

**RECOGNIZING USAGE PATTERNS FROM JORDAN
UNIVERSITY WEBSITE USING SELF ORGANIZING MAP**

By

Dania Ahmad Mohammad Hlayyel

Supervisor

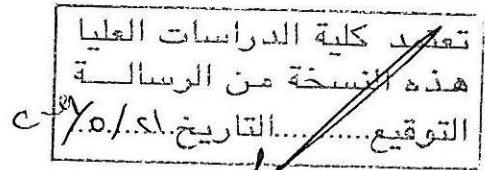
Dr. Ammar Al Huneiti

This Thesis was Submitted in Partial Fulfillment of the Requirements for the Master's
Degree of Computer Information system

Faculty of Graduate Studies

The University of Jordan

May, 2009



ا.م. / ٥ / ٢٠٠٩
د. عمار هنيدي

**The University of Jordan
Authorization Form**

I, Dania Ahmad Mohammad Hlayyel , authorize the University of Jordan to supply
copies of my Thesis/ Dissertation to libraries or establishments or individuals on request,
according to the University of Jordan regulations.

Signature: 

Date: 20/5/2009

COMMITTEE DECISION

This Thesis (Recognizing Usage Patterns From Jordan University Website Using Self Organizing Map) was Successfully Defended and Approved on ~~14/5/2009~~

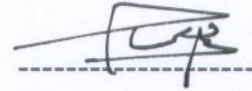
Examination Committee

Signature

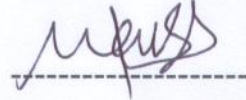
Dr. Ammar Huneiti, (Supervisor)
Assoc. Prof. of Hypermedia-Based Performance
Support Systems for the Web



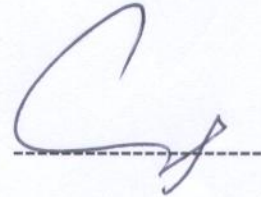
Dr. Amjad Hudaib (Member)
Assoc. Prof. of Software Engineering



Dr. Mousa AL-Akhras (Member)
Assoc. Prof. of Artificial Intelligence,
Artificial Neural Networks & Communications



Dr. Khalid A. Kaabneh (Member)
Assoc. Prof. of Image Processing
(Arab Academy For Banking and Financial Science)



تعتمد كلية الدراسات العليا
هذه النسخة من الرسالة
التوقيع.....التاريخ. ١٠/٥/٠٩

DEDICATION

*To whom lived in my heart
 Colored my life with love and beauty
 And gave it a great meaning
 Who spared no effort in help and support
 To the one who encouraged me through the duration of this thesis
 And stood all the time with me
 My husband Ziad*

*To the faithful wisdom and unmistakable foresight
 Who shed love generously to all my family
 Who stays close to me as stepping forward in the future
 My ever beloved parents*

*To those whom we grow together and learn
 Yet still hold one another close in mutual concern
 And face together, hand in hand
 The challenges life sends
 My brothers and Friends*

To all, I dedicate this work

ACKNOWLEDGMENT

This work would not have been completed in this way without the expertise, encouragement, patience, faith, and dedication of my Supervisor **Dr. Ammar Al Huneiti**.

I would like to thank my Supervisor **Dr. Ammar Al Huneiti** who guided me intellectually and spiritually and provided fruitful suggestions and sharp comments for the improvement of this thesis.

TABLE OF CONTENT

COMMITTEE DECISION	ii
DEDICATION	iii
ACKNOWLEDGMENT	iv
TABLE OF CONTENT	v
LIST OF FIGUERS	vii
LIST OF TABLES	viii
LIST OF ABBREVIATIONS.....	ix
ABSTRACT.....	x
1.1 INTRODUCTION.....	1
1.2 PROBLEM OVERVIEW.....	2
1.3 RESEARCH OBJECTIVES	3
1.4 MAIN CONTRIBUTION	4
1.5 THESIS ORGANIZATION.....	5
2. LITRATURE REVIEW.....	6
2.1 WEB DATA MINING	6
2.1.1 WEB CONTENT DATA MINING.....	8
2.1.2 WEB STRUCTURE MINING	9
2.1.3 WEB USAGE MINING	11
2.2 CLUSTERING ALGORITHMS AND TECHNIQUES	13
2.2.1 K-MEANS CLUSTERING TECHNIQUE.....	14
2.2.2 SOM TECHNIQUE.....	14
2.3 RELATED WORK OF CLUSTERING USING SOM.....	15
3. RECOGNIZING USAGE PATTERNS FROM JORDAN UNIVERSITY WEBSITE USING SELF ORGANIZING MAP	19
3.1 OVERVIEW.....	19
3.2 THE PROPOSED TECHNIQUE.....	20
3.2.1 DATA COLLECTION	21
3.2.2 DATA PREPROCESSING	24
3.2.2.1 LOG FILES CLEANSING.....	25
3.2.2.2 IDENTIFICATION OF VALID PAGES	29

3.2.2.3	USER IDENTIFICATION	30
3.2.2.4	CREATING USERS/PAGES RELATION.....	32
3.2.3	INITIAL CLUSTERING TO REDUCE DATA DIMENSIONALITY USING K-MEANS TECHNIQUE	33
3.2.4	SELF-ORGANISATION MAP OF USAGE PATTERNS.....	36
3.2.4.1	SOM TRAINING.....	39
3.2.5	PATTERN DISCOVERY AND ANALYSIS	39
3.2.5.1	Cluster Analysis.....	39
3.2.5.2	Pattern Discovery.....	40
4.	RESULTS AND ANALYSIS.....	41
4.1	EXPEREMENTIAL ENVIROMENT	41
4.2	MATLAB SCENARIOS.....	42
4.2.1	K-means Users Clustering Groups Scenarios.....	42
4.2.1.1	Two User Clusters Scenario.....	43
4.2.1.2	Three User Cluster Scenario	44
4.2.1.3	Four User Clusters Scenario	45
4.2.1.4	Five user clusters Scenario.....	46
4.2.1.5	Six user clusters Scenario	47
4.2.1.6	Seven User Clusters Scenario	48
4.3	RESULTS AND ANALYSIS	49
4.3.1	DISCUSSION OF RESULTS	51
4.3.1.1	Results of page clustering using SOM	52
5.	CONCLUSION AND FUTURE WORK	56
5.1	CONCLUSTION.....	56
5.2	FUTURE WORK.....	57
	REFERENCES	58
	APPENDICES	62
	ARABIC SUMMARY	77

LIST OF FIGUERS

<u>Figure</u>	<u>Page</u>
Figure 1: Co-occurrence of pages.....	5
Figure 2 : Web Mining Categories.....	7
Figure 3: High Level Web Usage Mining Process (Srivastava <i>et al.</i> 2000).....	12
Figure 4: Clustering process.....	14
Figure 5 : The block diagram of the general steps of the proposed technique	20
Figure 6 : A Sample of JU Web Server Log	22
Figure 7 : Users/Pages Matrix	32
Figure 8 : Represent page as vector of transaction groups	35
Figure 9 : SOM Input/output Layer.....	37
Figure 10: SOM Algorithm	38
Figure 11: Silhouette result of 2 user clusters	43
Figure 12: Silhouette result of 3 user clusters	44
Figure 13: Silhouette result of 4 user clusters	45
Figure 14: Silhouette result of 5 user clusters	46
Figure 15: Silhouette result of 6 user clusters	47
Figure 16: Silhouette result of 7 user clusters	48
Figure 17: Snapshot of Normalized Matrix.....	50
Figure 18: Result of SOM Matrix	50
Figure 19: Hierarchical Tree of selected JU website pages.....	51
Figure 20 : SOM Matrix Analysis.....	52
Figure 21: SOM Adjacent Cells (1, 2, 7, 8).....	53
Figure 22: SOM Adjacent Cells (13, 14, 19, 20, 25, 26, 31, 32).....	54
Figure 23: SOM Adjacent Cells (23, 24, 29, 30, 35, 36).....	55

LIST OF TABLES

<u>Table</u>	<u>Page</u>
Table 1: Basic Attributes of JU Server Log file	23
Table 2: Sample of Page Index.....	30
Table 3: Sample of User Index.....	31

LIST OF ABBREVIATIONS

1-D	One Dimension
2-D	Two Dimension
ANN	Artificial Neural Network
ASP	Active Server Pages
CSS	Cascading Style Sheets
DLL	Dynamic-link library
DOC	Document File
GIF	Graphics Interchange Format
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ID	Identification Number
IP	Internet Protocol
JPEG	Joint Photographic Experts Group
KD	Knowledge Discovery
NLP	Natural Language Processing
PDF	Portable Document Format
PPT	Power Point
SOM	Self Organizing Map
WAV	Waveform Audio Format
WWW	World Wide Web
XLS	Excel Sheet

RECOGNIZING USAGE PATTERNS FROM JORDAN UNIVERSITY WEBSITE USING SELF ORGANIZING MAP

By

Dania Ahmed Mohammad Hlayyel

Supervisor

Dr. Ammar Al Huneiti

ABSTRACT

Due to the rapid development and use of the Internet, the information content on the Web has become very rich and a problem of cognitive overload and hypermedia disorientation appeared. The huge amount of information presented on the web sites causes some difficulties in the search and information browsing over the internet. Accordingly, a need for a way or a technique to organize the layout and sequence of information as well as studying user navigations in order to discover the users interests has become particularly important. This thesis analyzes Jordan University JU website log data using an unsupervised learning technique and an Artificial Neural Networks ANN, namely self-organizing map (SOM) in order to recognize usage patterns left by user navigations. The work was over many procedures of preparing the data and creating a user/page relations as well as clustering of users and pages. Clustering was carried over two stage procedure, first an initial k-means clustering of users was applied to reduce the dimensionality of data, then SOM was used to cluster pages according to the users usage to provide a visual two dimensional map of user navigation. SOM was evaluated on a real web log file of one week from Jordan University Website; it shows that the resulting map was very meaningful and can easily produce meaningful visualizations of user navigation and detect usage patterns. It also shows the relationship between JU web pages based on the usage patterns of web users similar to themselves. Moreover, the result map can be used as an analysis tool for JU web masters to better understand the interests of visitors and the way users are browsing JU website.

1. INTRODUCTION

In this chapter, an overview of the problem is introduced. The research objectives and contribution are also presented. Finally, the thesis organization is shown.

1.1 INTRODUCTION

The exponential growth of the Web in terms of Web sites and their users during the last decade has generated huge amount of data related to the user's interactions with Web sites (Raju *et al.*, 2007). In addition, the websites continue to evolve, more and more information that are sometimes presented in a complex web designs. Therefore, a problem of cognitive overload and hypermedia disorientation appeared. The data result from user navigations over the web sites is recorded on web log files and is usually referred as web usage data.

The web usage pattern defines the common ways a user follows in browsing and navigating over a web site. Studying web usage data is a task of major importance, since this data can provide valuable information on analyzing and understanding users navigations and so how to better structure a Web site in order to create a more effective website for these users. The availability of huge web log files requires the development of tools to analyze usage patterns results from user navigations over the web site. Analyzing the user usage data and finding their common interest pages, documents, or links can help webmasters who are

responsible of the web site to find the most suitable and effective structure to their users and reduce their ambiguity.

Particularly, this thesis focuses on web usage mining, and applying data mining techniques over Jordan University (JU) web site in order to analyze and discover patterns left by user navigations. Kohonen's Self-Organizing Map (SOM) was used to create a map that allows the visualization of JU usage patterns.

1.2 PROBLEM OVERVIEW

“As more and more people facilitate web sites into their daily routines, it is important to create an interaction that is as easy as possible. As the web continues to evolve, more and more web site designs that may be complex and contains a huge amount of information a problem of cognitive overload appeared” (Janet *et al.*, 2006).

Unfortunately, the enormous size of hugely unstructured data on the web, has become a cause of ambiguity for consumers (Ahmed *et al.*, 2008). Moreover, a problem of hypermedia disorientation appeared when a designer build a website according to what s/he thinks is the best structure of links and presented information. However, the designer’s non clear view of what areas the users are interested in causes some difficulties for users to find their required information, since the designer structure may not be suitable for all users who are accessing the website. Therefore, designers need to find out their users interest in order to focus on them to improve user interaction with their website, this can be done by

discovering usage hidden patterns to organize the website structure according to user's preferences.

All of the above problems can be addressed using different techniques; one of these techniques is web usage mining, where an analysis of a website usage log file data may provide the web designers with patterns of users interest. The work in this thesis was based on analyzing the JU website (www.ju.edu.jo) access logs in order to find pages that are frequently accessed by users using SOM technique.

Analyzing the user usage data and finding their common interest pages, documents, or links can help webmasters or designers to find the most suitable and effective structure to their users. Jordan University website data would be very useful since different types of users are accessing the site on daily basis and they may have similar interests, objectives and needs.

1.3 RESEARCH OBJECTIVES

The objective of this thesis is to analyze Jordan University web log data and apply SOM technique in order to find overlapping profiles and discover hidden patterns, and finally apply page clustering to be effectively used for Web personalization and customization. Hidden patterns may be a frequent access from a group of users to a specific pages, the analysis of these patterns may led to the discovery of a relation between these frequent access pages. Such patterns could not be noticed without having some analysis over the web log files. Web log files contains

information about the browsing behavior of the web site visitors, particular the page navigation sequence (Velasquez *et al.*, 2003).

Moreover, the research aims to provide a clearer map of user navigations over JU web site by clustering users navigations into meaningful groups with shared interest, that can help webmasters or designers to get better idea of the users interests and therefore provide more suitable or customized services. This may provide the JU webmasters an insight into how to adapt the JU website according to the discovered usage patterns.

1.4 MAIN CONTRIBUTION

The main goal of this study is to capture, analyze, and model the behavioral patterns and profiles of users interacting with JU Web site. JU website needs to be analyzed according to the users usage as any other website in order to help JU webmasters to find the most suitable and effective structure to their users. Since JU website contains huge amount of information being presented for the users and different types of users access JU website in daily bases, it becomes important to find how users are accessing the website in order to find their interests and discover the relation between pages according to the user usage of the website. Studying users web usage provide a better understanding of the needs of the users and their browsing. Consequently this will allow better structuring of the website.

The study mainly was built on the assumption that different pages accessed frequently by different users may be similar or related. An example is given in Figure.1. The figure shows the navigational path followed by different users, where (U1) is accessing (P1) and (P4) over the web site, (U2) is accessing (P1) and (P4) as well. This shows that there is a strong relation between P1 and P4 especially if this pattern occurs frequently and this is called co-occurrence of pages or the concurrent user accessing of pages.

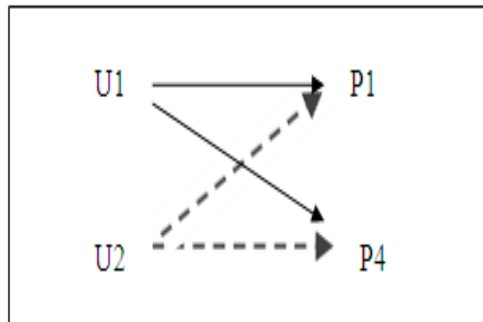


Figure 1: Co-occurrence of pages

1.5 THESIS ORGANIZATION

The rest of the thesis is organized as follows. In chapter 2, the related literature is reviewed; a summary of the previous works that are related to our research is presented. The background information of web mining and the previous works of web mining field are also presented. In chapter 3, the theory, the used technique, and the steps followed are presented. Chapter 4 summarizes the results of applying our technique over JU website log file data. Finally, conclusions and future work are presented in chapter 5.

2. LITRATURE REVIEW

In this chapter, summaries of some previous works that are related to the research are introduced. An overview of web mining and especially web usage mining is presented. Moreover, some of the web usage mining algorithms and as well as clustering techniques used in this field are discussed.

2.1 WEB DATA MINING

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information by extracting potential, unknown, and useful information and patterns from incomplete, noisy, or random data (Srivastava *et al.*, 2000). Data mining is part of Knowledge Discovery (KD) and can be used in many fields. KD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is mainly concerned about extracting patterns from specific data. Over the past few decades, data mining and more specifically web data mining has been extensively studied and many techniques were presented to improve the different types of web data mining.

Web mining as described in (Wel and Royackers, 2004) refers to the whole process of data mining and related techniques that are used to automatically discover and extract information from web documents and services. Web mining research focuses on applying data mining techniques to discover interesting

patterns of data from the Web (Zheng and Bouguettaya, 2009); it aims to discover useful information or knowledge from the Web hyperlink structure, page content and usage log.

Based on the primary kind of data used in the mining process, Web mining can be divided into three categories (Fernandez and Layos, 2003): Web content mining, Web structure mining and Web usage mining as shown in Figure 2.

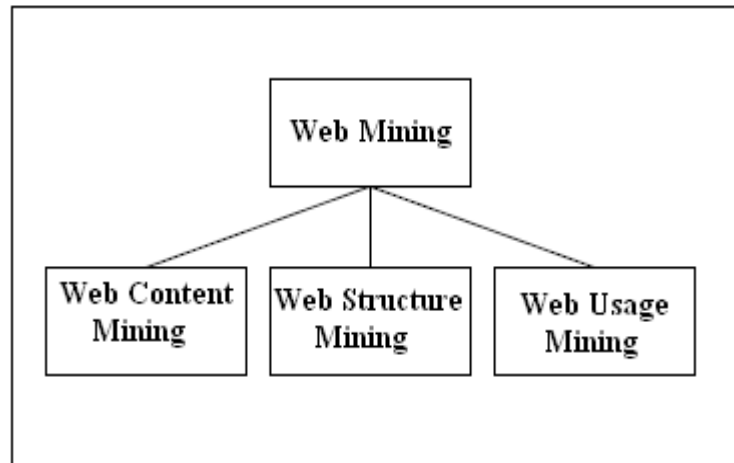


Figure 2 : Web Mining Categories

Web usage mining, which is our concern in this thesis, in general aims to capture, analyze, and model the behavioral patterns and profiles of users interacting with a Web site. Most of web usage mining information is usually gathered automatically by Web servers and collected in server access logs. Web Usage Mining as described in (Rossi *et al.*, 2005) is mainly about the analysis of the way a Web site is browsed by its users so as to improve it. Practical goals of the analysis of user usage data may include: improving the performances of Web

servers and the network; improving the structure of a site based on user typical navigation, building adaptive web sites and optimizing the information retrieval process.

