and prevent alterations. This is more important with respect to biometric identifiers

[1]..Mrs.D.Mathivadhani, Research Scholar, Department of Computer Science, Avinashilingam University for Women, Coimbatore-641 043,Tamilnadu,India,phone:94864 19356, e-mail:mathi_kirupha@yahoo.co.in.
[2].Dr.C.Meena, Head, Computer Centre, Avinashilingam University for Women, Coimbatore-641043, Phone: 9894269991 e-mail: meena_cc@avinuty.ac.in.

because of their uniqueness. A solution to this situation is watermarking. Several techniques exist for the protection of biometric data and this paper discusses two techniques that integrate digital fingerprint and digital watermarking for security reasons.

Combining digital watermark and biometric for data security and authentication is an emerging area and therefore, there have been only a few published papers. Irrespective of the technique or method used, the main objective of all these techniques is to produce a secure technique, which does not degrade the quality of the cover image and reduce recognition accuracy. Pankanti and Yeung (1999) proposed a fragile watermarking method for fingerprint image verification. The authors showed that their watermarking technique does not lead to a significant performance loss in fingerprint verification. Ratha et al. (2000) used WSQ (Wavelet Scalar Quantization) scheme to hide data in fingerprint images. Uludag et al. (2001) proposed watermarking methods that preserve the quantized gradient orientations at and around watermark embedding locations and singular points in the fingerprint image. Ahmed and Moskowitz (2005) used digital watermarking technique and Fourier transformation to protect fingerprint images used in forensic analysis. Noore et al. (2008) proposed a digital watermarking technique using face and demographic text data as multiple watermarks for verifying the chain of custody and protecting the integrity of a fingerprint image. Similar work was also done by Jain (2000). Zebbiche et al. (2006) proposed another method using watermarking to protect fingerprint data. They introduced an application of wavelet-based watermarking method to hide the fingerprint minutiae data in fingerprint images. A study on the implementation of fingerprint watermark for the protection of intellectual copyrights was conducted by Hui et al. (2008). In their work, the authors used spatial domain, DCT domain of multi-bits embedded watermark methods to embed and extract information of the fingerprint characteristics.

A careful analysis of the review identified the fact that even though several studies have been performed to protect fingerprint images using watermarking techniques, the performance comparison of techniques is sparse. This fact motivated the formation of this paper and compares two such techniques. The paper is organized as below. Section 1 provides an introduction to watermarking and biometric watermarking systems, Section 2 discusses the two algorithms selected for the study and Section 3 presents the experimental results. Section 4 concludes the work with future research directions.

II FINGERPRINT WATERMARKING SYSTEMS

This paper discusses two fingerprint watermarking systems, hereafter referred to as Model-1 (Zebbiche et al., 2009) and Model-2 (Vasta et al., 2006). The primary objective of both the models is to protect fingerprint images during transmission using watermarking techniques. This section explains the working of both the systems.

Model-1

Any fingerprint image has two regions, foreground or ridges area and background or the noisy area near the edges. In Model-1, as the first step, the fingerprint to be watermarked is segmented using an adaptive threshold technique to extract the fingerprint edges from the background. This step produces a binary mask image called "segmentation mask". In the second step, the segmentation mask is partitioned into non-overlapping blocks, where each block is classified as either "ridge block" or "background block" according to the number of foreground pixels belonging to that block. A block is considered to be a ridge block, if and only if, all the block's pixels are classified as a ridge pixel. In the final step, a binary "watermarking mask" is produced with a value of 1 if the block belongs to the ridges area and 0 otherwise. The watermark is embedded in only those layers, which belong to the ridges area. Discrete Cosine Transformation and Discrete Wavelet Transformation (DWT) domain are used to embed the watermark.

After the selection of blocks to carry the watermark, a DCT transformation is applied to the blocks. This results in 64 coefficients which are zigzag scanned (i.e., arranged in decreasing order) to obtain one dimensional vector X[N] representing the entire set of the DCT coefficients to be watermarked. The DC component for each block is not used. The watermark is embedded by steps described in Figure 1.

Algorithm - 1

(i) The sequence and watermark sequence are partitioned into nonoverlapping divisions.

(ii) Each division is multiplied by ±1, which is determined by its associated binary sequence information bit. This is performed in order to obtain the amplitude-modulated watermark (Uses secret key K2 to generate the sequence).

(iii) In the next step, the watermark is embedded using Equation (1) where $x_i[k]$ and $y_i[k]$ are the set of original and watermarked coefficients respectively and λ is the gain factor that is used to control the strength of the watermark.

$$Y_i[k] = (1 - \lambda w_i[k] b_i) x_i[k] \qquad (1)$$

The procedure used while using DWT domain to embed the watermark is explained below. First the blocks selected for embedding the watermark is transformed using the level 1 DWT, which results in (i) a low resolution subband (LL), (ii) high-resolution horizontal subbands (HLl ,HLl−1, ..., HL1), (iii) high-resolution vertical subbands (LHl, LHl−1, ..., LH1), and (iv) high-resolution diagonal subbands (HHl, HHl−1, . . . ,HH1). A watermark is embedded in the high-resolution subbands. The watermark is the same as Figure 1. The algorithm uses two secret keys, which are used during watermark sequence generation (K1) and in order to increase the security level, Model-1 introduces some uncertainty about the selected coefficients altered by permuting the transformation coefficients using a key K1. The advantage of this technique is that the hidden watermark can be retrieved only if the entire procedure through which the watermark has been generated along with the secret keys K1 and K2 is known. Thus, a three level security to the fingerprint is provided. A maximum-likelihood (ML) estimation scheme is used as watermark extraction to increase the performance and is explained in Zebbiche et al. (2009).

Model 2

In Model 2, the fingerprint protection is performed using watermarking technique that uses a combination of DWT and LSB. The watermark embedding and extraction algorithms are given in Figure 2 and Figure 3 respectively.

1. *A 2-D DWT is performed and the detailed subbands are divided into equal sized blocks (M x N) with its coefficients numbered in raster scan order.*
2. *From each block, the first wavelet positive phase coefficient whose value is greater than threshold is selected and the second LSB of this coefficient is replaced by one bit from the watermark image.*
3. *If the number of bits in the watermark image is less than the number of blocks in the fingerprint image, then all bits of the watermark image can be embedded. Otherwise, the following procedure is used.*

a. *For each block, I_r, a message block (MB_r) is formed by selecting few high order bits from each pixel of that block. A key K, which is sufficient large to prevent watermark removal attacks is appended to MB_r.*

b. *Using this K value, a cryptographic hash value is calculated using Equation (2) [.*

$$H_r = H(MB_r)\,k \qquad (2)$$

c. *The pixel position for embedding the watermark bit is calculated as the mod value of H_r and MxN, that is,*

$$[H_r \bmod M \times N]. \qquad (3)$$

During the actual embedding, the original watermark bit is embedded if MSB of H_r is equal to zero else the complement of the watermark bit is embedded.

4. *After successful embedding of all the bits from the watermark image, Inverse Discrete Wavelet Transformation (IDWT) is applied on the coefficients to generate the final secure watermarked image.*

Figure 2 : The Embedding Procedure of Model 1

1. *DWT is applied and the detailed subbands are divided into equal sized blocks of size (2M – 1) x (2N – 1).*
2. *For each block, the block boundaries are synchronized with the M x N blocks formed during the embedding process.*
   a. *For each block, I_r, a message block (MB_r) is formed by selecting few high order bits from each pixel of that block. A key K, which is sufficient large to prevent watermark removal attacks is appended to MB_r.*
   b. *The cryptographic has of the MB_r is computed using Equation (2)*
   c. *The synchronized block boundaries are identified by comparing the last few bits of H_r with the LSBs of pixels in every block and its neighboring blocks.*
3. *The first positive phase coefficient whose value is less than the threshold hold is identified from each block and the watermark bit is extracted from that coefficient.*
4. *The remaining bits are extracted by computing the pixel position of each block where the bit was embedded. The pixel positions are calculated using Equation (3) and the MSB of H_r is analyzed to determine if the actual value or its complement was inserted and the bit is thus extracted.*
5. *These extracted bits are arranged to form the watermark image and IDWT is applied on the watermark image to generate the watermark extracted fingerprint image.*

Figure 3 : Extraction Procedure of Model 2

The advantage with this technique, changes in Ir, results in different hash values and thus makes the watermark undetectable. It is very difficult to calculate the hash value Hr, without knowing the secret key K. Moreover, the usage of high order bits results in an image, ensures that the image quality is not degraded and thus maintains the biometric verification performance.

III RESULTS AND DISCUSSION

This paper evaluates the algorithms based on the verification Accuracy and quality of dewatermarked images of the two models. The models were tested using 10 attacks, namely, JPEG with Quality Factor 50%, JPEG 2000 with Quality Factor 50%, Gaussian Noise (3 x 3), Median Filter (3 x 3), Blurring (3 x 3), Gamma (0.5), Cropping (10 pixels), Resize (90%), Rotation (10o) and Affine Transform. The system is also evaluated when no attacks was performed.

The models were evaluated using a fingerprint database having 1000 fingerprints collected from 500 individuals. The size of the fingerprint images was set to 512 x 512. The matching algorithm was performed using the minutiae matching proposed by [5]. Tables 1 show the verification results when the models were subjected to various attacks.

It can be seen from the table, under normal condition, that is, when the watermark was subjected to no attacks, all the three models produced high recognition accuracy. Model 1 in DCT domain produced 98.64 %, Model 1 in DWT domain produced 99.39% and Model 2 produced 99.48% recognition accuracy. While comparing between the three algorithms, Model 1 in DWT domain performed better than the other two models. For all the other types of attacks, Model 2 seems to produce better results.

In this paper, the quality of the fingerprint image after extracting the watermark is computed using Peak Signal to Noise Ratio. The results obtained during experimentation are presented in Table 2. The table reports the results obtained for 10 images from the database and the same trend was obtained for all the images. The ten cover fingerprint images and the watermark image are shown in Figure 4.

Table 1 : Verification Results

Attack	Model 1 (DCT)	Model 1 (DWT)	Model 2
No attack	98.64	**99.48**	99.39
JPEG (50%)	98.53	**99.67**	99.25
JPEG 2000 (50%)	98.64	**99.62**	99.36
Gaussian Noise (3x3)	98.64	**99.40**	99.21
Median Filter (3 x 3)	72.66	**88.09**	85.33
Blurring (3 x 3)	97.89	**99.49**	99.02
Gamma (0.5)	99.87	99.19	**99.48**
Cropping (10 pixels)	79.12	84.32	**99.25**
Resize (90%)	75.43	81.88	**99.36**
Rotation (10^o)	71.76	78.68	**99.21**
Affine Transform	63.12	71.54	**85.33**

Table 2 PSNR OBTAINED FOR VARIOUS MODELS

Image	Model 1 (DCT)	Model 1 (DWT)	Model 2 (DWT + LSB)
Finger1	38.42	40.23	32.56
Finger2	38.26	41.99	33.65
Finger3	39.02	40.56	33.45
Finger4	39.12	41.44	32.88
Finger5	38.99	41.23	33.12
Finger6	38.67	40.39	34.23
Finger7	39.23	41.87	34.16
Finger8	39.41	41.65	32.75
Finger9	39.33	40.54	32.89
Finger10	39.76	40.77	33.01

From Table 2, it can be seen that Model 1 in DWT domain produces better quality images, followed by Model 1 in DCT domain. The quality of the image slightly degrades with Model 2.

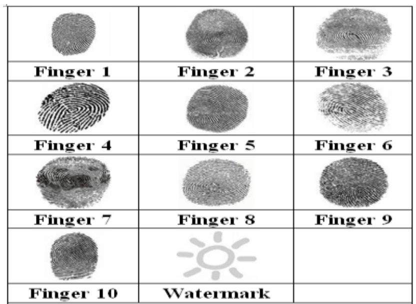


Figure 4: Cover and Watermark Images

From the results obtained, it can be concluded that the watermarking of fingerprint image is highly successful while using DCT, DWT or DWT + LSB techniques. All the three models produce high recognition results and produce a high quality image after dewatermarking.

IV CONCLUSIONS AND FUTURE WORK

Watermarking techniques are increased used in biometric security systems for authentication requirements and they use biometric characteristics such as face, voiceprint, fingerprint, etc. Out of these, fingerprint image is considered more reliable for personal authentication. They are considered good choice because of two very important characteristics, its uniqueness, and permanency. However, the major disadvantage with fingerprints is that it can be forged by hackers or criminals.

However, as the need for security increases, research for more permanent form of biometric, which is difficult to replicate, is considered. One such biometric is human iris. Iris recognition is based on visible features, i.e. rings, furrows, freckles, and corona and is considered very challenging, as they possess a high degree of randomness. The Iris is essentially formed by 8 months, and remains stable through life. Statistically more accurate than even DNA matching since the probability of 2 irises being identical is 1 in 10 to the power of 78 (Daugman and Downing, 2001). The future research direction is planned in the direction of protecting Iris watermarking for security and authentication. Here, biometric can act as an access granting mechanism in the source and destination place and upon successful identification, a watermark as an additional security can be used to authenticate an image.

V References

1) Ahmed, F. and Moskowitz, I.S. (2005) Composite Signature Based Watermarking for Fingerprint Authentication, ACM Multimedia and Security Workshop, New York, Pp. 1-8.
2) Coatrieux, G. Lamard, M. Daccache, W. Puentes, W. Roux, C. (2006) A Low Distorsion and Reversible Watermark: Application to Angiographic Images of the Retina, 27th Annual International Conference of the Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005, Pp. 2224-2227.
3) Daugman J and Downing C (2001) "Epigenetic randomness, complexity, and singularity of human iris patterns." Proceedings of the Royal Society, B, 268, Biological Sciences, Pp 1737 - 1740.).
4) Hui, K., Jing, L., Xiao-dong, Z. and Xiao-xu, Z. (2008) Study on Implementation of a Fingerprint Watermark, International Conference on Computer Science and Software Engineering (CSSE), Vol. 3, Pp.725-728.
5) Jain, A.K., Hong, L. and Bolle, r. (1997) Online fngerprint verification, IEEE Trans. Pattern Anal. Machine Intell., Vol. 19, No. 4, Pp. 302-314.
6) Jain, A.K., Nandakumar, K. and Nagar, A. (2008) Biometric Template Security, EURASIP Journal on Advances in Signal Processing, Vol. 2008, Article ID 579416.
7) Jain, S. (2000) Digital watermarking techniques: a case study in fingerprints & faces, Proceedings ICVGIP 2000, Pp. 139-144.
8) Lee, Y., Kang, H.J. and Ki, Y.H. (2005) Copyright Authentication Enhancement of Digital Watermarking Based on Intelligent Human Visual System Scheme, Knowledge-Based Intelligent Information and Engineering Systems, Intelligent Watermarking Algorithms and Applications, Volume 3682/2005, Pp. 567-572
9) Lin, C. (2000) Watermarking and Digital Signature Techniques for Multimedia Authentication and Copyright Protection, PhD Thesis, Columbia University, 2000.
10) Low, C.Y., Teoh, A.B. and Tee, C.(2009) Fusion of LSB and DWT Biometric Watermarking Using Offline Handwritten Signature for Copyright Protection, Proceedings of the Third International Conference on Advances in Biometrics, Lecture Notes In Computer Science, Vol. 5558, Pp. 786 – 795.
11) Maiorana, E., Campisi, P., Neri, A. (2007) Biometric Signature Authentication Using Radon Transform-Based Watermarking Techniques, IEEE Biometrics Symposium,2007, Pp. 1-6.
12) Noore, A., Singh, R., Vatsa, M. and Houck, M.M. (2009) Enhancing security of fingerprints through contextual biometric watermarking, Forensic Science International, Vol. 169, Issue 2, Pp. 188-194
13) Pankanti, S. and Yeung, M.M. (1999) Verification watermarks on fingerprint recognition and retrieval, Proc. SPIE EI 1999, San Jose, CA, Vol. 3657, Pp. 66-78.
14) Ratha, N.K., Connell, J.H. and Bolle, R.M. (2000) Secure data hiding in wavelet compressed fingerprint images, Proc.ACM Multimedia 2000 Workshops, Los Angeles, CA, Pp. 127-130
15) Ryoung, K., Jeong, D.S., Kang, B.J. and Lee, E.C. (2007) A Study on Iris Feature Watermarking on Face Data, Proceedings of the 8th international conference on Adaptive and Natural Computing Algorithms, Part II, Lecture Notes In Computer Science; Vol. 4432, Pp. 414-423.
16) Schaathun, H.G. (2006) On watermarking/ fingerprinting for copyright protection, Proceedings of the First International Conference on Innovative Computing, Information and Control, IEEE Computer Society, Vol. 3, Pp. 50-53.
17) Tzouveli, P. Ntalianis, K. Kollias, S. (20050 Human face watermarking based on Zernike moments, Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005, Pp. 399-404.
18) Uludag, U., Gunsel, B. and Ballan, M. (2001) A spatial method for watermarking of fingerprint images, Proceedings of. First International Workshop on Pattern Recognition in Information Systems, Setúbal, Portugal, Pp. 26-33.
19) Vatsa, M., Singh, R., Noore, A., Houck, M.M. and Morris, K. (2006) Robust biometric image watermarking for fingerprint and face template protection, IEICE Electronics Express, Vol. 3, No.2, Pp. 23-28.
20) Zebbiche, K. and Ghouti, L. et al. (2006) Protecting fingerprint data using watermarking. First NASA/ESA Conf. on Adaptive Hardware and Systems (AHS'06), Pp.451–456.
21) Zebbiche, K. and Khelifi, F. (2009) Region-Based Watermarking of Biometric Images:Case Study in Fingerprint Images, International Journal of Digital Multimedia Broadcasting, Vol. 2008, Article ID 492942, Pp. 1-13.

Improving and Maintaining Network Security Using MD5 Algorithm

GJCST Classifications:
E.3, D.4.6, K.4.2, K.6.5

Mr. Nitin A. Naik [1], Mr. Gajanan D. Kurundkar [2] and Dr. Santosh D. Khamitkar [3]

Abstract- **Networking devices and intrusion detection systems are capable of generating large volumes of audit information. This information should be considered sensitive. Privacy concerns must also be considered, as there are many legal and ethical issues with maintaining these types of data sets. Until now, little attention has been paid to protecting these data sets from attackers, both internal and external. While advances in computer and communications technology have made the network everywhere, they have also rendered networked systems vulnerable to malicious attacks orchestrated from a distance. These attacks, usually called cracker attacks or intrusions, start with crackers infiltrating a network through a vulnerable host and then going on to launch further attacks. Crackers depend on increasingly sophisticated techniques like using distributed attack sources. On the other hand, software that guards against them remains rooted in traditional centralized techniques, presenting an easily targetable single point of failure. With the free flow of routing data and the high availability of computer resources, possible threats to the networks can result in loss of privacy and in malicious use of information or resources that can eventually lead to large monetary losses. By applying MD5 an algorithm, which plays a major role network security and infrastructures built-in security constraints, are monitored properly.**

Keywords- MD5, Intrusion, Integrity, intrusion detection.

I INTRODUCTION

The recent past have witnessed an ever-growing confidence on computer networks for business Transactions. With the free flow of data and the high availability of computer resources, owners and managers of enterprise networks have to secure their resources from any possible threats to their networks. Although these threats take many forms, they all result in loss of privacy to some degree and in malicious use of information or resources that can eventually lead to large losses [1].

Data integrity is monitored with secure hashing functions. Access to data in the system is logged for audit purposes and restricted by way of a domain specific query language. We address the threats of both malicious insiders and external hackers by providing security measures for both scenarios. Network security and intrusion detection data is captured in real time and a MD5 checksum is generated.[2]. Intrusion detection systems can be classified based on a multitude of factors. Some significant ones, described in detail in are:

Response to Intrusion this can be passive or active. A passive system is content with just detecting intrusion, leaving its handling to a human agency. On the other hand, an active system takes action, for example terminating network connections to a suspected host. Obviously, active systems are much better scalable and responsive, but open themselves up to denial of service attacks by overreacting to deliberately triggered false alarms.

Source of audit data the data to be examined can be network data (network packets etc.) or host data (application logs, system call traces etc.)

