INTEGRATING NETWORK ANALYSIS AND DATA MINING TECHNIQUES INTO EFFECTIVE FRAMEWORK FOR WEB MINING AND RECOMMENDATION

VOLUME I

M. NAGI

PHD

2015

Integrating Network Analysis and Data Mining Techniques into Effective Framework for Web Mining and Recommendation

A Framework for Web Mining and Recommendation

Volume I

Mohamad NAGI

Submitted for the Degree of
Doctor of Philosophy

School of Electrical Engineering and Computer Science
University of Bradford
2015

Abstract

Mohamad Nagi

Integrating Network Analysis and Data Mining Techniques into Effective Framework for Web Mining and Recommendation

A Framework for Web Mining and Recommendation

keywords: behaviour analysis, network analysis, data mining, Web usage mining, Web structure mining, association rules mining, clustering, classification

The main motivation for the study described in this dissertation is to benefit from the development in technology and the huge amount of available data which can be easily captured, stored and maintained electronically. We concentrate on Web usage (i.e., log) mining and Web structure mining. Analysing Web log data will reveal valuable feedback reflecting how effective the current structure of a web site is and to help the owner of a web site in understanding the behaviour of the web site visitors. We developed a framework that integrates statistical analysis, frequent pattern mining, clustering, classification and network construction and analysis. We concentrated on the statistical data related to the visitors and how they surf and pass through the various pages of a given web site to land at some target pages. Further, the frequent pattern mining technique was used to study the relationship between the various pages constituting a given web site. Clustering is used to study the similarity of users and pages. Classification suggests a target class for a given new entity by comparing the characteristics of the new entity to those of the known classes. Network construction and analysis is also employed to identify and investigate the links between the various pages constituting a Web site by constructing a network based on the frequency of access to the Web pages such that pages get linked in the network if they are identified in the result of the frequent pattern mining process as frequently accessed together. The knowledge discovered by analysing a web site and its related data should be considered valuable for online shoppers and commercial web site owners. Benefitting from the outcome of the study, a recommendation system was developed to suggest pages to visitors based on their profiles as compared to similar profiles of other visitors. The conducted experiments using popular

datasets demonstrate the applicability and effectiveness of the proposed framework for Web mining and recommendation. As a by product of the proposed method, we demonstrate how it is effective in another domain for feature reduction by concentrating on gene expression data analysis as an application with some interesting results reported in Chapter 5.

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Mick Ridley and Prof. Reda Alhajj, for their professional guidance throughout the course of this PhD research. Their enthusiastic support and ability to motivate people were the instrumental factors in the development of this dissertation. I would also like to thank Dr. Tansel Özyer and Dr. Keivan Kianmehr for the useful discussions and suggestions they gave me from time to time to shape up my research. I am also thankful to the members of the research group of Prof. Reda Alhajj for their support and encouragement during the development of this thesis. I would also like to thank University of Bradford for giving me the opportunity to be enrolled in their graduate program and for providing the excellent environment and support during my enrollment in the program. Finally, I would also like to acknowledge my parents, my brother and sisters, my wife and kids for their unconditional love and support throughout my academic career without which I would not have achieved the target and reached this position.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The last decade of the 20th century witnessed the widespread usage of the Internet and the development of the World Wide web which allowed people to shift from physically communicating printed material to sharing electronic documents, e.g., emails. Further, mobile devices which have dominated since the end of the 20th century have gained tremendously increased popularity and almost every human has at least one mobile device, i.e., around seven billion phones exist worldwide (refer to Wikipedia for the distribution per country `http://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use`). Another development that has seriously influenced humanity is the introduction of web 2.0 which marks the evolution and revolution of social media platforms. As a result, people started to spend more time in the virtual world than in the physical world. Communication between people and the availability of electronic information lifted traditional borders and shortened the separating distances. All the cyber world is easily reachable and available 24/7 to everyone connected to the Internet.

Public and private sectors, large and small businesses, and even individuals collaborate and sometimes compete to best use and maximise the benefit from the new technology. This has been realised as the development of web sites which allow their owners to widely share their products, experience, even thoughts and concerns. Some web sites are open to the public and there are no restrictions on their access while other web sites provide limited and secured access. These cover a wide scope of domains from online banking to online shopping to blogging and review posting, etc. Further, these web sites are hosted by servers which keep track of a log to record all access patterns by all visitors. Further, every site has its own structure as a number of web pages which are linked by hyperlinks which can be represented by an arbitrary directed graph structure. Each web page has its own content which may consist of text, images, links to other pages, etc. As a result, the web turns into a rich source for knowledge discovery leading to valuable information, including web usage (log or access patterns) mining, web content mining, and web structure (hyperlinks) mining. All these three aspects are handled in this thesis by developing a framework which integrates network analysis, frequent pattern mining, clustering and classification leading to recommendations regarding restructuring of a web site and suggesting pages to be accessed based on historical access patterns. However, we concentrate only on the text (content) around hyperlinks in the web-content mining process. Detailed web content mining is outside the scope of this thesis.

1.1 Benefits of web sites for Online Shopping and Sharing

Electronic communication, sharing and filing of data is becoming the common trend in the community. People tend to use the web for various aspects of daily life, e.g., to find certain commodities, to search for a job, to communicate with family and friends, to read social media postings as well as traditional media, even to watch movies and live broadcasting TV stations, etc. This is a more common in developed countries and recently started to dominate in developing countries. I have developed this claim from my own travel experience based on visits to various countries in the Middle East, Europe, North America and North Africa. The whole trend is driven by the wide availability of Internet access at lower cost affordable to various sectors and levels of income in the society. In fact, from my interaction with people from different backgrounds in different countries and who have different levels of education ranging from elementary school graduates to PhD degree holders, I may claim that it is very uncommon to find someone who does not use the Internet on daily or at least weekly basis. It is anticipated that a very large percentage of sales transactions are mostly completed online. Commercial online transactions include the digital purchase of products and services and also the sale of text-based and banner-based advertising, among others. The success of online transactions is due to the fact that sellers and advertisers can capture and maintain online a large amount of information related to their customers, because customers' behaviour and digital footprint are traced without prior notice. Such valuable data is stored in log files, which consist of very large sets of data summarising the behaviour of web users during their visits to the specific web site; this is not dependant on cookies all just done on server.

Most of the traditionally available web sites follow a static trend in the sense that they treat all users the same, and will not smoothly adapt to users' specific needs, expectations or interests. This is a negative feature of most traditional online shopping web sites [64]. Fortunately, the trend is changing rapidly; certain online retailers and commercial portal web sites attribute their success greatly to their ability to anticipate and address the needs of their customers. Owners of these web sites depend on this advantage to attract more visitors to: (1) browse and maybe buy their products and services online, or (2) visit their portal more regularly. As companies today begin to focus their efforts on cost savings and effective sales, there is an increased interest in moving to online sales fulfilment in order to drive higher revenue. One of the main advantages of online sales is the 24/7 availability of the products and the increased connectivity to cover every worldwide location with Internet access. One of the main keys to maintain a successful business based on online sales is in the ability to capture, maintain and process access and utilisation patterns to predict and report the needs and expectations of online customers [88]. The analysis of users' behaviour would allow online shopping web site developers to increase users' experience and in return achieve an increase in revenue and overall customer satisfaction, even if not immediately. However, the current technology is still not capable of making full use of the log data for knowledge discovery. Some statistical measures maintained in the log data include page hits, frequent page sequence counts, etc. But the target should be understanding the insight of the browsing behaviour (i.e., what typically happened) and then translating this insight knowledge into foresight knowledge that would help Website owners to adjust their business policies with anticipation that some other things are likely to happen as an extension of the captured historical trend.

Explicitly speaking, customers of an online shopping web site typically exhibit individual patterns and behaviour as they shop online; not all customers follow the same path to search the web site or complete a sale [64]. However, it is possible to capture trends as the amount of data maintained in the log increases; researchers understand well the value of the captured data. In fact, there are many factors that influence how people act within the society in regards to the market, such as recession or market boom. Companies that wish to succeed in selling their products and services online need a way to analyse the behaviour of their existing online customers in order to further optimise and maximise the revenue capabilities of these customers. This need has been realised in the research community and several algorithms exist today that allow companies to analyse this online behaviour. However, we wish to analyse data and cluster customers as similar unique groups instead of a single group. One of the main aspects of the implementation of the research agenda as described in this dissertation will represent current market trends and help to predict and suggest alternative ways in which online shopping web sites or portals can maximise the revenue capabilities of their owners. This is investigated by applying web mining and social network analysis techniques. To be specific, one of the targets is to answer some typical commercial web site owner's concerns regarding where, when and what to promote. Properly and wisely handling these concerns would lead to more targeted promotion and hence will lead to huge cut on the unnecessary costs.