The discovered patterns from mining user navigations are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests. This means that the use of data mining over the web helps in discovering useful information at different levels of interest.

Part of the researches concentrates on Web Content Mining as in (Fernandez and Layos, 2003), (Liu, 2005). Others were interested in the Web Structure Data Mining for example in (Perelomov *et al.*, 2002) and (Yang and Lee, 2006). Many researches also focused on Web Usage Mining as presented in (Britos *et al.*, 2007) and (Srivastava *et al.*, 2000). Artificial Neural Networks and particularly the Self Organizing Map (SOM) technique was used to mine the web and extract usage patterns such as (Vesanto and Alhoniemi, 2000) and (Smith and Ng, 2003). Some of these researches will be discussed later in this chapter.

2.1.1 WEB CONTENT DATA MINING

The Web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. Web Content Mining describes the process of information discovery from millions of sources across the World Wide Web (WWW) (Fernandez and Layos, 2003). Web content mining is mining the content data in a site which is the collection of objects and relationships that are conveyed

to the user. As described in (Liu, 2005) Web content mining is mainly about: mining, extraction and integration of useful data, information and knowledge from Web page contents. Some of the previous researchers were basically interested in the web content mining as they consider it as a knowledge discovery task to find useful information within the web pages and its content. Some of the web content researches depend on a sensitivity mining approach where it consider the sensitivity of the web page instead of the traditional methods that filter the page content according to the keywords that have been mentioned in that page. Since many people may use the internet to disseminate violence or even spread rumors it is an important task for the Internet regulators to identify the sensitive pages and trace the corresponding IP addresses, and even block the IPs when necessary as done in (Wang *et al.*, 2008).

Other web content researches depend on some Natural Language Processing (NLP) techniques and use some pattern mining algorithms to discover knowledge based on the detailed meanings of the text where traditional techniques do not give knowledge about text semantics it only find group of keywords without getting the meaning as presented in (Jiang *et al.*,2007).

2.1.2 WEB STRUCTURE MINING

Web structure mining aims to discover knowledge hidden in the structures linking web pages where the structure of the data represents the designer's view of the content organization within the site. The target of Web structure mining is to tend the link structure of Web documents which reveals the personalized information

contained in the document structure. The type of data the web structure mining processes is the structured data of the web pages in a website (Zhang and Yin, 2008).

The effectiveness of using the Self Organizing Map in web structure data mining to structure some news websites is presented in (Perelomov *et al.*, 2002), it describes a South-East Asian centric system that integrates and organizes news articles from English news websites by the use of SOM. SOM provides such system with an automatic clustering of news documents into groups of related news articles from different websites. In addition, Automatic organization of news clusters in a 2D theme map, automatic extraction of meaningful labels for each cluster of news articles, and automatic generation of links to related news articles.

A machine learning approach was used to automatically construct a navigational structure for help users in information finding over the World Wide Web as done in (Yang and Lee, 2006) where an approach named NaviSOM was presented since they depend on SOM in their work. This approach depends on two maps of text mining namely document cluster map, and keyword cluster map. A hierarchical structure of web page clusters is then constructed. It basically applies a machine learning algorithm on a web pages to identify the topics and discover the relations among them.

2.1.3 WEB USAGE MINING

Web Usage Mining refers to the discovery and analysis of patterns in click stream and associated data collected as a result of user interactions with Web resources on one or more Web sites (Dai and Mobasher, 2003). The goal of web usage mining is to capture and model web user behavioral patterns and analyses the usage patterns of web sites in order to get an improved understanding of the users' interests and requirements

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications (Srivastava *et al.*, 2000). The analysis of server web access logs and user registration data can also provide valuable information on how to better structure a Web site in order to create a more effective, customized Web site.

Web usage mining depends mainly on capturing data from the server log files of a website. Unsupervised techniques like clustering are used in order to place data elements into related groups without advance knowledge of the group definitions according to the application requirements.

Web server log files are the primary data sources used in Web usage mining, which includes web server access logs and application server logs. A web log file contains information about the accesses of all visitors to a particular Web site (Velasquez *et al.*, 2003). These log files as mentioned in (Rossi *et al.*, 2005) consist of the list of all HTTP requests received by a web server with some

description of these requests. Interesting information contained in the log file include the IP address of the computer sending the request, the requested document, the date of the request and the User Agent that sent the request. Additional data sources that are also essential for both data preparation and pattern discovery include the site files and pages (HTM, HTML, etc.).

Web usage mining process consists of three main phases: (i) preprocessing, (ii) pattern discovery, and (iii) pattern analysis. Figure 3 shows these three main tasks for performing Web Usage Mining or Web Usage Analysis as described in (Srivastava *et al.*, 2000).

Preprocessing phase consists of converting the usage information contained in the various available data sources a data abstractions necessary for pattern discovery and to identify meaningful representations. Pattern Discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. In this stage the hidden useful patterns caused by user usage of a website can be discovered. Pattern analysis is the last step in the overall Web Usage mining process as shown in Figure 3.

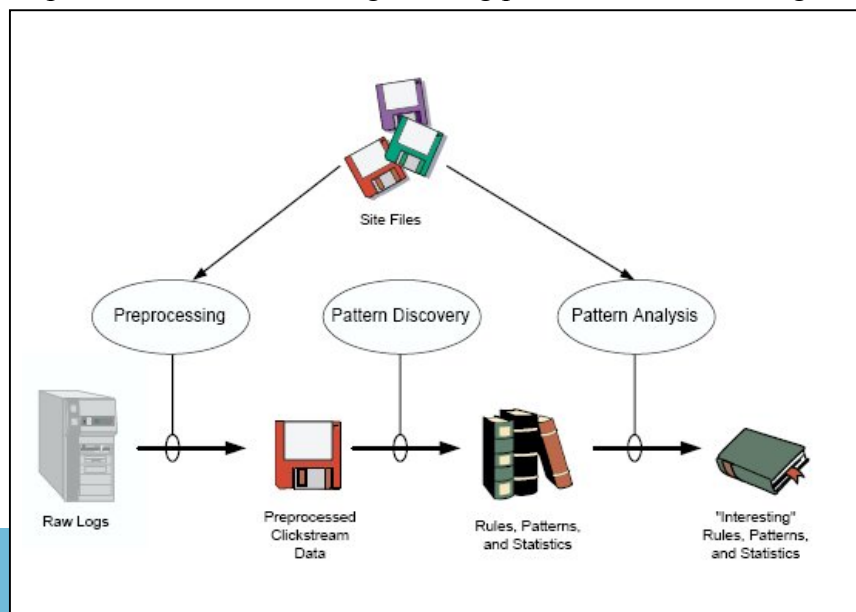


Figure 3: High Level Web Usage Mining Process (Srivastava *et al.* 2000)

2.2 CLUSTERING ALGORITHMS AND TECHNIQUES

Clustering is a division of data into groups of similar objects. It is an approach to identify natural groupings of similar entries in such sets of unclassified data often without any a priori knowledge as to what that similarity may involve (Berkhin, 2003). The principal idea is to partition the dataset into meaningful sub-classes, called clusters (Schatzmann, 2003). Accordingly, clustering can provide a helpful first impression of the way the data is distributed and it is an important stage in the data mining process.

One of the most successful algorithms for unsupervised learning is the self-organizing map (SOM) neural network developed by Kohonen. Part of the reasons of SOM success is the visual nature that it has. Much work has been done in many researches to create better visualizations for trained SOM networks.

Clustering as described in (Berkhin, 2003) is the division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. It represents many data objects by few clusters, and hence, it models data by its clusters. Clustering is often called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning. Many algorithms can be used in the clustering process but all clustering algorithms finally attempts to find natural groups of components, based on some similarity. The algorithm applied on a raw data and results in clear separate clusters from this data as shown in Figure 4.

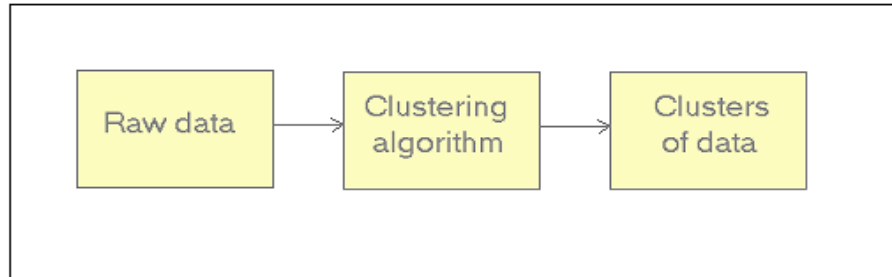


Figure 4: Clustering process

Many clustering techniques were developed and used in data mining field and especially in page clustering. In this thesis we are concerned in two clustering algorithms K-Means and Self Organizing Map (SOM) algorithms. Both algorithms will be discussed in brief in this section and more details will be given in the next chapters while presenting our work.

2.2.1 K-MEANS CLUSTERING TECHNIQUE

One of the most traditional clustering methods is the k-means. Its objective is to build a group of k groups starting from a mass of data, so that each unit belongs to just one group. It is a nonhierarchical method, so it does not make a linkage between the grouped data (Sperandio and Coelho, 2003). K-means is a partitional clustering algorithm. It is an algorithm that features quick clustering and easy operation, and is applied to the cluster analysis of several data such as Texts, images and others (Xinwu, 2008).

2.2.2 SOM TECHNIQUE

SOM has proven to be one of the most powerful algorithms in data visualization and exploration (Alhoniemi *et al.*, 2003). In addition, (SOM) is a fairly well-known neural network or ANN and indeed one of the most popular unsupervised

learning algorithms. SOM is a neural clustering technique. It is more sophisticated than K-means in terms of presentation; it does not only cluster the data points into groups, but also presents the relationship between the clusters in a two-dimensional space. SOM is also capable of presenting the data points in one, two or three-dimensional space. However, two dimensional spaces are most commonly used due to the trade-off between information content and ease of visualization. SOM also presents the degree of special autocorrelation where the degree of neighboring cluster features shares similar characteristics.

2.3 RELATED WORK OF CLUSTERING USING SOM

In (Velasquez *et al.*,2003) Velasquez et al. studied the effectiveness of combining web usage and content mining approaches in order to analysis the visitor behavior and show relevant information for the visitor to capture her/his attention. However, they extracted some useful variables from the web log files and the web site itself, using web usage and content mining. From all the possibly available data, three variables were analyzed: content, navigation sequence and time spent in each page visited. They suggested that combining these variables may provide some similarity between users according to their time sessions. Web content was filtered to represent the document or the web page by vector space model of words. Self Organizing Map (SOM) technique was used then in order to mine the visitor behavior vector and get knowledge about the visitor preferences.

Vesanto and Alhoniemi in (Vesanto and Alhoniemi, 2000) discussed the effectiveness of using SOM algorithm as a first step of data clustering and visualization. Since the visualization can only be used to obtain qualitative

information, they used hierarchical agglomerative clustering and partitive clustering using K-means in order to produce summaries—quantitative descriptions of data properties—interesting groups of map units was selected from the SOM by using such algorithms. Their goal was not to find an optimal clustering for the data but to get good insight into the cluster structure of the data for data mining purposes. Thus, the clustering method should be fast, robust, and visually efficient. Therefore, data was firstly clustered by SOM. The most important benefit of this procedure is that computational load decreases considerably, making it possible to cluster large data sets and to consider several different preprocessing strategies in a limited time.

SOM was also used in (Britos *et al.*, 2007) to identify the user's patterns. This research also compares the use of K-means algorithm in the process of web usage mining and the use of SOM algorithm by comparing two sites one of music and another of gastronomic. Web server log files were prepared and both methods, SOM and K-Means, were applied. Finally, they conclude that to identify common patterns in Web, the use of self-organized map (SOM) was better than K-Means in their work.

Mainly the research that is most relevant to our study is presented in (Smith and Ng, 2003) that developed a system (LOGSOM SYSTEM) that uses Kohonen's self-organizing map (SOM) to organize web pages into a two-dimensional map. The web page organization depends on the user navigation and usage of the web site and not on the content of the page. The system results in a map that can be

used to analyze user navigation visually which helped them in better understanding the behavior of the user using their website.

It was suggested that even if it is useful to have a system to organize the web pages in a content-driven manner; it may be more advantageous to organize the web pages in a web-user oriented manner. The LOGSOM was evaluated on a sample of data: the log files for 1 month of access to the School of Business Systems web server at Monash University. The web server logs were prepared and cleaned before applying the algorithm over it. The K-means algorithm was applied to reduce dimensionality of the huge data then the SOM was applied to visualize the clusters in 2- dimensional map. This research shows that the resulting map of this system is very meaningful and can be easily incorporated with a web browser to assist user navigation. The system also provides a visual tool to enable users to see the relationship between web pages based on the usage patterns of web users similar to themselves.

A validation scheme between the k-means statistical method and the SOM was developed in (Sperandio and Coelho, 2003). They exploit the knowledge discovery properties of the SOM for determine a good k-means clusters number estimation, and then visualize the k-means clusters over a trained map to compare and validate the result of both procedures, where data was clustered using SOM in order to determine the number of clusters that should be used in k-means and train it to finally find the best number of clusters to be used by k-means.

In (Merelo-Guervós *et al.*, 2004) SOM was used to create a community map which allows the visualization of community standing and relationship, and it can be used to discover which members of the community have similar interests.

A New Web Usage Mining and Visualization Tool was proposed in (Labroche *et al.*, 2007) which was based on a graphical representation of user activity on the web site. The graph contains user sessions that are associated to paths and called “web paths”. The web pages are the nodes of the graph and hyperlinks are the edges between the nodes in a directed graph. This tool relies on an Ant algorithm. This tool shows that it can easily produce meaningful visualization of user navigations; which was ensured after applying the tool on a real web log file from the French museum of Bourges website.

(Yu *et al.*, 2008) presented a method based on 1-D SOM to cluster the document collections. It was based on the one-dimensional array of Self-Organizing Map network (1-D SOM array). The main idea of this method is to obtain the clustering results by calculating the distances between every two adjacent most similar prototype to the input vector of well trained 1-D SOM. The benefit of this procedure is that in 2-D SOM, the clusters are distributed in a 2-D area and every cluster has more than 2 neighbors, the boundary lines among clusters are curves. Comparatively, the clusters in 1-D SOM are distributed in a line and every cluster has at most two neighbors, and the boundary between two clusters is a point. It is apparent 1-D SOM is much easier than that in 2D. However, 2-D may give more accurate clustering especially for large set of data.

3. RECOGNIZING USAGE PATTERNS FROM JORDAN UNIVERSITY WEBSITE USING SELF ORGANIZING MAP

In this chapter, an overview of the set of used techniques in the research will be presented. In addition, a complete discussion about the steps followed in this research will be outlined.

3.1 OVERVIEW

Analyzing the user web usage data and finding their common interest pages, documents, or links can help webmasters or web designers to find the most suitable and effective website structure to their users. Jordan University website is one of the websites that is being accessed by different types of users on a daily basis. JU website log data is very useful since it contains a huge amount of recourses, and information browsed by users of similar interests, objectives and needs.

Our proposed technique of using SOM aims to cluster JU web site pages according to the user usage of the web site in order to discover hidden usage patterns that may be useful for webmasters to better organize the website and help users to get to their required information more easily. The primary objective of this research will be on webpage clustering in order to separate unrelated pages and cluster related pages into meaningful groups according to their usage.

This research is mainly built on the assumption that different pages accessed frequently by the same users must be similar or related. This research aims to find the relations between different pages accessed by different users (Co-occurrence of pages), if there are some pages that are accessed frequently by the same users, this gives a sign that these pages are related, and accordingly the usage pattern relations between pages could be found.

3.2 THE PROPOSED TECHNIQUE

Our technique consists mainly of the following phases:

1. Data collection phase
2. Data pre-processing
3. Web page clustering
 - i. Building User/Pages Relation
 - ii. Initial Clustering to Reduce Data Dimensionality
 - iii. Page Clustering Using SOM
4. Pattern discovery and analysis

The block diagram of the general steps of our technique is shown in Figure 5.