Locus of data collection and processing Data-collection can be centralized or distributed. Again, this data can be processed centrally or at distributed locations. In recent times, there has been a lot of interest in distributed schemes for Intrusion detection's. While the research society has been active in this area, most existing schemes are inert in the sense that they only implement collecting information in a distributed manner. The controlling intelligence is centralized in the person of the system administrator managing the administrative domain. Malicious hackers will often install Trojan Horse files in place of standard utility programs, usually to help hide the evidence of their intrusion. For example, when looking at the file system on a UNIX box, you may notice that the 'ls' program seems to have an unusually large file size or an odd timestamp. You should then compare it with a clean copy of 'ls' from another machine or from installation media. A good tool for such quick comparisons is MD5, which makes cryptographic checksums of the target files. If there appears to be a problem, then you can run the suspect 'ls' program through a debugger such as gdb, and often find out what the Trojan Horse 'ls' was designed for. Typically, a Trojan horse 'ls' will not display any of the files or directories installed by the intruder, and may also have the SETUID and SETGID bits set to mode 6555.[4]

Precise and completely accurate copy. For any system administrator making copies of data collected, a MD5 hash value should also be calculated to ensure the integrity of the copies. It should be noted while RFC1321 (1992) states it is conjectured that it is computation infeasible to produce two messages having the same message digest, or to produce any Message having a given pre-specified target message digest,

Manuscript received "13th November 2009"

[1] Dept. of Computer Science , Yeshwant Mahavidyalaya, Nanded
(MS)(India) (+919881378420 email: naiknitin_2011@yahoo.com)
[2] Dept. of Computer Science , S.G.B. College Purna (MS)(India)
(+919850797966 email: gaju_k_2001@yahoo.com)
[3] School of Computational Sciences , S.R.T.M. University, Nanded
(MS)(India) (email : s_khamitkar@yahoo.com)

recent research has shown that a collision attack can be conducted against the MD5 algorithm using a standard home PC in a reasonably short amount of time [5]. This, combined with other similar research, has resulted in the use of MD5 digests being brought into question from an evidentiary viewpoint. Prior to any use of this utility, advice from legal personnel or proper authorities should be sought.
In this paper we will try to evaluate the impact of MD5 authentication on routing traffic in two contexts: Secured and un-secured.

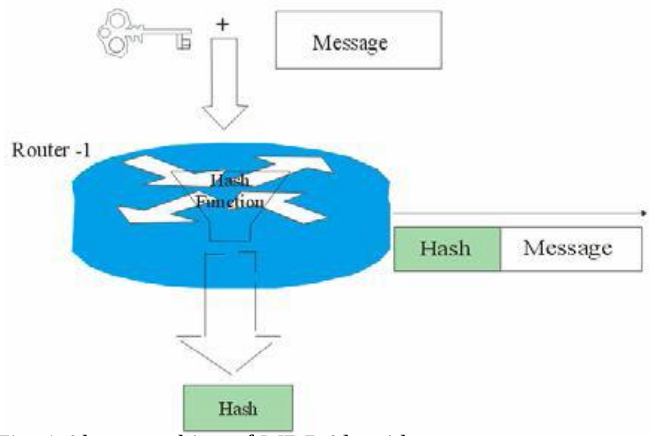


Fig. 1 About working of MD5 Algorithm

The MD5 algorithm takes the preconfigured shared secret key and the traffic data (or message) as inputs and returns a message digest (hash) that is appended to the message and sent through the appropriate interface. Above figure illustrates the sequence of events for routing protocol authentication at the destination router. EIGRP, RIPv2 and OSPF are supported keyed MD5 cryptographic checksums to provide authentication of traffic data including routing updates. Each key is represented by key number, key string, and key identifier, which are stored locally. EIGRP MD5 authentication supports multiple keys, which are grouped in one keychain. RIPv2 MD5 the basic RIPv2 message format provides for an 8-byte header with an array of 20-byte records as its data content. When keyed MD5 is used, the same header and content are used, except that the 16-byte authentication key field is reused to describe a Keyed Message Digest trailer. With MD5, all OSPF protocol exchanges are authenticated. The OSPF packet header includes an Authentication Type field and 64 bits of data for use by the appropriate authentication scheme. Each key has a lifetime period that validates the usage of this key for sending and receiving. The router selects one key from the keychain for sending an authentication packet. The key numbers are examined from the lowest to the highest, and the first valid key encountered is used [16] [17]. Murphy and Badger from TIS [8] proposed a public key signature scheme to protect the integrity of LSAs flooded through the network. With a public key infrastructure, the source router uses its private key to sign the MD5 value for every LSA created. Since the intermediate routers do not know the private key of the source router, they cannot tamper the LSAs without being detected. On the other hand, every receiver of LSAs must use the source router's public key to verify its integrity. Therefore, their scheme is very secure against compromised intermediate routers. The only problem is that public key systems (e.g., RSA) are usually very expensive.

II FILE LOCATIONS AND INTEGRITY

HIDS require different methods to avoid generating alarms. The vast majority of known HIDS vision methods attempt to exploit the signature-based products. Evading file monitoring HIDS has fewer options for evasion besides avoiding file changing exploits. All File HIDS use either MD5 or some variant of this algorithm to make a hash of monitored files. The MD5 algorithm is defined by its author as "It is conjectured that it is computationally infeasible to produce two messages having the same message digest, or to produce any message having a given pre specified target message digest." [9] The message digest is akin to a fingerprint. If as little as one-bit changes in the original file, the computed MD5 hash will change as well. While it is possible that two files will have the same digest, it would take an attacker an infeasible amount of time to develop a file with the same fingerprint and that file is not very likely to do what the attacker wanted. [9] The file monitor MD5 database is also usually password or key protected making substituting files difficult. The attacker must just exploit the weaknesses of the file monitoring method's design. Most file monitoring HIDS have directories that are excluded because they will often have files created and removed or modified. Temporary directories are a common location that can be used to evade file monitors. These can be used for the Initial exploit code until the attacker can determine where they can safely write files, such as home directories or other locations that are not likely to be monitored. Then, once the system is compromised, the attacker simply needs to remove the file monitor or be stealthy and recomputed file hashes if they can crack the password on the MD5 database. Most HIDS will check startup scripts, registry entries, and Operating System startup programs for changes so that an attacker cannot modify these files. Few check any tasks that are launched from these locations. For instance, the administrator installs a program that has a custom item that is added to the startup script. Such as Norton adding the Virus Protection system tray icon. An attacker can view the startup locations and then replace the binaries that the startup runs. As long as the replacement still executes the original binary, most HIDS will not alarm because the sensitive startup location has not changed and it is not checking third party application binaries. The last problem with File Integrity based HIDS is that they tend to alarm too late to be of much use. Since they depend on the host operating system, if the operating system has been compromised there is no way for the Administrator to know that the attacker did not disable the HIDS application. Or the use of root kits will allow the attacker to hide files from Operating System as well as other users, which would also prevent the HIDS from finding the files necessary to generate an alarm.

III INTRUSION DETECTION

Secure network control protocols sometimes come with a significantly higher price. For example, in the previous section, if we want to solve the faulty intermediate router problem (FIRP) in the OSPF protocol itself, we need to pay the cost of RSA/MD5 (2 x 106 uses per LSA) in software. Our implementation of OSPF/key-MD5, on the other hand, only takes 40 _ 70 usec per LSA. This explains why the RSA/MD5 approach is not included in the IETF standards. Thus, unless we can develop a secure and much more efficient public key system, the FIRP cannot be prevented by the standardized OSPF protocols. It is interesting to ask whether a unpreventable security problem like FIRP can be possibly detected and isolated by a distributed intrusion detection system. The short answer is "YES." A typical solution is to let every router log all the LSAs that it receives and forwards over a period. By analyzing and correlating distributed log files, hopefully we can identify which routers are not faithful in forwarding LSAs. Two problems for these approaches include:

The amount of data in the log files might be very huge such that the analysis and correlating task Will take a long time. Thus, by the time we nail down the evil routers, the network might have Been seriously damaged. In other words, "how fast can we detect and isolate faulty intermediate Routers?" The network bandwidth needed for transmitting these data and coordinating among IDS modules could be very high. Intrusion detection is a form of fault-diagnosis. Faults (in a security system) are not supposed to happen, but the fact is that they do happen. As with all fault diagnosis systems, IDS give the wrong answers from time to time. Because it is so difficult to define what intrusion actually means in a generic sense (it's political) intrusion detection systems tend to err on the side of caution and report many false positives, i.e. false alarms. This is a very difficult problem to do in real time. What does real-time mean? Some attacks are stealthy and occur over many hours or days. How can we make a prompt notification about such attempts? The intrusion detection will have to be fast to detect quick break-ins, but have a long memory in order to see slow ones (like the thief digging a tunnel into the bank with a teaspoon). How will we be alerted or notified about intrusions? By alarm on the screen? By E-mail or pager alert? What if the attacker first knocks out E-mail or the pager link? User privacy is also a problem. If an intrusion detection system examines everything going on within the system, looking for suspicious behavior, is that an intrusion of privacy? What if humans never see the data, but only the warnings? Where do we draw the line between justified and unjustified surveillance? Law enforcement agencies have been arguing about that one for years!

IV ABOUT WIRELESS NETWORKS

Wireless networks are very common, for both organizations and individuals. Many laptop computers have wireless cards pre-installed. The ability to enter a network while mobile has great benefits. However, wireless networking has many security issues· [10] Hackers have found wireless networks relatively easy to break into, and even use wireless technology to crack into wired networks]. As a result, it's very important that enterprises define effective wireless security policies that guard against unauthorized access to important resources·[11] Wireless Intrusion Prevention Systems are commonly used to enforce wireless security policies.

The risks to users of wireless technology have increased as the service has become more popular. There were relatively few dangers when wireless technology was first introduced. Crackers had not yet had time to latch on to the new technology and wireless was not commonly found in the work place. However, there are a great number of security risks associated with the current wireless protocols and encryption methods, and in the carelessness and ignorance that exists at the user and corporate IT level·[12] Cracking methods have become much more sophisticated and innovative with wireless. Cracking has also become much easier and more accessible with easy-to-use Windows or Linux-based tools being made available on the web at no charge.

V CONCLUSION

A strong information security plan will include a multiplicity of technical and administrative controls intended to prevent intrusions and unauthorized activities from both internal and external threat agents. However, even with a set of strong security products and security policies it is impossible to insure that a network is secure. For this reason, networks that are mission critical or contain sensitive information are deploying intrusion detection and cyber security monitoring systems.

VI REFERENCES

1) Jeff Doyle and Jennifer Carroll, 2005. CCIE, Professional Development Routing TCP/IP. 2nd Edn, Published by Cisco Press, pp: 936. ISBN: 1587052024
2) Johnny S.K. Wong ,Iowa State University , Benjamin D. Up off ,Los Alamos National Laboratory, ,"Securing network traffic and intrusion Detection data"
3) Axelsson. "Research in intrusion-detection systems: A survey. Technical Report 98–17", Dept. of Computer Eng. Chalmers Univ. of Tech, SE-412 96, Goteborg, Sweden.
4) Intrusion Detection Preliminaries: Sanitizing Your Ecommerce Web Servers," FOCUS on Intrusion Detection: Intrusion De...: Sanitizing Your E-Commerce Web Server"
5) Casey, E, 'Case study: Network intrusion investigation – lessons in forensic preparation', Digital Investigation vol 2, issue 4. December 2005. pp 254-260
6) Baker and R. Atkinson, RIP-2: MD5 Authentication. http:// portal.acm.org /citation.cfm?id=RFC2082.

7) F. Baker and R. Atkinson,. OSPF MD5 Authentication.http://tools.ietf.org /html/draft-ietf-ospf-md5-01.
8) S.L. Murphy and M.R. Badger. Digital Signature Protection of the OSPF Routing Protocol. In Internet Society Symposium on Network and Distributed Systems Security.
9) Rivest, R. "The MD5 Message- Digest Algorithm." http://www.ietf.org/rfc/rfc1321.txt
10) Robert McMillan. "Once thought safe, WPA Wi-Fi encryption is cracked". IDG.
11) Kevin Beaver, Peter T. Davis, Devin K. Akin. "Hacking Wireless Networks For Dummies".
12) "Wireless Security Primer (Part II)". Windowsecurity.com , http://www.windowsecurity.com / articles / Wireless_Security_Primer_Part_II.html

Rolled Fingerprint Segmentation

[1]Ms. S.Malathi, [2]Ms. S.Uma maheswari, [3]Dr. C. Meena

GJCST Classifications: I.4.9, G.3, K.6.5

***Abstract*- A critical step in automatic fingerprint recognition is the accurate segmentation of fingerprint images. The objective is to segment the foreground and background from the rolled fingerprint image. The foreground region contains features that can be used for recognition and identification, while the background region is the noisy area around the boundary of the image. In this paper, we propose a technique that improves segmentation by using wiener filtering as a preprocessing step to eliminate noisy background. Our proposed method was tested in NIST SD27 dual resolution database, which contains rolled and plain fingerprints. This method was also tested in FVC2004 database. The experimental results demonstrate the effectiveness of the proposed method, concerning the extraction of ROI, especially in Rolled fingerprint image.**

Keywords- Segmentation, Filter, ROI, Wiener filter, Mean, and Variance threshold, rolled fingerprint.

I INTRODUCTION

The segmentation is concerned with splitting the fingerprint image into two regions (foreground and background). Segmentation is one of the first and integral pre-processing steps for any Fingerprint Recognition System and one of the major research area of image processing and pattern recognition. At present, the fingerprint image segmentation in the non-ideal conditions captured is a key issue to need for fingerprint recognition technology. It is used to identify the region of interest within an image. It is necessary to identify the foreground and background regions in fingerprint image preprocessing in an automatic fingerprint identification system (AFIS)[B. M. Mehtre, et al., 1987]. This can improve the accuracy of feature extraction and save processing time, consequently reduce the system costs [B. M. Mehtre, et al., 1987] [B. M. Mehtre, et al., 1989].

There are two types of fingerprint segmentation algorithms: unsupervised and supervised. Unsupervised algorithms extract blockwise features such as local histogram of ridge orientation [B. M. Mehtre, et al., 1987] [B. M. Mehtre, et al., 1989]. gray-level variance, magnitude of the gradient in each image block [Bazen A.M., et al., 2001], Gabor feature [Alonso-Fernandez, F., et al.,2005, Bazen, A., Gerez, S., 2001,] . Practically, the presence of noise, low contrast area, and consistent contact of a fingertip with the sensor may result in loss of minutiae or more spurious minutiae Supervised method usually first extracts several features like coherence, average gray level, variance and Gabor response [Bazen, A., Gerez, S., 2001, Pais Barreto Marques, et al., A.C.,2005, Zhu, E et al.,2006], then a simple linear classifier is chosen for classification This method provides accurate results, but its computational complexity is higher than most unsupervised methods.

Segmentation process can become very complex and intricate because the boundary between the region of interest and the background blurs due to presence of noise. Various segmentation methods are developed. However, these methods are not satisfying. For example, if there is highly noisy background region surrounding the poor contrast foreground of the fingerprint image, these methods will fail to separate background with foreground. A robust segmentation method is required to deal with low quality images and to be insensitive to the contrast of the original images. In this paper, we analyzed the traditional mean and gray variance method gave satisfactory results only for noiseless background regions and not fingerprints lifted from fingerprint cards, which have noisy background. Hence we put forward wiener filtering as a preprocessing step before segmenting based on mean and variance method. On the base, a further improved gray mean and variance method is proposed accordingly to achieve more effective segmentation effect.

II FINGERPRINT IMAGE

Fingerprint images can be broadly classified into three categories, namely, (i) rolled/full, (ii) plain/flat and (iii) latent [V. N. Dvornychenko, et al.2006, P.Komarinski, 2001, D. Maltoni, et al.2003] as shown in Fig.1. Rolled fingerprint images are obtained by rolling a finger from one side to the other ("nail-to-nail") in order to capture all the ridge-details of a finger. Plain impressions are those in which the finger is pressed down on a flat surface but not rolled. Rolled and plain impressions are obtained either by scanning the inked impression on paper or by using live-scan device. In contrast, latent fingerprints are lifted from surfaces of objects that are inadvertently touched or handled by a person through a variety of means ranging from simply photographing the print to more complex dusting or chemical processing

a)Rolled b) Plain c)Latent

Fig.1 Types of Fingerprints

[1].Mrs.S.Malathi, Research Scholar, Department of Computer Science, Avinashilingam University for Women, Coimbatore-641 043,Tamilnadu India, phone:9994974487,e-mail:malathi_cc@avinuty.ac.in.

[2].S.Umamaheswari, Project Associate, Computer Centre,Avinashilingam, University for women, Coimbatore-641043, Tamilnadu, India, phone: 9003499881, e-mail: uma19@gmail.com

[3].Dr.C.Meena, Head, Computer Centre, Avinashilingam University for Women, Coimbatore-641043, Phone: 9894269991 e-mail: meena_cc@avinuty.ac.in.

III FINGERPRINT SEGMENTATION

A fingerprint image segmentation algorithm receives an input fingerprint image, applies a set of intermediate steps on the input image, and finally outputs the fingerprint foreground region where the ridge structure is coherent. The segmentation result should satisfy the following conditions

i. It should not be sensitive to the contrast in the image.
ii. It should detect smudged and noisy regions.
iii. The result of segmentation should be independent of whether the input image is an enhanced image or a raw image.
iv. It should give consistent result for a variety of images expected by the application

Fingerprint images have various resolution and normally 500 dpi fingerprint images are used [FVC2004 website]. The fingerprint image obtained from inked or live scan device, which includes noisy background, and smudge regions may cause problem in further steps of AFIS. So, some preprocessing is required to mitigate the effect of such unwanted region.

A. Variance And Mean Based Segmentation

Fingerprint image of the foreground region consist of fingerprint ridge and valley line, whose ridge and valley is a line of black-and-white texture, thereby the greater variance; while the gray background region do not hardly change, thereby the variance is relatively small. Based on the characteristic, we can make use of fingerprint local variance of fingerprint segmentation, this method is called variance

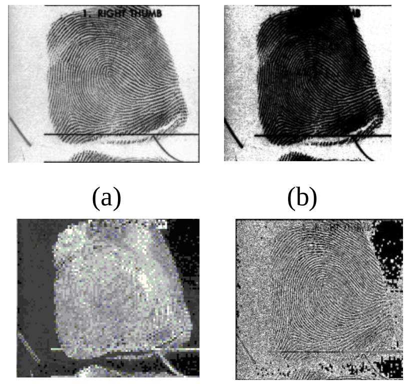
Here the fingerprint image is partitioned in blocks of 16 x 16 pixels. Then, each block is assigned to ROI according to criterion defined as

$$\frac{std_b}{m_m} \geq \alpha \frac{std_i}{m_i} \qquad (1)$$

Where std_b and m_b are corresponding standard Deviation and mean of each block, $stdi$ and m_i are Standard deviation and mean of whole area of fingerprint image, respectively. Also, α is a suitable threshold value. Then the gray scale variance for each block is calculated as

$$Var(I) = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j-0}^{N-1} \left(I(i,j) - M(I) \right)^2 \qquad (2)$$

where Var (I) the variance for block k, I(i, j) is the gray-level value at pixel (i, j), and M(I) is the mean gray-level value for the block I. If the variance is less than the global threshold, then the block is assigned to be a background region; otherwise, it is assigned to be part of the foreground. The result of segmentation is shown in Fig.2.



(a) (b) (c) (d)

Fig.2: Segmentation results for mean and variance threshold method (a) Original Image (b Normalized Image (c) Variance Image (d) Segmented Image

IV PROPOSED METHOD

The quality of fingerprint images captured by fingerprint scanner is influenced by many facts, such as dryness and wetness of finger, cleanness of scanner and dryness and wetness of the weather. In order to clean all the artifacts caused by the noise wiener filtering is applied to original image as preprocessing, to prepare images for further segmentation processing. The filtered image is then normalized because normalization improves the entire image contrast and not change fingerprint valley quality and finally the normalized image is fed to the mean and variance segmentation process.

I Image Binarization

The input image (I) is converted to a binary image (Ib).The threshold value is obtained using Global image threshold, which chooses the threshold to minimize the intraclass variance of the threshold black and white pixels.

II Wiener Filtering

The binary image Ib is subjected to Wiener filtering which has the form

$$W(f_1, f_2) = \frac{H^*(f_1, f_2) S_{xx}(f_1, f_2)}{|H(f_1, f_2)|^2 S_{xx}(f_1, f_2) + S_{\eta\eta}(f_1, f_2)}, \qquad (3)$$

Where $S_{xx}(f_1, f_2)$, $S_{\eta\eta}(f_1, f_2)$ are respectively power spectra of the original image and the additive noise, and $H(f_1, f_2)$is the Blurring filter. It is easy to see that the Wiener filter has two separate parts, an inverse filtering part and a noise smoothing part. It not only performs the deconvolution by inverse filtering (highpass filtering) but also removes the noise with a compression operation (lowpass filtering). Then Fourier transform of the image is taken and multiplied by the transform of the filter function (h). The process eliminates low frequency signals in the Fourier domain, which is the noise and emphasizes the high frequency signals which is the line. The output image is then created by taking the inverse Fourier transform of the product. The filter multiplies each pixel in the Fourier image by this filter to give the final filtered image (I_f). The results indicated that the Wiener filter not only increases the signal/noise ratio but also improves the contrast.