1.2 Overview of the Proposed Framework

Web-based applications are dominating every aspect of the daily life by extensively covering both public and private sectors. The web has become the main venue for communication and everything is available online from commodities to scientific research to public information, to games, etc. Every business, whether small or large, is trying to maximise its availability by going on the web by developing web sites ranging from very simple and static to very sophisticated and dynamic. The main theme is to attract the attention of the huge number of Internet users surfing the web. Once a web site is visited a trace is captured by the system and stored in a log which is not even noticed by the visitor. The log is a valuable resource for knowledge discovery leading to better satisfaction of the visitors with the hope of increasing their number in a way to achieve the target of maximising profit and/or benefit. On the other hand, a web site consists of a number of web pages which are linked together with the intention to facilitate for smooth surfing that attracts more users.

The work described in this thesis has been motivated by the fact that the structure of a web site may not satisfy a larger population of the visiting users who may jump between pages of the web site before they land on the target page(s); this is at least partially true because usually access patterns are not known when the web site is designed; they are rather anticipated and hence the design needs to be revisited after it is put in practice. We developed a robust framework that tackles this problem by considering both web log data and web structure data to suggest a more compact structure that could satisfy a larger user group. The study assumes the trend recorded so far in the web log reflects well the anticipated behaviour of the users in the future. We separately analyse web log and web structure data using various techniques, namely frequent pattern mining, clustering, text analysis, statistical analysis and network analysis.

1.2.1 Overview of the frequent pattern mining based method

Frequent pattern mining is a powerful technique capable of identifying in a set of objects (called items) those which demonstrate similar behaviour. More details are presented in Chapter 2, but a brief overview is presented here for the reader. To understand frequent pattern mining, consider a supermarket where customers purchase patterns are kept as transactions each includes a set of items purchased together. Analysing the set of transactions may lead to an understanding of items that are frequently purchased together. This information may be valuable in planning for promotions on items and for the layout of items on the shelves. The same methodology applies to pages of a web site by considering pages as items and users or IP addresses as transactions to determine sets of pages which are frequently accessed together. A set of items is said to be frequent if it includes items that exist together in a number of transactions larger than a predefined minimum threshold which is mostly specified by an expert or may be determined from the data by applying machine learning techniques like hill-climbing. A frequent set of items A is closed if and only if there does not exist another frequent set of items B such that $A \subseteq B$ and A and B have the same frequency. The number of closed frequent itemsets is small compared to the total number of frequent itemsets. Some closed itemsets are maximal: these are closed frequent itemsets that are not subsets of any frequent itemset; they are usually large in size and hence serve our purpose to produce sets of pages frequently accessed together.

To determine maximal closed frequent itemsets we proceed as follows. First we discretize the usage patterns of web pages into binary values such that an entry is set to 1 to indicate that the user visited the page and to 0 to indicate that the user either did not visit the web page or passed by it without spending time. Second, we determine frequent itemsets by applying the classical Apriori algorithm (refer to Chapter 2 to understand how Apriori works). Here it is worth mentioning that other frequent pattern mining algorithms may be applied to produce the same result. The main reason for using Apriori in our approach is its simplicity. Regardless of the algorithm used for frequent pattern mining, this task allows us to process web log data to determine all frequent itemsets of pages which have been mostly visited together by users. Third, from the determined frequent itemsets, we add to the set of closed frequent itemsets the largest frequent itemsets which are frequent itemsets that are not subsumed by other frequent itemsets. Forth, we concentrate on the remaining frequent itemsets which have not been added to the closed group yet. From the latter collection, we add to the closed group any itemset that satisfies the following three criteria: it is large compared to the frequent itemsets not yet added to the closed group; it is not subsumed by other frequent itemsets not yet added to the closed group;

and it has larger frequency than all the closed itemsets that subsume it. Finally, from the identified closed frequent itemsets we select the ones not subsumed by any other closed itemset; these are the target maximal closed frequent itemsets. After all maximal closed frequent itemsets are identified, they are used to guide the recommendation to be issued regarding the maintenance of the web site. Maximal closed frequent sets of pages form a valuable source of knowledge to enrich the outcome from the network analysis technique. Unless explicitly mentioned otherwise, we will use “frequent itemsets” in the thesis but we [are / will mean] concentrating on maximal closed frequent itemsets in the analysis.

1.2.2 Overview of the network analysis based method

Network analysis is a powerful model which could be effectively used to tackle a number of problems, including recommending restructuring of web sites as demonstrated in this thesis. A network could be used to model a given problem by identifying two constructs, actors and links. Refer to chapter 2 for more detailed coverage; here we provide a brief overview to allow the reader to understand better the proposed framework. Actors may all form one category leading to one-mode network. It is also possible to have actors from two (may be more) categories leading to a two (or higher) mode network. In the setting of the problem tackled in this thesis, there are two sets of actors, namely users or IP addresses and pages. A link between two actors reflects a kind of relationship. While links connect actors within the same group in one mode networks, in two (or higher order)-mode networks a link connects two actors if they are related and they belong to two different groups.

A network is represented as a graph which could be mathematically manipulated by considering the corresponding adjacency matrix. For one mode network the adjacency matrix is square where the diagonal reflects self references. On the other hand, the adjacency matrix that corresponds to a two mode network has one type of actors as rows and the other type as columns. Accordingly, for the web mining framework presented in this thesis we produce a two mode network where a row corresponds to a user or IP address and a column corresponds to a web page. An entry (*i, j*) indicates whether user *i* visited page *j*. This is extracted from the web log.

Folding is applied on the two mode network to produce a one mode network of the pages. The folding process works as follows. We get the transpose say *T* of the original adjacency matrix say *A* which has one row corresponding to each user and one column corresponding to each page. We multiply the two matrixes $T \times A$ to get a matrix *Y* where rows and columns are pages and hence it is the adjacency matrix of the one mode network between pages. Each entry (*i, j*) in matrix *Y* indicates how significant are the two pages by considering all users who accessed the web site.

The one-mode network of pages is preprocessed by discretising the weights of all edges such that zero weight (edge is eliminated) is assigned to each edge which has weight lower than the average weight of the edges in the graph; weight one is assigned to all other edges. The discretised graph is processed further to find communities of pages. A community of pages includes a set of pages which are more connected to each other than to other pages outside the community. We obtain communities by eliminating edges (links) from the graph by considering the betweenness centrality. For each edge, its betweenness centrality is determined as the number of shortest paths that pass through the edge. The higher the betweenness centrality the more becomes the edge a candidate to be removed. Edges are removed from a network based on two criteria, they should have high betweenness centrality and their removal should lead to more communities in the network. Our strategy for the network of web pages is to satisfy either of the following two constraints, the one that could be achieved first: (1) to have the number of communities equal to the number of clusters produced by the method presented in Chapter 4; (2) not to remove any edge whose betweenness centrality is below the average betweenness centrality of all the edges in the network. This will lead to reasonable number of communities where each community is somehow homogeneous. We select from each community a representative which has the highest average of two centrality measures, namely degree centrality and closeness centrality.

1.3 Main Contributions

The work described in this thesis covers some main aspects of web mining in a unified framework which analyses web log for access patterns and web structure by employing data mining and network analysis techniques. The engine of the proposed framework encapsulates modules for frequent pattern mining, association rules mining, clustering, classification and network analysis. These modules interact and complement each other to achieve a final goal of providing legitimate recommendation to web site users and owners regarding the suggested pages to visit next and better structure of the site, respectively. This is achieved by analysing the web log data captured over time to find trends in access patterns. The captured trend will benefit users and owners in various aspects. First, pages may be recommended to new users based on access patterns of other users who followed the same trend. Second, it helps, when combined with web structure mining, in restructuring the web site to facilitate better accessibility and hence the opportunity for increased profitability.

The recommendation of web pages to access next will depend on clustering of web users to help in building a classifier capable of determining the most convenient group similar to a given user. The network of users and pages will be analysed to find the most influential nodes combined with various other features to be extracted from the network. The test results reported in the thesis demonstrate the effectiveness and applicability of the proposed approach. We further illustrated how the proposed framework is powerful and effective as a general solution capable of tackling other domains by showing its applicability for feature reduction with encouraging test results from the gene expression data analysis domain where the target is to reduce the number of genes (or molecules) to be investigated further as potential disease biomarkers.

