Disjoint Web Document Clustering and Management in Electronic Commerce

Mei-Ling Shyu
Department of Electrical and Computer Engineering, University of Miami,
Coral Gables, FL 33124-0640, USA

Shu-Ching Chen*
Distributed Multimedia System Laboratory, School of Computer Science,
Florida International University, Miami, FL 33199, USA

Choochart Haruechaiyasak
Department of Electrical and Computer Engineering, University of Miami,
Coral Gables, FL 33124-0640, USA

Chi-Min Shu
National Yunlin University of Science and Technology,
Department of Environmental and Safety Engineering, Yunlin, Taiwan, R.O.C.

Sheng-Tun Li
National Kaohsiung First University of Science and Technology
Department of Information Management,
Yenchao, Kaohsiung, Taiwan 824, R.O.C.

Abstract

Due to the recent trend in electronic commerce, many companies provide their product related or any usable information on their Web sites for customer convenience. This company information is organized as separate Uniform Resource Locator (URL) pages. Each URL represents a Web document that can be linked to or from other documents via the hyperlinks. How to customize a company's Web site for its Web site layout so that it can target its potential customers to improve profits is important in electronic commerce. In this paper, we propose the Markov Model Mediator (MMM) mechanism to organize and manage the groups of related URLs into disjoint clusters for document management in electronic commerce. An experiment is conducted using a real data set and the experimental result shows that our proposed approach yields a better performance over all different tested cluster sizes in comparison with depth-first search (DFS), breadth-first search (BFS), and the random clustering strategies.

*This research was supported in part by NSF CDA-9711582.

1 Introduction

In order to be successful in the highly competitive industry of electronic commerce, companies must be able to precisely target their potential customers and convince them to purchase their products or services. It is very common that companies provide their product related or any usable information on their Web sites through separate *Uniform Resource Locator (URL)* pages. Each *URL* represents a Web document that can be linked to or from other documents via the hyperlinks. In the context of electronic commerce, the customer behavior is captured by analyzing the user navigation through the company's Web site and therefore Web site layout is very important for the company. The ability to automatic analysis of user actions is a useful input in the process. Therefore, Web usage mining has recently emerged as an analytical tool for management and decision-making [9].

Various data mining methods have been extended to Web usage mining including association rule mining [5], sequential pattern analysis [1], clustering [4, 10], etc. Clustering is a process of grouping physical or abstract objects into classes or similar objects [3]. Due to these dynamic behaviors of the Web documents, clustering based on only the static quantities, i.e., terms or keywords, does not very well

capture all characteristics of the Web documents. Thus, besides terms or concepts, Web document clustering can also incorporate some dynamic quantities such as the hyperlinks and the access patterns extracted from the user queries during the clustering process. In addition, most of the time the *URLs* can be categorized based on their high-level concepts, such as product-related and customer service-related pages. For example, clustering analysis of a software company's Web site might yield a result such as the programming development *URLs* are highly correlated to the software training *URLs*. The results from the clustering analysis can assist in managing the Web documents by reorganizing and customizing the company's Web site for its Web site layout. For example, to help the customer navigate through the Web site, some hyperlinks of programming development *URLs* can be added into the *URLs* of software training, and vice versa.

In our previous study, we have proposed the *Markov Model Mediator (MMM)* mechanism that employs the affinity-based data mining techniques to facilitate the functionality of a *database management system (DBMS)* by organizing a network of databases into clusters [6, 7]. In this paper, we extend the *MMM* mechanism to organize and manage the Web documents. The proposed framework considers the user access patterns to cluster groups of *URLs* in the clustering process, and is implemented using C++. An experiment is conducted using a real data set of Microsoft Web site – *Microsoft Anonymous Web Data*, which records how *www.microsoft.com* was visited by the users in a one-week timeframe in February 1998 [2]. The experimental result shows that the proposed Web document clustering approach yields a better performance in comparison with depth-first search (*DFS*), breadth-first search (*BFS*), and the *random* clustering methods.