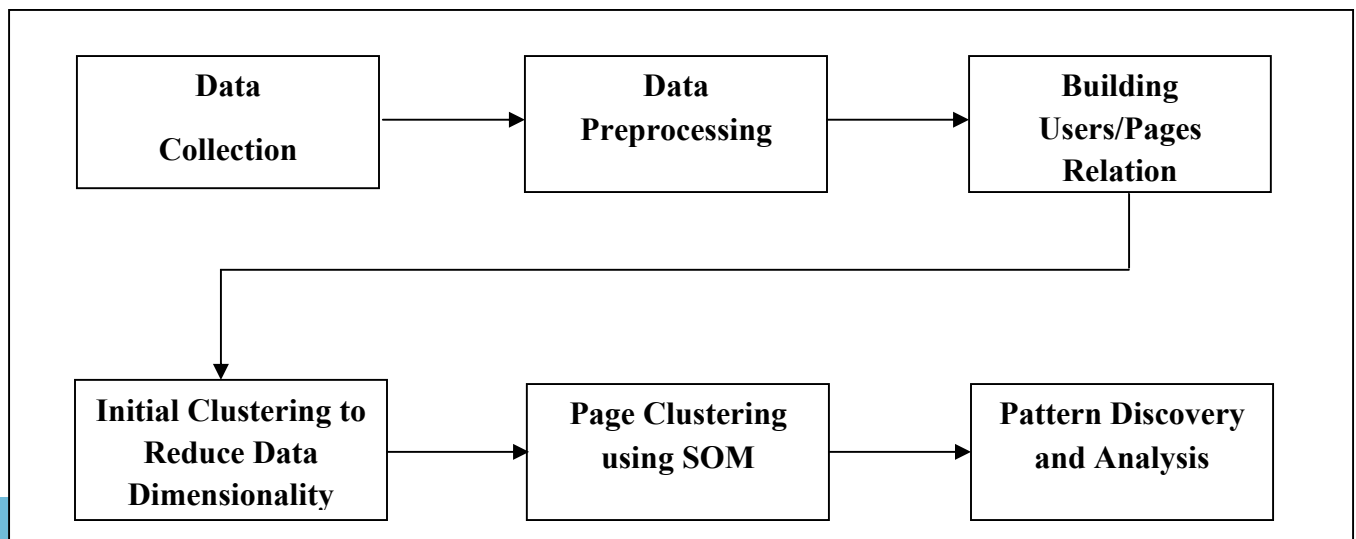


Figure 5 : The block diagram of the general steps of the proposed technique

3.2.1 DATA COLLECTION

The required data for this study were mainly collected from Jordan University Computer Center of the JU website (www.ju.edu.jo). The JU Computer Center has servers that contain all the JU Web Log Files that recorded internal and external accesses for the website. However, external accesses to the website do not have a unique IP addresses since all external accesses are given the same IP. Therefore, the concentration of this research was on analyzing the internal use of the JU website rather than analyzing the usage of external user's visits. A data of at least two weeks log from the JU web server log files were requested as a required data for this study.

The log files that were collected during the month of February 2009 where it was a start of a new semester, which means having different types of users (Students, Professors, and staff) accessing the website. Moreover; at February the computer center officially verbalized the service of adding a unique user name for each student or user accessing JU website. This operation helped us to easily identify users accessing the web site. The data we used was from 9/2/2009 till to 17/2/2009 or around one week of February.

As mentioned earlier, each log file contains information about all requests made on the server which may be a hit on JU website or on any other website over the internet. The JU log files records every click stream that occurs within JU servers. Basically the JU log files contain the following: (IP Address, User Name, Date &

Time, HTTP Request, Status Code, and Transfer Volume). A sample of such log file is shown in Figure 6. For more samples see Appendix A.

```

10.249.80.86 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://naim-22522/ HTTP/1.0" 407 0
10.249.80.76 - t.antary@ju.edu.jo [10/Feb/2009:08:57:59 +0000] "GET
http://www.google.com/supported_domains HTTP/1.0" 407 0
10.249.104.59 - rasha.oirp [10/Feb/2009:08:57:59 +0000] "GET http://tv.mehalla.net// HTTP/1.0" 200
34625 ALLOW "N2H2"
172.28.9.88 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://juhjk-4298b673e/ HTTP/1.1" 407 0
10.249.104.237 - l.jungae [10/Feb/2009:08:57:59 +0000] "GET
http://static.naver.com/mail4/0627_mail/btn_confirm.gif HTTP/1.0" 200 936 ALLOW "N2H2"
10.249.104.237 - l.jungae [10/Feb/2009:08:57:59 +0000] "GET
http://static.naver.com/mail4/1213/btn_st_mywrite_0223.gif HTTP/1.0" 200 1338 ALLOW "N2H2"
10.249.104.237 - l.jungae [10/Feb/2009:08:57:59 +0000] "GET
http://static.naver.com/mail4/ico_lt_mail2.gif HTTP/1.0" 200 211 ALLOW "N2H2"
172.28.9.191 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://medsh-2e082cec9/ HTTP/1.0" 407 0
10.249.24.213 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://alo2a/ HTTP/1.0" 407 0
10.249.88.47 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://jserver/ HTTP/1.0" 407 0
10.249.112.36 - wejdan.b [10/Feb/2009:08:57:59 +0000] "POST
http://www.elearning.jo/eduwave/elearningme.aspx HTTP/1.1" 200 24840 ALLOW "N2H2"
10.249.80.76 - t.antary@ju.edu.jo [10/Feb/2009:08:57:59 +0000] "GET
http://www.google.com/supported_domains HTTP/1.0" 407 0
10.249.80.86 - - [10/Feb/2009:08:57:59 +0000] "OPTIONS http://naim-22522/ HTTP/1.0" 407 0
172.28.9.173 - W.Khoury [10/Feb/2009:08:58:00 +0000] "GET
http://dc124.4shared.com/servlet/ProgressStatus?dcId=124&slId=dVV4T8FIHn7TDrPj&globSysLang=en
&random=0.006929787850189106 HTTP/1.0" 200 221 ALLOW "N2H2"

```

Figure 6 : A Sample of JU Web Server Log

In Table 1, the basic attributes of the JU server log file are presented with a sample data.

Table 1: Basic Attributes of JU Server Log file

IP Address	User Name	Date & Time	HTTP Request	Status Code	Transfer Volume
10.248.200.52	saj0077708	09/Feb/2009: 12:59:20 +0000	GET http://www.ju.edu.jo/faculties/facultyoflaw/sysimage/footer.jpg HTTP/1.0	200	14168 ALLOW "N2H2"
10.248.105.137	hbh0063873	09/Feb/2009: 12:59:20 +0000	GET http://blackboard.ju.edu.jo/javascript/validateForm.js HTTP/1.0	200	39272 ALLOW "N2H2"
10.249.104.243	--	14/Feb/2009: 19:58:41 +0000	GET http://www.ju.edu.jo/Home.aspx HTTP/1.0	407	0
10.248.168.51	ala0084496	14/Feb/2009: 19:43:53 +0000	GET http://gfx8.hotmail.com/mail/13.2.0260.1209/styles/Base/img/_12.jpg HTTP/1.0	200	9989 ALLOW "N2H2"
10.249.104.243	m.albaker	14/Feb/2009: 19:58:43 +0000	GET http://www.ju.edu.jo/Home.aspx HTTP/1.0	200	73331 ALLOW "N2H2"
10.249.80.76	tantary@ju.edu.jo	10/Feb/2009: 08:57:59 +0000	GET http://www.google.com/supported_domains HTTP/1.0	407	0
10.249.112.36	wejdani.b	10/Feb/2009: 08:57:59 +0000	POST http://www.elearning.jo/eduwave/elearningme.aspx HTTP/1.1	200	24840 ALLOW "N2H2"
10.248.200.58	sad8070895	17/Feb/2009: 17:42:16 +0000	GET http://www.maktoob.com/ HTTP/1.0	200	68927 ALLOW "N2H2"
10.248.113.5	ahm0086469	11/Feb/2009: 12:46:59 +0000	GET http://www.ju.edu.jo/_catalogs/masterpage/ARA/WebFiles/image/s/.gif HTTP/1.0	200	49 ALLOW "N2H2"

The collected data were a raw data, it was not suitable to apply the work over it directly since it contains many records and attributes that were beyond the scop of this study and not of the research interest. The data needed to be pre-processed first. Accordingly; some changes were needed in order to make data more suitable for this study.

As mentioned earlier the log files contain all user transactions (internal and external usage of the servers) and since our study is only concerned on the JU website data mining we do not need all the hits that the log files recorded. In addition, the data format of the log files needed also some changes since it contains some attributes that are not important and beyond the scope of this research.

3.2.2 DATA PREPROCESSING

Data pre-possessing is a critical step in the identification of user's patterns in web sites. It is usually the most time consuming phase in the web usage mining process. Since not all data being collected may be useful for the work, it is important to do some pre-processing on it in order to create a clean data set to apply data mining techniques. For example, many of the user hits may be on a page that contains an image or a flash animation or document or even videos. In these cases these resources are not necessary for the detection of user's patterns. Accordingly, log files needed to be cleaned from such recourses.

The following steps were followed in order to clean log files and prepare data:

1. Log Files Cleansing
 - i. Attributes Reduction
 - ii. Records Reduction
2. Identification of Valid Pages
3. Users Identification
4. Creating User/pages Relation

3.2.2.1 LOG FILES CLEANSING

1) Attributes Reduction

As mentioned earlier collected data were a raw data, it contains some attributes that is not needed in this study. Mainly the log files contain the following attributes:

- i. **IP Address:** Remote host that is being used by the user to access the web.
- ii. **User Name:** Three types of authenticated users were classified from the log files:
 - a) Students: the student's user name contains the first 3 characters of his/her first name followed by his/her university number, example: saj0077708

- b) Staff : both Professors and JU staff user names are the same as their JU email address, example: j.manaseer
 - c) General Users: the log files contain also some general users such as stu, reg, theses...etc. Such user names can be used by a group of users.
- iii. **Date And time:** date and time when the user accessed the web and requested a page, example: 09/Feb/2009:12:59:20 +0000
 - iv. **HTTP Request:** this field contains the http request that the user asked, example : GET http://www.ju.edu.jo/Home.aspx HTTP/1.0
 - v. **Status Code and Transfer Volume:** this field is concerned with the network response of the http request, example : 14168
ALLOW "N2H2" 200 or 400

Accordingly set of data that are applicable for this research needed to be prepared. Irrelevant file attributes should be deleted. Some of the log file attributes are not of our interest, so we delete the irrelevant attributes to leave only the following attributes:

- a. **IP address:** Remote host used to access the web.
- b. **User Name:** As mentioned before the computer center in JU have added a unique user name for each student or user accessing the JU website this helped us to easily identify users accessing the web site.

- c. **Date/Time:** The date and time the user made the HTTP request.
- d. **HTTP request:** The web page the user accessed.

2) Records Reduction

All users transactions that includes the JU website hits or any other website hit are recorded in the JU server log files, this means it may contain many of unnecessary records. Moreover; many of the JU hits may be for pages that have an image, documents, videos...etc. These resources are not necessary for the detection of usage patterns and were deleted from the log files.

A. Delete and discard all external and non JU website pages.

At this stage discard all external and non JU web site entries to reduce log records to only keep links that are related to the JU website as following URLs:

- www.ju.edu.jo
- www1.ju.edu.jo
- portal.ju.edu.jo
- reg.ju.edu.jo
- blackboard.ju.edu.jo
- acad.ju.edu.jo
- mail.ju.edu.jo

B. Remove redundant pages

In our work we need to define a clear set of pages to apply the study over them and since there may be more than one user requests to the same page some redundant pages were found within the log files, or even pages that have the same path but is written in capital letters and other time in a small letters. Accordingly, any redundant pages were removed to only have unique pages.

C. Remove non available pages

Sometimes a user may write the wrong path for a page, which means that the requested page is not available. Such pages are recorded in the log file as a normal request even that it is an error entry. Accordingly, this study is not interested in these error entries, which we have therefore deleted.

D. Remove some file extensions

As mentioned earlier many of the JU hits may be for pages that have an image, documents, videos...etc. These resources may not necessary for the detection of usage patterns; for this reason these records were deleted and from the log file as follows:

i. Remove image pages.

At this stage we search within the log files for an HTTP request that reach an image. Here a search for the image extensions within the HTTP request Path was done such as (.GIF, .JPEG, .JPG, .PNG) and

once such extensions are found they were deleted from the web log file.

- ii. Remove non significant extensions.

After removing image extensions, there still other extensions that were not significant or even are not considered as a page such as (.WAV, .ZIP, .CSS, .XLS, .DLL ...etc). But after removing the mentioned extensions the number of pages that remains still very big and to decrease them also (.DOC, .PDF, .PPT) files extensions were removed.

- iii. Keep Only (HTML, HTM, and ASPX).

After all of the above steps of filtrations we only kept files with the extension of (.HTML, .HTM, .ASPX) to be considered as a page hit over JU website.

3.2.2.2 IDENTIFICATION OF VALID PAGES

To identify pages we create an index of valid pages which were filtered to only 97 pages, this index contains the page ID, page description, page path. Sample of the page index is shown in Table 2. The full page index can be found in appendix B.

Table 2: Sample of Page Index

Page ID	Description	HTTP Request
1001	Arabic JU home	http://www.ju.edu.jo/arabichome
1002	JU Map	http://www.ju.edu.jo/UJ%20Map/UJMap.html
1003	Calendar Of JU	http://www.ju.edu.jo/calendar/index.html
1004	Announcements of JU	http://www.ju.edu.jo/announcements/uac/default.htm
1005	Contact us page	http://www1.ju.edu.jo/ContactUs.html
1006	Blackboard Page	http://blackboard.ju.edu.jo/
1007	JU News And Events	http://www.ju.edu.jo/Lists/NewsAndEvents/
1008	Student Information For Acad	http://acad.ju.edu.jo/
1009	E-Courses Page	http://www1.ju.edu.jo/e-courses/default.htm
1010	Portal Page	http://portals.ju.edu.jo/
1011	Regulation Documents	http://www.ju.edu.jo/Pages/Regulations/
1012	JU Documents Page	http://www.ju.edu.jo/documents
1013	JU Hospital Home page	http://www.ju.edu.jo/medical/hospital
1014	Providant Fund- Administration Page	http://www1.ju.edu.jo/providant-fund/administration.html
1015	Faculty of graduate studies " Arabic	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies
1016	Agreements JU Forms	http://www.ju.edu.jo/Agreements%20and%20MOUs/Forms
1017	Tenders Home page	http://www.ju.edu.jo/tenders
1018	Announcement on January	http://www1.ju.edu.jo/announcements/2008%20January/adv%2007-2-2008_3.htm
1019	Announcement on March about Jobs	http://www1.ju.edu.jo/announcements/2008%20March/jop.htm
1020	Center of Consultation Page	http://www.ju.edu.jo/centers/coc

3.2.2.3 USER IDENTIFICATION

To identify users accessing the JU website an index of unique users was also created. As mentioned before the computer center verbalized the service of adding a unique user name for each user accessing the JU website. This operation helps us to easily identify users accessing the web site. This index contains a list of

4377 users of all types (Students, Staff, and General users) each with his/her own ID created. Sample of user index is shown in Table 3. Larger sample is shown in Appendix C.

Table 3: Sample of User Index

User ID	User Name
10001	a.alhnity
10002	hbh0063873
10003	Amjadq
10004	-
10005	r.hamed
10006	m.yacoub
10007	Stu
10008	sos0075813
10009	Ayeshg
10010	abr0074296
10011	Weshah
10012	s.habahbeh
10013	s.abuhazeem
10014	mjd0085292
10015	Makash
10016	sam0057543
10017	any2040207
10018	rgd0059178
10019	e.domour
10020	kma2070179

3.2.2.4 CREATING USERS/PAGES RELATION

At this stage a matrix of users and visited pages was created using the MATLAB. This matrix shows all the hits that were made by the 4377 user to the 97 page a shown in Figure 7. This matrix only represents the visit or no visit of the user to a specific page, where “0” indicates a non visit and “1” indicates a visit to the page.

In this matrix the page vector for each 97 page was very long since we were dealing with 4377 user. Thus, k-means technique was used to reduce the dimensionality and decrease the page vector of the 97 page. Reducing dimensionality becomes very important, since the collected data were very huge and had long vectors where there are too many users and too many pages and the matrix will contain many zeros which mean we also will have sparse data. If $P = \{P1, P2, P3, \dots, P97\}$ and a set of 4377 users: $U = \{U1, U2, U3 \dots U4377\}$ and represent the user visit to a page by (1) and a non visit by (0) the result matrix will be as shown in Figure 7.

		Pages				
		P1	P2	P3	...	P97
users	U1	1	0	1		1
	U2	1	1	0		0
	U3	0	0	1		0
	.					
	.					
	.					
	U4377	1	1	0		0

1 - Visit
0 - No Visit

Figure 7 : Users/Pages Matrix

3.2.3 INITIAL CLUSTERING TO REDUCE DATA DIMENSIONALITY USING K-MEANS TECHNIQUE

K-means algorithm was used as one of the most traditional clustering methods, whose objective is to build a group of k groups starting from a mass of data, so that each unit belongs to just one group to reduce the dimensionality of data vector. Reducing dimensionality will make data small enough to apply SOM and large enough to have meaningful data that can provide valuable outcome.

Instead of having unique users to describe pages usage pages can be described by group of users. By using the K-means clustering algorithm, users will be clustered into groups. Then the pages will be described by the number of users visited any page from every user group. Basically K-Means algorithm is described as follows:

- A. Choose the number of k clusters
- B. Randomly assign items to the k clusters
- C. Calculate new centroid for each of the k clusters
- D. Calculate the distance of all items to the k centroids
- E. Assign items to closest centroid.
- F. Repeat until clusters assignments are stable

The K-means algorithm steps as described in [Smith and Ng, 2003]:

- i. Choose K initial cluster center (where is represented as k transaction groups) randomly from the center hypercube.
- ii. Assign all data points (representing transactions) to their closest cluster (measuring from cluster center). This is done by presenting a data point x and calculate the similarity (distance) d of this input to the weights w of each cluster center j . The closest center to a data point x is the cluster center with minimum distance to the data point x .

$$d_j = \|x - w_j\| = \sqrt{\sum_{i=1}^n (x_i - w_{ij})^2} \dots\dots\dots (1)$$

- iii. Recalculate the center of each cluster as the centroid (or the mean vector) of all the data in each cluster. The centroid C is calculated as follows :

$$\vec{c} = \langle w_1^c, w_2^c, \dots, w_n^c \rangle \dots\dots\dots (2)$$

Where

$$w_i^c = \frac{\sum_{j \in c} u_j^i}{N^c} \dots\dots\dots (3)$$

Where N^c is the number of data points in the cluster

$$u_i^t = \begin{cases} 1, & \text{if } \text{url}_i \in t \\ 0, & \text{otherwise} \end{cases} \dots\dots\dots (4)$$

- iv. If the new centers are different from the previous ones, repeat step 2,3 and 4. otherwise terminate the algorithm.