The main contributions of the work described in this thesis may be articulated as follows:

- Using frequent pattern and association rules mining for web usage and web structure analysis leading to effective recommendation system. Though first proposed for market basket analysis, the study described in this thesis demonstrates the effectiveness of frequent pattern and association rules mining in tackling a variety of domains. For the scope of this thesis, we conducted two studies. First we realised sessions as transactions and pages visited as items to discover pages which are frequently visited together. Second for a given web site we considered each page as transaction and the links (leading to pages) within the page as items to find pages which are linked to by most of the pages constituting the web site. As a result, from this contribution we learn how to map a given domain and prepare it for processing by frequent pattern mining and association rules mining algorithms. This adds another perspective to tackle real world problems and may draw the attention in some interesting directions not recognised earlier.

- Using clustering and classification for deciding on similarity between web users leading to enhanced recommendation by considering users with similar behaviour. Indeed clustering and classification are machine learning techniques intended to group similar users together. The most important step before applying clustering or classification is data preparation which shows how the specific domain to be investigated could benefit from such techniques. In this study we considered once users as objects and the pages they visited as their features and in a second setting we assumed objects are web pages and their features are the links (to pages) they contain. This way, we demonstrate once more the strength of clustering and classification in handling new domains. The clustering of users will lead to groups such that each group contains users who have similar access trends to pages of the given web site; these users mostly share the same interests. A classifier could capture the characteristics of each group of users to produce a model capable of determining the most appropriate group to host a new users based on his/her access pattern. By the same way, we produce clusters of pages from the second setting mentioned above such that within the same group we put clusters who have a large number of links in common. These are mostly pages with common or similar theme and hence any classifier to be build should be capable of determining the most fitting group for a new page by analysing the links it contains.

- Using social network analysis to further enrich the outcome from the data mining tasks leading to a more robust framework for analysis and recommendation. A network could be constructed for any domain which contains some related objects. The network will include one node for each object and links between object reflect their relationship. A network model is attractive because of the underlying strong mathematical foundation which could be employed for comprehensive and detailed analysis leading to knowledge discovery. In this study we demonstrated how web analysis and optimisation can benefit from the rich network model. In two different settings we realised users as nodes and web pages as nodes, respectively. In the first setting, two users are linked based on their access patterns to web pages, while in the second setting two pages are linked based on the high overlap between the links they contain. This allows us to apply some network clustering techniques to split the network into islands of related users or web pages. The outcome may be compared with the result of the employed clustering technique from machine learning. This may produce a degree of confidence in the utilised clustering algorithm.
- Demonstrating the effectiveness of the proposed framework for domains other than web mining; in particular feature reduction has been investigated with experiments conducted to show how it is possible to reduce the number of genes which could be considered as disease biomarkers. Actually finding disease biomarkers is an interesting research problem in bioinformatics. In the study described in this thesis we demonstrate how it is possible to use a combination of machine learning and network analysis techniques for feature reduction, in our case to reduce the number of molecules to be accused and further analysis as causing a specific disease. Such molecules could be analysed further in a web lab under the control of a domain expert. Understanding how the medical domain has been set and mapped for appropriate analysis by the proposed approach will guide

researchers to adapt the same approach to achieve feature reduction in other domains like image processing.

1.4 Thesis Organisation

This thesis is organised in six chapters, including this introduction chapter. Chapter 2 cover an overview of the related work and the basic background required to understand the various techniques used in this document. In particular, association rules mining, clustering, classification and network analysis and briefly presented. Chapter 3 presents how frequent pattern mining and network analysis can be combined into an effective approach for web usage mining. Chapter 4 describes an approach that combines data mining and network analysis for web structure analysis with emphasis on how the outcome from the analysis makes it possible to suggest web site restructuring. Chapter 5 shows how the developed approach is powerful enough to handle other domains. Specifically, the problem of feature reduction has been addressed to find a reduced set of genes or molecules which could be considered as disease biomarkers. This has been achieved by considering the relationship between molecules as equivalent to the relationship between web pages and hence the same methodology could be applied to both domains. Chapter 8 summarises the discoveries, the contributions and the future research directions.

Chapter 2

Literature Review

In this chapter, we provide a detailed survey from the literature about the works that are related to the contributions of this dissertation. Other than our already published work as listed in the front pages, we do not believe that there is anything described in the literature that overlaps or is similar to the work described in this thesis. However, there are some other efforts which apply various techniques in isolation and from other perspectives. Before describing the related work, it is important to provide the background necessary to understand the various components of the framework described in this thesis. Actually, the effectiveness and strength of the proposed framework are attributed to the integration of the various techniques from data mining and network analysis. These techniques complement and support each other leading to the novel integrated approach described in this thesis.

This chapter mainly covers the following aspects: association rule mining in Section 2.1, clustering and classification in Section 2.2, web usage and web structure mining in Section 2.4, social network analysis in Section 2.3, and the related work in Section 2.5.

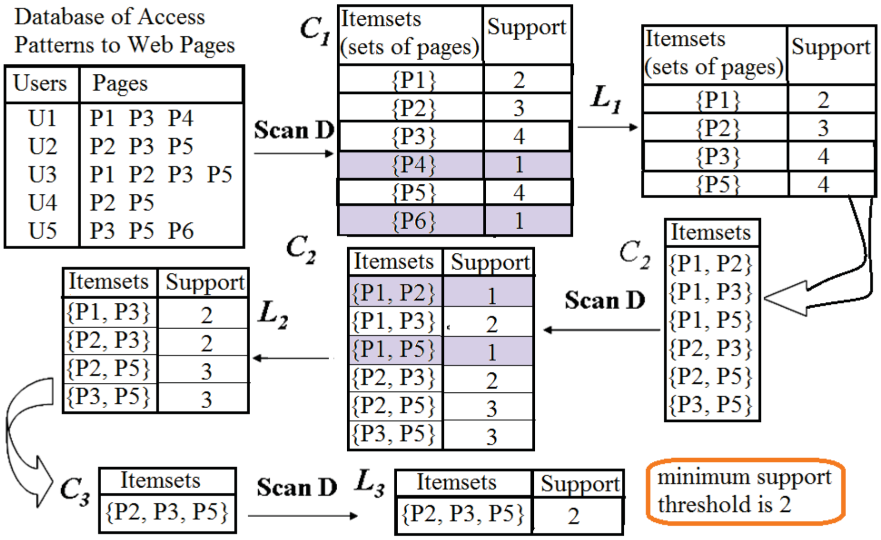
2.1 Association Rule Mining

Association rules mining is one of the main tasks under the general data mining umbrella. Data mining refers to knowledge discovery in data, i.e., it is the process of extracting knowledge mostly hidden in a given data. One of the main attractions of data mining is the shift from retrieving information explicitly stored in a data repository as provided by traditional query languages to finding correlations hidden in the data. The target is to maximise the benefit from data by turning it from passive into active. As a result the engine dealing with the data moves from being just a loading truck moving data in and out of the repository to a complete factory capable of processing the data to produce new discoveries which would be impossible to realise without encapsulating the factory within the engine. Association rules mining is one of the tasks incorporated within the factory. Other possible data mining tasks are clustering and classification which are described in Section 2.2.

Association rules mining (ARM) was first proposed by Agrawal in 1993 (see [6] for details) to process transactional data kept by retail stores. However it is a general method that works on any data which reflects a many to many relationship between two sets of entities, e.g., transactions and items. Using retail stores data the process leads to sets of items that are mostly purchased together by customers. Then from each frequent itemset rules are constructed using the general form X →Y, where X and Y are non-empty disjoined sets of items referred to as antecedent and consequent, respectively. The value of a rule depends on a confidence measure which is computed by concentrating on transactions that contain items in the antecedent and then find from them the percentage of transactions which contain the consequent. In other words, an association rule reflects the correlation between sets of items (hereafter called itemsets); items in our model are pages.

Association rules mining is a two-step process, first frequent itemsets are found and then association rules are generated. An itemset is frequent if it is supported by a certain percentage of the transactions larger than a prespecified minimum threshold value (mostly specified by a domain expert). In our settings, transactions are user sessions that include a listing of the pages visited by the user. Once all frequent itemsets are identified, each frequent itemset is used to derive all possible rules such that the antecedent and consequent of the rule are non-empty, disjoint and their union leads to the itemset. A rule is acceptable if it has high confidence, i.e., the support of all its items divided by the support of the items in the antecedent is larger than a predefined minimum confidence value (mostly specified by a domain expert).

Given a set of items, say *I*, and a transactional database where each transaction is subset from *I*, regardless of the number of transactions in the database, the maximum number of candidate itemsets to be checked is $2_{|I|}$, where $|I|$ is the cardinality of *I*. However, in general not all of these itemsets are frequent. Depending on the cardinality of *I* and the size of the available memory the brute force approach of generating and checking all possible itemsets may be feasible and efficient (requires only one database scan) for small databases with small number of items. However, it is not computationally or space efficient to generate and determine the frequency of a huge number of itemsets which is the general case in real application domains. This motivated researchers to investigate the possibility of pruning the number of database scans or the number of itemsets to be checked. For instance Apriori [6] enforces pruning by keeping only frequent itemsets.