III Normalization

The filtered image (If) is then normalized. Normalization is a process that changes the range of pixel intensity values. Let, (i, j) denote the gray-level value at pixel (i, j), M and Var denote the estimated mean and variance of I, respectively, and G(i,j) denote the normalized gray-level value at pixel (i, j). The filtered image is then normalized as follows:

$$G(i,j) = M_0 + \sqrt{\frac{Var_0 (I(i,j) - M)^2}{Var}} \quad \text{if } I(i,j) > M \qquad (2)$$

$$G(i,j) = M_0 - \sqrt{\frac{Var_0 (I(i,j) - M)^2}{Var}} \quad \text{Otherwise} \qquad (3)$$

Where M_0 and Var_0 are the desired mean and variance values, respectively

IV Segmentation

The normalized image is then subjected to mean and variance segmentation to produce the final output image. The result of the proposed segmentation method is shown in Fig. 3

(a) Original Image (b)Wiener filtered Image

(c)Normalized image (d) Segmented image

Fig. 3:Segmentation results for proposed method

V EXPERIMENTAL RESULTS

In this section, we present experimental result to demonstrate the performance of the proposed technique. The data set used in the experiments is the NIST SD27 dual resolution database, which contain rolled fingerprint image and tested in FVC 2004 database images. The Experimental results of the proposed method are shown in Fig.4.

Sample FVC2004 Database	Segmented Image
	
	
	

Fig.4: Experimental results for FVC2004 database images

VI CONCLUSION

In this paper, our goal is to produce a better segmentation to successful extract the boundary lines of the clear ridge area for the rolled fingerprint. The proposed method is employed as adding a preprocessing step in combination of mean and variance based method, to extract the ROI. Applying Wiener filtering as preprocessing step greatly improved the segmentation results. The experimental results demonstrate the effectiveness of the proposed method, concerning the extraction of ROI, especially in rolled fingerprint images and also robust, reliable and effective in identifying region of interest even in very low quality images.

VII REFERENCES

1) M. Mehtre, et al., 1987,Segmentation of Fingerprint Images Using the Directional Image, Pattern Recognition, (20) 4,429-435.
2) B.M.Mehtre and B.Chatterjee, 1989,Segmentation of Fingerprint Images-A Composite Method, Pattern Recognition, (22) 4,381-385.
3) Bazen A.M., Sabih H.G, 2001,Segmentation of fingerprint images, Proc. ProRISC .
4) Alonso-Fernandez, F., Fierrez-Aguilar, J., Ortega-Garcia, J.,2005,An enhanced gabor filter-based segmentation algorithm for fingerprint recognition systems, Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 239–244.

5) Bazen, A., Gerez, S., 2001, Segmentation of fingerprint images,Workshop on Circuits Systems and Signal Processing,. 276–280.
6) Pais Barreto Marques, A.C., Gay Thome, A.C., 2005, A neural network fingerprint segmentation method, Fifth International Conference on Hybrid Intelligent Systems,6.
7) Zhu, E., Yin, J., Hu, C., Zhang, G., 2006, A systematic method for fingerprint ridge orientation estimation and image segmentation. Pattern Recognition **39**(8), 1452–1472.
8) N. Dvornychenko and M. D. Garris., 2006 ,Summary of NIST latent fingerprint testing workshop. NISTIR 7377 retrieved from ,http://fingerprint.nist.gov/latent/ir_7377.pdf.
9) P.Komarinski, editor, 2001,Automated Fingerprint Identification Systems (AFIS), Elsevier Academic Press.
10) Maltoni, D. Maio, A. K. Jain, and S. Prabhakar, 2003,Handbook of Fingerprint Recognition, Springer-Verlag.
11) Fingerprint Verification Contest 2004:FVC2004: Retrieved from http://bias.csr.unibo.it/fvc2004.html.

Positional Approach for Alphabetic Sort Algorithm

Muhammad Bilal Qureshi, Muhammad Shuaib Qureshi
Department of Computer Science, Faculty of Basic & Applied Sciences,
International Islamic University Islamabad, Pakistan
{muhdbilal.qureshi, qureshi.shuaib} @gmail.com

GJCST Classifications:
F.2.2, E.5, E.1, F.1.3, G.2, H.2, H.3

***Abstract-* In This paper, the authors present a positional algorithmic approach for alphabetic sort. Results are achieved in linear time. Within this approach two embedded algorithms, Binary Search and Counting Sort are executed in parallel to achieve the goal. In this approach a Pre-Processor or Priority Queue is used, which minimizes time complexity. The algorithm is linear in speed. Time Complexity of this newly proposed algorithm is Θ (n). The interesting feature of this algorithm is that the order of alphabets is not change and the approach is too much simpler**

Keywords- Algorithm, Priority Queue, Sort, Search, Complexity, Analysis.

I. Introduction

Algorithm is any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output. An algorithm is thus a sequence of computational steps that transform the input into the output[1][2]. Algorithm is a tool for solving a well-specified computational problem. The algorithm describes a specific computational procedure for achieving the input/output relationship [1][2][3][9].

An algorithm is an orderly systematic procedure to solve a problem. The term algorithm is derived from the title Khowrizmi of ninth-century Persian mathematician Abu Musa al-Khowrizmi, who is credited with systematic study and development of important algebraic procedures. An algorithm is a sequence of unambiguous instructions for solving a problem in a finite amount of time.

A large variety of problems in computer science, mathematics and other disciplines depend on the use of algorithms for their solutions. The broad categories of applications types are:

i. Searching Algorithms (Linear and non-linear)
ii. Sorting Algorithms (Elementary and Advanced)
iii. Strings Processing (Pattern matching, Parsing, Compression, Cryptography)
iv. Optimization Algorithms (Shortest routes, minimum cost)
v. Geometric Algorithms(Triangulation, Convex Hull)
vi. Image Processing (Compression, Matching, Conversion)
vii. Data Mining Algorithms(Clustering, Cleansing, Rules mining)
viii. Mathematical Algorithms (Random number generator, matrix operations, FFT, etc)

II. Binary Search Algorithm

Binary Search is an algorithm for locating the position of an element in a sorted list by checking the middle, eliminating half of the list from consideration, and then performing the search on the remaining half. If the middle element is equal to the sought value, then the position has been found; otherwise, the upper half or lower half is chosen for search based on whether the element is greater than or less than the middle element. The method reduces the number of elements needed to be checked by a factor of two each time, and finds the target value, if it exists in logarithmic time. A binary search is a divide and conquer search algorithm. [1][3][9][4][5].

Pseudo Code for Binary Search Algorithm

BINARY-SEARCH

i. start = 1; end = n ;
ii. while (start < end)
iii. {
iv. middle = (start + end) / 2 ;
v. if s > a middle then start = middle + 1 ;
vi. else end = middle – 1 ;
vii. }
viii. if (s == a start) location = start ;
ix. else location = 0;

Time & Space Complexities of Binary Search Algorithm: [3][4][5]

Worst case performance	=	O (log n)
Best case performance	=	O (1)
Average case performance	=	O (log n)
Worst case space complexity	=	O (1)

III. Counting Sort Algorithm

Counting Sort (sometimes referred to as Ultra Sort or Math Sort) is a sorting algorithm, which takes advantage of knowing the range of the numbers in the array to be sorted (array A). It uses this range to create an array C of this length. Each index i in array C is then used to count how many elements in A have the value i; then counts stored in C can then be used to put the elements in A into their right position in the resulting sorted array. The algorithm was created by Harold H. Seward [1][3][6].

Pseudo Code for Counting Sort Algorithm:

COUNTING -SORT (A)

i. n=length[A]
ii. for j← 0 to n do
iii. C[j]← 0

iv. for j ← 1 to n do
v. k← A[j]
vi. C[k]← C[k] +1
vii. for j ← 1 to n do
viii. C[j]← C[j] + C[j-1]
ix. for j ← n downto 1 do
x. i ← A[j]
xi. k← C[i]
xii. B[k]← A[j]
xiii. C[i] ← C[i]-1
xiv. return B

Time & Space Complexities of Counting Sort: [5][6][1]

Worst case performance = O (n + k)
Best case performance = O (n + k)
Average case performance = O (n + k)
Worst case space complexity = Θ (n + k)

IV. Proposed Alphabetic Sort Algorithm

The proposed algorithm uses two arrays named as array A and array B. Array A is an input array while as Array B is a Pre-Processed array or Priority Queue. In the whole process two well-known algorithms, Binary Search and Counting Sort are used in parallel. The Binary Search algorithm is used for searching the corresponding positions while as Counting Sort algorithm is used for sorting purposes.
Steps involved in the proposed alphabetic sort algorithm.

i. Comparison is made between array A and pre-processed array B for searching the corresponding positions of alphabets in Priority Queue B. For this purpose Binary search algorithm is used.
ii. Substituting the alphabets in array A with corresponding positional digits.
iii. For sorting array A, Counting Sort algorithm is used.
iv. Again fetching corresponding elements of the positional digits held in Array A from Priority Queue B. So Binary Search algorithm is used.
v. Replacing the digits in the array A by the fetched elements.

V. Graphical Depiction of the Proposed Algorithm

Input Array A:

1	2	3	4	5	6	7	8	9	10	11
a	t	c	b	w	z	a	b	g	k	h

Preprocessed Array or Priority Queue B:

1	2	3	4	----------------	24	25	26
a	b	c	d	---------------	x	y	z

Take alphabet from array A, fetch its corresponding location/position from Priority Queue B and then replace alphabet in array A by this fetched positional digit.

	1	2	3	4	5	6	7	8	9	10	11
A:	1	20	3	2	23	26	1	2	7	11	8

Now sort this array A by using sorting algorithm. The sorted array will be as,

	1	2	3	4	5	6	7	8	9	10	11
A:	1	1	2	2	3	7	8	11	20	23	26

Now again fetch corresponding elements of these new positional digits held in Array A from Priority Queue B. Replace the digits in this updated array A by the fetched elements. Now the array A will be as,

	1	2	3	4	5	6	7	8	9	10	11
A:	a	a	b	b	c	g	h	k	t	w	z

Sorted alphabets list.

Here a pre-processor is also introduced, due to which time complexity is minimized.

VI. Analysis of the Proposed Algorithm

For the whole process twice Binary Search algorithm, once sorting algorithm and twice substitution is used.
Time Complexity of the proposed algorithm is,

$$T(n) = \Theta(\lg n) + \Theta(\lg n) + \Theta(n) = \Theta(n)$$

(Ignoring constant terms & values)

n	$lg\ n$	$\sqrt{n}$	$n\ lg\ n$	n^2	n^3	2^n	$n!$	n^n
2	1	1.4	2	4	8	4	2	2
4	2	2	8	16	64	16	24	256
8	3	2.8	24	64	512	256	40,320	16,777,216
16	4	4	64	256	4,096	65,536	20,922,789,888,000	1.845×10^{19}
32	5	5.7	160	1,024	32,768	4,294,967,296	2.631×10^{35}	1.461×10^{48}
64	6	8	384	4,096	262,144	1.8×10^{19}	1.269×10^{89}	3.940×10^{115}
128	7	11	896	16,384	2,097,152	3.4×10^{38}	3.856×10^{215}	5.283×10^{269}

The logarithm (lg n) has lower growth rate and exponential function nn has the highest growth rate.
The symbolic representation of the relationship of functions growth rates is,

$$lg\ n < \sqrt{n} < n < n\ lg n < n^2 < n^3 < 2^n < n! < n^n$$

Therefore, Time Complexity of the proposed algorithm is Θ (n), Linear Time Complexity.

VII. Conclusion

The linear Time Complexity T(n) = Θ (n) is achieved by this algorithm. The flavor of this newly proposed algorithm is its speed and simplicity. The interesting feature here is that the order of alphabets is not changed.

VIII. References

1) Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald L.; Stein, Clifford, Introduction to Algorithms (3rd ed.), MIT Press.
2) A.Levitin, Introduction to Design and Analysis of Algorithms, Pearson Edition, 2006.
3) R.E.Neapolitan, K.Naimipour, Fundamentals of Algorithms, Heath and Company.

4) http://mathworld.wolfram.com/ BinarySearch.html , 2009.
5) http://en.wikipedia.org/wiki/ Binary_search_ algorithm , 2009.
6) Anthony Ralston, Edwin D. Reilly, David Hemmendinger, ed (2003). "Ultrasort". Encyclopedia of Computer Science (4th ed.). Wiley. pp. 1660-1661.
7) http://www.shannarasite.org/, 2009
8) D.F. Stubbs and N.W. Webre, Data Structures with Abstract Data Type and Ada, PWS-KENT Publishing Company, pp.301-341, 1993.
9) S.Basse, V.A.Gelder, Computer Algorithms, Pearson Edition Inc.

Advanced Natural Language Translation System

Johnson Manuel Devadoss

GJCST Classifications:
I.2.7, I.2.1, H.5.2, D.2.2, H.1.2, I.3.6

Abstract- **The present invention is an Advanced Natural Language Translation System (ANLTS). It discloses a method to address the most common variation in the world, which is communication gap between people of different ethnicity. Typically, communication is said to be successful between two people if someone speaks and opponent party can understand. In other words, the intended recipient's brain language area can comprehend the speech. The problem of not understanding the speech of others is the cause of language barriers. Therefore, this invention discloses a method to solve the language barrier problem where it is capable of interpreting meaning of speech in one language to a language native to another – to a language the recipient brain can comprehend.**

Imagine a world where we can communicate with our native language to everyone without the need of human translators, interpreters, hand-held device, and language translation books. In order to facilitate language translation, this present invention recognizes the speech, collects the language comprehensive information from every recipient's brain language area within the audible range, and sends it to voice processing center for analyzing. Then, it translates the collected speech to intended recipient(s) native language by using more than 6,700 language dictionaries database. The translated language is retransmitted in audible frequency to the language area of each recipient(s) brain.

I. Field of the Invention

The present invention relates generally to a speech translating method, and more particularly, to automatically translate speech from one language to a language native to another which is understandable by the language (Wernicke/Broca) area of intended recipients' brain.

II. Background of the Invention

Languages are humankind's principle tools for interacting expressing ideas, emotions, knowledge, memories, and values. Languages are also primary vehicles of cultural expressions and intangible cultural heritage, essential to the identity of individuals and groups. Safeguarding endangered languages is a crucial task in maintaining cultural diversity worldwide. According to researchers, more than 6,700 languages are spoken in 228 countries. For example, in India more than 250 languages are used for speech. People like to speak in their native language and prefer to communicate with others in their native language. This makes it difficult for people to travel to foreign states or countries as they need to learn the foreign language.

In field of entertainment, if someone wants to watch a foreign movie / performance, they experience problems in clearly understanding the event. Obviously, many electronic translator equipments are available in the world, but it only supports popularly spoken languages.

Language barriers and misunderstandings can get in the way of effective communication and create complications in the workplace, including problems with safety. A recent Business Journal article on the rising number of foreign national workers in Charlotte-Mecklenburg's construction industry pointed out - those workers who speak little or no English are at much greater risk of having an accident on the job because of not having a full grasp of safety standards.

Approximately 22% of the Sheraton Corporation's workforce is Hispanic, primarily Mexicans. Language is the main barrier here. To help its employers deal with the language challenge, the company has bilingual employees to serve as translators and mentors. In addition, all printed material is provided in both the essential languages Spanish and English. Another example is Woonsocket Spinning Company - Woonsocket is one of the few remaining woolen mills in the United States. 70% of their employees are foreign-born. Overcoming language barriers is the greatest challenge for both workers and the employer. To help with this, the company hires interpreters or has other employees who speak the language help the non-English speaking employees, particularly during orientation and training. Studies like this suggest companies spend a lot of time and effort to overcome language barriers among employees.

Patients from under developing countries seeking medical care always need to be accompanied with human translators to explain their medical problems and to understand physician's advice. Results from a survey of leading physician organizations, medical groups and other health care associations in California suggest that nearly half (48%) of the 293 respondents knew of an instance in which a patient's limited English proficiency impacted his or her quality of care. The three biggest complaints were difficulty of history talking, wrong diagnosis, and a general frustration with the lack of nuance in physician-patient communication with patients who have Limited English Proficiency (LEP).

In the ever-growing IT industry, people from various nationalities collaborate in meetings and conferences. Due to language barrier, they cannot communicate freely resulting in business people investing lot of time and money learning new languages.

Even in marketing, due to language as barrier quality retail and consumer product owners struggle to market their products on international market.

There are number of language translation systems available in the world designed and developed to translate an inputted language to another language. All these methods/systems require a device to capture the voice and deliver. Such systems are known in the prior patents as disclosed in U.S. Pat. No 4,882,681 to Brotz et al for Remote Language Translating Device. This prior patent disposes the translation of conversation between the users by transmitting/receiving speech using external hardware device. However, people would not prefer to carry or even remember to carry the hardware device all the time. Also the disadvantage of such system is that it can be used to convert only a certain number of languages which are pre-programmed on the device.

U.S Patent. No 6,161,082 to Goldberg et al for Network based language translation system performs a similar task. It disposes a network based language translation system - has a translation software installed on the network. It proves that software over network can do speech translation, but user still has to set their language preferences. More than 67% of world's population do not or have limited computer knowledge, so they cannot set their language preferences and operate high-tech gadgets. Another recent patent is US Pat. No US 2009/0157410 to Donohoe et al for speech translating system. This recent patent discloses a system for translating speech from one language to a language selected from a set of languages. It can be applicable only for limited amount of users but more than 6,700 languages are being used by people to express their thoughts around the world.

Another patent is U.S. Pat. No. 4,641,264 to Nitta et al for a Method of Automatic Translation between Natural Languages - this discloses a system for the translation of entire sentences. Then again it also requires an input and output device to capture and deliver the speech. It is not capable to determine the recipients' understandable language. We have to manually set the targeted language or select from pre-defined languages (as target) in the device.

Therefore to overcome all the above language barriers, there is a need for a system to perform automatic translation of speech wherein when one speaks in a native language others are able to comprehend in their own native languages without interpreters, hand-held device and language translation books.

III. SUMMARY OF THE INVENTION

Speech translation is converting to a language that the language area of recipient human brain can understand. Recipient(s) may not be able to comprehend the speech because their brain language area is not tuned to understand the spoken language. In medical terms, it is called ―Wernicke's Aphasia‖.

The language area of human brain is called ―Wernicke‖ which is nothing but a neuron in human brain capable to interpret words that we hear or read. Wernicke then relays this information via a dense bundle of fibers to Broca's area that generates words that we speak in response. Wernicke/Broca together has all the language comprehensive information needed for understanding speech.

This invention disposes a process where humans are not going be aware a translation is happening in the background. They will be able to speak their own native language but others surrounding them can automatically understand the speech in their own native language. This system therefore bridges all communication gaps among people.

The main object of the present invention is to provide an Advanced Natural Language Translation System that is capable of providing a translation of speech in one language to a language native to another which is understandable by the language (Wernicke's/Broca's) area of the recipients' brain. The present invention thereby replaces interpreters, hand-held device and language translation books.

The present Advanced Natural Language Translation System (ANLTS) invention has two main logical processing units – the Intelligent Natural Language Program (INLP) and the Voice Processing Center. The human ear can hear frequencies at ~70 decibels. When we talk our thoughts are converted into voice signals and transmitted into the surrounding regions. This system employs a data broadcasting technique to broadcast the Intelligent Natural Language Program (INLP) over a wide area using radio waves.

The Intelligent Natural Language Program (INLP) is like a Pico-planner program on the network that looks for human voice signals. It further comprises of an Intelligent Speech Recognition Algorithm and the Language Area Acquisition Algorithm. The Intelligent Speech Recognition Algorithm provides phoneme-level sequence to the parser where each has a probability of being correct. The Language Area Acquisition Algorithm collects information from the language area of the human brain and transmits it to Voice Processing Center. Radio waves are used to transfer signals to and from the Voice Processing Center.

Voice Processing Center receives the signals having language comprehensive information and several competitive phoneme or word hypotheses each of which are assigned the probability of being correct. Voice Processing Center operates using a Language Area Inference Engine. The Language Area Inference Engine is an artificial intelligence program that tries to derive native language information from a knowledge base. Language Area Inference Engine is considered to be a special case of reasoning engines, capable of employing both induction and deduction methods of reasoning.