By using the K-means clustering algorithm, JU user transactions were clustered into five groups as shown in Figure 8. K-means cluster the data in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized.

		Pages				
		P1	P2	P3	...	P97
User Groups	UG1	0	4	0		0
	UG2	0	3	1		5
	UG3	1	21	0		0
	UG4	0	3	0		0
	UG5	14	0	0		3

Total Visits of group 2 to page 97

Figure 8 : Represent page as vector of transaction groups

The number of clusters $K=5$ was chosen after applying the silhouette to get an idea of how well-separated the resulting clusters are. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring

clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster.

It is defined as:

$$S(i) = (\min(b(i,:),2) - a(i)) ./ \max(a(i), \min(b(i,:),2)) \dots \dots \dots (5)$$

$a(i)$ is the average distance from the i th point to the other points in its cluster, and $b(i,k)$ is the average distance from the i th point to points in another cluster k . The distance function that was used in silhouette testing was 'Hamming'. Hamming is a distance function that is only suitable for binary data. In hamming distance function each centroid is the median of points in that cluster.

3.2.4 SELF-ORGANISATION MAP OF USAGE PATTERNS

Kohonen's SOM was used in this work as an unsupervised clustering algorithm. The SOM is a neural network method that produces a similarity map of input data. The maps comprehensively visualize natural groupings and relationships in the data. SOM has only two layers (Input and Output) where SOM does not have any hidden layers. Each neuron in the input layer receives a component of input vector. Its neurons are arranged in various forms, e.g. one-dimensional linear or two-dimensional planar array matrix (Yu *et al.*, 2008).

The Inputs for SOM consist of a set of pages presented as vectors of user groups as shown in Figure 9, according to the usage pattern of transaction groups classified by k-means algorithm. The desired output is a two-dimensional map or matrix of M nodes.

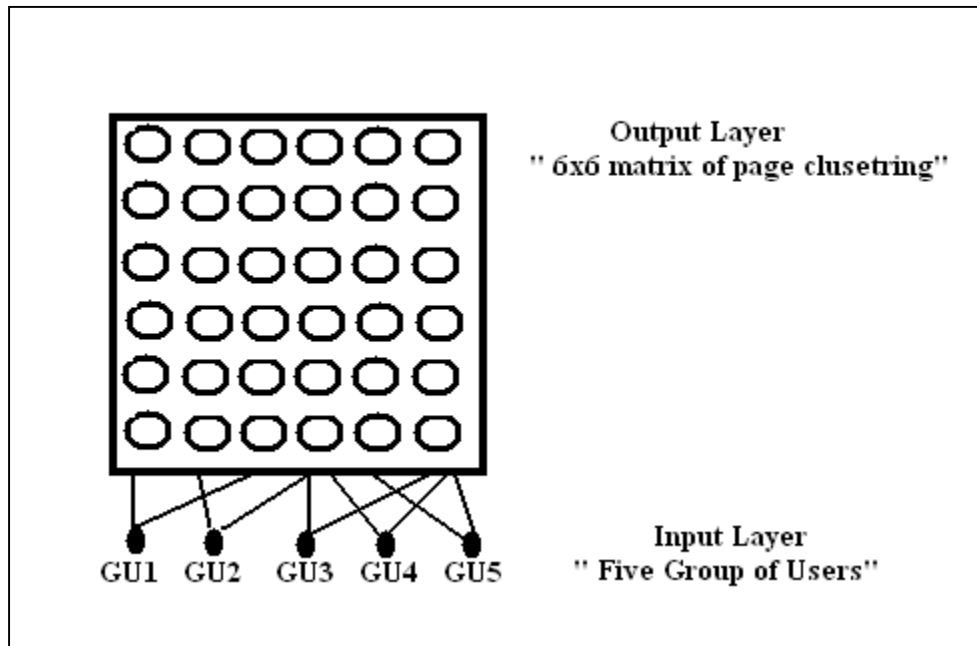


Figure 9 : SOM Input/output Layer

SOM technique was used in clustering since it has the following strength points:

1. SOM is unsupervised learning technique.
2. SOM can present clusters virtually as 2-D matrix
3. SOM provide spatial autocorrelation which means the closer clusters to each other the more they are similar.
4. SOM is flexible enough to present large amount of data points in a hierarchical clustering manner.

SOM algorithm that was presented in (Smith and Ng, 2003) is shown in Figure 10.

1. Initialize:

- Weights W_{ij} to small random values.
- Neighbourhood size $N_m(0)$ to be large (but less than the number of nodes in one dimension of the array).
- Parameter function $\alpha(t)$ $\sigma^2(t)$ to be between 0 and 1.

2. Present an input pattern x through the input layer and calculate the similarity (distance) d of this input to the weights w of each node j .

$$d_j = \|x - w_j\| = \sqrt{\sum_{i=1}^n (x_i - w_{ij})^2} \dots\dots\dots (6)$$

3. Select the node with minimum distance as winner m .

4. Update the weights connecting the input layer to the winning node its neighboring nodes according to the learning rate.

$$w_{ij}(t+1) = w_{ij}(t) + c[x_i - w_{ij}(t)] \dots\dots\dots (7)$$

Where $c = \alpha(t) \exp(-\|r_i - r_m\| / \sigma^2(t))$ for all nodes j in $N_m(t)$

Where $r_i - r_m$ is the physical distance (number of nodes) between node i and the winning node m .

5. Continue from step 2 for Ω epochs (in our case 1000 epoch); increase t by 1, then decrease the neighborhood size $\alpha(t)$ and $\sigma^2(t)$ such as

$$\alpha(t) = \alpha(0) N_m(t) / N_m(0) \dots\dots\dots (8)$$

Repeat until the weights have stabilized.

6. After the network is trained through repeated presentations of all URLs, present unite input vectors of every URL to the trained network and assign the winning node the URL address. Update the number labeling the nodes as the number of URLs allocated to the node

Figure 10: SOM Algorithm

3.2.4.1 SOM TRAINING

SOM algorithm can learn to detect regularities and correlations in its input and adapt their future responses to that input accordingly. Self-organizing maps learn to recognize groups of similar input vectors in such a way that neurons physically near each other in the output layer respond to similar input vectors. Data points lying near each other in the input space are mapped onto nearby map units (spatial autocorrelation). SOM is trained iteratively where at each training step, a sample vector is randomly chosen from the input data set. Distances between vector and all the prototype vectors are computed.

A self-organizing map learns to categorize input vectors. It also learns the distribution of input vectors. Feature maps allocate more neurons to recognize parts of the input space where many input vectors occur and allocate fewer neurons to parts of the input space where few input vectors occur.

3.2.5 PATTERN DISCOVERY AND ANALYSIS

3.2.5.1 Cluster Analysis

Clustering of users tends to establish groups of users exhibiting similar browsing patterns, and cluster of pages establishes pages as groups where they may be so related to each other according to the user transactions analysis. Such knowledge

is especially important to provide personalized Web content to the users with similar interests. In this research we cluster pages according to the user's usage of specific page. Analyzing such clusters may provide the webmasters of an insight of how the users are accessing the website, what are their interests, as well as finding if there are some pages that are so related to each other.

SOM has the capability of graphic representations that provides the analyst with a presentation of how neighboring map features are related to each other (spatial autocorrelation) which helps in better understanding for the result patterns.

3.2.5.2 Pattern Discovery

After SOM learning and cluster have been analyzed the hidden patterns from user usage of the web site can be discovered and accordingly find if there are any action should be taken to better organize the web site to be more customized.

4. RESULTS AND ANALYSIS

This chapter first introduces the environments that were used for experiments and shows the results that were found after attempting different scenarios of K-means clustering and SOM training.

4.1 EXPERIMENTAL ENVIRONMENT

In this thesis, MATLAB implementation for k-means and SOM algorithms was utilized in order to cluster JU web pages. The JU log files were filtered and cleaned by using Python programming language.

Python is an easy to learn, powerful programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. As mentioned before, log files were filtered to only have the attributes and records that we need by applying python code over it in order to get the needed data to be entered to SOM clustering. Preprocessed data and filtered log files were imported to MATLAB to apply k-means and SOM clustering techniques.

MATLAB was used since it has the following features:

- An environment to develop own functions and scripts
- The ability to import and export too many types of data files
- Object-oriented programming capabilities.

4.2 MATLAB SCENARIOS

In order to get the best clustering results many scenarios of k-means clustering and SOM trainings were attempted. In the first step many scenarios were attempted to reach the best k-means cluster of JU user transactions. To get an idea of how well-separated the resulting clusters are we used silhouette plot.

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distant from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. Moreover, the thickness of the cluster data used to indicate the distribution of the data over all clusters.

In addition, different k-means clusters were used in order to get the best scenario. A replicate indicates the number of times to repeat the clustering, each with a new set of initial cluster centroid positions. To be more accurate all tested scenarios were done over 5 replicates.

4.2.1 K-means Users Clustering Groups Scenarios

The number of inputs of the SOM will need to be equivalent to the number of transactions, and because this number was so large k-means was used to reduce the huge number of transactions. K-means was used as an initial clustering, scenarios as follows:

4.2.1.1 Two User Clusters Scenario

First, k-means was tested over 2 user clusters. The silhouette result in this case was as follows:

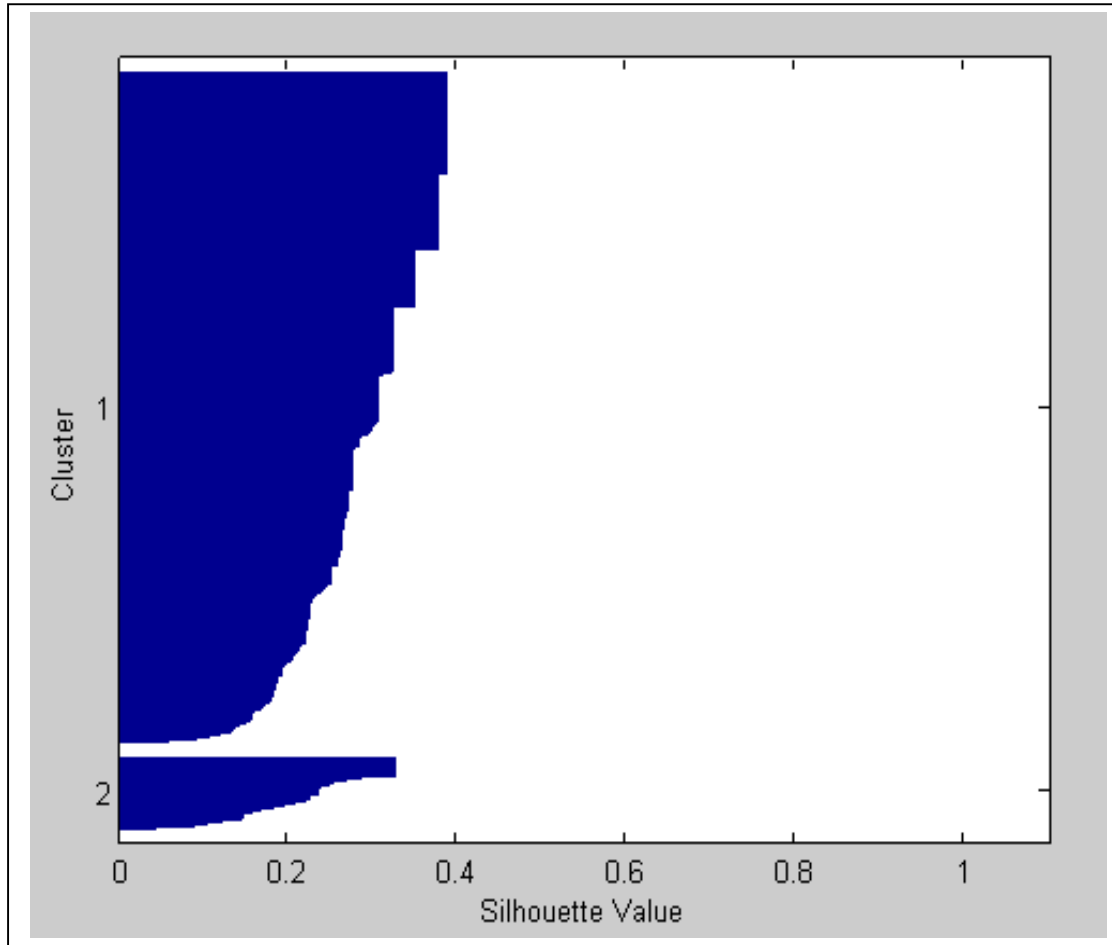


Figure 11: Silhouette result of 2 user clusters

This scenario was ignored because most of the data were grouped in one cluster. And as mentioned before the thickness of the result cluster indicates a good or bad clustering.

4.2.1.2 Three User Cluster Scenario

In the second scenario, k-means was attempted over 3 transaction groups. The silhouette result in this case was as shown in Figure 12:

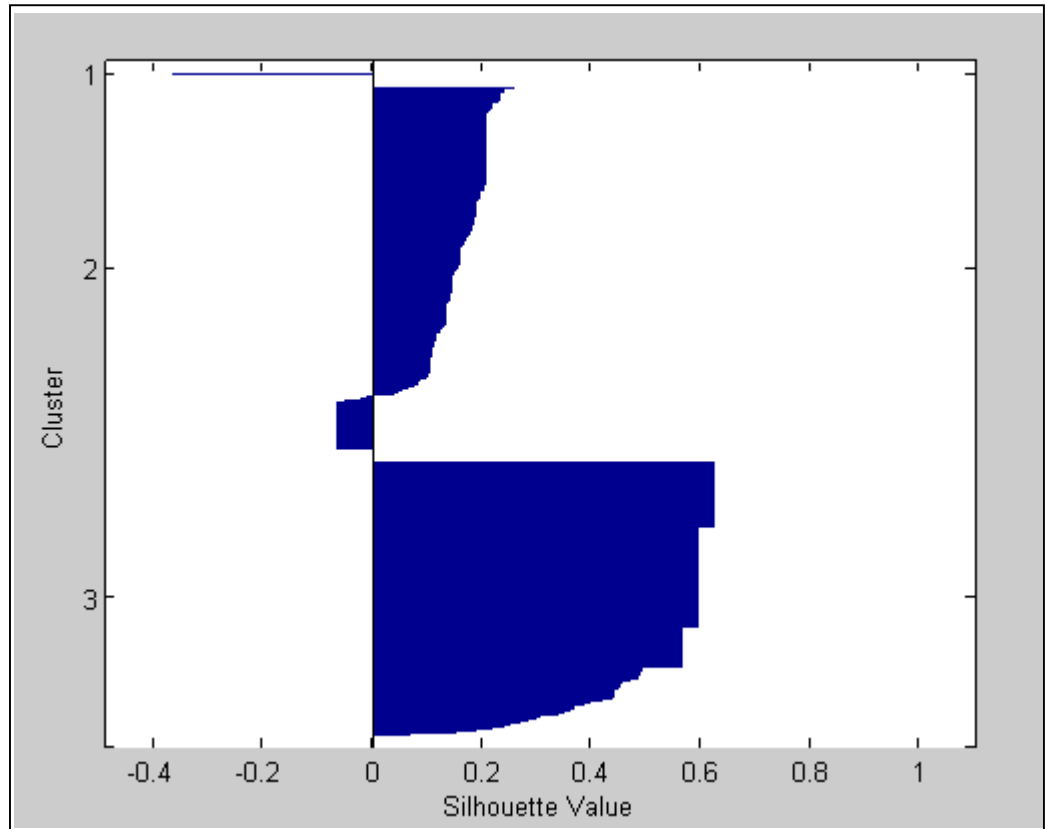


Figure 11: Silhouette result of 3 user clusters

This scenario was ignored since the silhouette shows that some set of data got the silhouette value under 0 which means it was assigned to the wrong cluster. Moreover, data were grouped in one group more than the others.