Database of Access Patterns to Web Pages
Scan D
C_1
L_1
C_2
C_2
Scan D
L_2
C_3
Scan D
L_3
minimum support threshold is 2

Users	Pages
U1	P1 P3 P4
U2	P2 P3 P5
U3	P1 P2 P3 P5
U4	P2 P5
U5	P3 P5 P6

Itemsets (sets of pages)	Support
{P1}	2
{P2}	3
{P3}	4
{P4}	1
{P5}	4
{P6}	1

Itemsets (sets of pages)	Support
{P1}	2
{P2}	3
{P3}	4
{P5}	4

Itemsets
{P1, P2}
{P1, P3}
{P1, P5}
{P2, P3}
{P2, P5}
{P3, P5}

Itemsets	Support
{P1, P2}	1
{P1, P3}	2
{P1, P5}	1
{P2, P3}	2
{P2, P5}	3
{P3, P5}	3

Itemsets	Support
{P1, P3}	2
{P2, P3}	2
{P2, P5}	3
{P3, P5}	3

Itemsets
{P2, P3, P5}

Itemsets	Support
{P2, P3, P5}	2

Figure 2.1: Example to illustrate how the Apriori algorithm finds frequent patterns of pages based on some given access patterns by a number of users

The interest in frequent pattern mining in the machine learning and database community is attributed to the work by Agrawal in 1993 when he proposed the Apriori [6] algorithm for market basket analysis. Apriori applies an iterative process to find all frequent itemsets in a given transactional database of items starting with the singleton itemsets until the largest frequent itemsets are determined. First one database scan is required to determine the frequency of each singleton itemset. Then for each subsequent iteration $k > 1$, only frequent itemsets from iteration $k - 1$ are used to determine candidate itemsets of size k and a database scan follows to determine and keep only frequent itemsets of size k. In other words, each iteration of the algorithm requires a complete database scan to compute the frequencies of the candidate itemsets. Considering only frequent itemsets at level k to produce candidate itemsets of size k+1 is the main property which turns Apriori into a feasible and attractive algorithm because it helps to reduce the search space for the itemsets. The whole process is illustrated in the example shown in Figure 2.1.

Here it is worth emphasising that the Apriori algorithm may become infeasible in case it faces the worst case scenario where all or most of the generated candidate itemsets are frequent. This may be seen similar to the brute force scenario. In other words, almost all possible itemsets will be generated and kept when the minimum support threshold is low and the size of the largest frequent itemset(s) is close to the total number of the available items. To overcome this shortcoming, several other Apriori-like algorithms, such as DHP [76], DCP [73], and DCI [74], which mainly focus on improving the performance of mining by reducing the candidate generation and/or by introducing special data structures that reduce the time for computing the support of candidates. On the other hand, algorithms like FP-growth [48] try to improve the performance by reducing the number of database scans.

FP-growth is based on a novel data structure, namely, the Frequent Pattern (FP)-tree [48] which is a compact data structure that can be constructed by scanning the database only twice. The first scan is used to determine the frequency of each item in the database. Only items that satisfy the minimum support threshold are kept and sorted in descending order to form what is called the header table. Each entry in the header table includes a frequent item and a link to its first occurrence in the FP-tree. Each path in the FP-tree from the direct children of the root to a specific leave is shared totally or partially by $n \geq 1$ transactions. Other than the root node, each node in the FP-tree contains information related to a specific frequent item in the database. This information includes the name of the item, its frequency and a link to the next occurrence of the item in the tree, if any. The second scan of the database is used to construct the FP-tree as follows. Each transaction is read, sorted by the same order as the header table, truncated by eliminating all items not in the header table and the sorted remaining sequence of items are inserted in the FP-tree. Once the tree is built, the frequent patterns can be mined by the FP-growth algorithm [48].

Another approach that constructs a tree structure from the database was developed by Amir *et al.* [9]. Their approach produces a trie structure from the transactions in the database by employing the following two steps. For each transaction, elements of its powerset are determined, then each nonempty element of the powerset is inserted in the tree. Each node in the tree includes the name of an item and the frequency of the itemset which contains the elements from the specific item up to the root. Thus, the trie structure is infeasible for low support threshold and large number of items. In other words, the trie structure based approach would work for small data and hence it is not practical in our case.

All the above mentioned algorithms suffer from the memory bottleneck because they expect all the data structures used for determining the frequent itemsets to fit in main memory. As the number of items and the database size increases the amount of memory required will increase and hence scalability becomes an issue. Handling the scalability problem is outside the scope of this thesis.

Eclat [21] is another frequent pattern mining algorithm that works well for databases with small number of items. Instead of using horizontal database format, the Eclat algorithm uses the vertical database format: associated with each item there is a cover which contains transaction identifiers (i.e., TID)-list of the transactions where the item appears. The vertical database format has the advantage that it improves the computation of the support by reducing the effective portion of the database to be scanned. In fact, the support of an itemset *P* can be computed from the intersection of the covers, *cover*(*Q*) ∩ *cover*(*R*), of its two subsets *Q* and *R* from which candidate itemset *P* was generated. For example, if the cover (TID-list) of an itemset **CD** is *{*t_2, t_4, t_5, t_6*}* and the cover of another itemset **CW** is *{*t_1, t_2, t_3, t_4, t_5*}*, then the cover of the resulting candidate itemset **CDW** is *{*t_2, t_4, t_5*}*, with a support count of 3. Although the Eclat algorithm reduces the number of database scans, it has to keep item covers in main memory, i.e., Eclat is very much memory intensive. Moreover, the number of candidates generated using this approach is typically larger than the breadth first approaches because Eclat does not fully exploit the monotonicity property.

2.2 Clustering and Classification

Clustering and classification are learning techniques [23, 45], which attempt to capture the characteristics of a given data and develop a process capable of classifying data instances based on their characteristics. While clustering is known as unsupervised learning technique, classification refers to supervised learning. In other words, clustering is expected to capture the characteristics and classify the given data instances by depending only on the characteristics without external input.

Given a set of data objects, each has its own features, and a similarity measure, clustering refers to the process of categorising a set of objects into groups such that objects in each group are characterised by high similarity while the groups are characterised by high dissimilarity. There are two main trends known as crisp and fuzzy clustering [103]. The latter allows the groups to overlap and the former produces disjoint groups such that each group is a collection of homogeneous objects and the various groups are heterogeneous. Thus the process involves the optimisation of two objectives: maximising the homogeneity within each group and maximising the heterogeneity across the groups. Further, to produce balanced clusters it is recommended to consider as a third objective maximising the number of objects per group. Similarity between the objects and between groups can be measured using an appropriate function including Euclidean or non-Euclidean distance functions [23, 59]; the choice is driven by the characteristics of the objects to be clustered.

There are a wide variety of clustering algorithms, each has its own advantages and disadvantages. One of the most commonly used and may be the oldest known clustering algorithm is k-means [59]. It is a popular partitioning based clustering algorithm because it is easy to implement and it produces a convenient result fast. The algorithm works as follows. Given a set of objects, a similar measure and the anticipated number of clusters, say k, as input, select k initial seeds (called hereafter centroids) one per cluster and proceed to iteratively execute the following steps until the clusters become stable, i.e., none of the objects changes its cluster: (1) find the similarity of each object with each of the k centroids; (2) add each object to the cluster of the most similar centroid; (3) recompute the centroids. Though simple, k-means is criticised for three deficiencies [59]: (1) requires the number of clusters to be known in advance and supplied as input parameter; (2) can get stuck in local minima; and (3) the final clusters depend on the selected initial seeds to start with, i.e., different distribution of the objects into clusters will be produced when the initial seeds change. As an alternative to k-means, it is possible to use multi-objective genetic algorithm based process to cluster a given set of objects by trying to find an optimal solution that satisfies multiple objectives, i.e., high similarity within the clusters and high dissimilarity across the clusters [75].

Classification is in general similar to clustering. However, for classification the target is to build a model that inspires characteristics of known objects or instances from a set of prespecified classes in order to successfully predict the class of new instances (forming the test set) not part of the set of instances used in building the model (the training set). In other words, given a set of objects such that the class of each object is known, the classification process involves two steps: (1) model construction, and (2) model testing. The first step is to divide the given data into two disjoined subsets, one of them is called the training subset and is used in constructing the classifier model. The other subset is called the test set and it is used in testing the classifier to decide on its accuracy. One of the common ways to decide on the split between training and test subsets is to divide the given set of objects into *k* subsets, and iteratively use (*k*−1) subsets as the training set and the remaining subset as the test set. After *k* iterations, the accuracy of the classifier is determined as the average accuracy of the *k* runs. This is called *k*-fold cross validation [89], with 10-fold as the ideal case preferred. However, in practice [89] people go up to 5-fold because the process is time and effort consuming. Finally, as a by-product of the association rules mining task, it is possible to build a classifier called associative classifier. This is possible by considering only rules where the consequent is only the class variable and keep in the classifier model only such rules when they have high confidence.