The paper is organized as follows. Next section gives the proposed Web document clustering approach. In Section 3, the experiment on a real data set is conducted and the result is presented. The paper is concluded in Section 4.

2 The Proposed Web Document Clustering Method

The proposed Web document clustering method is based on the *Markov Model Mediator (MMM)* mechanism. In the proposed method, each *URL* is assumed to be categorized into a high-level concept group of *URLs*. Thus, each *URL* group in the Web site is modeled by an *MMM* and contains a set of conceptually related *URLs*. An *MMM* is represented by a 6-tuple $\lambda = (\mathcal{S}, \mathcal{F}, \mathcal{A}, \mathcal{B}, \Pi, \Psi)$. The definitions and the constructions of the components of the *MMM* mechanism can be found in [8].

2.1 Affinity-Based Measures

The relative affinity measures are used to indicate how frequently two *URLs* are accessed together. Two *URL* groups whose *URLs* are accessed together more frequently are said to have a higher relative affinity relationship. For a given Web site, its *URL* groups and user queries (user access patterns) are defined as follows.

- $G = \{g_1, g_2, \ldots, g_g\}$ = a set of *URL* groups in the Web site
- n_i = number of *URLs* in g_i
- $Q = \{1, 2, \ldots, q\}$ = a set of user queries
- $use_{m,k}$ = usage pattern of *URL* m with respect to query k per time period

$$use_{m,k} = \begin{cases} 1 & \text{if } \textit{URL } m \text{ is accessed by query } k \\ 0 & \text{otherwise} \end{cases}$$

- $access_k$ = access frequency of query k per time period
- $aff_{m,n}$ = affinity measure of *URLs* m and n

$$aff_{m,n} = \sum_{k=1}^{q} use_{m,k} \times use_{n,k} \times access_k$$

These affinity measures are used to construct the probability distributions for an *MMM*. The constructions of the probability distributions $\mathcal{A}$, $\mathcal{B}$ and Π are shown in [8].

2.2 Similarity Measures

A similarity measure shows how well two *URL* groups match the observations generated by the sample queries. Let $N_k = k1 + k2$, $\mathcal{OS}$ be a set of all observation sets, and $S(g_i, g_j)$ be the similarity measure between two *URL* groups g_i and g_j. The similarity values are computed for the pairs of *URL* groups that are connected in the browsing graph.

$$S(g_i, g_j) = \sum_{O^k \in \mathcal{OS}} P(O^k \mid X, Y; g_i, g_j) P(X, Y; g_i, g_j) F(N_k),$$

where

- $O^k = \{o_1, \ldots, o_{N_k}\}$ is an observation set with the attributes belonging to g_i and g_j and generated by query k
- $X = \{x_1, \ldots, x_{k1}\}$ is a set of *URLs* belonging to g_i in O^k
- $Y = \{y_1, \ldots, y_{k2}\}$ is a set of *URLs* belonging to g_j in O^k

- A_i, B_i, and Π_i are the state transition probability distribution, the observation symbol probability distribution, and the initial state probability distribution for each *URL* group g_i, respectively.
- $P(O^k \mid X, Y; g_i, g_j)$ = the probability of occurrence of O^k given $X \in g_i$ and $Y \in g_j$
  $= \prod_{u=1}^{k1} B_i(o_u \mid x_u) \prod_{v=k1+1}^{N_k} B_j(o_v \mid y_{v-k1})$
- $P(X, Y; g_i, g_j)$ = the joint probability of $X \in g_i$ and $Y \in g_j$
  $= \prod_{u=2}^{k1} A_i(x_u \mid x_{u-1})\ \Pi_i(x_1)\ \prod_{v=k1+2}^{N_k} A_j(y_{v-k1} \mid y_{v-k1-1})\ \Pi_j(y_1)$
- $F(N_k) = 10^{N_k}$ = a factor to adjust the variable observation set lengths

The resulting similarity values can be constructed as a matrix of n by n, where n is the number of *URL* groups. The similarity matrix is symmetric which means $S(g_i, g_j)$ is equal to $S(g_j, g_i)$.