4.2.1.3 Four User Clusters Scenario

Figure 13 shows k-means clustering of 4 groups. And the silhouette result in this case was as follows:

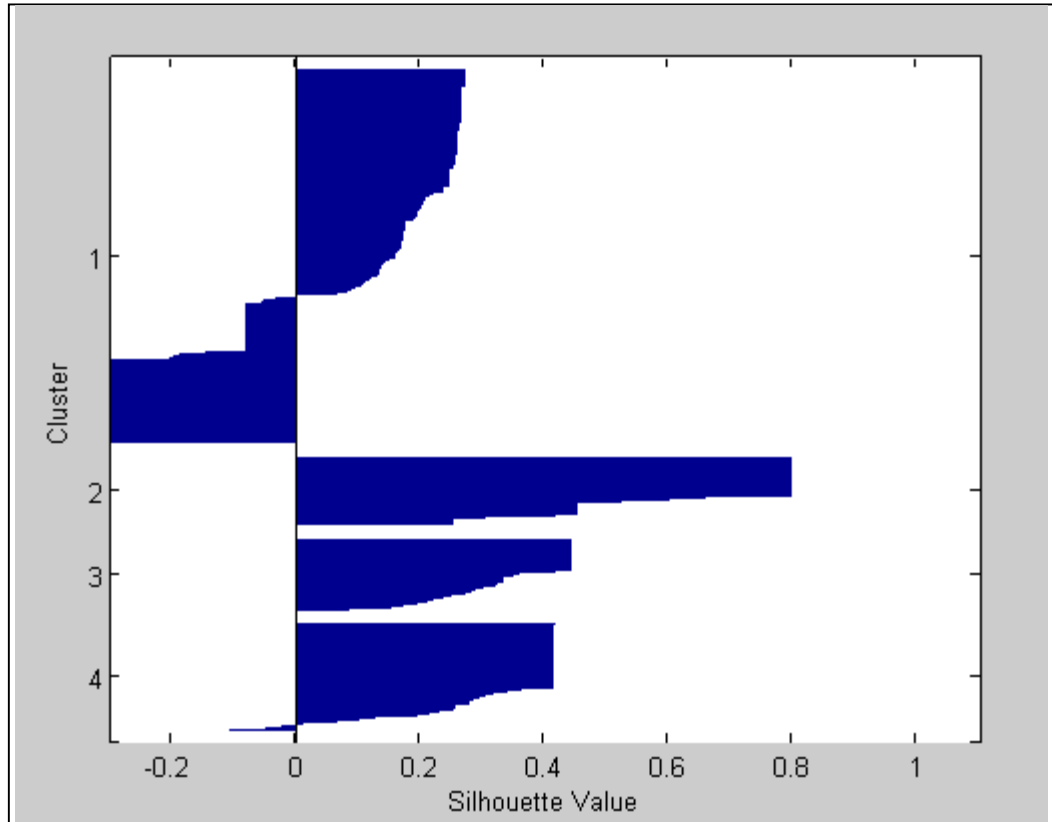


Figure 13: Silhouette result of 4 user clusters

This scenario was ignored since the silhouette shows that a large amount of data got silhouette value under 0 which means it was assigned to the wrong cluster.

4.2.1.4 Five user clusters Scenario

The data was also tested to be clustered over 5 k-means clusters. The result of the silhouette in this case was the best of all other scenarios; Figure 14 shows the silhouette result.

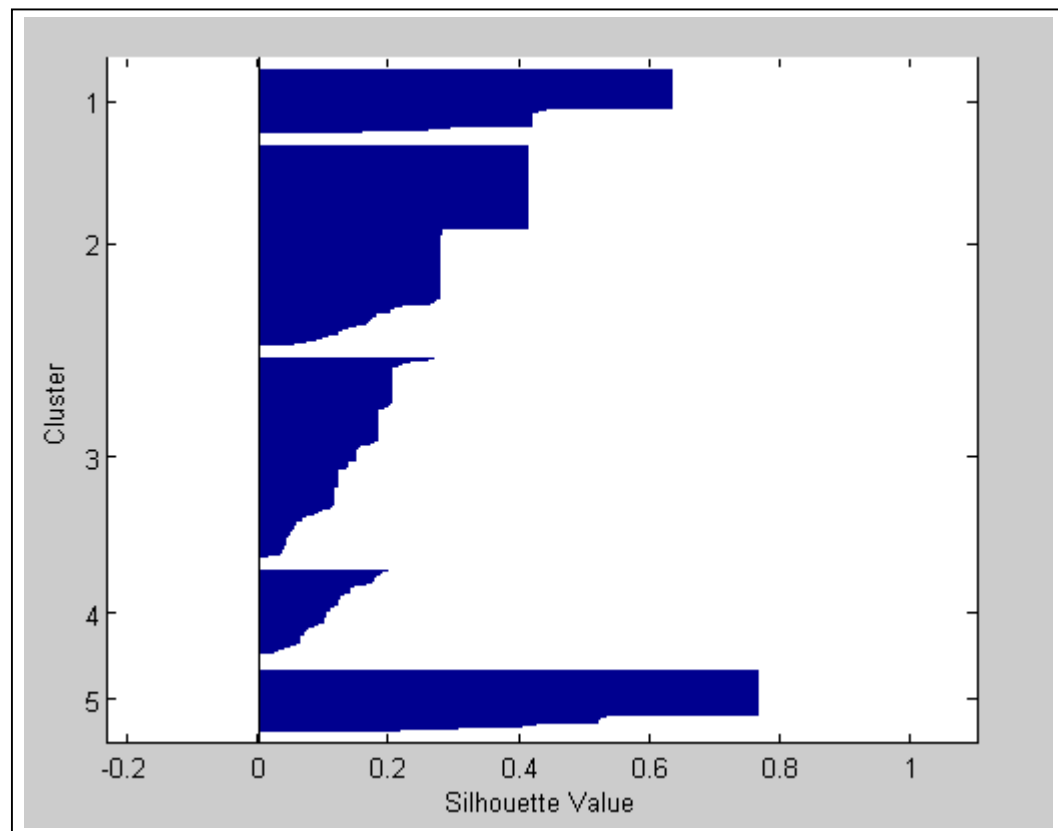


Figure 14: Silhouette result of 5 user clusters

This scenario was adopted since the data are nicely distributed on all clusters and there is no data value under 0 which means that it is clustered in a good way.

4.2.1.5 Six user clusters Scenario

In this scenario k-means was attempted over 6 clustering groups. The silhouette result in this case was as follows:

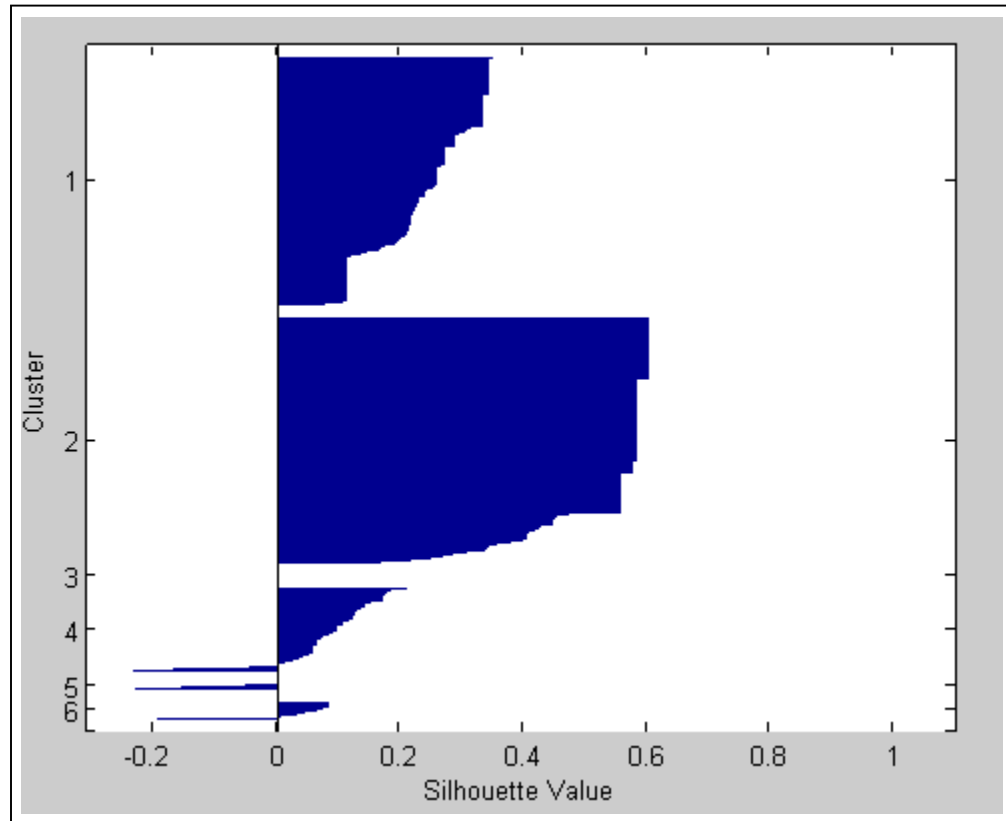


Figure 15: Silhouette result of 6 user clusters

This scenario was also ignored since the silhouette shows that some set of data got the silhouette value under 0. Moreover, data were grouped in some cluster more than the others.

4.2.1.6 Seven User Clusters Scenario

Finally k-means was tested over 7 user cluster groups. The silhouette result in this case was as follows:

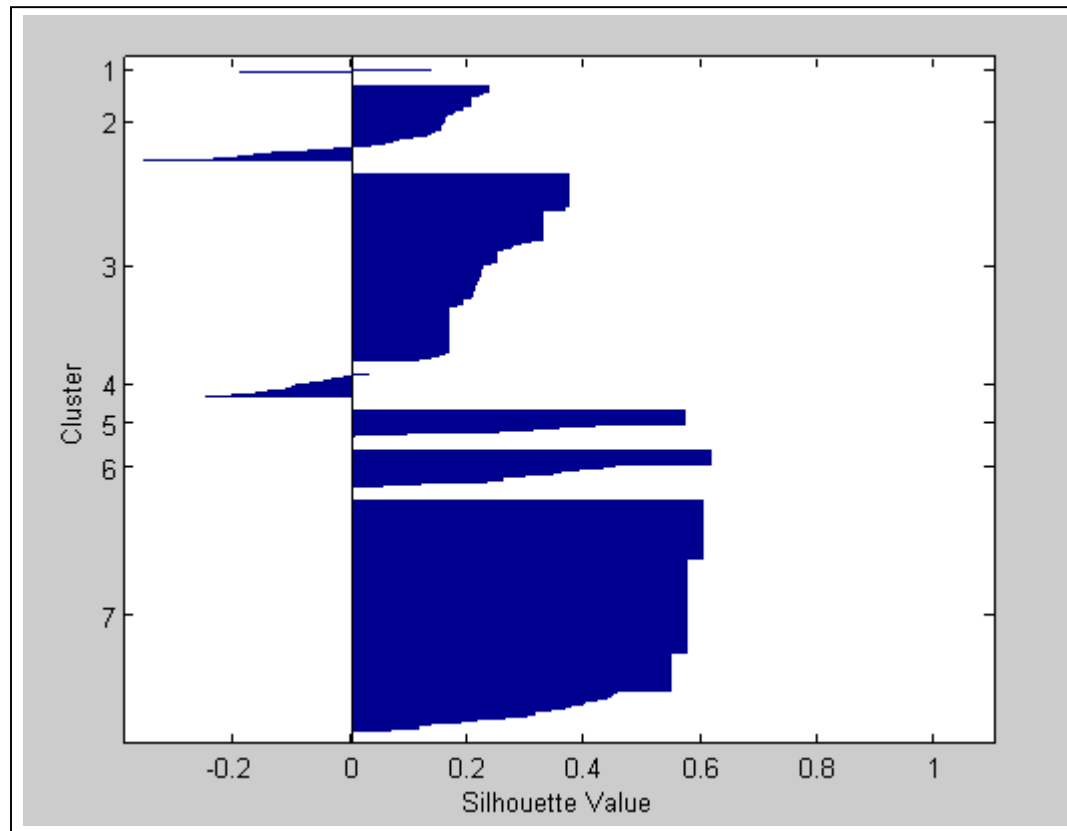


Figure 16: Silhouette result of 7 user clusters

The silhouette shows that some set of data got the silhouette value under 0.

Moreover, data were not separated in a good way in all clusters.

4.3 RESULTS AND ANALYSIS

As mentioned in the previous section users were clustered into 5 user group clusters according to the silhouette result which shows that 5 clusters with 5 replicates gives the best silhouette. After creating the initial 5 user clusters, these clusters become the input neurons to the SOM input layer. Next, a Two-dimensional Self-organizing Map was learned to represent different regions of the input space where input vectors occur. The neurons will arrange themselves in a two-dimensional grid.

The maps produced with the SOM algorithm are very much influenced by our choice of parameters. This may includes the map width and height and the value of the learning rate. In our case a five input neurons and an output of layer of 36 neurons spread out in a 6 by 6 grid (lattice). Where each neuron responds to a different region of the rectangle, and neighboring neurons responds to adjacent regions.

Then map was trained over 1000 epoch. After training, the layer of neurons begun to self-organize so that each neuron now classifies a different region of the input space, and adjacent (connected) neurons respond to adjacent regions that shows the spatial autocorrelations. The result of transaction groups and pages results with a matrix of pages that has a huge page vector hits and others with a very small page vector hits. This causes the need to do some normalization over this matrix to decrease the margins between the page vectors. Accordingly, a

calculation of each row of page vector and divide every cell of each specific page on the total number that results from the summation of the page vector was made. Thus data becomes between 0 and 1 which provides us with better results. A snapshot of the resulting matrix is shown in Figure 17. The full normalized matrix is shown in Appendix D.

0	0.10811	0	0
0	0.081081	1	0
1	0.56757	0	1
0	0.081081	0	0
0	0.16216	0	0

Figure 17: Snapshot of Normalized Matrix

The above normalized matrix was trained over the 1000 epoch to finally produce the SOM matrix that was created of five input neurons results from k-means clustering and an output of layer of 36 neurons spread out in a 6 by 6 grid (lattice). The data was distributed as shown in Figure 18.

7	4	1	3	5	2
9	0	1	3	3	6
0	4	0	2	2	1
3	2	2	3	0	12
1	1	1	0	2	0
10	0	3	2	2	0

Figure 18: Result of SOM Matrix

4.3.1 DISCUSSION OF RESULTS

In order to understand the result of the SOM lattice the numbers in each cell were replaced with the real pages that are grouped in that cell and find their relation by using Hierarchical Tree of selected JU website pages shown in Figure 19.

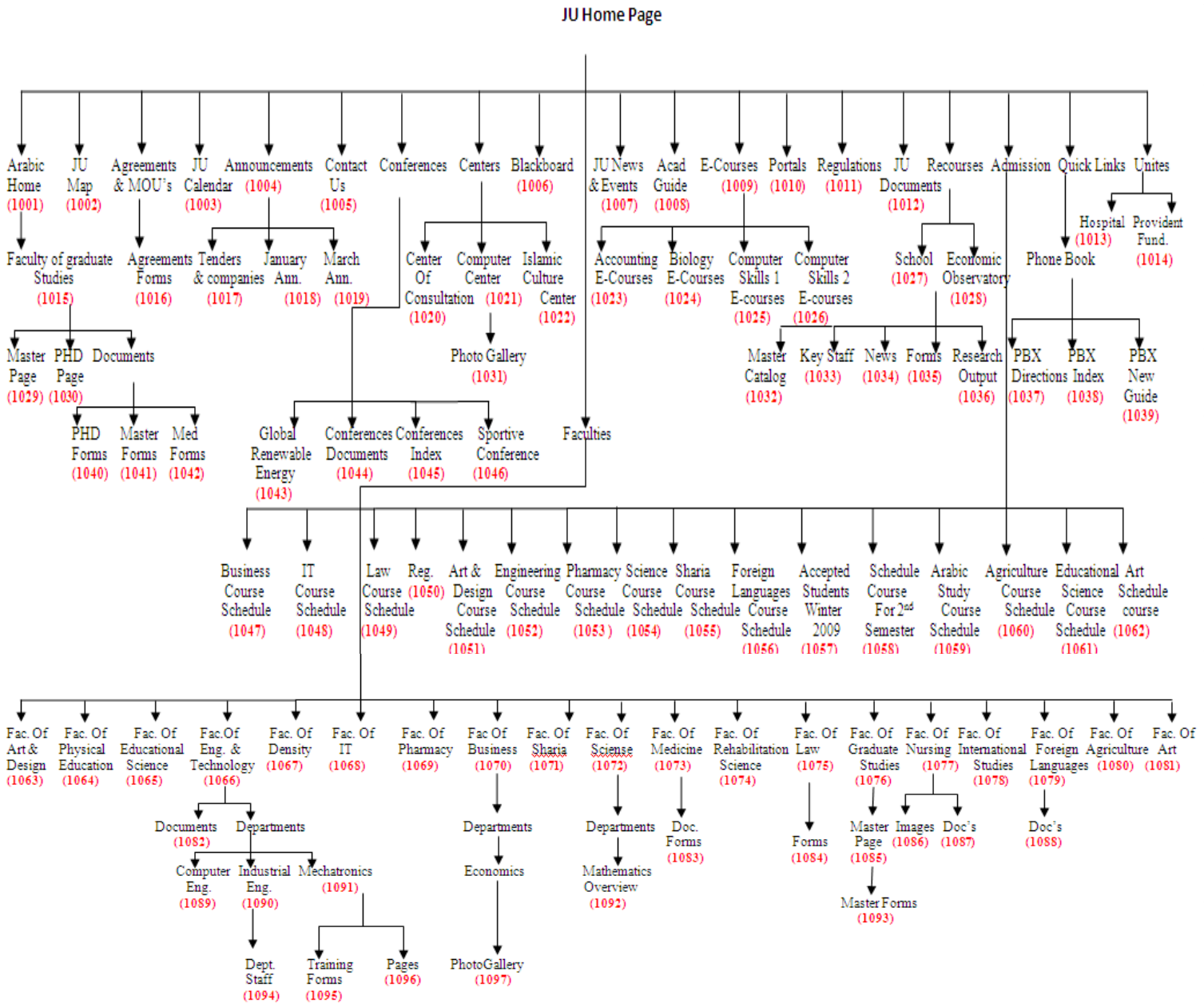


Figure 19: Hierarchical Tree of selected JU website pages

4.3.1.1 Results of page clustering using SOM

To better understand the SOM matrix that results, the matrix was analyzed by giving each group of cells a title that describes the content of these adjacent cells. Even that the results do not have a clear cut between cells, mainly the matrix contains the following groups as shown also in Figure 20:

1. Course Schedules
2. Registration /portals
3. Faculties
4. E-courses/Blackboard and document forms
5. Generic Pages as (Contact us page, Announcements page, Arabic home page...etc)

Course Schedules	Reg/Portals	Faculties
E-courses/Blackboard and Document Forms		Generic Pages

Figure 20: SOM Matrix Analysis

I. SAMPLE 1

Figure 21 defines that most of the course schedules were grouped within 4 neighboring cells. Moreover; it also grouped for example (Faculty of Engineering Industrial Eng department staff in cell 1 with Faculty of Engineering Industrial Eng. Overview in cell 2) which means that SOM discovered that these pages are strongly related and place them in two adjacent cells.