2.3 web Usage and web Structure Mining

The web is growing rapidly into a major source for information and knowledge discovery [5]. A web site may consist of a number of web pages structured and linked such that it is possible to navigate between the various pages to seek more specialised or generalised information related to the specific domain covered by the web site. Three main types of information are associated with any web site; these are content each page expressed as text whether structured, semi-structured or unstructured; links which connect the various pages of a web site and also connect a web site to other external sites; and web access patterns captured by the web serve and kept in a web log file. Thus, it is possible to apply data mining techniques on each of the three mentioned types of information. In other words, it is possible to talk about web content, web usage and web structure mining [55]. The result from each source will reveal certain interesting aspects related to the specific web site. Once combined, the three results may form a rich source for guiding visitors and owners of web sites to help them perform better. The outcome may be used to solve the web site structure and content optimisation problem.

Web structure mining requires locating all links within the pages constituting a given web site. The extracted links will then be used to generate a directed graph representing the overall structure of the web site. Each node within the produced graph represents a page from the web site; thus the number of nodes is the same as the number of existing pages. And each edge is a link between two pages to represent a direct hyperlink from the first to the second page. The complexity of the graph and its structure depend on the flexibility and professionalism employed in developing the web site. Some pages may be buried deep in the structure though they might contain some important information mostly targeted by visitors of the web site. Not all visitors are experienced and patient enough to dig deep and surf the various pages leading to their target page(s). Thus it is essential to identify such important pages and pull them up to the very front close to the main page of the web site in order to maximise the number of satisfied visitors.

Web usage mining concentrates on the log captured by the servers dedicated to watch out a given web site. In other words, all clicks by a user as recorded in the log file include details like the IP address of the Internet connection from which the user accessed to web site, the pages accessed, the time, etc.

Thus web usage mining involves the processing of the logs captured and maintained by the web server to study and track individual user behaviour. For each visit from a particular IP address, the pages visited, total visits and total time spent looking at the page, also known as “think time”, are captured by the server and added to the web log. Then these values are parsed from original web server logs, or could be taken from preprocessed logs as well.

The log file is divided into sessions and each IP address may be considered as a different user. The target is to find the pages visited by each user. This leads to a two dimensional table where each session or IP address forms a row and each page forms a column. The table can be analysed to cluster users together based on the similarity of their access patterns in terms of pages accessed. The same table may be processed to find the sets of pages frequently accessed together by a considerable number of users. Both results would reveal some interesting knowledge which will lead to informative decision making when a recommendation is to be issued.

Web content mining, on the other hand, helps in determining the anticipated correlation between the context around the link and the page to which the link points. Also, web content mining may help in classifying the various components of a given web page and how they are related. This may help in restructuring the whole web site into a more cohesive and concise counterpart in terms of content per page and links. However, this is outside the scope of the study described in this dissertation which rather concentrates on web usage and structure mining. We touch on web content by only looking around the links within the page and we ignore all the other content of a page.

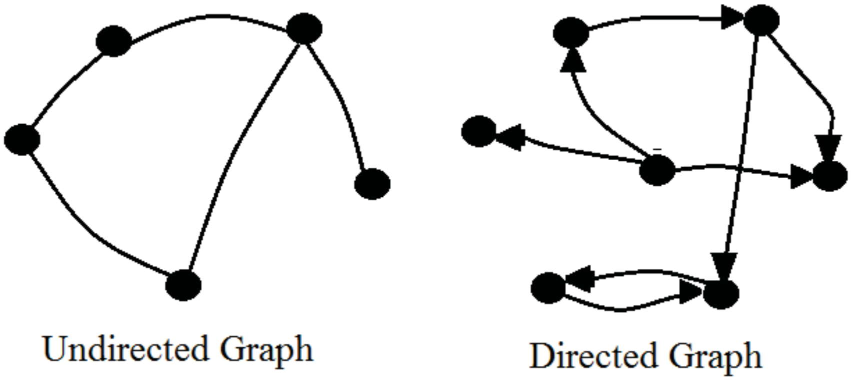


Figure 2.2: Examples of directed and undirected graphs

2.4 Social Network Analysis

A social network is a connected structure which consists of nodes and links. Nodes represent actors and links reflect a relationship between actors that exist in the network. In a social structure actors are humans and semantics of the relationship depend on the specific domain and scope to be covered in the model. Thus, a graph is the most attractive representation for a social network. Graphs may be classified in various ways based on the types of links and the types of nodes. Based on links, two main categories of graphs are directed and undirected graphs as shown in Figure 2.2. In directed graphs each link has a source and destination, and each node is characterised by two types of degree: indegree of a node, say *n*, is computed based on the total number of links having node *n* as destination and outdegree of node *n* is determined based on the total number of links having node *n* as source. On the other hand, source and destination are not specified in an undirected graph. Accordingly, a node *n* in an undirected graph has a degree which is computed as the total number of links connected to it.

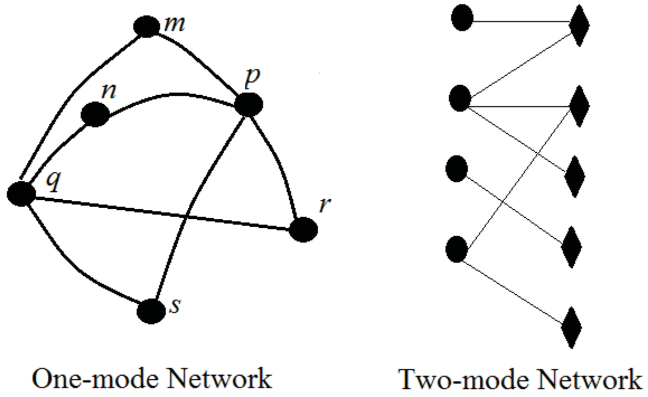
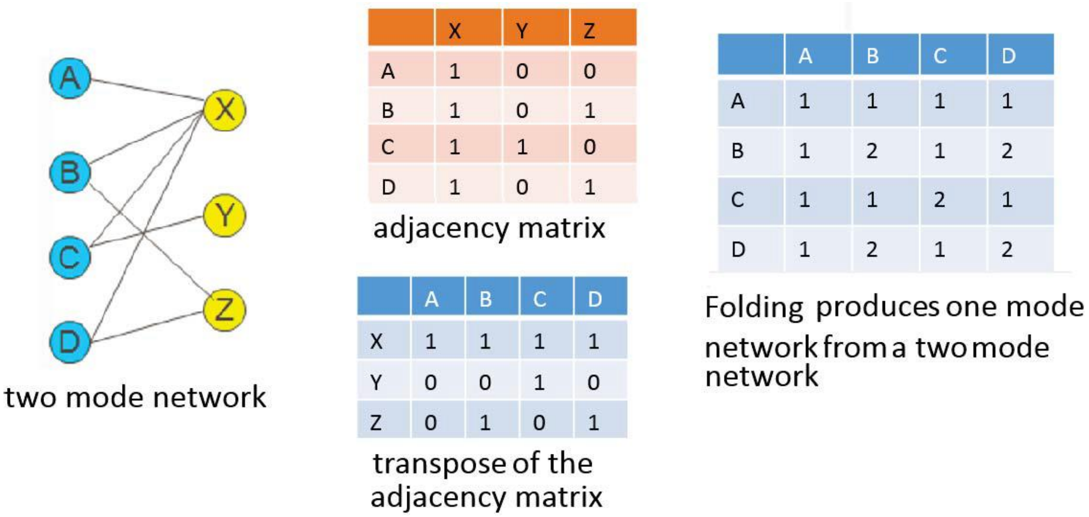


Figure 2.3: Examples of one-mode and two-mode networks

By considering nodes, one way to classify graphs is based on the number of sets of nodes which satisfy the property: whether no two nodes in the same set are connected, instead nodes in a given set are only connected to nodes in other sets. The most common graph type contains a homogeneous set of nodes which may be arbitrarily connected. Another common graph type is a bipartite graph which includes two sets of nodes and connections are allowed only between two nodes if they do not belong to the same set. The number of sets of nodes a graph may contain is the factor used to decide on the mode of the corresponding network; one-mode and two-mode networks are the most commonly used as shown in Figure 2.3.

Using a graph to represent a social network enriches the model because such representation facilitates using graph theory measures in the analysis and in processing the network structure.