This invention facilitates tourism. People are now free to travel to any corner of the world. They do not have to carry any hand-held devices. This invention facilitates people to enjoy foreign movie/performances without need of friends as human translators or sophisticated translation devices. Patients can be provided with the right care that they require. This invention also eliminates all miscommunications and reduces death totality in industries. Employers can hire people from any ethnicity, as language will no longer be a barrier.

This invention also facilitates businesspersons from any country to expose their quality products worldwide within a less budget. Everyone can continue to effectively communicate in their own native language in meetings and conferences while employers can save money on language translation books.

All these put together with other aspects of the present invention, along with the various features that describe the present invention, especially those pointed out in the claims section form a part of the present invention. To gain more knowledge of the present invention understanding of the drawings attached and the detailed description is highly essential.

IV. Brief Description of the Drawings

Fig 1.a illustrates two people of the system speaking in their native language using Advanced Natural language Translation System.

Fig 1.b illustrates a group of five people of the system exchanging conversation in their native language using Advanced Natural Language Translation System.

Fig 1.c illustrates a group of business people of the system exchanging their business conversation in their native language using Advanced Natural Language Translation System

Fig 1.d illustrates representative of the system addressing a crowd in his native language using Advanced Natural Language Translation System.

Fig 2 illustrates the detailed operation of this invention.

Fig.3 is a partially schematic, isometric illustration of a human brain illustrating areas associated with language comprehension

Fig 4 illustrates a processing flow of this invention

V. Detailed Description of the Invention

Reference will now be made in detail to the preferred embodiments of the present invention, examples of which are illustrated in the accompanying drawings, wherein like reference numerals refer to like elements throughout.

Communication is said to be effective between two people, if one speaks and opponent party can understand. In other words, the intended recipients' brain language area can comprehend the words/sentence/speech. The present invention does that –interpreting meaning of word(s) in a language understandable by Wernicke's of intended recipient brain.

In human beings, the left hemisphere usually contains the specialized language areas. While this holds true for 97% of right-handed people, about 19% of left-handed people have their language areas in the right hemisphere and as many as 68% of them have some language abilities in both the left and the right hemisphere. Both the two hemispheres are thought to contribute to the processing and understanding of language: the left hemisphere processes the linguistic of prosody, while the right hemisphere processes the emotions conveyed by prosody.

Fig. 3 is an isometric, left side view of the brain 300. The targeted language areas of the brain 300 can include Broca's area 308 and/or Wernicke's area 310. Sections of the brain 300 anterior to, posterior to, or between these areas can be targeted in addition to Broca's area 308 and Wernicke's area 310. For example, the targeted areas can include the middle frontal gyrus 302, the inferior frontal gyrus 304, and/or the inferior frontal lobe 306 anterior to Broca's area 308. The other areas targeted for stimulation can include the superior temporal lobe 314, the superior temporal gyrus 316, and/or the association fibers of the arcuate fasciculus 312, the inferior parietal lobe 318 and/or other structures, including the supramarginal gyrus, angular gyrus, retrosplenial cortex and/or the retrosplenial cuneus of the brain 300.

There are four distinct cortical language-related areas in the left hemisphere. These are: (1) a lateral and ventral temporal lobe region that includes superior temporal sulcus(STS) 316, middle temporal gyrus (MTG), parts of the inferior temporal gyrus (ITG) and fusiform and parahippocampal gyri; (2) a prefrontal region that included much of the inferior and superior frontal gyri, rostral and caudal aspects of the middle frontal gyrus, and a portion of the anterior cingulate; (3) angular gyrus; and (4) a perisplenial region including posterior cingulate, ventromedial precuneus, and cingulate isthmus. These regions were clearly distinct from auditory, premotor, supplementary motor area (SMA), and supramarginal gyrus areas that had been bilaterally activated by the tone task. The other large region activated by the semantic task is the right posterior cerebellum.

The first language area within the left hemisphere is called Broca's area 308. The Broca's area 308 doesn't just handle getting language out in a motor sense it is more generally involved in the ability to deal with grammar itself, at least the more complex aspects of grammar. The second language area is called Wernicke's area 310. Wernicke's Aphasia is not only about speech comprehension People with Wernicke's Aphasia also having difficulty in naming things. They often respond with words that sound similar, or the names of related things, as if they are having a very hard time with their mental "dictionaries." For example, hearing the difference between —bad‖ and —bed‖ is easy for native English speakers. The Dutch language however, makes no difference between these vowels, and therefore the Dutch have difficulties hearing the difference between them. This problem is exactly what patients with Wernicke's aphasia have in their own language: they can't isolate significant sound characteristics and classify them into known meaningful systems.

By analyzing data from numerous brain-imaging experiments, researchers have now distinguished three sub-areas within Wernicke's area 310. The first sub-area responds to spoken words (including the individual's own) and other sounds. The second sub-area responds only to words spoken by someone else but is also activated when the individual recalls a list of words. The third sub-area is more closely associated with producing speech than with perceiving it. All of these findings are still compatible, however, the general role of Wernicke's area 310, relates to the representation of phonetic sequences, regardless of whether the individual hears them, generates them, or recalls them from memory.

Fig.1 illustrates the broad structure of this present invention. Fig.1.a shows a woman 102 saying her name in her native language French——‖ 106. The present invention employs a data broadcasting technique to broadcast the Intelligent

Natural Language Program (INLP) 110 over a wide area using radio waves. Intelligent Natural Language Program 110 is a Pico-program which is advanced version of natural language processing programs i.e., ELIZA, SHRDLU, A.L.I.C.E, written in special kind of Pico-Planner programming language. The Intelligent Natural Language Program 110 has two Algorithms: - Intelligent Speech Recognition Algorithm 112 and Language Area Acquisition Algorithm 114. Intelligent Speech Recognition Algorithm 112 captures and improves the recognition rate of the spoken dialog in three ways. First, generate phoneme sequence from recognized voice pitches. This phoneme sequence contains substitution, insertion, and deletion of phonemes, as compared to a correct transcription, which contains only expected phonemes. Second, activate a hypothesis as to the correct phoneme sequence from noisy phoneme sequence by filtering out false first choices of the hypotheses and selecting grammatically and semantically plausible best hypotheses. Third, provide a phoneme and word hypotheses to the parser which consist of several competitive phoneme or word hypotheses each of which are assigned the probability of being correct. The Intelligent Speech Recognition Algorithm captures the spoken sentence of woman ——‖ 106 and provides phoneme-level sequence i.e., phoneme and word hypotheses.

Intelligent Natural Language Program 110 initiates the Language Area Acquisition Algorithm 114 to gather the language comprehensive information from the single listener man 104 who is in the audible range of the woman's 102 voice. The Language Area Acquisition Algorithm 114 is capable of collecting the language area comprehensive information like Language Comprehension, Semantic Processing, Language Recognition, and Language Interpretation from Wernicke's area 310 and Broca's area 308. It will collect the information from Wernicke's area 310 of single listener man's brain namely Superior Temporal Sulcus and Middle Temporal Gyrus, Inferior Temporal Gyrus, Fusiform Gyrus, Angular Gyrus, Inferior Frontal Gyrus, Rostral and Caudal Middle Frontal Gyrus, Superior Frontal Gyrus, Anterior Cingulate, and Perisplenial Cortex/Precuneus. The language comprehensive, phonemes and word hypotheses are collected and sent to Voice Processing Center over datacasting network.

In Fig.2 Voice Processing Center 210 receives the signals having language comprehensive information and phoneme-level sequence -- each of which is assigned the probability of being correct. The language comprehensive information is compared with cache database. In Fig.4 Cache database 408 is a collection of native language data. Retrieval of original native language is expensive owing to longer access time; the cache is a cost effective way to store the original native language or other computed languages. It acts like a temporary storage area where frequently accessed native language data can be stored for rapid access. Once the data is stored in the cache, it can be used in the future by accessing the cached copy rather than re-fetching or re-computing the original native language data. Cache database 408 is thus an effective approach to achieve high scalability and performance.

Voice Processing Center 210 is operated by a Language Area Inference Engine 412, which includes a knowledge base of all possible language area information. An artificial intelligence program tries to derive native language information from a knowledge base for woman's 104 and man's 102-language comprehensive information. It tries to derive reasoning from the knowledge base. The separation of Language Area Inference Engine 412 as a distinct software component stems from the typical speech translating system architecture. This architecture relies on a data store, or working memory, serving as a global database representing facts or assertions about the Wernicke's 310 and Broca's 308 areas of human brain; on a set of rules which constitute the program, stored in a rule memory of production memory; and on an inference engine, required to execute the language comprehensive rules. The Language Area Inference Engine 412 must determine which language comprehensive rules are relevant to a given language comprehensive data store configuration and choose which one(s) to apply. This control strategy is used to select native languages.

The Language Area Inference Engine 412(In Fig 4) can be described as a form of finite state machine with a cycle consisting of three action states: match, select, and execute language comprehensive rules.

In the first state, match language comprehensive rules, the Language Area Inference Engine 412(In Fig 4) finds all of the language comprehensive rules that are satisfied by the current contents of the data store. When language comprehensive rules are in the typical condition-action form, the next step is to test the conditions against the working memory. The language comprehensive rule matching are all candidates for execution: they are collectively referred to as the conflict set. Note that the same language comprehensive rule may appear several times in the conflict set if it matches different subsets of data items. The pair of a language comprehensive rule and a subset of matching data items are called an instantiation of the language comprehensive rule.

The Language Area Inference Engine 412 (In Fig 4) then passes along the conflict set to the second state, select language comprehensive rules. In this state, the Language Area Inference Engine 412 (In Fig 4) applies LEX strategy to determine which language comprehensive rules will actually be executed. The selection strategy can be hard-coded into the engine or may be specified as part of the model. The LEX strategy orders instantiations based on recency of the time tags attached to their language comprehensive data items. Instantiations with language comprehensive data items having recently matched language comprehensive rules in previous cycles are considered with higher priority. Within this ordering, instantiations are further sorted on the complexity of the conditions in the rule.

Finally, the selected language comprehensive instantiations are passed over to the third state, execute language comprehensive rules. The Language Area Inference Engine 412 (In Fig 4) executes or fires the selected language comprehensive rules, with the language comprehensive instantiation's data items as parameters. Usually the actions in the right-hand side of a language comprehensive rule change the data store, but they may also trigger further processing outside of the Language Area Inference Engine 412(In Fig 4). Since the data store is usually updated by firing rules, a different set of rules will match during the next cycle after these actions are performed. The Language Area Inference Engine 412 then cycle back to the first state and are ready to start over again and it stops either on a quiescent state of the data store when no rules match the data.

The selected native languages are then compared 414 (In Fig 4) with the source native language. If both native language information are same then translation will not take place otherwise a translation will take place. The accurate translation of input speech is done by sophisticated parsing 420 (In Fig 4) and generation 422(In Fig 4). The translation module has parsing 420(In Fig 4) and generation 422(In Fig 4) which is capable of interpreting the woman's 102 spoken dialog. The parsing 420 (In Fig 4) module performs the process of prediction including complete semantic interpretations, constraint checks, and ambiguity resolution and discourse interpretations. This system uses the fuse constraint-based and case-based approaches to perform syntactic/semantic and discourse interpretations. The parser 420 (In Fig 4) handles multiple hypotheses in parallel rather than a single word sequence.

A generation 422 (In Fig 4) module is designed to generate the appropriate spoken sentences with correct articulation control. It generates the appropriate spoken sentences using language dictionaries knowledge base. The Language Dictionaries Knowledge Base 424 (In Fig 4) is used for keeping track of more than 6,700 language discourse and world knowledge established during the conversation, and is continuously up-dated during processing. Thus, the appropriate sentence has been generated for woman's spoken sentence to man's 104 (In Fig 1.a) native language – "*สวัสดีฉันชื่อ susan*" 108 (In Fig 1.a) where man's brain language area (i.e., Wernicke's 310/Broca's 308 area) can comprehended.

This system performs real-time translations, which is far better performance than text-based machine translation systems. Unlike traditional methods of machine translation in which a generation 422 (In Fig 4) process is invoked after parsing 420 is completed; this system concurrently executes the generation 422 (In Fig 4) process during parsing 420 (In Fig 4). It employs a parallel incremental generation scheme, where the generation process and the parsing process run almost concurrently. This enables the system to generate a part of the woman's 102 (In Fig 1.a) vocal expression during the parsing of the rest of the woman's 102 (In Fig 1.a) vocal expression. Thus this system stimulates a live feeling – where one speaks and instantaneously the listeners can comprehend the speech in their native languages.

The advanced natural language system also handles two-way conversations. This system provides the bi-directional translation with an ability to understand interaction at the discourse knowledge level, predict possible next vocal expression, understand what particular pronouns refer to, and provides high-level constraints for the generation of contextually appropriate sentences involving various context-dependent phenomena.

Fig.1.b illustrates the conversation between friends who are all foreign-language speaking people. Vietnamese speaking person is saying —This food is delicious‖ in his native language such as "*Đây là món ăn ngon*" 116, this sentence is comprehended as "*Aquest és un deliciós plat*" 118 by the Cantalan speaking person, as "*Tämä on herkullinen ruokalaji*" 120 by Finnish speaking person, and as ;"*זהו מאכל טעים*" 122 by Hebrew speaking person and also as"*This food is delicious* "124 by English speaking person. The Finnish speaking person acknowledges back to them in his native as" *Kyllä, rakastan sitä!* 126. Others comprehend the Finnish sentence as "*Sí, m'encanta!* " 128, "*אני אוהב את זה, כן!*" 130, "*Yes, I love it!*" 132 respectively using Advanced Native Language Translation System.

Similarly, Fig.1.c illustrates a business conversation. A boss 134 is asking "*How we can market our product internationally?*" 136 to his subordinates. His subordinates are a Chinese woman 138, Bulgarian man 140, and Danish woman 142. The boss's 134 spoken dialog is comprehended as "*我們如何才能在國際市場我們的產品？*" 144 by Chinese speaking woman, as "*Как може да ни продукт на пазара в международен план?* 146 by Bulgarian speaking man, and as "*Hvordan vi kan markedsføre vores produkt internationalt?* 148 Danish-speaking women using Advanced Native Language Translation System.

Fig.1.d illustrates the spokesman 150 is giving a speech in his native language Spanish as "*Espero que nuestra sabiduría crecerá con nuestro poder*" 152 to a crowd. There are Slovenian, Korean, Hindi, Hungarian, and Portuguese speaking people in the crowd. So, the spokesman's Spanish speech is automatically comprehended by Slovenian speaking person as "*Upam, da bo naša modrost v naši moči,*" 154, by Korean speaking person as "*가진자라기바란다*" 156, by Hindi speaking person as "मैं अपने ज्ञान हमारी शक्ति के साथ बड़ा हो जाएगा उम्मीद है" 158, by Hungarian speaking person as "*Remélem bölcsesség fog nőni a teljesítmény*" 160, and by Portuguese speaking person as "*Espero que a nossa sabedoria irá crescer com o nosso poder*" 162 using Advanced Native Language Translation System.

As described above, the present invention discloses a system for translating a speech in one language to a language native to the intended recipient(s). Accordingly, the present invention discloses a system of comprehending native languages without the use of any hand-held translators. This invention employs a system where there will no longer be a need to learn new language. Effective communication is now feasible between people speaking different languages. This system explores the capabilities of the human brain, utilizes the language information of the brain, and performs the automatic translation in the background. It should be noted that with all the reading of language area of the human brain - the human brain will not be affected or caused any harm during this process.

The foregoing descriptions of specific embodiments of the present invention have been presented for purposes of illustration and description. They are not intended to be exhaustive or to limit the present invention to the precise forms disclosed, and obviously many modifications and variations are possible in light of the above teaching. The embodiments were chosen and described in order to best explain the principles of the present invention and its practical application. Although the present invention has been described with reference to particular embodiments, it will be apparent to those skilled in the art that variations and modifications can be substituted without departing from the principles and spirit of the invention.

What is claimed is:

i. The *Advanced Natural Language Translation System* is an intelligent method to translate native language spoken from one person into a language that is understood by language area of listener's brain without use of any intermediate device.
ii. The Advanced Natural Language Translation System of claim 1 comprises of an Intelligent Natural Language Program and a Voice Processing Center.
iii. The Intelligent Natural Language Algorithm of claim 2 is capable of recognizing and capturing the voice spoken by a person and collecting language comprehensive information from the language area of the intended recipient's brain.
iv. The Intelligent Natural Language Algorithm of claim 2 further comprises of Intelligent Speech Recognition Algorithm and Language Area Acquisition Algorithm
v. The Intelligent Speech Recognition Algorithm of claim 4 is an intelligent algorithm that recognizes human voice spoken in a native language digitizes the voice into signals, transports the digital signal to Voice Processing Center for further processing.
vi. The Intelligent Speech Recognition Algorithm of claim 4 is also capable of synthesizing the human voice eliminating noise and transmitting only the required voice signals
vii. The Language Area Acquisition Algorithm of claim 4 is capable of scanning the language area of the human brain (Wernicke's and Broca's areas) and collects the language associated data like Language Comprehension, Semantic Processing, Language Recognition, and Language Interpretation.
viii. The Language Area Acquisition Algorithm of claim 4 is capable of relaying language comprehensive details for all people in the audible range of the source human voice.
ix. The Language Area Acquisition Algorithm of claim 4, wherein the information collected is later used for understanding the languages understood by all humans in the audible range.
x. The Voice Processing Center of claim 2 is capable of identifying the native language from received language comprehensive information and translating voice signals into other languages which is native to intended recipient(s)
xi. The Voice Processing Center of claim 2 further comprises of Language Area Inference Engine, Language Dictionaries Knowledge Base.
xii. The Language Area Inference Engine of claim 11, is an exhaustive, comprehensive, obsessively massive list of samples of language area information called a knowledge base. This information is collected from experimental data of brain's language areas (Wernicke's, Broca's) and information from neurologists.
xiii. The Language Area Inference Engine of claim 11, is an artificial intelligence program that tries to derive native language information from a knowledge base.
xiv. The Language Area Inference Engine of claim 11, wherein performs matching, selecting and executing possible set of language comprehensive rules and arrives with the native language for the listener(s).The Language Dictionaries Knowledge Base of claim 11, is also an exhaustive, comprehensive, obsessively massive dictionaries of all words from each of the 6,700 languages spoken around the world. This database is used for translating the spoken word to any of the other 6,700 languages.
xv. The Language Area Inference Engine of claim 11, is capable of parsing, generating and synthesizing the final translated voice using the Language Dictionaries Knowledge Base of claim 11.
xvi. The Advanced Natural Language Translation System of claim 1 must at least comprise of the:

- System having an Input human voice spoken in a native language
- System having a listener individual or a group of individuals unable to understand the native language.

xvii. The Language Area Acquisition Algorithm of claim 4 wherein during the scanning process for language information the human brain is not affected/harmed in any way.

VI. References

1) "How the brain learns to read" By David A. Sousa
2) "Natural Language Generation in Artificial Intelligence" and Computational Linguistics‖ By Cécile L. Paris, William R. Swartout, William C. Mann
3) "Artificial intelligence methods and applications"By Nikolaos G. Bourbakis
4) T. Morimoto et al., "Spoken Language Translation," Proc. Info Japan, Tokyo, 1990.