2.3 URL Group Clustering Strategy

The similarity values are transformed into the branch probabilities $P_{i,j}$ for pairs of nodes i and j (*URL* groups) in a browsing graph. The transformation is done by normalizing the similarity matrix per row to indicate the branch probabilities from a specific node (*URL* group) to all its accessible nodes (*URL* groups). Then the stationary probability ϕ_i for each node i in the browsing graph can be obtained from the branch probabilities, and the weights of the nodes in the browsing graph can be calculated as follows.

$$\sum_i \phi_i = 1 \qquad \phi_j = \sum_i \phi_i \times P_{i,j} \quad j = 1, 2, \cdots$$
$$W_{i,j} = \phi_i \times P_{i,j} + \phi_j \times P_{j,i}$$

The stationary probability ϕ_i denotes the relative frequency of accessing node i (*URL* group g_i) in the long run. Once all the branch weights are calculated, the browsing graph can be constructed by considering the n highest weights between pairs of nodes, where n is a variable which specifies the graph complexity.

The proposed document clustering strategy is traversal-based and greedy-oriented. *URL* groups are partitioned with the order of their stationary probabilities. For a cluster of size c, the *URL* group which has the largest stationary probability is selected to start a cluster. Then we consider another $(c-1)$ *URL* groups whose connecting weights with the starting group are the highest among the rest of the groups. The *URL* group which has the largest stationary probability is selected and the process continues until the current cluster fills up. At this point, the next un-partitioned *URL* group from the sorted list starts a new cluster. The whole process is repeated until no un-partitioned *URL* group remains.

3 The Experiment

We implemented our proposed framework in C++. For the data set, the *Microsoft Anonymous Web Data* from the Knowledge Discovery in Databases Archive at University of California, Irvine *UCI KDD* archive [2] is used. The data set records which areas of *www.microsoft.com* each user visited in a one-week timeframe in February 1998. The sample data set used in our experiment contains *5,000* of anonymous users. There are a total of *294 URLs* covered in the data set. From these *URLs*, we construct *39* attributes based on their usage and contents. For example, the attribute *country* is assigned for those *URLs* whose content are written in non-English languages and the attribute *programming* is assigned for the *URLs* whose contents are related to the programming languages. We then categorized these *URLs* into *13* groups based on these predefined attributes.

Once the above information is preprocessed, we test the data set for different cluster sizes and compare our approach with breadth-first search (*BFS*), depth-first search (*DFS*) and the random clustering methods. For *BFS* and *DFS*, the node ordering is obtained by traversing the browsing graph obtained from our *URL* group clustering strategy. To compare the clustering performance, we use a metric based on the total number of inter-cluster accesses calculated from each individual user. For example, consider a user session which contains the following set of *URLs*,

URL_1: "Mexico" which belongs to *URL* group country,

URL_2: "Internet Explorer" which belongs to *URL* group Internet,

URL_3: "Internet Development" which belongs to *URL* group Internet, and

URL_4: "Mail Support", which belongs to *URL* group Service and Support.

Suppose we have a cluster size of one, therefore, URL_1 belongs to one cluster, URL_2 and URL_3 belong to the same cluster, and URL_4 belongs to another cluster. We consider the first *URL* in the list for the base cluster. Then for any *URL* that belongs to the clusters other than the base cluster, the number of inter-cluster accesses is incremented by one. From the above example, the user session has the inter-cluster accesses of three. Thus the lower number of inter-cluster accesses, the better the clustering strategy.