For instance, a network may be represented as an adjacency matrix which is one of the most common methods to represent a graph. Once a network is represented as a matrix, all matrix algebra may be applied to study and analyse the content in link to the semantics underlying the corresponding network model.



	X	Y	Z
A	1	0	0
B	1	0	1
C	1	1	0
D	1	0	1

	A	B	C	D
X	1	1	1	1
Y	0	0	1	0
Z	0	1	0	1

	A	B	C	D
A	1	1	1	1
B	1	2	1	2
C	1	1	2	1
D	1	2	1	2

Figure 2.4: Example of a two-mode network and its folding into a one mode network

A two-mode network may be transformed into two one-mode networks by applying a process called folding [3], which depends on two operations, namely matrix transpose and multiplication. This process is illustrated in Figure 2.4. Explicitly speaking, given the adjacency matrix of a network that connects two sets of entities, say users and web pages, the folding process multiplies the adjacency matrix of (users, pages) with its transpose represented as (pages, users) to produce a new adjacency matrix represented as (users, users). The latter adjacency matrix reflects the strength of the links between users based on the number of pages they have in common. On the other hand, multiplying the transpose (pages, users) by the original matrix (users, pages) will produce a new adjacency matrix of a one-mode network which reflects the strength of connectivity between each pair of pages based on the frequency of users who accessed the two pages.

To analyse a social network, a wide variety of measures may be used, e.g., degree, closeness, betweenness, density, diameter, cliques, etc. Degree reflects the number of links connected to a node and may be classified into indegree and outdegree for directed networks, as detailed above. Further, normalised degree is computed by dividing the actual degree by (N−1), where $N \geq 2$ is the number of nodes or actors in the network. For instance, the degree of node *p* in the one-mode network shown in Figure 2.3 is 4 and the normalised degree is $\frac{4}{5}$.

Closeness centrality of a node *n* represents the minimum total number of links that must be travelled in order to visit each node in the network. In a complete graph of *N* nodes, the closeness of each node is (N −1) or the normalised closeness is 1. On the other hand, the closeness of node *p* in the one-mode network shown in Figure 2.3 is 6 (because two links are needed to visit node *q* which is the only node without direct link to *p*).

Betweenness centrality of a node or link depends on the number of shortest paths passing through the node or link. A shortest path between two nodes *u* and *v* is the minimum number of consecutive links that must be traversed between *v* and *u* such that no intermediate link or node is visited more than once. For instance, the shortest path between the two nodes *m* and *s* has the length 2 because at least two links must be traversed between the two nodes. The two node *p* and *q* in Figure 2.3 have the same importance because every pair of the four other nodes have two alternative shortest paths, one passes through *p* and the other passes through *q*. Finally, the diameter of this graph is 2 because the longest shortest path in the graph consists of two links.

In a directed graph, it is possible to distinguish nodes into hubs and authorities. A hub is a node with high outdegree and an authority is a node with high indegree. These two measures are important when it comes to analyse the importance of specific nodes especially in graphs derived from the pages of web sites.

2.5 Related Work

The web is a very rich source of information. However, there is always a gap between the way web sites are designed and how well they are received by visitors. This turned web data analysis into a very attractive research area due to its importance in knowledge discovery to guide web visitors and web site owners. web related data comes from three sources, namely web usage, web structure and web content. The former two could be utilised to build a recommendation system for visitors to allow them to surf smoothly by suggesting to them pages already visited by others who followed the same trend. The same two sources could be utilised to guide web site owners in their effort to keep their web site more attractive to visitors who may migrate to competitor web sites when they are less satisfied. This motivated researchers to invent a variety of techniques. Indeed, in a number of recent efforts, e.g., [1, 33, 70] researchers have studied methods that help in finding important web pages and also in determining page quality. For instance, Abiteboul et al. [1] proposed an algorithm, named OPICS, for computing the page importance in a dynamic graph. Their algorithm works online and does not require storing the link matrix. It is online in the sense that it continuously refines its estimate of page importance while the web/graph is visited. In the work described in [33], the authors propose a new ranking function, called page quality that measures the intrinsic quality of a page. The proposed framework investigates search engine bias in more concrete terms; it provides clear understanding on why PageRank is effective in many cases; it also highlights when PageRank is problematic. They propose a practical way to estimate the intrinsic page quality in order to avoid the inherent bias of PageRank.

The work described in [5] is an overview of different issues in web mining. They categorise different groups of web mining concepts and review different techniques used in each of the categories. Another work done by Zhang et al. [112] provides an inclusive overview of different works done in this area of study. They go over the works by categorising them based on their focus on different tasks in web usage mining. These tasks include preprocessing, knowledge discovery, and pattern analysis. The work of Ardissono *et al.* [10] describes a framework for the dynamic revision of user models in a web store shell, which tailors the suggestion of goods to the characteristics of the individual user. The behaviour of the user is monitored by segmenting it on the basis of the focus spaces explored in the browsing activity, and the actions performed by the user are summarised into abstract facts. These facts are used for revising the user model via the interpretation process performed by a Bayesian network which relates user features to user behaviour. Borodin *et al.* [24] developed an approach that uses hyperlink structures to determine the relative authority of a web page and produce improved algorithms for the ranking of web search results. In particular, it works within the hubs and authorities framework. The algorithms proposed in [24] use a Bayesian approach as opposed to the usual algebraic and graph theoretic approaches.

Borges *et al.* [22] proposed a different approach to capture user navigation behaviour patterns. User navigation sessions are modelled as hypertext probabilistic grammar whose higher probability strings correspond to the user's preferred trails. The algorithm to mine such trails makes use of the N-gram model which assumes that the last *N* pages browsed affect the probability of the next page to be visited. In the work described in [91], the authors detect users navigation paths by implementing a profiler which captures client's selected links and pages order, page viewing time and cache references. The information captured by the profiler is then utilised by a knowledge discovery technique to cluster users with similar interests. A path clustering method based on the similarity of the history of user navigation is used in the proposed framework. Fu *et al.* [41] take a different approach than the approach described in this thesis. They clustered web users based on their access patterns. In the proposed algorithm, first the server log data is similarly processed to identify sessions of web usage. The sessions are then generalised using the attribute-oriented induction method; and finally they are clustered using a hierarchical clustering algorithm, namely BIRCH. Li *et al.* [67] investigated the interaction between usability and a web site structure. They discuss a web structure mining algorithm which allows the automatic extraction of navigational structures in a web site without performing hypertext analysis. They also perform several usability experiments to correlate the usability of a web site and its structural design.

Peng [78] gives an overview of how FP-Growth could be applied in web usage mining. He defines a measure for confidence of an association rule. The measure is called Browse Interestingness and takes into consideration the time that a user spends on the pages and navigation as well. Ivancsy and Vajk [54] give an introduction on web log mining and show how frequent pattern discovery tasks can be applied on the web log data in order to obtain information about the user's navigational behaviour. They find frequent patterns of users' accesses by using an implementation of the Apriori algorithm, namely ItemsetCode algorithm. They find association rules in access patterns of users. Hsu *et al.* [52] discussed the problem of link recommendations in Weblogs and in similar social networks. They have used both collaborative recommendations and content based recommendations by exploiting the link structures and mutually declared interests. They argue that their strategy allows more in-depth analysis of the structures combined with the content. The proposed hybrid approach, *LJMiner* mines web log data of people from the social network service LiveJounal (www.livejournal.com), which people use mainly as personal publishing tool. On the other hand, the research described in this thesis concentrate on mining hyperlink usage behaviour to better structure the hyperlinks of a web site rather than mining an actual friendship network.

Lee *et al.* [66] predicted that next generation web portals would place an increased emphasis on web personalisation. This will result in an automatic user profiling and recommendations made to users based on their web usage. In order to achieve this automatic recommendation, they suggest mining web usage or click stream data in order to discover interesting web usage patterns and statistical correlations between web pages and user groups. By seamlessly matching these recommendations with users' social behaviour, a better web personalization can be achieved. Adnan *et al.* [3] presented a social network modelling technique that uses frequent close patterns to construct the social network. The authors argue that frequent closed patterns have the advantage that they successfully grab the inherent information content of the dataset being analysed; the model is applicable to a broader application domain. Entropies of the frequent closed patterns are used to collect the best set of features for the social network construction process. The work described in this thesis uses sequential closed patterns to construct the social network of hyperlinks. This can be considered as the reciprocal view of the approach presented in [3], where the relationships among the users of the data are identified based on the data usage; while the work in this thesis identifies relationships among key data items (hyperlinks) based on how the users used them. Finally, there exist some other systems for web data mining, e.g., AxisLogMiner, and WebMiner [36].