5) K. Kita, T. Kawabata, and H. Saito, "HMM Continuous Speech Recognition using Predictive LR Parsing," Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, 1989.
6) "Natural language processing technologies in artificial intelligence" By Klaus K. Obermeier
7) "Advances in artificial intelligence: natural language and knowledge-based" By Martin Charles Golumbic

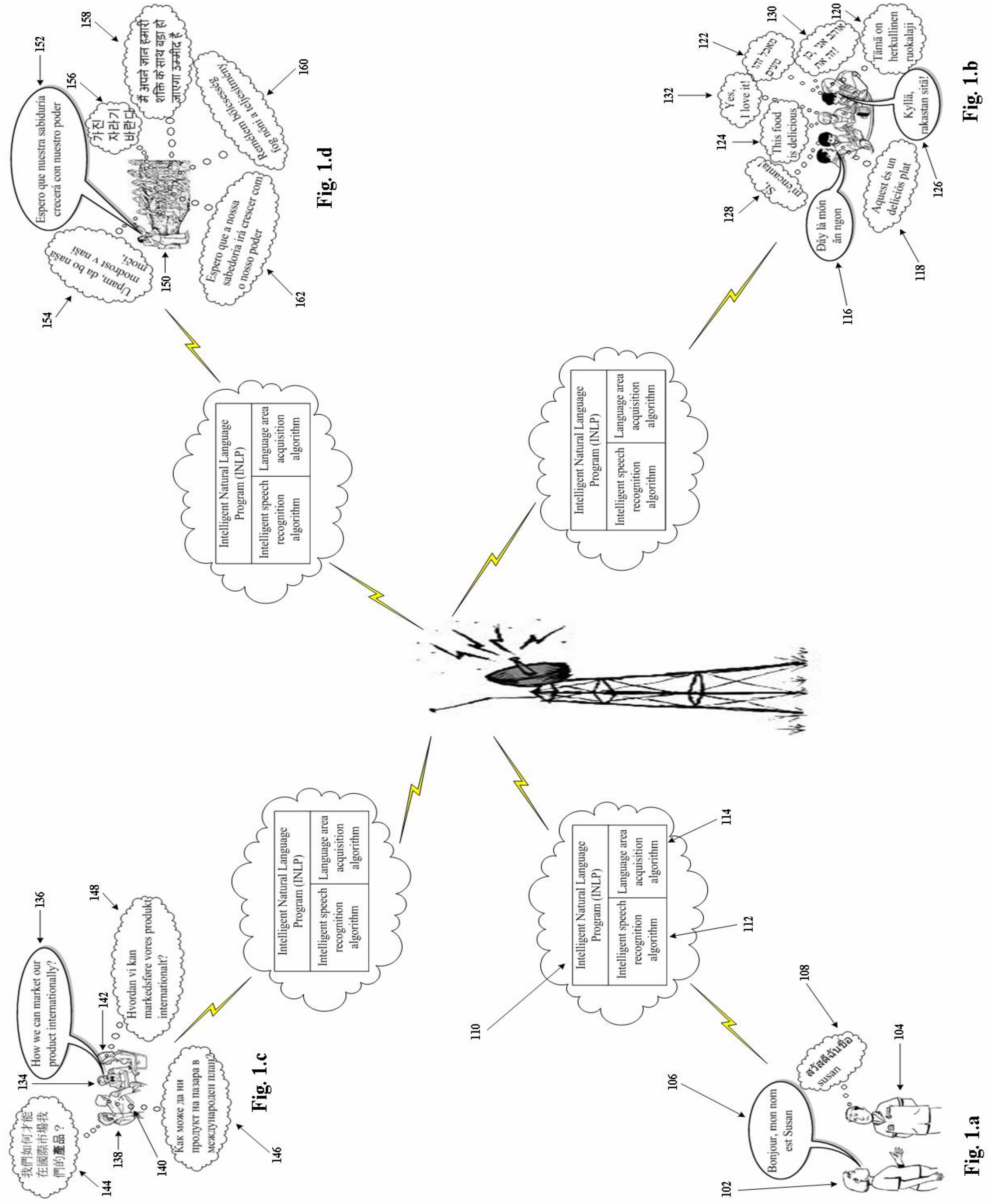


Fig. 1.a Fig. 1.b Fig. 1.c Fig. 1.d

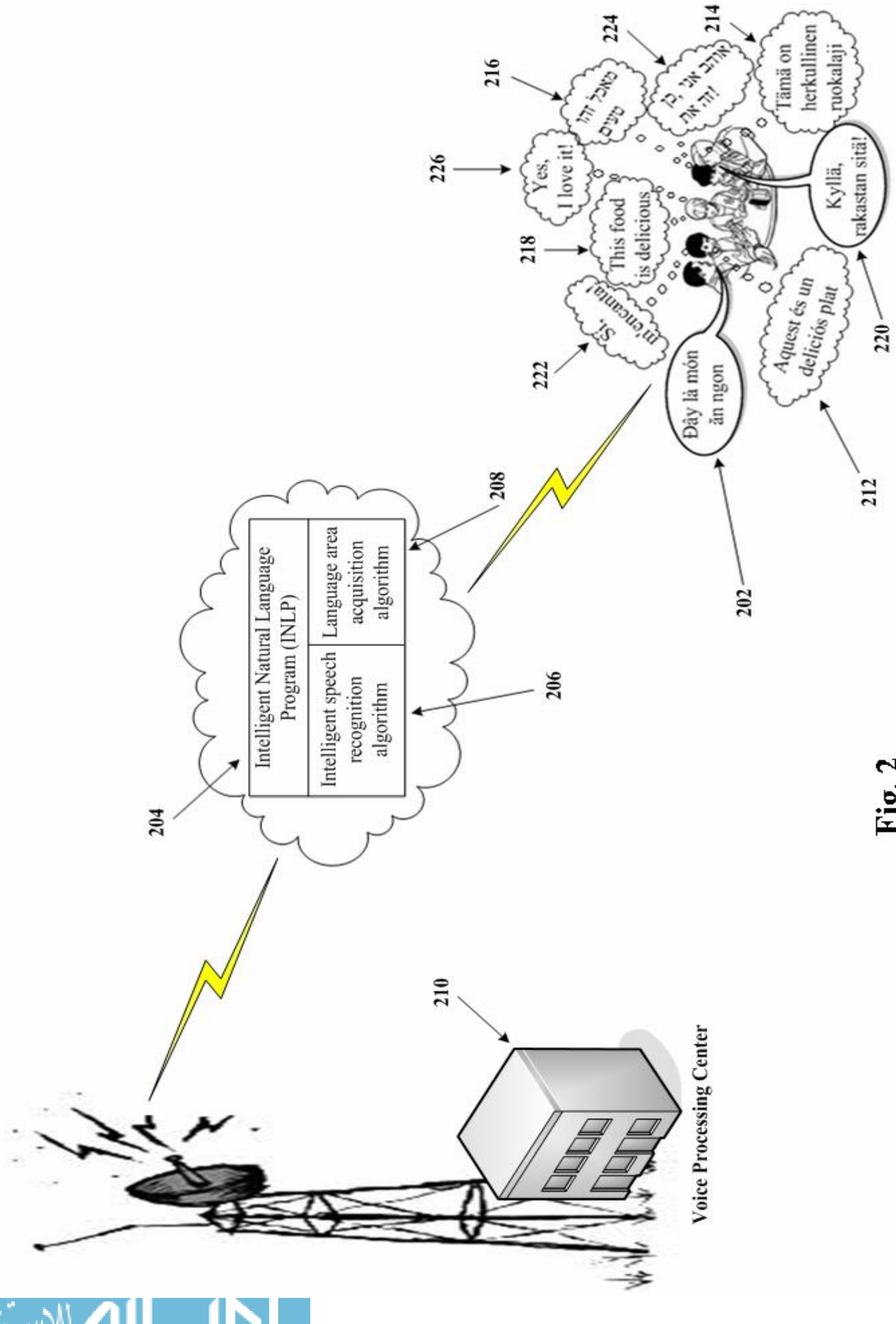

204
Intelligent Natural Language Program (INLP)
Intelligent speech recognition algorithm
Language area acquisition algorithm
206
208
210
Voice Processing Center
202
212
214
216
218
220
222
224
226
Đây là món ăn ngon
Sì, m'incanta!
This food is delicious
Yes, I love it!
Tämä on herkullinen ruokalaji
Kyllä, rakastan sitä!
Aquest és un deliciós plat

Fig. 2

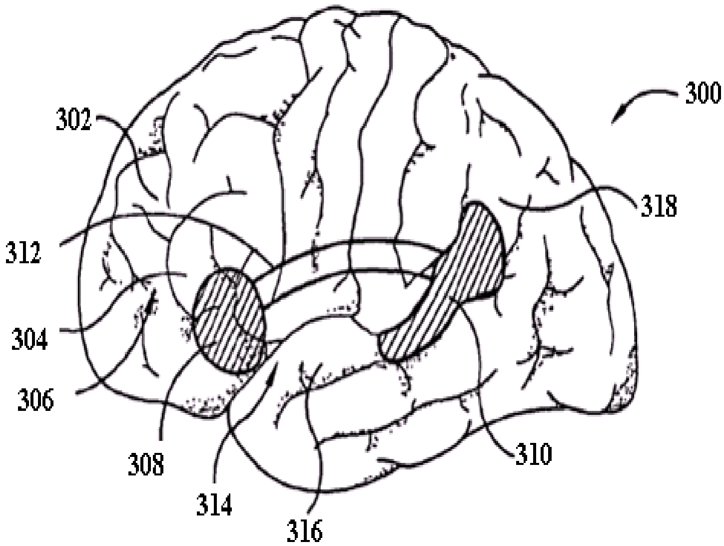
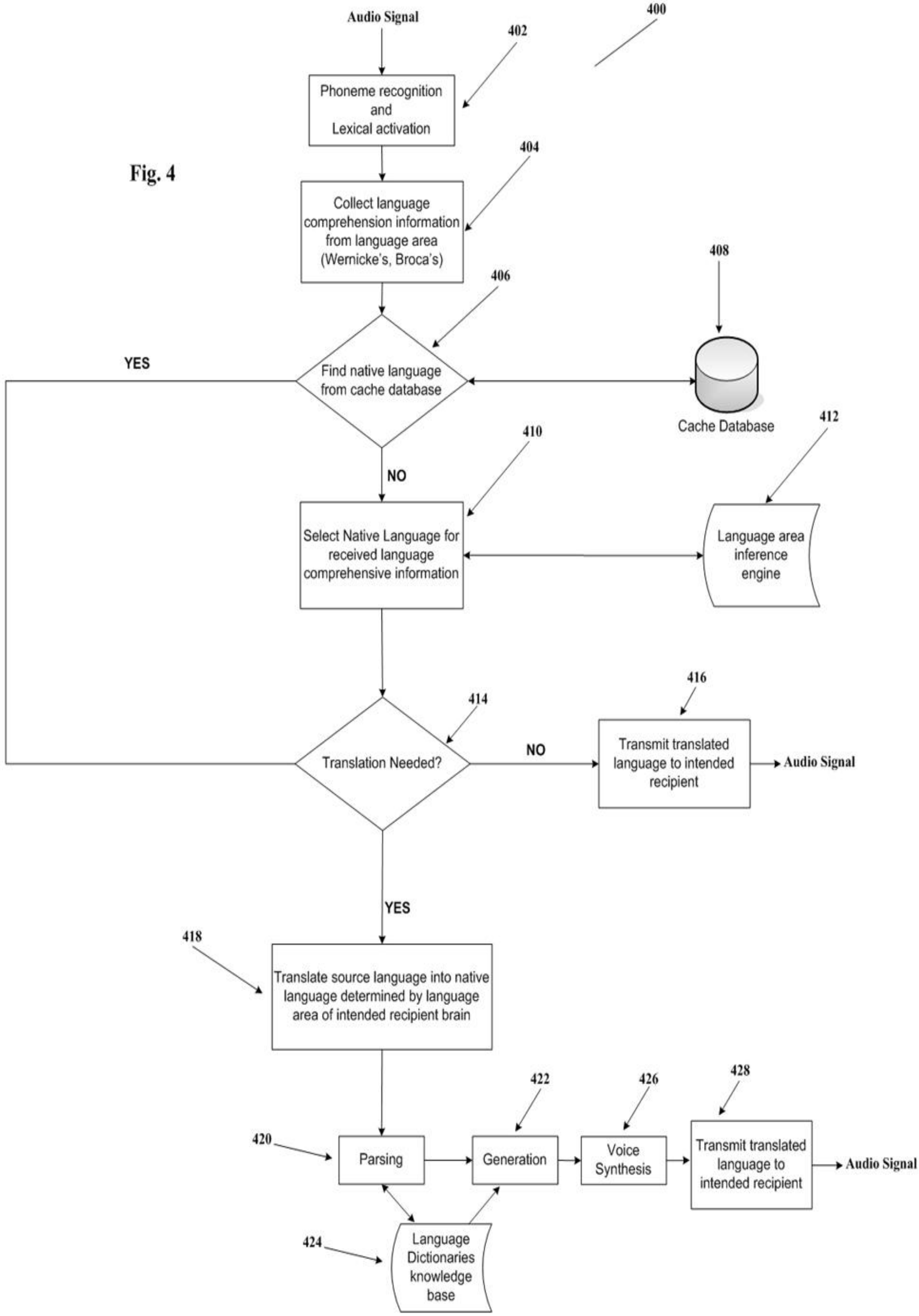


Fig. 3

Fig. 4


Audio Signal
400
402
Phoneme recognition and Lexical activation
404
Collect language comprehension information from language area (Wernicke's, Broca's)
406
408
Find native language from cache database
YES
Cache Database
410
412
NO
Select Native Language for received language comprehensive information
Language area inference engine
414
416
Translation Needed?
NO
Transmit translated language to intended recipient
Audio Signal
YES
418
Translate source language into native language determined by language area of intended recipient brain
420
422
426
428
Parsing
Generation
Voice Synthesis
Transmit translated language to intended recipient
Audio Signal
424
Language Dictionaries knowledge base

Modified DSR (Preemptive) to reduce link breakage and routing overhead for MANET using Proactive Route Maintenance (PRM)

[1] V.Ramesh [2] Dr.P.Subbaiah [3]K.Sangeetha Supriya

GJCST Classifications: C.2.2, C.2.m

***Abstract*- Dynamic Source Routing (DSR) algorithm is simple and best suited for high mobility nodes in wireless ad hoc networks. Due to high mobility in ad-hoc network, route may not exit for long time. Hence, DSR algorithm finds an alternative route when the existing communicating route goes down. It becomes a time consuming process if the communicating route fails frequently. In order to avoid this, we propose a modification to the existing DSR protocol. In this paper, we add a link breakage prediction algorithm to the Dynamic Source Routing (DSR) protocol. The mobile node uses signal power strength from the received packets to predict the link breakage time, and sends a warning to the source node of the packet if the link is soon-to-be-broken. The source node can perform a pro-active route rebuild to avoid disconnection. Intermediate nodes in the route continuously monitor the signal strength at the time of communication, based on a predefined threshold signal value. Intermediate node sends a message to the source node that the route is likely to be disconnected, if signal strength falls below the threshold value. If source receive this message it starts using backup route and if back route also fails then it finds alternative route. The backup route will minimize the time consuming process of finding an alternative route to some extent. Experiments demonstrate that adding link breakage prediction to DSR can significantly reduce the total number of dropped data packets (by at least 25%). Simulation results shows the probability of the communication breakage decreases when parallel routes are used and comparisons between DSR and Modified DSR(Preemptive Version) with respective to no of broken paths and routing overhead.**

KeyWords-Ad-Hoc Networks, Preemptive, Dynamic Source Routing, Proactive, age of path.

I. Introduction

There are currently two variations of mobile wireless networks. The first is known as infrastructure network. The bridges for these networks are known as base stations.

Manuscript received 14.12.2009

[1] Research Scholar at Sathyabama University, Chennai & Assoc.Prof, CMR CET, Hyderabad, AP, India.
(telephone:09959116287 email: v2ramesh634@yahoo.co.in)
[2] Principal, Veerabrahmendra Institute of Technology, Badvel, Kadapa, AP, India.
(telephone: 09703245276 email: subbaiah_nani@sify.com)
[3] Asst.Professor,BVRIT,Narsapur,AP, India.
(telephone:09705870357email: sangeethasupriyak@gmail.com)

A mobile unit within these networks connects to and communicates with, the nearest base station that is within its communication radius. As the mobile unit travels out of range of one base station into the range of another, a "handoff" occurs from the old base station to the new, allowing the mobile to be able to continue communication seamlessly throughout the network. Typical applications of this type of network include office wireless local area networks (WLANs). The second type of mobile wireless network is the mobile ad-hoc network or MANET. Unlike infrastructure network, this type of network needs no base station. Mobile nodes communicate to each other by either directly or through intermediate nodes. Ad-hoc network becomes popular since it can be applied in many situations, such as emergency search-and-rescue operations, classroom, meetings or conference and many more. To facilitate communication within the network, routing protocols used to discover routes between nodes. Building a MANET routing protocol is not an easy job, since efficiency and correctness becomes the main concern. Some approach had been proposed to make routing protocol becomes efficient and correct. Routing protocols in MANET, generally, can be categorized as table-driven and on-demand. In table-driven (also called proactive protocol), like in most routing protocol for wired network, each node is required to maintain routing table keep updated whether there is or not a request for routes. In on-demand (also called as reactive protocol), each node seeks for routes only when there is need to do so. [1]

II. Dynamic Source Routing

The Dynamic Source Routing (DSR) protocol is a simple and efficient routing protocol designed specifically for use in multi-hop wireless ad hoc networks of mobile nodes. It is based on the concept of source routing, a routing technique in which the sender of the packet determines the complete sequence of the nodes through which to forward the packet. The sender explicitly lists this route in the packet's header, identifying each forwarding "hop" by the address of the next node to which to transmit the packet on its way to the destination host.

The DSR protocol consists of two mechanisms: Route Discovery and Route Maintenance. When a mobile node wants to send a packet to some destination, it first checks its route cache to determine whether it already has a route to the destination. If it has one, it will use this route to send the packet. Otherwise, it will initiate route discovery by broadcasting a route request packet. When receiving a request packet, a node appends its own address to the route record in the route request packet if it did not receive this request message before, and re-broadcasts the query to its neighbors. Alternatively, it will send a reply packet to the source without propagating the query packet further if it can complete the query from its route cache. Furthermore, any node participating in route discovery can learn routes from passing packets and gather this routing information into its route cache.

When sending or forwarding a packet to a destination, Route Maintenance is used to detect if the network topology has changed such that the link used by this packet is broken. Each node along the route, when transmitting the packet to the next hop, is responsible for detecting if its link to the next hop has broken. When the retransmission and acknowledgement mechanism detects that the link is broken, the detecting node returns a Route Error packet to the source of the packet. The node will then search its route cache to find if there is an alternative route to the destination of this packet. If there is one, the node will change the source route in the packet header and send it using this new route. This mechanism is called “salvaging” a packet. When a Route Error packet is received or overheard, the link in error is removed from the local route cache, and all routes which contain this hop must be truncated at that point. The source can then attempt to use any other route to the destination that is already in its route cache, or can invoke Route Discovery again to find a new route.[4][9]

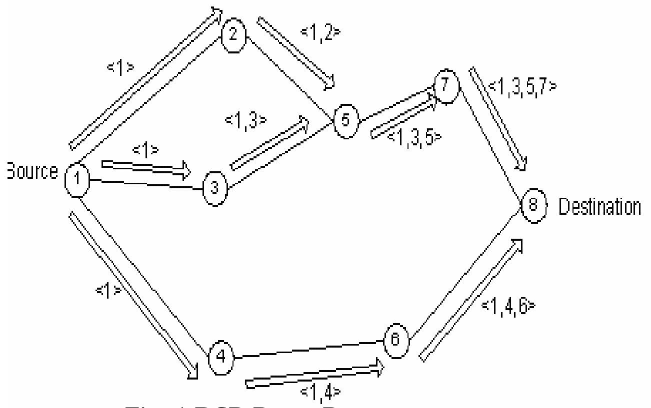


Fig: 1 DSR Route Request

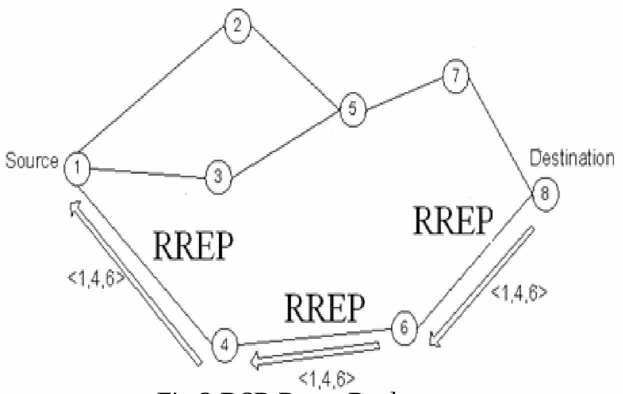


Fig:2 DSR Route Reply

III. PROACTIVE ROUTE MAINTENANCE

We assume that all nodes wishing to communicate with other nodes within the ad hoc network are willing to participate fully in the protocols of the network. Each node participating in the network should also be willing to forward packets for other nodes in the network. We refer to the minimum number of hops necessary for a packet to reach from source to destination. We assume that he diameter of an ad-hoc network will be small(5 to 10 hops), but greater than 1. Packets may be lost or corrupted in transmission on the ad-hoc wireless network. A node receiving a corrupted packet can detect the error and discard the packet.

The GPS and signal strength methods both use physically measured parameters to predict the link status. The node with GPS can know the position of itself directly. But GPS currently is not a standard component of mobile devices, and in the metropolitan area and indoor, the signal can be too weak to be received. The signal strength method only consumes receiving node’s computing power, and does not depend on any additional device. It is used in this paper. At first we assume that the sender power level is constant. Received signal power samples are measured from packets received from the sender. From this information it is possible to compute the rate of change for a particular neighbor’s signal power level. Because the signal power threshold for the wireless network interface is fixed, the time when the power level drops below the acceptable value can be computed.[7] Characteristics of PRM include:

Freshness. All nodes near an active route have the up-to-date routing information. Broken paths are eliminated, new paths recognized, and non-optimal paths replaced by optimal ones.