Figure 1 shows the number of inter-cluster accesses for the *MMM*, *DFS*, *BFS* and the *random* clustering approaches. The comparison test was done by varying the cluster size from *1* to the maximum number of *URL* groups (i.e., *13*). The result shows that the clustering of the *URL*

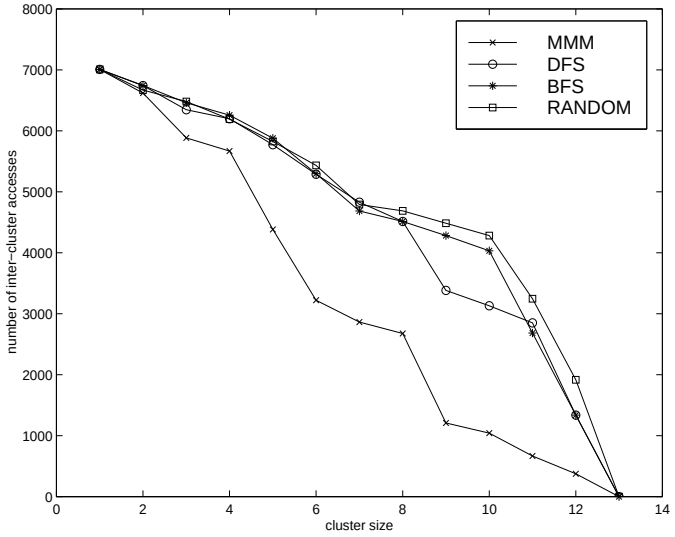

Figure 1. The clustering performance among the MMM, DFS, BFS, and the RANDOM clustering approaches.

groups constructed from our *MMM* mechanism and the proposed stochastic clustering strategy gives the best performance among all the approaches, while the *random* clustering strategy gives the worst performance, in most cases. *DFS* and *BFS* yield similar performance except for the cluster size of *9* and *10*. *DFS* and *BFS* methods yield relatively worse performance than the *MMM* approach since they are based solely on the static structure of the *URL* groups; while our *MMM* approach considers both the static structure and the user access patterns.

4 Conclusions

In this paper, the *Markov model mediator (MMM)* mechanism is presented to assist in organizing and managing the Web documents into disjoint clusters. The approach can provide a useful view of user access behavior on the different groups of *URLs*. An experiment was performed to compare the performance of our *MMM* mechanism with the *BFS*, *DFS*, and the *random* clustering approaches on the number of total inter-cluster accesses in the test data. The experimental result shows that the *MMM* mechanism performs better than the other three clustering approaches. Good clustering results can help manage the Web documents by reorganizing and customizing a company's Web site for its Web site layout so that the company can precisely target its potential customers to increase the company's profits. Hence, the proposed framework can be applied as one of the Web usage mining technique for Web site management in electronic commerce.

References

[1] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *Proceedings of the Int'l Conference on Data Engineering (ICDE)*, Taipei, Taiwan, March 1995.

[2] S.D. Bay, The UCI KDD Archive [http://kdd.ics.uci.edu]: Department of Information and Computer Science, University of California, Irvine, CA, 1999.

[3] M. Chen, J. Han, and P. Yu, "Data Mining: An Overview from Database Perspective," *IEEE Transactions on Knowledge and Data Engineering*, 8(6): pp. 866-883, December 1996.

[4] Y. Fu, K. Sandhu, and M. Shih, "Clustering of Web Users Based on Access Patterns," *International Workshop on Web Usage Analysis and User Profiling (WEBKDD99)*, San Diego, CA, 1999.

[5] J. Moore, et al., "Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering," *Proceedings of the 7th workshop on information technologies and systems*, 1997.

[6] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "A Probabilistic-Based Mechanism For Video Database Management Systems," *IEEE International Conference on Multimedia and Expo (ICME 2000)*, pp. 467-470, July 30-August 2, 2000, New York City, U.S.A.

[7] M.-L. Shyu, S.-C. Chen, and R. L. Kashyap, "Organizing a Network of Databases Using Probabilistic Reasoning," *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1990-1995, Nashville, Tennessee, USA, October 8-11, 2000.

[8] M.-L. Shyu, S.-C. Chen, and C.-M. Shu, "Affinity-Based Probabilistic Reasoning and Document Clustering on the WWW," *The 24th IEEE Computer Society International Computer Software and Applications Conference (COMPSAC)*, pp. 149-154, Taipei, Taiwan, October 25-27, 2000.

[9] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, Vol. 1, Issue 2, 2000.

[10] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration," *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pp. 46-53, 1998.