Numerous other methods described in the literature tackle the structural aspect of web sites and the results are correlated with usability for more informative knowledge discovery, e.g., [24, 25, 31, 32, 33, 56, 70]. For instance, the work described in [67] developed a spatial frequent pattern mining method to efficiently extract navigational patterns from the hyperlink structure of a web site. A navigational pattern is defined as a set of links commonly shared by most of the pages in a web site. The method depends on Eclat [21] to mine hyperlinks in a subset of the given database of hyperlinks, i.e. a subset of the adjacency matrix which corresponds to the graph linking the various pages constituting a web site. The subset to be used in the mining process is specified by a window of adaptive size. The window slides along the diagonal of the web site's adjacency matrix. To evaluate their method, the author conducted a user study where users were assigned tasks to accomplish, e.g., finding a certain piece of information on a web site. They recorded both the time needed to accomplish a task and failure ratios. The results recorded were compared with their algorithm. They found a correlation between the size of the navigational structure set and the overall usability of a web site, specifically the more navigational structure a web site has, the more usable it is as a general rule of thumb. In the work described in this thesis, we use a frequent pattern data mining algorithm, namely Apriori to serve a purpose different than that described in [67]; the latter mines hyperlinks inside a window; it is a kind of web structure mining for a partition of the web site (specific number of pages that are closely linked together). On the other hand, the study described in this thesis is comprehensive and covers the whole web site; we mainly derive a social network between the pages based on their appearance in the association rules. Further, Eclat could be effective for small scale mining tasks; but it is not scalable because it inverts the data and does operations directly on the columns; this may be the reason for using Eclat as an appropriate method for the window-based approach described in [67].

The work described in [87] predicts the pages users are expected to visit by analysing the web log using frequent pattern mining. The outcome is used to guide web site owners to improve the web site structure. This method may look somehow related the method described in this thesis because both approaches employ data mining techniques for knowledge discovery in web related data. However, the details of the two methods vary tremendously; while the work described in [87] uses data mining techniques to predict future access patterns, the work described in this thesis combines data mining and network analysis to recommend potential pages to be visited based on historical access pattern. Further, they do not measure the importance of a particular page based on the time visitors spend on the page. Their approach applies frequent patterns mining to discover navigation preferences of the visitors based on the most frequently visited pages and the frequent navigational visiting patterns. However, we argue that there may exist some intermediate pages along the navigational pattern leading to the target page that a user is actually interested in. A visitor passes through the intermediate pages and spends more time on important pages. Therefore, the time spent by a visitor on a page should be considered as an important measure to quantify the significance of a page in a web site structure. Further, we do not limit our investigation to individual paths navigated by users, we rather use the outcome from the mining process to decide on the links between the pages in the constructed network; the more rules favour a specific link the higher weight it gets and hence the more central it becomes in the analysis of the social network. Therefore, our results reflect a global picture of the web site usage and navigational patterns.

The work described in [51] finds pages relevant for a given web page by employing two hyperlink analysis-based algorithms. The first algorithm analyses hyperlinks between web pages by mapping and utilising an existing technique which was originally developed for citation analysis to web page hyperlink analysis. To realise the mapping web pages were assumed authors and the hyperlinks between web pages were considered as equivalent to citations between authors. The second algorithm decides on relevant pages by extracting and analysing the relationships between web pages using linear algebra theory. This way, the authors were able to integrate the topological relationships among the pages into the process in order to identify deeper relations among pages for finding relevant pages.

The work described in [38] presents a technique which considers the neighborhood of pages in order to identify more potentially relevant pages. The work described in [110] proposes a method of preparing web logs for mining specific data on a per session basis. This way, the time and page data gathered may be used to investigate an individual's browsing behaviour. The authors also discuss how to preprocess the log file to identify specific information such as stripping entries left by robots. Joshi et al. [57] defined a new measure to calculate the similarity between two different sessions. They defined the similarity measure by taking into account the structure of the URL. This measure takes a value between 0 and 1 (instead of being a binary value). Then they applied a fuzzy clustering algorithm in order to find similar user sessions.

In the approach described in [107], the authors decided to avoid treating all pages equally by modifying the standard PageRank algorithm to distribute the rank amongst related pages with respect to their weighted importance. This resulted in a more accurate representation of the importance of all pages within a web site. We used the weighted PageRank derivation described in [107] to enrich the web structure mining component of the proposed framework, with the target of returning more accurate results than the standard PageRank algorithm. The result reported by the weighted PageRank algorithm is first validated by applying HITS [63] to check the consistency of the results, and then it is confirmed or complemented by the outcome from the employed social network methodology. HITS is another document ranking algorithm developed by Kleinberg; it also ranks documents based on link information.

Chapter 3

The Proposed web Usage Mining Approach: Highlighting Important Pages

The amount of data maintained by web sites to keep track of the visitors is growing exponentially. Benefitting from such data is the target of the study presented in this chapter. The described approach investigates and explores the process of analysing log data of web site visitors' traffic in order to assist the owner(s) of a web site in understanding the behaviour of web site visitors. We developed an integrated approach that involves statistical analysis, association rules mining, and social network construction and analysis. First we analyse the statistical data on the types of visitors that access the web site, as well as the steps they take to reach and satisfy the goal of their visit. Second, we derive association rules in order to identify the correlation between the web pages. Third, we study the links between the web pages by constructing a social network based on the frequency of access to the web pages such that web pages get linked in the social network if they are identified as frequently accessed together. The value added from the overall analysis of the web site and its related data should be considered valuable for online shopping and commercial web site owners. The owners will get the information related to ranking and relevance

The content of this chapter is mainly based on the following published papers:

M. Adnan, M. Nagi, K. Kianmehr, R. Tahboub, M. Ridley and J. Rokne, "Promoting where, when and what?: An analysis of web logs by integrating data mining and social network techniques to guide eCommerce business promotions", *Social Networks Analysis and Mining*, 1(3):173-185, 2011.

M. Nagi, A. Guerbas, K. Kianmehr, P. Karampelas, M. Ridley, R. Alhajj, and J. Rokne, "Employing Social Network Construction and Analysis in web Structure Optimisation", in *Social Networks Analysis and Mining: foundations and applications*, by Springer, pp.13-34, 2010.

of web pages in order to display targeted advertisements or messages to their customers. Such an automated approach gives advantage to its users in the current competitive cyberspace. In the long run, this is expected to allow for the increase in sales and overall customer loyalty.

In general, a reasonable amount of advertisement of the right product, at the right place and the right time may boost the sales of the organisation immensely. This is also beneficial to visitors of the web site in the sense that it helps them find the right product quickly, thus enriching the browsing experience. In other words, the main target of maximising user satisfaction once achieved, as discussed in this chapter, will be to the benefit of the users and the web site owners. The users will save the time spent shopping around for products and sites that satisfy their needs, and web site owners will enjoy keeping their current customers with the hope to gain more customers who share the same expectations with the already satisfied customers. We try to make this feasible by an integrated approach that combines the power of data mining techniques with the strengths of social network construction and analysis. We derive association rules to find correlations between web pages constituting a web site. We use the outcome from the first step of the association rules mining process, namely the closed frequent itemsets in order to construct a social network which is extensively analysed for knowledge discovery. The reported testing results are very promising.

The rest of the chapter is organised as follows. Section 3.1 is an overview of the data preparation process; this is necessary for the employed approaches to function on the clean data. Section 3.2 presents the data analysis method used to study the different aspects of the data from statistical point of view. Section 3.3 describes the association rules mining approach to study the links between pages constituting web sites. Section 3.4 presents a novel social network extraction and analysis technique that uses frequent closed sequences to establish relationships between each two correlated pages. Section 3.5 covers how these individual pieces of information produced by the different methods may guide the site owner in handling promotional decision making concerns as posted earlier in this document.

3.1 Data Preparation

The web log is in general a summary of all the activities done by all users visiting the web site. It is necessary to clean the web log before it is used; this step is referred to as preprocessing. The target of the preprocessing step is to “clean” the web log in order to minimize interference from robots; this step will make the web log ready to generate the actual output of this stage. Actually this is the preprocessing step for the whole system. So, the input to the association rules mining and the social network construction and analysis methodology is the clean web log data produced after this preprocessing stage.

The preprocessing step converts the raw format into a form that can be accepted as input by a specific implementation of an algorithm. Thus, it is important to understand the raw data and the new environment to which the data is to be mapped so that the preprocessing step would produce the targeted outcome. One of the issues with analysing the data is figuring out how to distinguish unique users [36]. The problem is that certain users may use a proxy server or share a data connection, which means we cannot use the Internet Protocol (IP) address of the user as a unique identifier; this is one of the main issues in web usage mining. Cooley *et al.* [36] prioritises suggestions among many that can be found in the literature to get around this problem. These include: combining machine name, browser information or temporal information with IP addresses to uniquely identify users. There are also other methods described in the literature, which consult site-topology [82], or session timeouts [28] to determine legitimate page traversals by a single user. Once the users have been semi-uniquely identified, we can parse the data into the format we require for the next steps.