Robustness. An active node that is forwarding data packets usually maintains several fresh alternative paths. After one path fails, the data packet is usually forwarded via another path without causing packet loss or extra delay. PRM will resort to a route discovery operation only after all alternative paths have failed.

Lightweight maintenance. Unlike in existing proactive routing protocols, the route maintenance is confined to those small areas surrounding active routes, where control packets make only a small portion of data transmission. As the lifetime of a route is lengthened, the overhead of the proactive route maintenance can be compensated by the less frequent route discovery operations.[10]

The proposed Concept is illustrated using the following example.

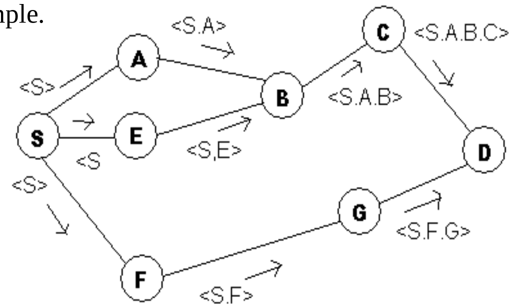
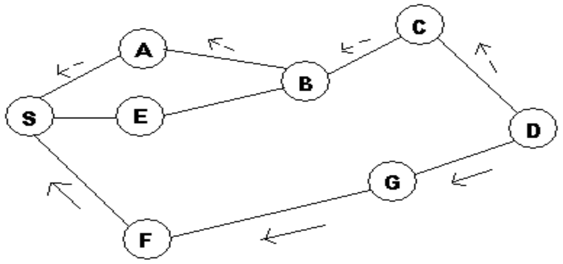


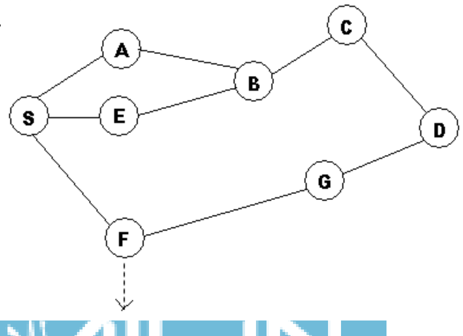
Fig. 3

When a source node S want to send message to the destination node D, it initiates route discovery by broadcasting the RREQ packet to its neighbors (A, E, F) as shown in Fig 3. The intermediate nodes (A, E, F) on receive the RREQ packet rebroadcast the packet to its neighbors by appending its id in the route record of the RREQ packet. Similarly, other intermediate nodes also forward the RREQ packet to the destination. When the destination node D receives two or more RREQ packets from the same source through different routes, it finds the two best routes based on the no of hopes. The route, which has least number of hopes. The route which has least number of hops it becomes primary<S, F, G>, and second least number of hops route becomes backup route<S, A, B, C>. The destination node D sends Route Reply (RREP) packet using the Primary (<S, F, G>) and Backup(<S, A, B, C>) route as shown in the following Fig. Each RREP packet contains the Primary as well as the Backup route information. When source node S receives first RREP packet form destination, it treats this is the primary route and wireless communication is more error prone compared to wired network. To improve the reliability we are sending route reply (primary + backup routes information) through the primary and the secondary route. If any one packet gets corrupted at the time of transmission, source must be able to use the other packet.[6]



Primary Route ⟶ {<S.F.G> + {<S.A.B.C>}}
Backup Route ⇢ {<S.A.B.C>+{<S.F.G>}}

The communication between the source node S and destination node D commence using the primary path<S, F, G>. During communication, the node F starts moving away from S. When the signal strength of node F falls below threshold T, it sends a warning message "Path likely to be disconnect" to source node S. As soon as S receives the warning message, it starts using the Backup route along with primary route. Whenever destination node receives the data packets from the source node through two different paths (Primary + Backup), it sends acknowledgement through both the paths. If source node S receives an acknowledgement from the destination node through the Backup route, it makes preemptive switch over to the Backup route; otherwise S initiates the route discovery process.



A. Generating the Warning Message based on the Signal Strength

Let us consider the following scenario while using the Backup route.

Case 1:

Node C is moving toward node G, as shown in Fig 4

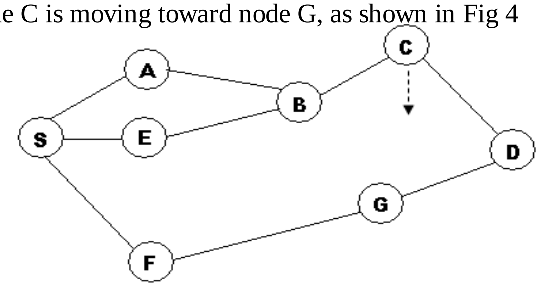


Fig. 4

As node C is moving towards node G, the signal strength increases and Backup route become more stable.

Case 2:

Node C is moving away from node G, as shown in Fig 5.

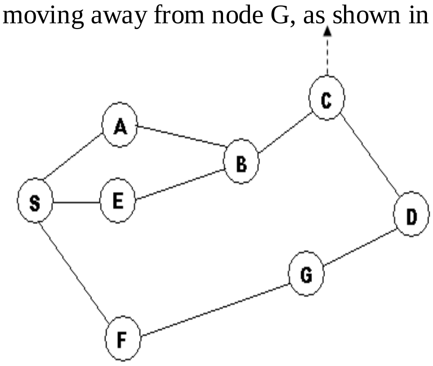


Fig. 5

As node C is moving away from node G, the signal strength of C falls below the threshold T and as a result the Backup route fails. Let $p(0 \leq p \leq 1)$ is the probability of the route failure in case of DSR. In the best-case p=0 and in the worst case p=1. Hence on an average case the probability of route failure p=0.5 (50%). Similarly in the proposed Proactive routing in Dynamic Source Routing Protocol for Wireless Ad-hoc Networks with Backup Route.

The probability of Primary route failure is p=0.5 (50%) ------------- (1)

The probability of backup route failure is p=0.5 (50%) ----------- (2)

Form (1) and (2) we conclude that the probability of both the route failure p=0.25 (25%). Therefore, Modified Dynamic Source Routing Protocol for Wireless Ad-Hoc Networks with Backup Route has a significant effect on the performance as it improves the reliability form 50% to 75% with minimal control overhead.

The threshold value plays an important role for control packet overhead.

Case 1:

If threshold T is large:

It may send false warning to source node to use backup route.

Case 2:
If threshold T is small:
Source node may not get sufficient time to discover new route, if backup route fails.
Therefore threshold T value is set moderate, to overcome above-mentioned drawbacks.
A Preemptive region is defined around every node as shown in the figure 6 for node A. As soon as node C enters the preemptive region, a warning message is sent to the sender node A. Then the node A initiates a route discovery process. With the establishment of a new route, data transmission is continued along this new route. The time required to discover a new path can be termed as recovery time Trec. Hence the time between the warning and the path break Twarn should be atleast or slightly greater than Trec.

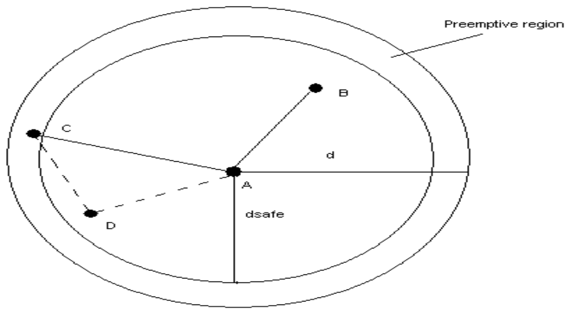


Fig: 6 Preemptive Region

In order to determine the optimal range, it is necessary to exchange the location and velocity information of the nodes amongst all the nodes depending on the receiver signal power. The receiver signal power,

$$Pr = P0 / r^n$$

at a distance r from the transmitter, where P0 is the transmitted power and path loss exponent n is typically between 2 and 4. The minimum power receivable by the device is the power at the maximum transmission range,

$$Pd = P0 / d4$$

Similarly, the preemptive signal power threshold is the signal power at the edge of the preemptive region. In addition, for a preemptive region of width of w, the signal power threshold is

$$Psafe = P0 / d4\ safe$$

Where dsafe is equal to (d- *w*) and *w*=relative speed**Twarn*
The preemptive ratio α is defined as α =Psafe / Pd = range/ (range- w)

In reality, the received signal power may experience sudden fluctuations due to channel fading and multipath effects, which will trigger a false warning, causing unnecessary route request floods. This may result in lower quality routes being initiated and also increasing the routing overheads. In cellular networks, an exponential average of the signal power is used to verify that the signal power drop was not due to fading. However, if the traffic is bursty or infrequent, the preemptive region may be fully crossed by the time enough packets are received to drop the average below the threshold. Therefore quicker power estimates can be achieved by sending a warning whenever the instantaneous power drops below the threshold and checking the warning packet received power when it is received by the source. If the warning packet power is also below the threshold, there is a good probability that the warning is real.

B.Generating the Warning Message based on 'Age of the Path'

With transmissions being done along the same path, relay nodes will experience a continuous drain of their battery power for the same source destination pair, which may result in path failure. Therefore alternate route discoveries are required before the onset of failure.
Nodes keep a record of their most recent encounter times with all other nodes. With a path discovery being made, the source node sets a timer. The preemptive warning is generated based on two parameters- Age of the path defined as the time difference Tage between the transmissions of two consecutive route discovery packets from the source to the same destination and threshold value Γ is defined for the age of the path. As long as Tage is lesser than Γ, data transmission can be continued on the same path. When the timer value exceeds the threshold Γ, a warning message is generated leading to a new path discovery. However, this new path may or may not be the shortest path to the destination. The choice of the threshold depends on node density of the network. If the node density is small with lesser number of paths available, Γ must be large.

IV. MULTIPLE ROUTES

Use of multiple routes simultaneously, instead of a single route at a time, would help to improve the ongoing communication between the two ends. The source node will use each of these routes alternatively to send packets to the destination node. Use of multiple routes reduces the dependency on a single route, which results in more stable communication. This is because, if a single route fails, we need to again initiate the Route Discovery process. However, if multiple routes are used, when one route fails, another route can be used. Only when all the routes fail, the Route Discovery is to be done to search a new route. We note that the use of multiple routes is different from the backup route theory of DSR. In the backup route approach, the source node uses the primary route for communication and keeps a backup (secondary) route in its route cache. Whenever the primary route fails, the backup route is used.[6] The problem with this approach is that, while the source is still using the primary route, the backup route might fail and the source would remain unaware of that. If after some time the primary route fails and the source node switches to the backup route, it discovers that the backup route has been already broken. But if multiple routes are used in parallel, the source node will be informed of the route failure immediately whenever it occurs. Thus, the source node will never attempt to use a stale route.

V. Simulation Study

The discrete event network simulator NS-2 has been used for analysis and comparison of the adhoc routing protocols. The mobile node movement is restricted to a square cell of 600 X 600m containing 70 nodes. Random waypoint model was used here. Figure 9 shows the plot of Pk (k =no. of parallel routes) with respect to time. The different parameters of the plot are listed in Table.

No. of parallel routes (k)	1-5
No. of intermediate nodes (n)	5
Motion time	10s
Pause time	2s after every 10s
Total time	200s

Figure 9 shows that when multiple parallel routes are used, the communication between the source node and the destination node reduces exponentially. That is, greater the number of parallel routes, the more stable the communication becomes. This is because when multiple routes are used, dependency on a single route is reduced. Therefore, even if a single route fails, we have other routes in hand to use for transmitting packets. If a very long time is considered, the fluctuation in the probability values stops and reaches a saturation level.

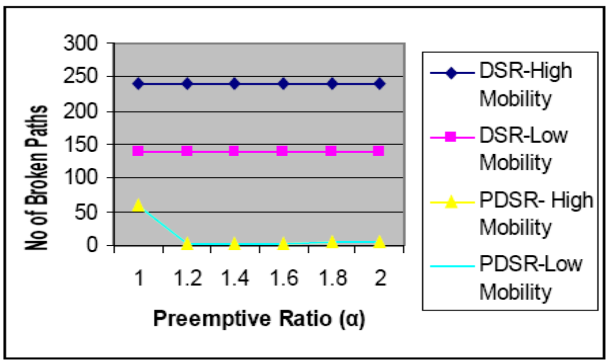


Fig: 7 Comparison based on broken paths

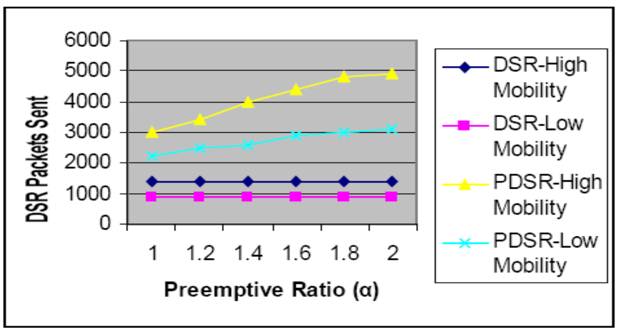


Fig: 8 Comparison based on Packets sent

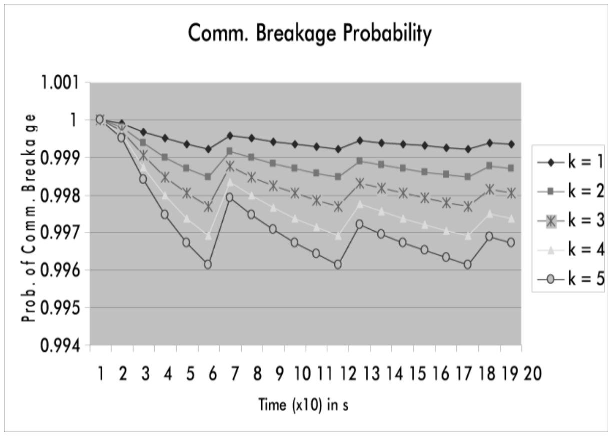


Fig: 9 Probability of communication breakage decreases when parallel routes are used

VI. Conclusion

Reactive ad hoc routing algorithms initiate route discovery only after a path breaks, it has significant control overhead for detecting the disconnection and re-construction of a new route. DSR with PRM mechanism detects early about the link that is likely to break soon, and hence it uses a backup path before the existing link fails.

The paper explains the preemption of Primary to Backup route by the source node S, whenever the signal strength of the primary route falls below the threshold value T. The modified DSR will improve the communication reliability between the source and destination node even if the mobility is high. In addition, Proactive routing improves the overhead of rediscovering route whenever the primary route fails.

VII. References

1) Siva Ram Murthy, B.S, Manoj, "Routing Protocols for Ad Hoc wireless Networks," in Ad Hoc wireless networks: Architectures and Protocols, Chapter 7. Pearson Publication.
2) Hongbo Zhou, "A Survery on Routing Protocols in MANETs," Technical. Note March 2003.
3) Elizabeth M. Royer, University of California, Santa Barbara Chai-Keong Toh, Georgia Institute of Technology "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," IEEE personal communications, April 2007.
4) David B. Johnson, Davis A. Maltz, "The Dynamic Source Routing Protocol for Mobile Ad Hoc Networks" October 1999 IETF Draft.
5) C.R. Dow, P.J.Lin, S.C.Chen, J.H.Lin, S.F.Hwang. "A study of Recent Research Trends and Experimental Guidelines in Mobile Ad Hoc Networks," aina, oo. 72-77, 19th International Conference on Advanced information networking and applications (AINA'05) volume 1(AINA papers) 2005.

6) T. Goff, N.B. Abu-Ghazaleh, D.S. Phatak and R. Kahvecioglu, “Preemptive Maintenance Routing in Ad Hoc Networks”, journal of parallel and Distributed Computing, Special Issue on Wireless Mobile Communication and Computing 2003.
7) SM. Jiang, DJ. He and JQ. Rao, “A Prediction-Based Link Availability Estimation for Mobile Ad-Hoc Networks. Proceedings of IEEE INFOCOM, pages 1745-1752, Vol.3, April 2001.
8) Nasipuri, R. Casta˜neda, and S. R. Das, “Performance of multipath routing for on-demand protocols in ad hoc networks,” ACM/Kluwer Mobile Networks and Applications (MONET) Journal, vol. 6, no. 4, pp. 339–349, Apr. 2001.
9) M C Domingo, D Remondo and O. Leon, “A Simple Routing Scheme for Improving Adhoc Network survivability”, GLOBECOM, IEEE, 2003.
10) Mohammad Al-Shurman and Seong-Moo Yoo, Seungjin Park, “A Performance Simulation for Route Maintenance in Wireless Ad Hoc Networks”, ACM,2004

Performance Analysis and Enhancement of IEEE 802.11 Wireless Local Area Networks

Ms. Amardeep Kaur[1], Dr. Sandip Vijay[2], Dr. S.C.Gupta[3]

GJCST Classifications:
C.2.2, C.2.m

***Abstract*- The wireless networks can be employed to provide network connectivity almost anywhere, it provides large companies the option to connect the current wired networks to the new wireless network without any problems and gives user the option to use any kind of applications regardless of its source or vendors. However, the WLAN performance is a key factor in spreading and usage of such technologies. Also the Wireless local area networks (WLAN) are more bandwidth limited as compared to the wired networks because they rely on an inexpensive, but error prone, physical medium (air).Hence it is important to improve their performance.**

In this paper the performance optimization methods have been presented using an advanced network simulator, OPNET modeler 9.1. The previous research mainly focused on improving the performance via network layer [1-3]. Here performance optimization has been shown via a series of simulation tests with different parameters such as Data rate, Fragmentation threshold, RTS/CTS threshold, buffer size and the physical characteristics. The different quality of service parameters are chosen to be the throughput, media access delay, the retransmission attempts, dropped data packets and Queue size etc.

Then finally the results are compiled to improve the performance of wireless local area networks.

Keywords- Wireless LAN, IEEE 802.11, OPNET

I. Introduction

Future Wireless local area networks(WLAN) enable people on the move to communicate with anyone, anywhere at anytime with a range of multimedia services The exponential growth of cellular telephones and mobile systems coupled with spreading of laptops and palmtops indicate a bright future for such networks both as standalone as well as network infrastructures[2]. However, the WLAN performance is a key factor in spreading and usage of such technologies. So this paper deals with the optimization techniques based on the advanced network simulator, OPNET. The OPNET (Optimized Network Engineering Tool) can be best described as a set of decision support tools, providing a comprehensive development environment for the specification, simulation and performance analysis of communication networks, computer systems, and applications and distributed systems. Discrete event simulations are used as the means of analyzing the system performance and behavior. This sophisticated package comes with a range of tools, which allows us to specify models in detail, identify the elements of the model of interests, execute the simulation, and analyze the generated output data. The OPNET simulator has many features such as object orientation and hierarchical modeling.

II. OPNET Implementation and Simulation

The OPNET implementation (Fig-1) consists of six nodes with one Access Point (AP), forming a wireless infrastructure network. Simulation environment was set as per the Table 1(a) and Table 1(b). All the simulations in this chapter are done for 600 simulation-second(s). The packet size distribution is exponential with a mean of 92 bytes. The inter arrival time is exp (0.02) for all the nodes unless otherwise specified. Since the packet size is exponentially distributed with mean of 92 bytes, RTS/CTS exchange is required for most of the packets. All the wireless station nodes and the access point use Frequency Hopping Spread Spectrum at the physical layer. All the nodes employ the PCF basic CSMA/CA access mechanism. The nodes transmit at a maximum data rate of 11 Mbps. Packets received at a node with power less than 7.33 E-14 Watts will find receiver to be busy. In this implementation all the nodes are static receiving and forwarding the packets through the access point. In the OPNET implementation, the effects of following parameters are analyzed based on throughput (bits/sec), the media access delay and the retransmission attempts, load, queue size etc.

i. Data rate
ii. Fragmentation threshold
iii. RTS/CTS exchange
iv. Physical characteristics
v. Buffer size

Table 1(a) Parameter Setting of WLAN Network

WLAN environment	Office
Workspace area	100m x 100m
Node model	Wireless_LAN

Manuscript Received "19th Nov.2009"
1.Lecturer, ECE department, Dehradun Institute of Technology, Dehradun, (Telephone no.: 09760974124, Email-id: amardeep05enator@gmail.com) 2.Prof. and Head, Drona College of Engineering, Dehradun, (Telephone no.: 09557429979, Email-id: vijaysandip@rediffmail.com) 3.Eminent Professor, IIT, Roorkee. (Telephone no.: 09719339553, Email-id: sureshprem1938@gmail.com)

Table 1(b) Parameter Setting of WLAN Network

Number of nodes	7 (node_0 to node_6)
Access Point	1 (AP)

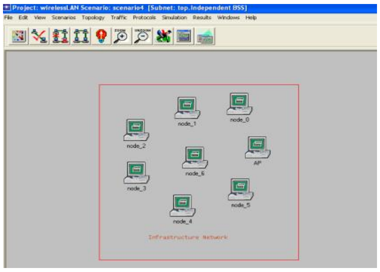


Fig-1: OPNET Implementation

A. *Data Rate*

The WLAN model in OPNET modeler supports data transfer at 1, 2, 5.5, and 11 Mbps. These data rates are modeled as the speed of transmitter and receiver connected to the WLAN.A station can only transmit data packets at the data rate specified by the attribute. However, it can receive data at any data rate.