1	2
JU School Home Page	Conferences Page Documents
Faculty Of Graduate Studies PHP Page	Faculty Of Nursing Images
Faculty of Engineering Industrial Eng department staff	Faculty of Engineering Industrial Eng. Overview
Faculty of Business photo gallery	Schedule Courses for IT Faculty
Schedule Courses for Business administration Faculty	
Schedule Courses for Educational Science Faculty	
Schedule Courses for Art Faculty	
7	8
Schedule Course for Faculty of Art and Design	
Schedule Courses for Law Faculty	
Schedule Courses for Engineering Faculty	
Pharmacy Courses schedule	
Schedule Courses for Sharia Faculty	
Schedule Courses for Foreign Languages Faculty	
Schedule Courses for the Second Semester 2008-2009	
International Arabic study Schedule Course	
Schedule Courses for Agriculture Faculty	

Figure 21: SOM Adjacent Cells (1, 2, 7, 8)

II. SAMPLE 2

Figure 22 shows another sample where SOM clustering discovered the relation between Computer Center page and its photo gallery and so grouped them in two adjacent cells. SOM also put e-courses related pages in four neighboring cells as well as having the blackboard adjacent to the e-courses. This indicates a strong relationship based on users navigation.

13	14
	Faculty of graduate Studies Master Page
	JU News And Events
	Computer Center Photo Gallery
	E-Courses Computer skills 2 content
19	20
Computer Center Page	E-Courses Page
Schedule Courses for Science Faculty	E-Courses Computer skills 1 content
Faculty Of Science Mathematics Overview	
25	26
E-Courses Accounting page	E-Courses Biology page
31	32
Calendar Of JU	
Blackboard Page	

Figure22: SOM Adjacent Cells (13, 14, 19, 20, 25, 26, 31, 32)

III. SAMPLE 3

This sample shows that SOM has grouped some generic JU pages in adjacent cells. As JU announcements, contact us page, and the agreements as shown in Figure 23.

23	24
	Arabic JU home
	Announcements of JU
	Contact us page
	Agreements JU Forms
	Tenders Home page
	Announcement on January
	Economic Observatory Master Catalog Page
	Economic Observatory Key Staff
	Faculty Of Graduate Studies Document PHD Forms
	Faculty of Engineering computer Eng. Overview
	Faculty of Engineering Mechatronics Training Forms
	Faculty of Engineering Mechatronics pages
29	30
Provident Fund- Administration Page	
Faculty of Dentistry Page	
35	36
Announcement on March about Jobs	
Faculty Of Law Forms	

Figure 23: SOM Adjacent Cells (23, 24, 29, 30, 35, 36)

5. CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

This research was based on using the Kohonen's self-organizing map (SOM) to organize JU web pages in a domain onto a two-dimensional map. The organization of these JU pages is based on the users' usage behavior on the web site. It has been demonstrated that the resulting map of this system is very meaningful and can be easily understood to find relations between the result map and user navigation over the website.

K-means was used as an initial clustering to reduce dimensionality of data to be ready to be used by SOM, we used 5 k-means clusters in order to cluster users, so the input neurons of SOM was 5. SOM provides us with visual map to enable finding the relationship between web pages based on the usage patterns of web users and visually see them. It also provides an analysis tool for web masters and web authors to better understand the interests of visitors to their pages.

This study was a testing for a technique to cluster JU web pages, SOM was used because of its property of special autocorrelation to help in easily discover user patterns and capture how users are accessing the website. The limitation of the current system is that it has only been evaluated on a sample of data: the log files for about one week of February access of JU website. Moreover, pages were reduced to 97 which means that there may be some relations that could be found if we have more pages.

5.2 FUTURE WORK

In future SOM can be tested over more than 97 pages to get better and more accurate results of the user navigations over JU website. In order to make recommendations to the user that may help in further navigation through the web site.

Furthermore, different data set may be analyzed since user interests may differ from one month to another, so that recommendations can be made to JU web site administrators and designers, regarding structural changes to the site in order to enable more efficient browsing. Moreover, different SOM dimensions can be attempted.

REFERENCES

- [1] Ahmed S., Halim Z., Baig R., and Bashir S. (2008), Web Content Mining: A Solution to Consumer's Product Hunt. **PROCEEDINGS OF WORLD ACADEMY OF SCIENCE, ENGINEERING AND TECHNOLOGY, VOLUME 27 FEBRUARY 2008.**
- [2] Alhoniemi E., Himberg J., Hollmen J., Laine S., Lehtimäki P., Raivio K., Simila T., Simula O., Sirola M., Sulkava M., Tikka J., and Vesanto J. (2003), **SOM in data mining**, Chapter 14, pp. (171-177).
- [3] Berkhin P. (2003), **Survey of Clustering Data Mining Techniques.**
- [4] Britos P., Martinelli D., Merlino H., and García-Martínez R. (2007), Web Usage Mining Using Self Organized Maps. **International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007.**
- [5] Chanchary H. F., Haque I., and Khalid Md. S. (2008), Web Usage Mining to Evaluate the Transfer of Learning in a Web-based Learning Environment. **IEEE Workshop on Knowledge Discovery and Data Mining.**
- [6] Dai H., and Mobasher B. (2003), Integrating Semantic Knowledge with Web Usage Mining for Personalization.
- [7] Fernandez V. F., and Layos L. M. (2003), Text Content Approaches in Web Content Mining. **Madrid Research Agency, under project 07T/0030/2003-1.**
- [8] Han J., Cheng H., Xin D., and Yan X. (2007), Frequent pattern mining: current status and future directions. **Springer Science Data Min Knowledge Disc**, pp.55–86.

- [9] Janet P. T., and Dr. Albers M. J.(2006), Measuring Cognitive Load to Test the Usability of Web Sites.
- [10] Jiang T., Tan A. , and Ke Wang (2007), Mining Generalized Associations of Semantic Relations from Textual Web Content. **IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING**, VOL. 19, NO. 2, FEBRUARY 2007.
- [11] Liu B. (2005), Web Content Mining. **14th International World Wide Web Conference (WWW-2005)** pp.(2-42)
- [12] Labroche N., Lesot M. , and Yaffi L. (2007), A New Web Usage Mining and Visualization Tool. **19th IEEE International Conference on Tools with Artificial Intelligence**.
- [13] Merelo-Guervós J. , Prieto B. , Prieto A., Romero G., Castillo V. P., and Tricas F. (2004), CLUSTERING WEB-BASED COMMUNITIES USING SELFORGANIZING MAPS. **International Conference Web Based Communities 2004**.
- [14] Mobasher B. , Dai H. , Luo T. , and Nakagawa M. (2002), Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. **Kluwer Academic Publishers. Data Mining and Knowledge Discovery**, 6, 61–82, 2002.
- [15] Mobasher B., **Web Usage Mining, chapter 12 , Springer**, 2006 pp.(449-483).
- [16] Perelomov I., Azcarraga A. P., Tan J. ,and Seng C. T. (2002), Using Structured Self-Organizing Maps in News Integration Websites.
- [17] Raju G T, Kunal, and Satyanarayana P S (2007), Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network based Clustering Algorithm. **IEEE International Conference on Computational Intelligence and Multimedia Applications 2007**.

[18] Rossi F., Aicha E.G., and Lechevallier Y. (2005), Usage Guided Clustering of Web Pages with the Median Self Organizing Map. **European symposium on Artificial Neural Networks.**

[19] Schatzmann J. (2003), **Using Self-Organizing Maps to Visualize Clusters and Trends in Multidimensional Datasets.**

[20] Smith K. A., and Ng A. (2003), Web page clustering using a self-organizing map of user navigation patterns. **Elsevier Science Decision Support Systems** 35 (2003) 245–256.

[21] Sperandio M., and Coelho J. (2003), **K-MEANS AND SOM, A CONCURRENT VALIDATION SCHEME FOR DATA MINING.**

[22] Srivastava J., Cooley R., Deshpande M., and Pang N. T. (2000), Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. **ACM SIGKDD**, Jan 2000. Volume 1, Issue 2.

[23] Tomsich P., Rauber A., and Merkl D. (2001), Optimizing the par SOM Neural Network, **Implementation for Data Mining with Distributed Memory Systems and Cluster Computing.**

[24] Vesanto J., and Alhoniemi E. (2000), Clustering of the Self-Organizing Map. **IEEE TRANSACTIONS ON NEURAL NETWORKS**, VOL. 11, NO. 3, MAY 2000.

[25] Velasquez J., Yasuda H. and Aoki T. (2003), Combining the web content and usage mining to understand the visitor behavior in a web site. **Proceedings of the Third IEEE International Conference on Data Mining.**

[26] Wang C., Liu Y., Jian L., and Zhang P (2008), A Utility-based Web Content Sensitivity Mining Approach. **IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.**

- [27] Wel V. L. and Royackers L. (2004), Ethical issues in web data mining. **Kluwer Academic Publishers. Ethics and Information Technology** 6: 129–140, 2004.
- [28] Xinwu L. (2008), Research on Text Clustering Algorithm Based on K_means and SOM. **IEEE International Symposium on Intelligent Information Technology Application Workshops.**
- [29] Yang H. C., and Lee C. H.(2006), Mining Unstructured Web Pages to Enhance Web Information Retrieval. **IEEE First International Conference on Innovative Computing, Information and Control.**
- [30] Yu Y., He P., Bai Y , Yang Z. (2008), A Document Clustering Method Based on One-Dimensional SOM. **Seventh IEEE/ACIS International Conference on Computer and Information Science.**
- [31] Zhang X., and Yin X. (2008), Design of an Information Intelligent System based on Web Data Mining. **IEEE International Conference on Computer Science and Information Technology** 2008.
- [32] Zheng G. and Bouguettaya A.(2009), Service Mining on the Web. **IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 2, NO. 1, JANUARY-MARCH 2009.**

APPENDICES

Appendix A. Sample of JU web Log files

```

10.248.136.51 - h.rawabdeh [09/Feb/2009:16:35:13 +0000] "GET
http://rad.live.com/ADSAdClient31.dll?GetSAd=&DPJS=4&AP=1390&PG=WLMUS4&UC=14
2 HTTP/1.0" 200 2352 ALLOW "N2H2"

10.248.16.186 - roa0063561 [09/Feb/2009:16:35:13 +0000] "GET
http://mirhanstar.com/vb/private.php HTTP/1.0" 200 60965 ALLOW "N2H2"

10.248.64.61 - sls0030652 [09/Feb/2009:16:35:13 +0000] "GET
http://cdn.maktoob.com/images/news/images/news2.css HTTP/1.0" 200 17181 ALLOW "N2H2"

10.249.136.111 - o.dandan [09/Feb/2009:16:35:13 +0000] "GET
http://www.aljazeera.net/custom/customScript.js HTTP/1.1" 304 0 ALLOW "N2H2"

10.249.97.94 - r.zubi [09/Feb/2009:16:35:13 +0000] "GET
http://ammonnews.net/advertizing/bank.swf HTTP/1.1" 500 35532 ALLOW "N2H2"

10.249.24.210 - j.kalifa [09/Feb/2009:16:35:13 +0000] "GET
http://exchange.jo/ticker/index_today.php HTTP/1.0" 200 1726 ALLOW "N2H2"

10.248.44.25 - a.zubaidi [09/Feb/2009:16:35:13 +0000] "GET
http://i1.ytimg.com/vi/hl8vWsM7XYw/default.jpg HTTP/1.0" 304 2770 ALLOW "N2H2"

10.248.16.69 - wfa0064179 [09/Feb/2009:16:35:13 +0000] "GET
http://regweb2.ju.edu.jo:7777/regapp/images/Menu_3.gif HTTP/1.0" 304 0 ALLOW "N2H2"

10.248.64.61 - sls0030652 [09/Feb/2009:16:35:13 +0000] "GET
http://cdn.maktoob.com/images/flvPlayer/swfobject.js HTTP/1.0" 200 9759 ALLOW "N2H2"

10.248.112.67 - - [09/Feb/2009:16:35:13 +0000] "OPTIONS http://ahmad-khawaldeh/
HTTP/1.0" 407 0

10.248.8.84 - aro8060599 [09/Feb/2009:16:35:13 +0000] "GET
http://clients1.google.com/complete/search?hl=ar&gl=jo&q=upper%20basic%20stage%20e
HTTP/1.0" 200 46 ALLOW "N2H2"

172.28.9.80 - G.ALHAJJAJ [09/Feb/2009:16:35:13 +0000] "GET
http://www.ammonnews.net/ammoneNewsImage/200928mid33944.jpeg HTTP/1.0" 304 0
ALLOW "N2H2"

10.248.112.48 - bto0076420 [09/Feb/2009:16:35:13 +0000] "POST
http://regweb4.ju.edu.jo:7777/regapp/FreeTimeNewControlS HTTP/1.0" 302 267 ALLOW
"N2H2"

```

0.249.104.147 - nqaruoti [11/Feb/2009:17:52:56 +0000] "POST
<http://address.mail.yahoo.com/index.php> HTTP/1.0" 200 21905 ALLOW "N2H2"

10.249.97.3 - compe.head [11/Feb/2009:17:52:56 +0000] "GET
<http://groups.google.jo/group/ro0o3ah/icon?v=2&hl=ar> HTTP/1.0" 200 1281 ALLOW "N2H2"

10.249.24.251 - mmjafar [11/Feb/2009:17:52:56 +0000] "GET
http://www.elsevier.com/legacy_products/p29/229/229_01_article_recent.html HTTP/1.0" 200
 6708 ALLOW "N2H2"

172.28.9.152 - - [11/Feb/2009:17:52:56 +0000] "GET http://u31.eset.com/nod_eval/update.ver
 HTTP/1.0" 407 0

10.249.80.50 - myamani [11/Feb/2009:17:52:56 +0000] "CONNECT <https://mail.ju.edu.jo:443/>
 HTTP/1.0" 400 0 ALLOW "N2H2"

10.249.104.45 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://user-018acfl61c/> HTTP/1.0"
 407 0

10.248.112.78 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://fes2-wk22/> HTTP/1.0" 407 0

172.28.9.11 - jmasad [11/Feb/2009:17:52:56 +0000] "GET
http://www.aljazeera.net/mritems/images/2008/12/2/1_874356_1_59.jpg HTTP/1.0" 304 0
 ALLOW "N2H2"

10.248.136.35 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://ebraheem/> HTTP/1.0" 407 0

10.249.97.3 - compe.head [11/Feb/2009:17:52:56 +0000] "GET
http://groups.google.jo/groups/img/gsecs/discussions_24.png HTTP/1.0" 304 0 ALLOW "N2H2"

10.249.24.89 - hbt8080657 [11/Feb/2009:17:52:56 +0000] "GET http://apps.facebook.com/----cgcda/taker/more_quizzes?xlinker=28&page_num=2&seed=0.46656776761 HTTP/1.1" 200
 20011 ALLOW "N2H2"

10.249.40.125 - m.jaber [11/Feb/2009:17:52:56 +0000] "GET
<http://view.atdmt.com/SCD/view/ynxxxgen0010000003scd/direct/01/> HTTP/1.0" 302 0 ALLOW
 "N2H2"

10.249.104.45 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://user-018acfl61c/> HTTP/1.0"
 407 0

10.249.136.156 - theses [11/Feb/2009:17:52:56 +0000] "GET
<http://www.islamonline.net/servlet/Trick.jpg> HTTP/1.0" 404 103 ALLOW "N2H2"

10.248.112.78 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://fes2-wk22/> HTTP/1.0" 407 0

10.248.136.35 - - [11/Feb/2009:17:52:56 +0000] "OPTIONS <http://ebraheem/> HTTP/1.0" 407 0

10.248.112.62 - mam0082756 [16/Feb/2009:15:23:24 +0000] "GET
http://www.ju.edu.jo/_catalogs/masterpage/UJFiles/icon_04a.gif HTTP/1.0" 200 79 ALLOW
 "N2H2"

10.248.44.84 - ary0082800 [16/Feb/2009:15:23:24 +0000] "GET
<http://www.dramjad.net/header.swf> HTTP/1.0" 200 824542 ALLOW "N2H2"

10.248.105.44 - msa2051179 [16/Feb/2009:15:23:24 +0000] "GET
<http://blackboard.ju.edu.jo/webapps/login> HTTP/1.0" 302 0 ALLOW "N2H2"

10.248.8.215 - rym0068539 [16/Feb/2009:15:23:24 +0000] "GET <http://www.hajr-network.net/hajrvb/images/icons/icon7.gif> HTTP/1.0" 200 1058 ALLOW "N2H2"

172.28.9.22 - hosp [16/Feb/2009:15:23:24 +0000] "GET
http://www.ibtesama.com/images/ster_C7_R1.jpg HTTP/1.0" 200 1274 ALLOW "N2H2"

10.248.48.87 - mhm0072519 [16/Feb/2009:15:23:24 +0000] "GET
<http://get.adobe.com/flashplayer/> HTTP/1.0" 200 22339 ALLOW "N2H2"

192.168.5.70 - - [16/Feb/2009:15:23:24 +0000] "GET
<http://www.download.windowsupdate.com/msdownload/update/v3/static/trustedr/en/authrootseq.txt> HTTP/1.1" 407 0

172.28.9.43 - mkhateeb [16/Feb/2009:15:23:24 +0000] "GET
http://i.cdn.turner.com/cnn/.element/img/2.0/global/nav/footer/corner_footer_bl.gif HTTP/1.1" 304 0 ALLOW "N2H2"

10.248.44.104 - h.natourea [16/Feb/2009:15:23:24 +0000] "GET
http://xsltcache.alexa.com/site_stats/gif/t/a/YXJiMy5tYWt0b29iLmNvbQ==/s.gif HTTP/1.0" 200 2855 ALLOW "N2H2"

10.248.24.98 - sah0068777 [16/Feb/2009:15:23:24 +0000] "GET <http://www.anime-plus.com/support> HTTP/1.0" 200 16778 ALLOW "N2H2"

10.248.112.62 - mam0082756 [16/Feb/2009:15:23:24 +0000] "GET
http://www.ju.edu.jo/_catalogs/masterpage/UJFiles/towblue-r.gif HTTP/1.0" 200 73 ALLOW
 "N2H2"

10.248.193.17 - rza0076789 [16/Feb/2009:15:23:24 +0000] "GET
<http://www.medicimaging.com/images/logo1.png> HTTP/1.0" 200 5933 ALLOW "N2H2"

10.248.105.44 - msa2051179 [16/Feb/2009:15:23:24 +0000] "GET
<http://blackboard.ju.edu.jo/webapps/login/> HTTP/1.0" 200 14530 ALLOW "N2H2"

172.28.9.43 - mkhateeb [16/Feb/2009:15:23:24 +0000] "GET
http://i.cdn.turner.com/cnn/.element/img/2.0/global/nav/footer/corner_footer_br.gif HTTP/1.1" 304 0 ALLOW "N2H2"

10.249.96.184 - a.kanan [16/Feb/2009:15:23:24 +0000] "CONNECT https://edit.yahoo.com:443/ HTTP/1.0" 400 0 ALLOW "N2H2"

10.248.136.130 - - [16/Feb/2009:15:23:24 +0000] "OPTIONS http://alya24312/ HTTP/1.0" 407 0

10.248.136.35 - - [16/Feb/2009:15:23:24 +0000] "OPTIONS http://lab3admin1/ HTTP/1.0" 407 0

10.248.112.62 - mam0082756 [16/Feb/2009:15:23:24 +0000] "GET http://www.ju.edu.jo/_catalogs/masterpage/UJFiles/QLS.jpg HTTP/1.0" 200 13008 ALLOW "N2H2"

0.248.16.63 - abd8080486 [16/Feb/2009:15:40:04 +0000] "GET http://www.ju.edu.jo/_catalogs/masterpage/ARA/WebFiles/images/blank.gif HTTP/1.0" 200 49 ALLOW "N2H2"

10.248.16.133 - mhm0074021 [16/Feb/2009:15:40:04 +0000] "GET http://www.5ater.com/vb/images/ehdaa_smilies/eh_s(9).gif HTTP/1.0" 304 0 ALLOW "N2H2"

10.249.97.101 - - [16/Feb/2009:15:40:04 +0000] "GET http://10.1.1.2/wpdad.dat HTTP/1.1" 401 0

10.248.112.231 - jmy0068922 [16/Feb/2009:15:40:04 +0000] "GET http://www.charlierose.com/js/hoverIntent.js HTTP/1.0" 200 3174 ALLOW "N2H2"

10.248.16.63 - abd8080486 [16/Feb/2009:15:40:04 +0000] "GET http://www.ju.edu.jo/_catalogs/masterpage/ARA/WebFiles/images/blank.gif HTTP/1.0" 200 49 ALLOW "N2H2"

172.28.9.101 - ali.mansour [16/Feb/2009:15:40:04 +0000] "GET http://eur.a1.yimg.com/java.europe.yahoo.com/eu/any/paramount/20090216wm430x80.jpg HTTP/1.0" 200 26280 ALLOW "N2H2"

10.248.200.63 - mhm2050391 [16/Feb/2009:15:40:04 +0000] "GET http://images.jordan.gov.jo/wps/wcm/resources/image/468fed90a1547e07/triangle-7-link.gif HTTP/1.0" 200 82 ALLOW "N2H2"

10.248.16.133 - mhm0074021 [16/Feb/2009:15:40:04 +0000] "GET http://www.5ater.com/vb/w/w_12.jpg HTTP/1.0" 304 0 ALLOW "N2H2"

172.28.9.122 - a.yaseen [16/Feb/2009:15:40:04 +0000] "GET http://www.alrai.com/images/top.jpg HTTP/1.0" 200 9066 ALLOW "N2H2"

172.28.9.122 - a.yaseen [16/Feb/2009:15:40:04 +0000] "GET http://www.alrai.com/images/print.jpg HTTP/1.0" 200 779 ALLOW "N2H2"

172.28.9.122 - a.yaseen [16/Feb/2009:15:40:04 +0000] "GET http://www.alrai.com/images/home.jpg HTTP/1.0" 200 9176 ALLOW "N2H2"

10.249.112.38 - - [16/Feb/2009:15:40:04 +0000] "OPTIONS http://do3a22581/ HTTP/1.0" 407 0

10.248.48.104 - mna0051647 [16/Feb/2009:15:48:35 +0000] "GET
http://fetweb.ju.edu.jo/staff/cpe/asarhan/ HTTP/1.0" 304 0 ALLOW "N2H2"

10.249.104.228 - m.khawaldeh [16/Feb/2009:15:48:35 +0000] "GET
http://mail.opi.yahoo.com/online?u=hanan_2006@msn.com&m=g&t=0 HTTP/1.1" 200 100
ALLOW "N2H2"

10.248.112.121 - - [16/Feb/2009:15:48:35 +0000] "OPTIONS http://actlab23/ HTTP/1.0" 407 0

172.28.9.143 - h.alahmad [16/Feb/2009:15:48:35 +0000] "GET
http://webhosting.yahoo.com/ps/sb/index.php?redirect_uri=tt&&& HTTP/1.1" 500 3748
ALLOW "N2H2"

10.249.136.67 - WLY0078256 [16/Feb/2009:15:48:35 +0000] "GET http://www.m-
al Hassanain.com/main/scp/onlin/rased.php/ HTTP/1.1" 206 1005 ALLOW "N2H2"

10.248.44.85 - aby0078656 [16/Feb/2009:15:48:35 +0000] "GET
http://www.alrai.com/pages.php?opinion_id=9974 HTTP/1.0" 200 30162 ALLOW "N2H2"

10.248.48.80 - ahm0068612 [16/Feb/2009:15:48:35 +0000] "GET
http://www.3alarasi.com/images/screen_blue/icon-sidebarnavigationlinks.gif HTTP/1.0" 200 46
ALLOW "N2H2"

10.248.48.129 - mhm0052066 [16/Feb/2009:15:48:35 +0000] "GET
http://l.yimg.com/a/i/us/tr/gr/tvly_suitcase_25x25.gif HTTP/1.0" 200 1251 ALLOW "N2H2"

10.248.48.80 - ahm0068612 [16/Feb/2009:15:48:35 +0000] "GET
http://www.3alarasi.com/images/plus.gif HTTP/1.0" 200 236 ALLOW "N2H2"

10.249.8.85 - - [16/Feb/2009:15:48:35 +0000] "OPTIONS http://atef96511bdc/ HTTP/1.0" 407 0

10.248.48.144 - ala0057978 [16/Feb/2009:15:48:35 +0000] "GET
http://top9.mail.ru/counter?id=1291136;t=69;l=1;FTID=0;VID=0uWnp2241AGc HTTP/1.0" 200
885 ALLOW "N2H2"

10.248.48.80 - ahm0068612 [16/Feb/2009:15:48:35 +0000] "GET
http://www.3alarasi.com/js/functions.js HTTP/1.0" 200 9683 ALLOW "N2H2"

192.168.5.30 - i.qado [16/Feb/2009:15:48:35 +0000] "GET
http://www.afmelzem.com/vb/clientscript/yui/connection/connection-min.js?v=374 HTTP/1.1"
200 11602 ALLOW "N2H2"

10.249.136.246 - - [16/Feb/2009:15:48:35 +0000] "OPTIONS http://ju-cb4f7b83b5fa/ HTTP/1.1"
407 0

10.248.105.69 - and0084896 [16/Feb/2009:15:48:35 +0000] "GET http://ar-
hp.com/vb/RaidArena/buttons/collapse_tcat.gif HTTP/1.0" 200 206 ALLOW "N2H2"

Appendix B. List of Page Index

Page ID	Description	Path
	JU home Page	http://www.ju.edu.jo/
1001	Arabic JU home	http://www.ju.edu.jo/arabichome
1002	JU Map	http://www.ju.edu.jo/UJ%20Map/UJMap.html
1003	Calender Of JU	http://www.ju.edu.jo/calendar/index.html
1004	Announcements of JU	http://www.ju.edu.jo/announcements/uac/default.htm
1005	Contact us page	http://www1.ju.edu.jo/ContactUs.html
1006	BlackBoard Page	http://blackboard.ju.edu.jo/
1007	JU News And Events	http://www.ju.edu.jo/Lists/NewsAndEvents/
1008	Student Information For Acad	http://acad.ju.edu.jo/
1009	E-Courses Page	http://www1.ju.edu.jo/e-courses/default.htm
1010	Portal Page	http://portals.ju.edu.jo/
1011	Regulation Documents	http://www.ju.edu.jo/Pages/Regulations/
1012	JU Documents Page	http://www.ju.edu.jo/documents
1013	JU Hospital Home page	http://www.ju.edu.jo/medical/hospital
1014	Providant Fund-Administration Page	http://www1.ju.edu.jo/providant-fund/administration.html
1015	Faculty of graduate studies " Arabic Faculties "	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies
1016	Agreements JU Forms	http://www.ju.edu.jo/Agreements%20and%20MOUs/Forms
1017	Tenders Home page	http://www.ju.edu.jo/tenders
1018	Announcement on January	http://www1.ju.edu.jo/announcements/2008%20January/adv%207-2-2008_3.htm
1019	Announcement on March about Jobs	http://www1.ju.edu.jo/announcements/2008%20March/jop.htm
1020	Center of Consultation Page	http://www.ju.edu.jo/centers/coc
1021	Computer Center Page	http://www.ju.edu.jo/centers/computercenter
1022	Islamic Cultural Center Home Page	http://www.ju.edu.jo/centers/icc
1023	E-Courses Accounting page	http://www1.ju.edu.jo/ecourse/acc102/index.htm
1024	E-Courses Biology page	http://www1.ju.edu.jo/ecourse/biology351/banner.htm
1025	E-Courses Computer skills 1 content	http://www1.ju.edu.jo/ecourse/cskills1/contents.htm
1026	E-Courses Computer skills 2 content	http://www1.ju.edu.jo/ecourse/cskills2/index.htm

1027	JU School Home Page	http://www.ju.edu.jo/resources/school
1028	Economic Observatory Home page	http://www.ju.edu.jo/resources/economicobservatory
1029	Faculty Of Graduate Studies Master Page	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies/Master
1030	Faculty Of Graduate Studies PHP Page	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies/Phd
1031	Computer Center PhotoGallary	http://www.ju.edu.jo/centers/ComputerCenter/PhotoGallary/thumbs0.html
1032	Economic Observatory Master Catalog Page	http://www.ju.edu.jo/resources/economicobservatory/catalogs/masterpage
1033	Economic Observatory KeyStaff	http://www.ju.edu.jo/resources/economicobservatory/KeyStaff
1034	Economic Observatory News	http://www.ju.edu.jo/resources/EconomicObservatory/Lists/News/
1035	Economic Observatory Forms	http://www.ju.edu.jo/resources/EconomicObservatory/Pages/Forms/
1036	Economic Observatory ResearchOutput	http://www.ju.edu.jo/resources/EconomicObservatory/Lists/ResearchOutput/
1037	Phone Directory Directions For Using PBX	http://www1.ju.edu.jo/results/PBX/direction.htm
1038	Phone Directory Numbers By Index	http://www1.ju.edu.jo/results/PBX/index.htm
1039	Phone Directory Numbers New Guides	http://www1.ju.edu.jo/results/PBX/NewGuide.htm
1040	Faculty Of Graduate Studies Document PHD Forms	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies/Documents/PhDForms/Word
1041	Faculty of graduate studies Master Forms Documents List	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies/Documents/MasterForms/Word/
1042	Faculty Of Graduate Studies Document Med Forms	http://www.ju.edu.jo/arabicfaculties/facultyofgraduatestudies/Documents/MedForms/Word/
1043	Conference of a Global Renewable Energy	http://www1.ju.edu.jo/conferences1/gcreader/index.htm
1044	Conferences Page Documents	http://www.ju.edu.jo/Pages/Conferences/Documents/
1045	Conferences index page	http://www1.ju.edu.jo/conferences1/index.html
1046	Sportive Conference	http://www1.ju.edu.jo/conferences1/6thPhys/default.htm
1047	Schedule Courses for Buisness administration Faculty	http://www1.ju.edu.jo/results/courses/FBA.htm
1048	Schedule Courses for IT Faculty	http://www1.ju.edu.jo/results/courses/IT.htm
1049	Schedule Courses for Law	http://www1.ju.edu.jo/results/courses/Law.htm
1050	Regestration Page	http://reg.ju.edu.jo/

1051	Faculty of Art and Design Course Schedul	http://www1.ju.edu.jo/results/courses/Design.htm
1052	Schedule Courses for Engineering Faculty	http://www1.ju.edu.jo/results/courses/Eng.htm
1053	Pharmacy Courses shedual	http://www1.ju.edu.jo/results/courses/Pharm.htm
1054	Schedule Courses for Science Faculty	http://www1.ju.edu.jo/results/courses/Sci.htm
1055	Schedule Courses for Sharia Faculty	http://www1.ju.edu.jo/results/courses/Sharia.htm
1056	Schedule Courses for Foreign Languages Faculty	http://www1.ju.edu.jo/results/courses/FL.htm
1057	University accepted students for winter 2009	http://www1.ju.edu.jo/uacwinter2009
1058	Schedual Courses for the Seconed Semester 2008-2009	http://www1.ju.edu.jo/results/courses/index.htm
1059	International Arabic study Scedual Course	http://www1.ju.edu.jo/results/courses/Ara.htm
1060	Schedule Courses for Agreculture Faculty	http://www1.ju.edu.jo/results/courses/Agr.htm
1061	Schedule Courses for Educational Science Faculty	http://www1.ju.edu.jo/results/courses/Fes.htm
1062	Schedule Courses for Art Faculty	http://www1.ju.edu.jo/results/courses/Art.htm
1063	Faculty Of Arts and Design Page	http://www.ju.edu.jo/faculties/FacultyofArtsandDesign
1064	Faculty of Physical Education Page	http://www.ju.edu.jo/faculties/FacultyofPhysicalEducation
1065	Faculty Of Educational Sciences	http://www.ju.edu.jo/faculties/FacultyOfEducationalSciences
1066	Faculty of Engineering and Technology	http://www.ju.edu.jo/faculties/FacultyofEngineering
1067	Faculty of Dentistry Page	http://www.ju.edu.jo/faculties/FacultyofDentistry
1068	Faculty of IT Page	http://www.ju.edu.jo/faculties/FacultyofIT
1069	Faculty of Pharmacy Page	http://www.ju.edu.jo/faculties/FacultyofPharmacy
1070	Faculty of Buisness	http://www.ju.edu.jo/faculties/FacultyofBusiness
1071	Faculty of Sharia Page	http://www.ju.edu.jo/faculties/FacultyofSharia
1072	Faculty Of Science Page	http://www.ju.edu.jo/faculties/FacultyofScience
1073	FacultyofMedicine Home page	http://www.ju.edu.jo/faculties/FacultyofMedicine
1074	Faculty of Rehabilitation Sciences page	http://www.ju.edu.jo/faculties/FacultyofRehabilitation
1075	Faculty of Law	http://www.ju.edu.jo/faculties/FacultyofLaw
1076	Faculty of graduate Studies	http://www.ju.edu.jo/faculties/facultyofgraduatestudies
1077	Faculty of Nursing	http://www.ju.edu.jo/faculties/FacultyofNursing