Two scenarios are created one is with data rate of 1Mbps and another scenario is created with data rate of 11Mbps, rest of the parameters is same in both the scenarios. The data rate is the parameter signifies the speed of the nodes connected in the network.

TABLE 2: SIMULATION PARAMETERS

Attribute	Scenario-1	Scenerio-2
Rts Threshold	None	None
Fragmentation	None	None
Data Rate	**1 Mbps**	**11 Mbps**
Physical Characteristics	FH	FH
Buffer Size (bits)	256000	256000

Simulation Results: The first set of simulation scenarios show the effect of data rate (Fig-2) on the performance of WLAN. From the simulation of above two scenarios, it has been observed that as we increase the data rate from 1Mbps to 11Mbps, the throughput increases by about 27-30%. This is also predictable from the theoretical viewpoint that as we increase the data rate, the number of bits received increases. Also the average media access delay reduces from scenario1 to scenario 2. Initially it was near to the value of media access delay of scenario1 and after that it increases linearly and then it will end up at the difference of about 90-95% to scenario1 at the end of simulation period of 600 simulation seconds. This is very encouraging result and it is understandable because the data will stay for less time in media (buffer) for higher data rate scenarios. The retransmission attempts also decreases by about 69% from scenario 1 to scenario 2 that is as we increase the data rate the packets are delivered accurately and there is less requirement of retransmissions.

Fig-2: The effect of data rate on
(a) Throughput (b) Media access delay
(c) Retransmission attempts

Fig-2(a) Fig-2(b)

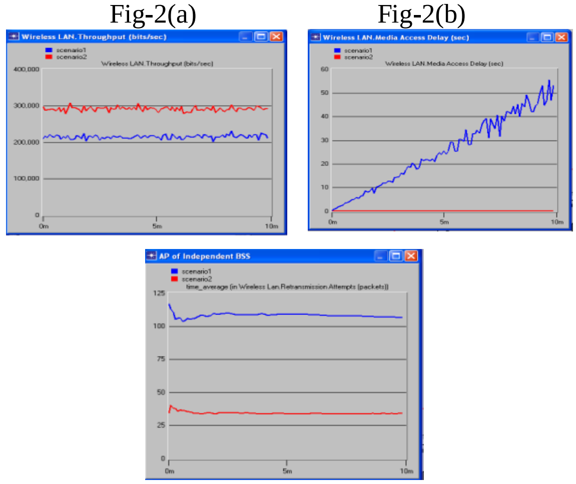


Fig-2(c)

B. *Fragmentation Threshold*

Fragmentation threshold specifies the value to decide if the MSDU received from the higher layers needs fragmentation before transmission. The number of fragments to be transmitted is calculated based on the size of the MSDU and the fragmentation threshold. In OPNET the default value is none which means that no fragmentation will take place regardless of the MSDU size. The destination station received these fragments and stores them in the reassembly buffer until all fragments are received. This fragmentation and reassembly is implemented using the built-in SAR (segmentation and reassembly) package in OPNET.

The two scenarios are created, one with no fragmentation of incoming packets and one with fragmentation of 1024 bytes packets. The fragmentation threshold decides whether the packets received from higher layers need fragmentation before transmission. The default value is none that means there is no need of fragmentation before transmission .If there is fragmentation of packets before transmission, this will definitely increase the load on both the transmitter and the receiver.

TABLE 5: SIMULATION PARAMETERS

Attribute	Scenario-7	Scenerio-8
Rts Threshold	None	None
Fragmentation	None	None
Data Rate	11 Mbps	11 Mbps
Physical Characteristics	FH	FH
Buffer Size (bits)	6400	256000

Simulation Results: If the buffer size is increased, (Fig-6) then the number of Retransmission attempts would be reduced, at the starting time of simulation period it is about 33-35% lesser than scenario7, but till the end of simulation the retransmission attempts become approximately equal to the retransmission attempts of scenario8. Also the size of the queue will be decreased for larger buffer due to the fact that the larger buffer will take less time to send the packets, so the queue size will not build up continuously for larger buffer. This reduction difference is about 1-1.5% throughout the simulation duration of 600 simulation seconds.

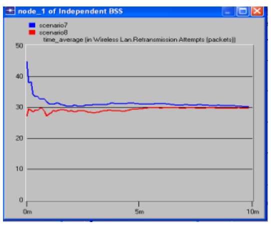


Fig-6(a)

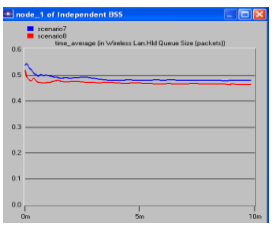


Fig-6(b)

Fig-6: Effect of the Buffer Size on (a) Retransmission attempts (b) Queue Size

C. *Physical Characteristics*

Local radio networks use radio or infrared waves in order to transmit data. The technology used for sending radio transmissions is called narrowband transmission, which runs different communication signals through different channels. However, radio transmissions are often subject to numerous limitations, which makes this type of transmission difficult. Here the effect of WLAN Physical layer (PHY) characteristics will be analyzed. The OPNET supports three pre-defined physical Layer characteristics that is “Direct-sequence,” Frequency-hopping” and” Infrared”.

These physical characteristics will have appreciable effect on the throughput, media access delay but it will have a significant effect on the number of retransmission attempts

TABLE 6: SIMULATION PARAMETERS

Attribute	Scenario-9	Scenerio-10	Scenario-11
Rts Threshold	None	None	None
Fragmentation	None	None	None
Data Rate	11 Mbps	11 Mbps	11 Mbps
Physical Characteristics	FH	DS	Infrared
Buffer Size (bits)	256000	256000	256000

Simulation Results: Using the above simulation parameters, the simulation results (Fig-7) show that the infrared coding proves to be the best technique, while the frequency hopping method proves to be the worst among the three and the direct sequence coding proves to be lying between the infrared and frequency hopping techniques for the three parameters that are the throughput, number of retransmission attempts and the Delay. The difference in the DS and FH coding is very less for throughput and for the delay the difference in these coding techniques is not very high which is about 15-17%, but for retransmission attempts the DS is performing much better than the FH coding by about 40-43%. Moreover, for all the three parameters the infrared coding is the best among three available options. For throughput, the performance of infrared is better by about 30-33% as compare to the DS and FH coding, also at the end of simulation there is a transition in the throughput of infrared coding by about 70-75%. For retransmission attempts, the performance of infrared coding is performing well by about 60-62% as compare to DS coding while this reduction is about 78-80% for the FH coding. For the overall delay in WLAN, the infrared performs better than DS and FH by about 85-90% for entire simulation duration of 600 simulation seconds. However, at the end there is a transition in the delay of infrared coding.

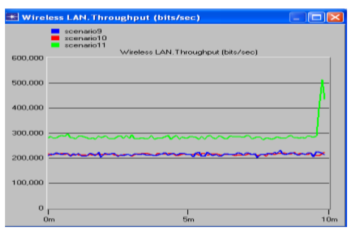


Fig-7(a)

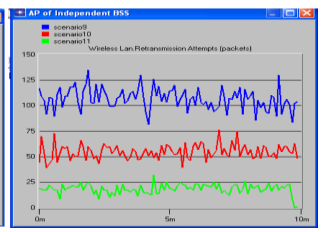


Fig-7(b)

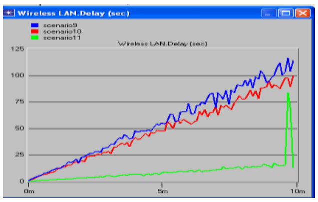


Fig-7(c)

Fig-7: Effect of Physical characteristics on (a) Throughput (b) Retransmission attempts (c) WLAN Delay

III. Conclusion

The overall performance of the IEEE 802.11 Wireless Local area networks has been analyzed in detail with the help of OPNET Modeler. The performance has been analyzed with the help of the parameters like throughput, media access delay, the number of retransmission attempts, dropped data packets etc. for data rate, fragmentation threshold, RTS/CTS threshold, physical characteristics and the buffer size. These different parameters reveal the different methods to optimize the performance of wireless local area networks such as by increasing the data rate the performance can be optimized in terms of throughput, media access delay and the number of Retransmission attempts or by incorporating the RTS/CTS exchanges the number of retransmission attempts can be reduced or by increasing the buffer size the number of dropped data packets can be reduced and the infrared coding at the physical layer proves to be the best if performance criteria is defined in terms of throughput, the media access delay and the number of retransmission attempts while the frequency hopping is not preferred coding method at the physical layer for the above performance parameters.

IV. References

1) Enhancements and Performance Evaluation of Wireless Local Area Networks" Jiaqing Song and Ljiljana Trajkovic Communication Networks Laboratory Simon Fraser University, Burnaby, BC, Canada
2) "Simulation-Based Performance Analysis of 802.11a Wireless LAN" Mohammad H.Manshaei and Thierry Turletti INRIA, 2004 Route des Lucioles, BP-93,
06902 Sophia-Antipolis Cedex, France
3) "Wireless network performance optimization using OPNET Modeler" by Sameh H. Ghwanmeh and Abedel Rehman Al-Zoubidi, department of computer Engg. Jordan
4) Hallani, S. Shahrestani,(2004), Wireless mesh networking: Implementation issues and analysis, in Proc. 3rd International Business Information Management Conf., 200-205.
5) "Performance of Infrastructure Mode Wireless LAN Access Network based on OPNET Simulator"Saeed A. Bawazir,Saleh H. Al-Sharaeh, Department of Computer Science,Normal, AL 35762,USA
6) www.opnet.com
7) Tanenbaum, A.S.,"Computer Networks" , IIIrd Edition,Prentice hallof India Pvt.ltd.,New Delhi

Sequential & Parallel Algorithms for Big-Integer Numbers Subtraction

GJCST Classifications:
C.1.4, C.1.2, F.1.3

Youssef Bassil, Aziz Barbar

Abstract- **Many emerging computer applications require the processing of large numbers, larger than what a CPU can handle. In fact, the top of the line PCs can only manipulate numbers not longer than 32 bits or 64 bits. This is due to the size of the registers and the data-path inside the CPU. As a result, performing arithmetic operations such as subtraction on big-integer numbers is to some extend limited. Different algorithms were designed in an attempt to solve this problem; they all operate on big-integer numbers by first converting them into a binary representation then performing bitwise operations on single bits. Such algorithms are of complexity O(n) where n is the total number of bits in each operand.**

This paper proposes two new algorithms for performing arithmetic subtraction on big-integer numbers. The two algorithms are different in that one is sequential while the other is parallel. The similarity between them is that both follow the same concept of dividing the big-integer inputs into several blocks or tokens of 60 bits (18 digits) each; thus reducing the input size n in O(n) by a factor of 60. Subtraction of corresponding tokens, one from each operand, is performed as humans perform subtraction, using a pencil and a paper in the decimal system.

Both algorithms are to be implemented using MS C#.NET 2005 and tested over a multiple processor system. Further studies can be done on other arithmetic operations such as addition and multiplication.

Keywords- Computer algorithm, Large numbers subtraction, Sequential algorithm, Parallel algorithm.

I INTRODUCTION

Contemporary PCs usually handle and operate on numbers not longer than 32 bits and 64 bits (Maxfield & Brown, 2004). The reason behind this is that PCs' CPUs can only accommodate and manipulate numbers of that length (Hennessy & Patterson, 2006). The real problem arises when certain applications require performing computer arithmetic such as addition, subtraction, multiplication, and division on numbers larger than 64 bits, at high speed. For instance, in cryptography cipher keys can be as large as 512-bits and 1024-bits. In banking systems, customer's balances can be sometimes larger than 64-bits taking into consideration the difference between currencies.

Manuscript received "28th December"
Youssef Bassil, Department of Computer Science, American University of Science & Technology, Beirut, Lebanon,
Aziz Barbar, Department of Computer Science, American University of Science & Technology, Beirut, Lebanon,
Telephone: 961 1 218716 ext 311

In some scientific and mathematical applications, performing precise and accurate real-time computations demand the use of numbers larger than 64 bits.

Various solutions were proposed to solve this problem; the majority of them carry out arithmetic operations on the bit level using bitwise operations (Knuth, 1997), (Koren, 2001). Furthermore, none of them is designed to exploit multiprocessor systems and shared memory architecture. The complexity of such algorithms is usually O(n) where the basic operation is executed n times, that is eventually equal to the number of bits in the big-integer input.

In this paper, we are proposing a sequential and a parallel algorithm for handling arithmetic subtraction on big-integer numbers. Both algorithms carry out subtraction on 60 bits unit tokens, and not on individual bits as other existing approaches. Accordingly, the algorithm is of best-case complexity O(n) with n reduced by a factor of 60; hence, the computation is 60 times faster. In general, both algorithms emulate the elementary pen and paper method used to perform subtraction in the decimal system in that they start by dividing the big-integer operands into tokens or blocks of 60 bits each. Then each two corresponding tokens are subtracted from each other while the borrows are handled properly. In the sequential algorithm, this whole process is executed on a single processor system; while in the parallel algorithm, each two corresponding tokens are assigned to a particular processor in a multi-processor system to be subtracted from each other. Experiments showed a momentous improvement over other existing techniques and approaches.

II EXISTING SOLUTIONS

Many programming libraries were developed to solve the problem of performing arithmetic calculations over big-integer numbers. Some of them are proprietary third party dynamic link libraries (DLL), either available for free or sold at a given cost; while others are shipped as a part of the programming language application programming interface (API). For instance, the MS .NET Framework 4.0 provides the BigInteger class in the namespace System.Numerics (MSDN, 2009). The Java programming language provides another BigInteger class in the java.math package (Java Documentation, 2008). Both carry out arithmetic operations on big-integer numbers using bitwise operations (Java BigInteger Source-code, 2006). They first convert the base-10 big-integer input to a base-2 binary representation, then they employ the bitwise operators OR and XOR to perform binary subtraction over string of bits.

The algorithm behind these libraries is of complexity O(n) where n is the total number of bits constituting each operand. In terms of time efficiency, the number of times the basic operation is executed, is equal to the number of bits in the big-integer operands. Moreover, most of these libraries are not designed to work in a parallel fashion; they are only optimized to operate over single-processor systems.

III THE PROPOSED SEQUENTIAL ALGORITHM

The sequential algorithm proposed in this paper is based on the same principle humans use to perform subtraction, using a pencil and a paper in the decimal system. Generally speaking, inputs of big-integer numbers are chopped to several smaller tokens each made out of 60 bits (18 digits). Afterwards, each two corresponding tokens are treated as single units and aligned on top of each others. Then they are subtracted from each others while handling correctly the borrows. It is worth noting that no conversion to base-2 is to occur, the computation is totally done in the base-10 decimal system

Below is a list of steps executed by the sequential algorithm to subtract two big-integer numbers:

i. Two big-integer operands of type string a and b, such as a is greater or equal to b, are fed to the algorithm. SubtractBigInteger(a , b)

ii. Both string operands a and b are then parsed and divided from right to left into smaller chunks or tokens ti(p) where i is the token index and p is the operand to which ti belongs. Consequently, operand a = $t_{n-1}(a)\ldots t_0(a)$ and operand b = $t_{m-1}{}^{(b)\ldots}$ $t_0(b)$, where n and m are the total number of tokens constituting each of the operands. The length of each single produced token ti is less than or equal to 18. (In the C# programming language, the largest integer data type is long (signed by default) which can store up to 19 digits or 2^{63}= 9223372036854775808. Since in mathematical subtraction there exist the concept of a borrow, it is crucial to reserve 1 digit for a possible borrow, resulting in 19-1=18 digits represented by 60 bits). The resulting tokens will be stored as strings in two arrays, each for a particular operand.

iii. The tokens contained in the two arrays are to be converted from *string* to *long* data type. In other words, each single token, now representing an array element with a maximum length of 18 digits, is to be converted to an integer value of type *long*. The conversion is required because arithmetic subtraction cannot be performed over *string* types

iv. Both arrays, now containing long type tokens, are aligned on top of each other. Starting from the rightmost token, each two corresponding tokens are subtracted from each other as in performing subtraction using a pencil and a paper: $t_i(c) = t_i(a) - t_i(b)$. If $t_i(a) < t_i(b)$, then a borrow of 1 must be subtracted from $t_{i+1}(a)$. t_{i+1} is the next token on the left of the two tokens being currently subtracted. Consequently, $t_{i+1}(a)$ would be equal to $t_{i+1}(a)$-1 and a 1 representing the borrow is appended as the 19th digit to $t_i(a)$. It is important to mention that in case $t_{i+1}(a)$ is equal to 0, a borrow is taken from $t_{i+2}(a)$ then propagated to t_i+1(a) and then to $t_i(a)$. Under special cases, a borrow can be propagated from $t_{n-1}(a)$ to ti(a). Since operand a is always greater or equal to operand b, tn-1(a) must be able to provide a borrow in a way or another. Now $t_i(a)$ >= $t_i(b)$ and $t_i(a) - t_i(b)$ is feasible to be calculated.

v. Finally, all the produced ti(c) are to be concatenated together to attain result = $t_{r-1}(c)\ldots t_0(c)$. It is important to note that this algorithm can handle operands of different sizes, in a sense that excessive tokens, which should logically belong to the largest operand are just appended to the final result. Figure 1 summaries the different steps performed by the sequential algorithm in order to subtract two operands a and b.

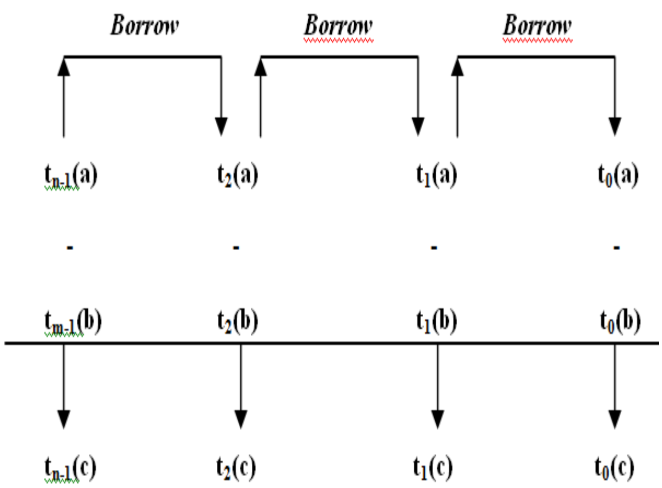


Fig.1. Subtracting two operands using the proposed sequential algorithm.

A. *Implementation*

Below is the code of the sequential algorithm implemented using MS C#.NET 2005 under the .NET Framework 2.0 and MS Visual Studio 2005.

```
private string SubtractBigInteger(string a, string b)
{
   long[] tokens_A = ParseOperand(a);
   long[] tokens_B = ParseOperand(b);

   int length = tokens_A.Length ;

   long[] result = new long[length];

   int i, j;
   for (i = length - 1, j = length - 1; j != -1; i--, j--)
   {
      // we must borrow a 1 from the token on the left
      while (tokens_A[i] < tokens_B[j])
      {
```

```
            int k = i - 1;

            while (true)
             {
              if (tokens_A[k] != 0)
              {
               tokens_A[k]--;
                 // Adding the 19th digit to the left of the
                              token
                  // that needs borrow
                  tokens_A[k + 1] = tokens_A[k + 1] + (1 *
                                       1000000000000000000);
                 break;
              }
              else k--;

           }
        }

          // Performing the subtraction
            result[i] = tokens_A[i] - tokens_B[j];
         }

        return ConvertToString(result);
      }

private long[] ParseOperand(string operand)
   {
       ArrayList list = new ArrayList();

      for (int i = 0; operand.Length > 18; i++)
        {
          list.Add(operand.Substring(operand.Length - 18));

        operand = operand.Substring(0, operand.Length -18))
}
        list.Add(operand);
         list.Reverse();

         long[] tokens = new long[list.Count];

         for (int j = 0; j < tokens.Length; j++)
         {
            tokens[j] = Convert.ToInt64(list[j]);
         }
         return tokens;
      }
```

B. Experiments and Results

We will be comparing in our testing the execution time of the proposed sequential algorithm with the System.Numercis.BigInteger class included in MS .NET Framework 4.0, and the java.math.BigInteger class included in Java SE 1.6.

Below are two code segments that illustrate how to use the methods of the built-in classes System.Numercis.BigInteger and java.math.BigInteger in order to subtract two big-integer numbers using the C#.NET and the Java language.

```
using System.Numerics;
public class BigIntegerTest_Csharp
  {
      public static void Main(string args[])
{
       String operandA =
      "12345678909876543211234567890987654321" ;
       String operandA =
     "12345678909876543211234567890987654321" ;
       BigInteger a = BigInteger.Parse(operandA) ;
       BigInteger b = BigInteger.Parse(operandB) ;

       BigInteger results = BigInteger.Subtract(a, b);
      Console.WriteLine(results.ToString());
     }
}

   import java.math.BigInteger;
    public class BigIntegerTest_Java
    {
      public static void main(String args[])
      {

           String operandA =
           "12345678909876543211234567890987654321" ;
           String operandA =
           "12345678909876543211234567890987654321" ;
          BigInteger a = new BigInteger(operandA) ;
          BigInteger b = new BigInteger(operandB) ;

         System.out.print("" + a.subtract(b)) ;
     }
   }
```

As a testing platform, we are using a desktop IBM-compatible PC with Intel Core single core processor with 1.66 MHz clock speed, 256KB of cache, and 512MB of RAM. The operating system used is MS Windows XP Professional SP2.

It is worth noting that the execution time obtained for all different algorithms is an average time obtained after five consecutive runs of the same test.

Test Case	Operands	Value	Value Length
1	A	X	20,000 base-10 digits
1	B	Y	20,000 base-10 digits
2	A	X	100,000 base-10 digits
2	B	Y	100,000 base-10 digits
3	A	X	500,000 base-10 digits
3	B	Y	500,000 base-10 digits
4	A	X	1000,000 base-10 digits
4	B	Y	1000,000 base-10 digits

Table 1: Test cases.

Test Case	Operation	Results	Execution Time in Seconds
1	A-B	X-Y	2.01
2	A-B	X-Y	17.19
3	A-B	X-Y	514.33
4	A-B	X-Y	2487.05

Table 2: Results obtained for the .NET class.

Test Case	Operation	Results	Execution Time in Seconds
1	A-B	X-Y	0.51
2	A-B	X-Y	8.02
3	A-B	X-Y	165.34
4	A-B	X-Y	601.90

Table 3: Results obtained for the Java class.

Test Case	Operation	Results	Execution Time in Seconds
1	A-B	X-Y	0.18
2	A-B	X-Y	2.06
3	A-B	X-Y	66.04
4	A-B	X-Y	318.71

Table 4: Results obtained for our sequential algorithm.

From the obtained results, delineated in tables 1-4, it became clear that our sequential algorithm outsmarted all other algorithms in all different test cases. When big-integer numbers were in length respectively 20,000 and 100,000 digits, our algorithm beat the .NET and Java classes with few little seconds. However, when numbers became as large as 500,000 digits, our algorithm surpassed the Java class by around 100 seconds (1.6 minutes) and the .NET class by around 450 seconds (7.5 minutes). Furthermore, our proposed algorithm increased the pace between its rivals when the length of operands reached the 1,000,000 digits. It surpassed the Java class by around 283 seconds (4.7 minutes), and the .NET class by around 2169 seconds (36.1 minutes).

C. *Algorithm Analysis*

The sequential algorithm showed a real speed improvement over other existing approaches. It outperformed the .NET and Java built-in classes by several seconds and this gap exponentially increased as the length of the big-integer operands became larger. This speed improvement is due to the reduction of the input size n in $O(n)$. The .NET, Java, and our proposed algorithm are all of best-case complexity $O(n)$. However, the n in the .NET and Java algorithms represents the total number of bits in each operand; while in our proposed algorithm, n represents the total number of tokens in each operand. For instance, the decimal number 999999999999999999 (18 digits) is represented in base-2 as 110111100000101101101011001110100111011000111111 111111111111 (60 bits). This makes $n=60$ and thus the basic operation is executed 60 times. On the other hand, in our proposed algorithm, this whole 999999999999999999 is treated as a single unit token. This makes $n=1$ and thus the basic operation is executed only 1 time. As a result, the time efficiency of our algorithm is supposedly to be 60 times faster than the other algorithms. However, this is not the case, since handling the borrows requires various extra operations to be executed; a fact that imposes further processing overhead, and increases the computation time. Accordingly, the best-case is when no borrows are needed throughout the execution of the whole algorithm, then the basic operation is executed n times where n is the total number of tokens and this makes $C_{Best}(n)=n$ belonging to $O(n)$.

IV THE PROPOSED PARALLEL ALGORITHM

The parallel algorithm proposed in this paper is a multithreaded parallel algorithm designed to be executed over multi-processor shared memory architecture. It is based on the principle of performing arithmetic subtraction as humans perform subtraction, using a pencil and a paper in the decimal system. Ordinarily, the algorithm starts by breaking down big-integer numbers into blocks or tokens of 60 bits each. Then subtraction starts in a sequence of multiple iterations. On the first iteration, each two corresponding tokens are assigned to a particular thread, which subtracts them from each others using a particular microprocessor. When a borrow is needed, the algorithm assumes that the borrow is there. For instance, if token 99 is to be subtracted from token 88, the algorithm will directly subtract 99 from 188 as if a borrow had occurred. Then, a value of 1 representing the borrow is stored in a shared array. On the second iteration, that borrow will be subtracted accordingly from the token on the left of the previous result. Iterations continue until no more borrows are generated from a previous iteration.

Below is a list of steps executed by the parallel algorithm to subtract two big-integer numbers:

i. Two very large string numbers operand a and operand b, such as a is greater or equal to b, are fed to the algorithm. SubtractBigInteger_Parallel(a , b)

ii. Both *string* operands a and b are then parsed and divided from right to left into smaller chunks or tokens $t_i(p)$ where i is the token index and p is the operand to which $\mathbf{t_i}$ belongs. Consequently, operand a = $t_{n-1}(a)\ldots t_0(a)$ and operand b = $t_{m-1}(b)\ldots$ t0(b) where n and $_{\mathbf{m}}$ are the total number of tokens constituting each of the operands. The length of each single produced token t_i is less than or equal to 18. (In the C# programming language, the largest integer data type is *long* (signed by default) which can store up to 19 digits or 2^{63}= 9223372036854775808. Since in mathematical subtraction there exist the concept of a borrow, it is crucial to reserve 1 digit for a possible *borrow*, resulting in 19-1=18 digits represented by 60 bits). The resulting tokens will be stored as *string* in two arrays, each for a particular operand.

iii. The tokens contained in the two arrays are to be converted from string to long data type. In other words, each single token, now representing an array element with a maximum length of 18 digits, is to be converted to an integer value of type long. The conversion is required because arithmetic subtraction cannot be performed on string types

iv. Each processor pi in a multiprocessor system is assigned two tokens, one from each operand. Therefore the processor pi is assigned tokens ti(a) and ti(b) with the purpose of calculating ti(c) = ti(a) - ti(b). For instance, p0 will calculate t0(c), p1 will calculate t1(c), p2 will calculate t2(c) and so on. We are to assume that the number of processor is equal to the number of tokens; otherwise, tokens are distributed equally among processors. For instance, if the number of processors is half the number of tokens, each processor will be assigned 4 tokens (2 from each operand) to be calculated as in sequential approach. $t_i(c) = t_i(a) - t_i(b)$ and then $ti_{+1}(c) = t_{i+1}(a) - t_{i+1}(b)$

v. A borrow is handled using multiple processing iterations and a shared array called borrows[0...n-1] used to store all the produced borrows. For that reason, we have added a new variable called T as in ti(c,T) to represent the iteration into which ti(c) is being calculated. T=1 is the first iteration and T=n is the nth iteration. In this approach, in case a borrow was needed, the algorithm assumes that the borrow is there. For instance, if token 99 is to be subtracted from token 88, the algorithm will directly subtract 99 from 188 as if a borrow had occurred, and borrows[i+1] is set to 1. It is i+1 so that on the next iteration T=2, borrows[i+1] will be correctly subtracted from the previously calculated t_i+1(c,1). Likewise, if another borrow is needed for $t_i(c,2)$, borrows[i+1] is set to 1 overwriting any previous value. Consequently, on the next iteration (T=3) borrows[i+1] will be correctly subtracted from $t_{i+1}(c,2)$. This will keep on looping until no more borrows are generated (borrows[0...n-1] contains no 1's). As an example, if on the first iteration (T=1), a borrow was needed for t4(a,1), then a borrow is assumed to have occurred, t4(c,1) is calculated and borrows[5] is set to 1, p5 (processor 5) starts a second iteration (T=2) in an attempt to calculate $t_5(c,2)$ = t5(c,1) - borrows[5]. In the meantime, all other pi where borrows[i]=0 will refrain from executing. If after T=2 no more borrows are needed, the loop process stops.

vi. Finally, all the ti(c) produced after many iterations are to be concatenated together: result = $t_{n-1}(c)\ldots t_0(c)$. Figure 2 summaries the different steps performed by the parallel algorithm in order to subtract two operands a and b.

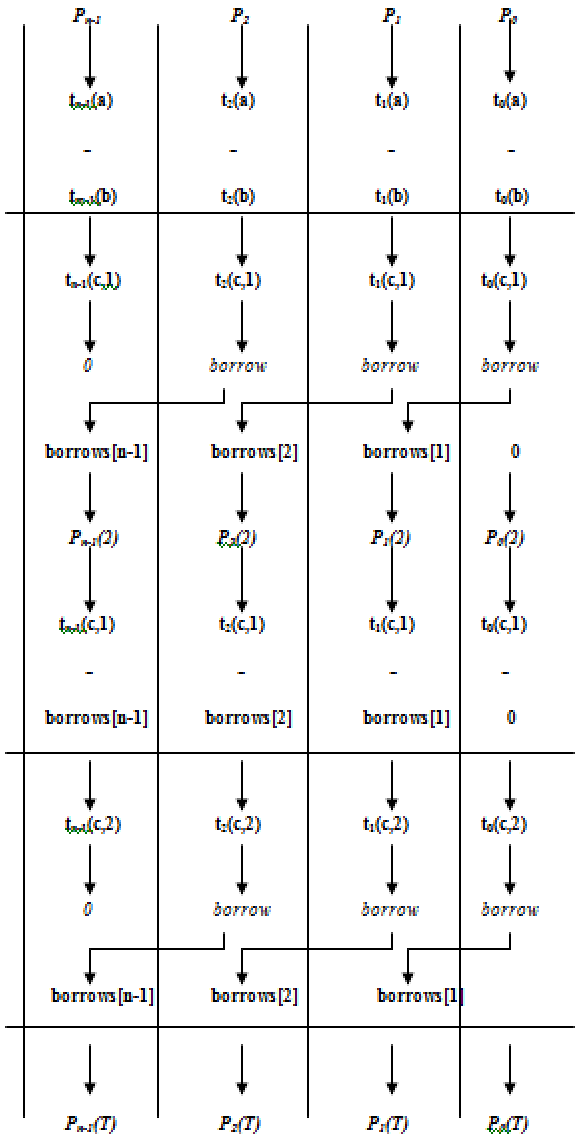


Fig. 2. Subtracting two operands using the proposed parallel algorithm.

A. *Implementation*

Below is the code of the proposed parallel algorithm implemented in MS C#.NET 2005 under the .NET Framework 2.0 and MS Visual Studio 2005. It uses classes and methods from *System.Threading* namespace to create, destroy and execute threads. All threads can read and write to a shared memory space where tokens, carries, flags and other variables are stored and shared.

```
long[] tokens_A ;
long[] tokens_B ;

long[] result;
int[] borrows;

int numberOfProcessors;

int sharedIndex;
```

```
int terminatedThreads=0;
int T=1;
 Thread[] threads;

public void SubtractBigInteger_Parallel(string a,
                stringb)
{
   tokens_A = ParseOperand(a, 18);
   tokens_B = ParseOperand(b, 18);

  result = new long[tokens_A.Length];


   // By default borrows is populated with 0s
   borrows = new int[tokens_A.Length];

   numberOfProcessors = GetNumOfProcessors();

   threads = new Thread[numberOfProcessors];

   CreateThreads();
}

private void CreateThreads()
{
   sharedIndex = numberOfProcessors;

   for (int i = 0; i < numberOfProcessors; i++)
   {
      threads[i] = new Thread(new
       ThreadStart(Process));
      threads[i].Start();
   }
}

private void Process()
{
   int index = sharedIndex--;

   if(T==1) // First iteration
   {
         if(tokens_A[index] < tokens_B[index])
         {
            // Add a borrow
            tokens_A[index] + (1
               * 1000000000000000000);

             borrows[index-1] = 1 ;
   }

   else borrows[index-1] = 0 ;

  result[index] = tokens_A[index] - tokens_B[index];
 }
 else
{
    if(result[index] == 0) // result[index] < borrows[index]
    {
            // Add a borrow
                result[index] + (1 * 1000000000000000000);

             borrows[index-1] = 1 ;
         }
      else borrows[index-1] = 0 ;

      result[index] = result[index] - borrows[index];
   }

      terminatedThreads++;

      IsProcessingDone();
   }

   private void IsProcessingDone()
   {
      if (terminatedThreads == numberOfProcessors)
      {
         if (AreMoreBorrows())
         {
            T++ ;

            // Creates new set of threads in the next iteration
            CreateThreads();
         }
         else DisplayResults();
      }
   }

   private bool AreMoreBorrows()
   {
      for (int i = 0; i < borrows.Length; i++)
      {
         if (borrows[i] = = 1)
            return true;
      }

      return false;
   }

   private string DisplayResults()
   {
      return ConvertToString(result);
   }

  private long[] ParseOperand(string operand)
  {
      ArrayList list = new ArrayList();

      for (int i = 0; operand.Length > 18; i++)
      {
         list.Add(operand.Substring(operand.Length - 18));

 operand = operand.Substring(0, operand.Length-      18);
}
      list.Add(operand);
```

```
        list.Reverse();

        long[] tokens = new long[list.Count];

        for (int j = 0; j < tokens.Length; j++)
        {
            tokens[j] = Convert.ToInt64(list[j]);
        }

        return tokens;
    }
```

B. *Experiments and Results*

In this section, a comparison of the execution time between the sequential and the parallel algorithm, both proposed in this paper, is undertaken using a desktop IBM-compatible PC with 4 processors of type Intel Core single core with 1.8 MHz clock speed, 512KB of cache, and 2GB of RAM. The operating system used is MS Windows Server 2003 SP1.
It is important to note here that the execution time obtained for all different algorithms is an average time obtained after five consecutive runs of the same test.

Test Case	Operands	Value	Value Length
1	A	X	20,000 base-10 digits
1	B	Y	20,000 base-10 digits
2	A	X	100,000 base-10 digits
2	B	Y	100,000 base-10 digits
3	A	X	500,000 base-10 digits
3	B	Y	500,000 base-10 digits
4	A	X	1000,000 base-10 digits
4	B	Y	1000,000 base-10 digits

Table 5: Test cases.

Test Case	Operation	Results	Execution Time in Seconds
1	A+B	X+Y	0.10
2	A+B	X+Y	2.12
3	A+B	X+Y	64.51
4	A+B	X+Y	311.66

Table 6: Results obtained for our sequential algorithm.

Test Case	Operation	Results	Execution Time in Seconds
1	A+B	X+Y	0.03
2	A+B	X+Y	0.92
3	A+B	X+Y	23.17
4	A+B	X+Y	91.96

Table 7: Results obtained for our parallel algorithm.

The results delineated in tables 5-7 show that the parallel algorithm outperformed the sequential algorithm by an average factor of 3.2. At the beginning, when operands were in length 20,000 and 100,000 respectively, the difference was not that evident. However, when numbers became larger, the gap increased and the execution time was speeded up by around 320%.

C. *Algorithm Analysis*

The parallel algorithm improved on the sequential algorithm and boosted its execution time by around 320%. In terms of algorithm complexity, and assuming that every token is exactly assigned to a particular processor, the best-case efficiency is when no borrows are generated after the first iteration; a fact that achieves the best performance, and thus $C_{Best}(n)=1$, that is each processor executes the basic operation only one time. The worst-case efficiency is when a new borrow is generated after each iteration, this would require n-1 iterations in order to propagate and subtract all the borrows and thus $C_{Worst}(n)=n-1$. Consequently, The average-case efficiency is $C_{Average}(n)=(n-1)/2$

V FUTURE WORK

Future research can improve upon our proposed algorithms so much so that other arithmetic operations such as addition, multiplication and division are added. Besides, a distributed version of the same algorithms is to be designed so that it can be executed over a network of regular machines, making the implementation less expensive and more scalable.

VI ACKNOWLEDGMENTS

1) Maxfield, C., & Brown, A. (2004). The Definitive Guide to How Computers Do Math: Featuring the Virtual DIY Calculator. Wiley-Interscience.
2) Hennessy, J., & Patterson, D. (2006). Computer Architecture: A Quantitative Approach (4th ed.). Morgan Kaufmann.
3) Knuth, D. (1997). Art of Computer Programming, Volume 2: Seminumerical Algorithms. (3rd ed.). Addison-Wesley Professional.
4) Koren, I. (2001). *Computer Arithmetic Algorithms* (2nd ed.). AK Peters Ltd.
5) Microsoft Developer Network, MSDN. (2009) http://msdn.microsoft.com/en-us/library/system.numerics.biginteger(VS.100).aspx
6) Sun Java SE Documentation 6.0. (2008) http://java.sun.com/javase/6/docs/api/java/math/BigInteger.html
7) Java BigInteger source code (2006) http://developer.classpath.org/doc/java/math/BigInteger-source.html

FELLOW OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (FICCT)

- FICCT' title will be awarded to the person after approval of Editor-in-Chief and Editorial Board. The title 'FICCT" can be added to name in the following manner e.g. Dr. Andrew Knoll, Ph.D., FICCT,
  Er. Pettor Jone, M.E., FICCT
- FICCT can submit two papers every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free unlimited Web-space will be allotted to 'FICCT 'along with subDomain to contribute and partake in our activities.
- A professional email address will be allotted free with unlimited email space.
- FICCT will be authorized to receive e-Journals - GJCST for the Lifetime.
- FICCT will be exempted from the registration fees of Seminar/Symposium/Conference/Workshop conducted internationally of GJCST (FREE of Charge).
- FICCT will be an Honorable Guest of any gathering hold.

ASSOCIATE OF INTERNATIONAL CONGRESS OF COMPUTER SCIENCE AND TECHNOLOGY (AICCT)

- AICCT title will be awarded to the person/institution after approval of Editor-in-Chef and Editorial Board. The title 'AICCTcan be added to name in the following manner:
  eg. Dr. Thomas Herry, Ph.D., AICCT
- AICCT can submit one paper every year for publication without any charges. The paper will be sent to two peer reviewers. The paper will be published after the acceptance of peer reviewers and Editorial Board.
- Free 2GB Web-space will be allotted to 'FICCT' along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted with free 1GB email space.
- AICCT will be authorized to receive e-Journal GJCST for lifetime.

ANNUAL MEMBER

- Annual Member will be authorized to receive e-Journal GJCST for one year (subscription for one year).
- The member will be allotted free 1 GB Web-space along with subDomain to contribute and participate in our activities.
- A professional email address will be allotted free 500 MB email space.

PAPER PUBLICATION

- The members can publish paper once. The paper will be sent to two peer reviewer. The paper will be published after the acceptance of peer reviewers and Editorial Board.

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJopurnals.org website. There are basically 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.

Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, then email the paper directly to dean@Globaljournal.org.

Offline Submission: Author can send the typed form of paper by Post. But Online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION <u>(Must be strictly followed)</u>

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Times New Roman.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, “Abstract” word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be two lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript’s Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global Journals are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals and Editorial Board, will become the *copyright of the Global Journals.*

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications

Research letters: The letters are small and concise comments on previously published matters.

5.STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers*:* These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)*Title* should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, *"Abstract"* (less than 150 words) containing the major results and conclusions.

(c) Up to *ten keywords*, that precisely identifies the paper's subject, purpose, and focus.

(d) An *Introduction*, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; *conclusions* should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals, ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the *Harvard scheme* of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals.

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals (Publication Prior to Print)

The Global Journals are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

Index

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at
http://www.GlobalJournals.org or http://www.ComputerResearch.org
Email us at helpdesk@globaljournals.org

32011
6 586986 142797
.com