1078	Faculty of International Studies	http://www.ju.edu.jo/faculties/internationalstudies
1079	Faculty of Foreign languages Page	http://www.ju.edu.jo/faculties/FL
1080	Faculty of Agriculture Page	http://www.ju.edu.jo/faculties/FacultyofAgriculture
1081	Faculty Of Arts Page	http://www.ju.edu.jo/faculties/FacultyOfArts
1082	Faculty of Engineering Documents	http://www.ju.edu.jo/faculties/FacultyofEngineering/Documents
1083	FacultyofMedicine Documents forms	http://www.ju.edu.jo/faculties/facultyofMedicine/Documents/
1084	Faculty Of Law Forms	http://www.ju.edu.jo/faculties/facultyofLaw/Documents/
1085	Faculty of graduate Studies Master Page	http://www.ju.edu.jo/faculties/facultyofgraduatestudies/Master
1086	Faculty Of Nursing Images	http://www.ju.edu.jo/faculties/facultyofNursing/PublicingImages/
1087	Faculty of Nursing Documents	http://www.ju.edu.jo/faculties/facultyofNursing/Documents/
1088	Faculty of Foreign languages Documents	http://www.ju.edu.jo/faculties/fl/Documents/
1089	Faculty of Engineering computer Eng. Overview	http://www.ju.edu.jo/faculties/FacultyofEngineering/computer/
1090	Faculty of Engineering Industrial Eng. Overview	http://www.ju.edu.jo/faculties/FacultyofEngineering/Industrial
1091	Faculty of Engineering Mechatronics Overview	http://www.ju.edu.jo/faculties/FacultyofEngineering/Mechatronics/
1092	Faculty Of Science Mathematics Overview	http://www.ju.edu.jo/faculties/FacultyofScience/Mathematics
1093	Faculty of graduate Studies Master Forms	http://www.ju.edu.jo/faculties/facultyofgraduatestudies/Documents/MasterForms/Word
1094	Faculty of Engineering Industrial Eng department staff	http://www.ju.edu.jo/faculties/FacultyofEngineering/Industrial/departmentstaff
1095	Faculty of Engineering Mechatronics Training Forms	http://www.ju.edu.jo/faculties/FacultyofEngineering/Training%20Forms
1096	Faculty of Engineering Mechatronics pages	http://www.ju.edu.jo/faculties/facultyofengineering/mechatronics/pages
1097	Faculty of Buisness photo gallery	http://www.ju.edu.jo/faculties/FacultyofBusiness/Economics/PhotoGallery

Appendix C. Sample of User Index

User ID	User Name
10001	hbh0063873
10002	Amjadq
10003	-
10004	r.hamed
10005	m.yacoub
10006	Stu
10007	sos0075813
10008	Ayeshg
10009	abr0074296
10010	Weshah
10011	s.habahbeh
10012	s.abuhazeem
10013	mjd0085292
10014	Makash
10015	sam0057543
10016	rgd0059178
10017	e.domour
10018	syr0057454
10019	bna0054194
10020	s.alaqqad
10021	sda0087400
10022	s.herzallah
10023	ala0069030
10024	Reg
10025	MAT2051677
10026	o.toot
10027	j.alhasanat
10028	Salhieh
10029	r.saaideh
10030	hsn0045054
10031	nsr0064606
10032	dan0054007
10033	rza0067279
10034	daa0076155
10035	zyd2050958
10036	j.manaseer
10037	r.farraaj

User ID	User Name
10038	rmz0086713
10039	rym0082475
10040	e.hamarsheh
10041	r.azzah
10042	amr0068367
10043	Asleit
10044	jan0054002
10045	rasafady
10046	kal0060101
10047	a.afifi
10048	tsn0073170
10049	ama0083593
10050	a.suliman
10051	mousa.akhras
10052	r.atour
10053	saj0077708
10054	b.hamoudeh
10055	may2060946
10056	m.abufarah
10057	wad0075838
10058	rsa8060385
10059	AHM0070638
10060	t.haddad
10061	nhamzehn
10062	mra2060370
10063	nsm0076309
10064	a.hayajneh
10065	z.alzboon
10066	mhm9060061
10067	khouryhn
10068	w.alazhari
10069	aljarrah
10070	jarrahs
10071	m.muzaaffar
10072	Khalis
10073	lan0086390
10074	sroran

User ID	User Name
10075	FAt2070487
10076	mra0079301
10077	Hilow
10078	frh0076944
10079	laestej
10080	mhm0040755
10081	mhm0084609
10082	wsc.officemanager
10083	reg.kasit
10084	rba2060590
10085	m.aldmour
10086	zyn0075060
10087	wly0062617
10088	ama0081250
10089	Altahat
10090	sra0063851
10091	Ksakarna
10092	f.alzoub
10093	Tareq
10094	e.juber
10095	mhm0083050
10096	aoy0056880
10097	mhm0081360
10098	a.fannoon
10099	ziad.abuwaar
10100	hny0082411
10101	l.najdawi
10102	t.omer
10103	r.zubi
10104	i.jumaa
10105	d.abu-eid
10106	sba0050023
10107	dym0087591
10108	gdy8080566
10109	dla0059264
10110	sharia.diwan
10111	aml0074191

User ID	User Name
10112	m.hamza
10113	a.hajaya
10114	mhm0060482
10115	M.HUSSAIN
10116	b.alsmadi
10117	e.zyadat
10118	maysaa
10119	theses
10120	ni.khairy
10121	Army
10122	mhm2070739
10123	hbt0077085
10124	reg.tech.sup
10125	r.khalid
10126	roa0062742
10127	h.dasouqi
10128	s.badran
10129	mnd0076055
10130	fad0075080
10131	b.jaber
10132	a.almaaytah
10133	a.sutary
10134	linguistics
10135	ahm0067880
10136	a.zghoul
10137	sarhan
10138	zahdeib
10139	dal0078212
10140	ahm0052933
10141	fat0076753
10142	b.obeidat
10143	s.damrawi
10144	Sfhaddad
10145	Hattarb
10146	d.ammari
10147	r.tarawneh
10148	hbh0056566

User ID	User Name
10149	hny0076839
10150	w.hammouri
10151	zyd2051047
10152	rasha.oirp
10153	adn0057629
10154	a.suyyagh
10155	tma0076835
10156	ALI.HAQ
10157	Ideal
10158	Nkhairy
10159	roa0081681
10160	s.kamal
10161	h.khadrawi
10162	i.shraah
10163	r.yaghan
10164	s.dahiyat
10165	m.omari
10166	hda0086216
10167	Aaaldu
10168	Samia
10169	rsa0063729
10170	lyt0082754
10171	Juhosp
10172	asa0065789
10173	t.abukhalaf
10174	noh0066406
10175	Rshannak
10176	s.khateeb
10177	y.shafout
10178	asa0082502
10179	lc054
10180	reg.social
10181	mry0075295
10182	m.alyaman
10183	mhm0082609
10184	s.oran
10185	mohd.ali

User ID	User Name
10186	b.khatib
10187	R.DAABES
10188	Ataimah
10189	mhm0076405
10190	a.aloun
10191	mhm0062017
10192	l.alsaudi
10193	eslam.sallem
10194	s.shawish
10195	sda0067341
10196	h.jalghoum
10197	Ismeik
10198	rihab33
10199	A.HAFIZ
10200	Farahmajali
10201	m.khresat
10202	m.alzboon
10203	n.almousa
10204	frh0082372
10205	Malrababah
10206	sla0085019
10207	hla0087671
10208	bsa0060202
10209	e.naser
10210	shy2070047
10211	Khaledz
10212	Gabuerei
10213	f.braik
10214	Sweilehb
10215	kh.karaki
10216	rak0057085
10217	Karablie
10218	rna0061667
10219	kyt8081380
10220	n.alshawabkah
10221	s.midor
10222	amr0076445

User ID	User Name
10223	Dnft
10224	g.hamad
10225	Jbakri
10226	Mmubarak
10227	sla0075162
10228	aaa0075168
10229	b.khalidi
10230	Rima
10231	Aleisawi
10232	Awalid
10233	mro0077077
10234	w.demeh
10235	mhm0078474
10236	Tahani
10237	Search
10238	rnd0067393
10239	i.taharwa
10240	s.ajalil
10241	asm7080005
10242	asr0058248
10243	f.dardas
10244	m.asfoar
10245	s.odeh
10246	nazeeh.almanasyeh
10247	kal0038826
10248	aly0076750
10249	a.mkahal
10250	dar.diwan
10251	Obein
10252	n.alassaf
10253	m.qadri
10254	hmz0079008
10255	Toxico
10256	Rihamm
10257	s.abudahab
10258	sch01004
10259	sch01007
10260	m.akhras

User ID	User Name
10261	haz0061398
10262	k.masri
10263	Hosp
10264	Akamil
10265	STU
10266	a.quzmar
10267	k.abusaleem
10268	mhm0057561
10269	m.aldalahmeh
10270	Rahaf
10271	Snaa
10272	ahmad.masadeh
10273	maa0067004
10274	h.dajeh
10275	a.shhab
10276	Samirhab
10277	aly0080774
10278	Agbakri
10279	sumaya.oirp
10280	a_azzam
10281	Mhapip
10282	Ibrdb
10283	lay0085933
10284	Ocu
10285	Idries
10286	m.alfuqaha
10287	aya0068956
10289	Akram
10290	Abunima
10291	z.hawari
10292	w.bajali@ju.edu.jo
10293	reg.sci.agr
10294	i.bajes
10295	tam8062088
10296	a.aldoami
10297	a.asad
10298	a.mashaqbeh
10299	amr2060128

User ID	User Name
10300	z.alzubi
10301	Hanayneh
10302	Abueid
10303	d.abughunmi
10304	Lawad
10305	sam0079079
10306	hmz0046883
10307	Nabeelmgh
10308	f.al-jbour
10309	heba.saadeh
10310	m.alhaj
10311	Pathol
10312	Almomani
10313	Mhshayyab
10314	s.abuhamdah
10315	maha.sobeh
10316	mra0040423
10317	a.mansour
10318	sma9040110
10319	Faisalrub
10320	Linaw
10321	a.majaly
10322	ju.club
10323	A.hamad
10324	a.sabaileh
10325	t.shawashi
10326	m.tawfiq
10327	sro0056898
10328	m.nassar
10329	i.mosleh
10330	t.shatarat
10331	m.kadhim
10332	Hhammad
10333	t.awad
10334	r.kakish
10335	m.saadeh
10336	Hawa
10337	g.abunamous
10338	kly0062084

User ID	User Name
10339	z.darweesh
10340	t.antary@ju.edu.jo
10341	ahm8080577
10342	hna0057500
10343	Hiary
10344	ahm8080578
10345	e.fawzi
10346	Khwaileh
10347	s.abushmas
10348	arlette.nejmeh
10349	h-hodali
10350	roa0057537
10351	hna9020246
10352	Naja
10353	maa0084441
10354	Jjba
10355	o.abughneim
10356	dr.adnanassaf
10357	Hazemh
10358	s.al-jaber
10359	Fbakri
10360	r.momani
10361	j.halabi
10362	abd0087464
10363	b.alkaraki
10364	a_saleh
10365	Aburayan
10366	e.mahasneh
10367	m.khatib
10368	r.rahahleh
10369	w.ramadan
10370	amr0076861
10371	m.khanfar
10372	m.zitawee
10373	aya0072612
10374	ary0064632
10375	shera.oirp
10376	tam0062623

User ID	User Name
10377	Naserm
10378	nda0077404
10379	s.abawi
10380	Yaserkhas
10381	m.al-naim
10382	s.dabbas
10383	Fl
10384	Kabuloum
10385	ahm0077693
10386	Abjaber
10387	g.musleh
10388	Ghasaff
10389	Masalem
10390	e.melhem
10391	Abudahab
10392	sad0061591
10393	ala0072646
10394	maz0046341
10395	Wahid
10396	mat8062475
10397	mhm0065573
10398	bdr0068580
10399	r.arabiat
10400	eye.bank
10401	syr0065543
10402	amr0044056
10403	h.khasawneh
10404	reg.art.phy
10405	MAZ0046341
10406	a.latif
10407	wly0066804
10408	Sanar
10409	rob0059070
10410	HLA0087671
10411	l.shabeeb
10412	haddad66
10413	mhm8080587
10415	sam0038643

User ID	User Name
10416	lbn8061921
10417	Sumayash
10418	h.mousa
10419	s.qatami
10420	abd0060473
10421	Mosnuman
10422	m.odeh
10423	y.shdifat
10424	m.saheb
10425	tma0057687
10426	m.olimi
10427	m.musleh
10428	reg.mgr.tech
10429	Tanash
10430	f.al_shaar
10431	sla0066707
10432	a.karadsheh
10433	reg.business
10434	m.tawalbeh
10435	r.noufal
10436	ahm0036534
10437	asa0041067
10438	Drabeer
10439	l.baniata
10440	emad_fares
10441	b.freihat
10442	Lakhalil
10443	Anasghaith
10444	m.hajtas
10445	Adnan
10446	yas0056027
10447	Aarwa
10448	Yanals
10449	Hassan
10450	Suhaharb
10451	mhd2060028
10452	Alshboul
10453	aam0064725

User ID	User Name
10454	mhn0052824
10455	Deoaah
10456	Mhadidi
10457	a.mohammad
10458	samia.oirp
10459	m.sulaiman
10460	Ramimali
10461	dym8080086
10462	nsr8071410
10463	v.kasabri
10464	a.saedi
10465	fl.diwan
10466	mhm0056914
10467	m.suyagh
10468	m.obidat
10469	Fouadam
10470	h-khlaif
10471	y.alyounes
10472	Badarneh
10473	n.ramadan
10474	a.barhoum
10475	Nora
10476	Samar
10477	Daradkeh
10478	Mshteivi
10479	AHM0064828
10480	a.hamarsheh
10481	aro0079371
10482	mhm0083801
10483	aym0061020
10484	s.arabeyyat
10485	Rahaf
10486	sar0080114
10487	Ymubarak
10488	Fadik
10489	asr0080036
10490	Snaimat
10491	w.sulyman

User ID	User Name
10492	f.hayajneh
10493	Lalbanna
10494	ala0054447
10495	reg.med
10496	a.alawneh
10497	r.barakat
10498	Mkmajali
10499	f.mustafa
10500	s.alnsoor
10501	Amahasneh
10502	fl.da3
10503	mon.saideh
10504	r.abdelkader
10505	a.akhorshaid
10506	muhsen_m
10507	roa0086344
10508	Nfayoumi
10509	aym8071264
10510	Mamoun
10511	z.makhamreh
10512	a.abualees
10513	m.albanna
10514	Shtaywy
10515	e.jallad
10516	s.khawaldeh
10517	ala0077092
10518	a.alrabadi
10519	soq0080398
10520	gdy0068847
10521	Khalilmo
10522	k.mustafa
10523	nda0072520
10524	hla0080431
10525	smr0071510
10526	hna0042974
10527	dan0081464
10528	nan0052884
10529	asy0060821

User ID	User Name
10530	Musaaz
10531	Ahadidy
10532	abt8080936
10533	Ateyyeha
10534	Tjalil
10535	d.abulail
10536	f.jamaeen
10537	Ashrim
10538	Abandah
10539	abd0060515
10540	m.rayyan
10541	i.shara
10542	yo_khl
10543	Eman
10544	Nbanna
10545	b.malkawi
10546	L.swad
10547	Sarie
10548	t.antary
10549	n.kharabsheh
10550	Montaha
10551	DAA0074248
10552	mohammad_ameen
10553	Reham
10554	sky0057383
10555	Qtaishat
10556	daa0074248
10557	ayh0068439
10558	Dmourh
10559	ans0060734
10560	raa8080725
10561	hbh0078743
10562	n.rabadi
10563	mhm0067124
10564	ran0075421
10565	mna0064329
10566	Dnadi
10567	Reema

Appendix D. Sample of Normalized Matrix

0	0.10811	0	0	0	0.032362	0	0.15538	0.009544	0.11765	0	0	0.083333
0	0.081081	1	0	0	0.94741	0	0.073705	0.41941	0.17647	1	0.33333	0.18333
1	0.56757	0	1	1	0.014563	0	0.73705	0	0.5098	0	0.33333	0.56667
0	0.081081	0	0	0	0.005663	0	0.015936	0.006363	0.078431	0	0	0.016667
0	0.16216	0	0	0	0	1	0.017928	0.56469	0.11765	0	0.33333	0.15

الكشف عن نمطية الاستخدام لموقع الجامعة الأردنية باستخدام الخارطة ذاتية التنظيم

إعداد

دانية أحمد محمد هليل

المشرف

الدكتور عمار الحنيطي

ملخص

نظرا لسرعة تطوير واستخدام شبكة الإنترنت و غنى الشبكة بالمعلومات وتشابكها،ظهرت بعض المشكلات والتي منها ربط الوسائط المتعددة بوثائق الشبكة المعرفية و حدوث بعض الالتباس لدى مستخدم الشبكة من كثرة المعلومات و زخمها. الكميات الضخمة من المعلومات على مواقع الإنترنت سببت لدى مستخدم الإنترنت بعض الصعوبات في البحث والتصفح بالإضافة لبعض التعقيدات الموجودة في بعض المواقع. بناءا على ذلك أصبح من الضروري وجود وسيلة أو أسلوب لتخطيط وتنظيم سلسلة المعلومات ودراسة اسلوب استخدام المستخدمين لمواقع الشبكة.في هذه الرسالة نحاول تحليل موقع الجامعة الأردنية من خلال استخدام الخارطة ذاتية التنظيم في دراسة اسلوب تصفح المستخدمين لموقع الجامعة الأردنية بغية التعرف على أنماط الاستخدام لصفحات الموقع و تجميعها و تحليلها. الدراسة بنيت على مدى عدة إجراءات منها إعداد البيانات وتجهيزها ثم بناء و تجميع العلاقات بين المستخدمين والصفحات.ومن ثم القيام بعملية تجميع المستخدمين حسب علاقاتهم وبعدها تجميع الصفحات المترابطة بناء على كيفية استخدام المستخدمين لهذه الصفحات. تم اختبار الدراسة على بيانات لمدة اسبوع من ملفات تصفح المستخدمين لموقع الجامعة الأردنية وأظهرت النتائج أن الخارطة ذاتية التنظيم تسهل عملية الكشف عن أنماط استخدام المستخدمين بشكل أوضح وأكثر سهولة.