A sequence of page views separated into transactions can be very useful for web usage mining activities. At this preprocessing stage, each user session is separated into a transaction [36]. In other words, each transaction contains the sequence of pages a particular user viewed on a given web site. A new transaction is created if the time between page views exceeds some predefined limit. For instance, Catledge *et al.* [28] found that 25.5 minutes is a good timeout period based on their user experiments. Computing these sequences also allows us to easily calculate the time a user spends on each page, which can be correlated to the interest or relevance of the page to the user. It should be noted that we only extract the relevant lines from the log file. A user accessing images or other media not directly related to the information they are accessing can be discarded at the beginning of the preprocessing stage.

We also perform what we call path compression on the sequences before they are handed off for analysis. Basically, this means we ensure that consecutive views of the same page do not exist in the sequence. In other words, a sequence A,A,B,C,C,D,E,F will be compressed to A,B,C,D,E,F. This allows our algorithms to be written much cleaner and should not place much impact on the final results. While it is true that a user viewing the same page multiple times consecutively may be much more interested in the data, we feel that does not come into play with our basis for the analysis. In fact, a user may visit the same page several times in a session or sequence of sessions because it contains some information that caught his/her attention and hence may come back to refresh his/her mind or check more details. At the end, we are interested in the fact that the page was visited as a destination not a junction along the path leading to a destination.

We applied this data preparation methodology on the data used in this chapter and also in Chapter 4. Further, in order to fruitfully present our approach, we adopted an example driven approach where we study a specific web log of catalogue browsing. To be precise, we have used the *Music Machines*[1] web log from the University of Washington's web log data repository[2]. This web log in particular exhibits users' musical instrument catalogue browsing behaviour and consists of around 2500 distinct pages including HTML pages, images, media, and text files. For our analysis, we have considered only the portion of the data related to the year 1997. In order to make things simple, we also did not consider dynamic pages, images, and media files; thus we only considered html files. The web log was sessionised using Chen *et al.*'s approach [30] with a session interval of 30 minutes.

One approach which may be effective for this preprocessing step is described in [110]. This approach concentrates on identifying and discarding sessions which may be characterised by the following access patterns that are likely to be robots' characteristics:

- Trying to avoid time latency by visiting after midnight, i.e., during light traffic load periods.
- Using "HEAD" instead of "GET" as the access method to verify the validity of a hyperlink; the "HEAD" method performs faster in this case as it does not retrieve the web document itself.
- Doing breadth search rather than depth search; robots do not navigate down to low-level pages because they do not need to access detailed and specific topics
- Ignoring graphical content; robots are not interested in images and graphic files because their goal is to retrieve information and possibly to create and update their databases.

It is possible to identify sessions from the cleaned log file. Sessions are used to compute the total number of visits vi and the total time spent by visitors, ti for each page pi. It is necessary to confirm that the number of user sessions extracted from the web log are of sufficient size to provide a realistic ranking of popular pages. Finally, the outcome from this preprocessing step is the main input to our web log analysis approach described in this chapter.

[1] http://machines.hyperreal.org/
[2] http://www.cs.washington.edu/ai/adaptive-data/

3.2 web Log Analysis

This section describes all the steps taken to extensively analyse the clean web log data and produce the intended results.

Table 3.1: Significant Backtracks related to the "Music Machines" web log

Page	Percentage of total backtracks related to the "Music Machines" web log
/new/index.html	79.85%
/manufacturers/moog/multimoog/index.html	73.30%
/manufacturers/yamaha/tx-81z/samples/index.html	71.73%
/manufacturers/roland/tr-626/index.html	70.26%
/manufacturers/roland/mc-303/samples/index.html	70.12%
/manufacturers/moog/minimoog/index.html	69.57%
/features/first-synth.html	68.92%
/categories/software/mac/performer/index.html	67.72%
/manufacturers/korg/x-synths/index.html	67.63%
/manufacturers/roland/d-synths/index.html	67.43%
/manufacturers/moog/mg-1/index.html	67.33%
/dr-660/index.html	67.14%
/manufacturers/casio/cz/samples/index.html	66.61%
/manufacturers/ensoniq/asr-10/index.html	64.96%
/categories/drummachines/ samples/deepsky kicks/index.html	64.17%

3.2.1 Identifying Backtracks

Backtracks can be directly found by looking at the sequences produced by the preprocessing stage. A backtrack page is a page that is reached by a user before they start to proceed to the previous page they viewed. To clarify this, consider a case where a user views pages A,B,C,B; in this case, page C is considered as a backtrack page. This helps us to determine pages that may contain little relevant information or of no interest to the user. In the real world, a product that is often backtracked may be selected to appear lower down on a list of products, so that products of more interest are displayed first. Pages of high backtrack count may be undesirable to users or may be priced too high. Flagging pages that are highly backtracked allow a business to target specific products that may be performing undesirably. When performing this analysis on the music machines web log, we have filtered out pages that are backtracked less than 100 times (arbitrarily chosen value which is considered small compared to the 2500 total number of pages available in the processed web site), as we consider these actions to be normal browsing behaviour, thus insignificant to our analysis of the data. Some significant pages that are backtracked often are summarised in Table 3.1.

3.2.2 Employing Probability Measures in Sequence Analysis

It is necessary to investigate various types of sequences in order to predict and decide on the probability of the next page to be viewed. There are three possible cases to investigate: (1) Consider only one step back, i.e., the last two pages accessed will lead to the probability of accessing the next page. For instance, the sequence of views A,B and A,C would lead to the case where from page A there is 50% probability of viewing page B and 50% probability of viewing page C. (2) Consider the complete sequence of previous pages already viewed. This requires matching the whole prefix sequences up until the point where we predict the next page. For instance, viewing two sequences in different sessions such as A,B,C,D in one session and A,B,C,E in another session, will lead to the prediction that a user who views the prefix pages A,B,C is 50% likely to view page D and 50% likely to view page E. (3) Consider the set of previous pages viewed. This case is similar to the previous case but here the order of the previously visited pages is not important. It is possible to combine these probability measures to decide whether it is essential to move links of pages that will be more likely viewed next into a more prominent section of the web page.

Table 3.2: Summary of Next Page Probability values from Previous Page Views

Previous Page	Next Page	Probability
/analogue-heaven/index.html	/index.html	0.78
/images.html	/images/index.html	0.73
/categories/software/windows /index.html	categories/software/index.html/	0.66
/categories/drum-machines /index.html	/index.html	0.62
/manufacturers/roland/juno /index.html	/manufacturers/roland/index. html	0.62
/links/sites.html	/links/index.html	0.61
/links/links.html	/links/index.html	0.61
/links/misc.html	/links/index.html	0.60
/links/manufacturers.html	/links/index.html	0.59
/links/machines.html	/links/index.html	0.59
/dr-660/index.html	/index.html	0.54
/manufacturers/roland/tr-909/index.html	/manufacturers/roland/index. html	0.53
/links/dealers.html	/links/index.html	0.53
/prices/index.html	/manufacturers/index.html	0.46
categories/software/index.html/	/categories/software/windows /index.html	0.45

Computing Probabilities Based on Previous Page

The probability of the next page to be viewed based on the previous page viewed is one of the most important measures within this grouping of probability measures. This allows us to recommend products on the current page that the vast majority of users will likely browse to on their own. This clearly allows users to find information faster than following manual navigation. Another method a web site might deploy using this information is the ability to purchase the current product as a bundle with the product that is most likely to be viewed next by a majority of the web site visitors. This does not only increase revenue, but also provides increased user satisfaction by reaching their goals without wasting their precious time. A summary of the top probability values are reported in Table 3.2.

Table 3.3: Summary of Next Page Probability values from Previous Sequence

Sequence	Next Page	Probability
905, 185, 905, 185, 905, 185, 905, 185	905	0.43
185, 905, 185, 905, 185, 905, 185	905	0.43
185, 905, 185, 905, 185, 905, 185, 905	185	0.40
905, 185, 905, 185, 905, 185, 905	185	0.39
906, 905, 185, 905, 185, 905	185	0.35
906, 905, 185, 905, 185, 905	185	0.33
906, 185, 905, 185, 905 185	905	0.30
906, 185, 905, 185, 905	905	0.29
133, 135,	906	0.22
910, 504,	910	0.22
135, 133, 135	906	0.22
910, 242, 910	910	0.22
910, 426, 910	910	0.22
102, 906, 102	906	0.21
910, 100, 99, 100	910	0.21

Computing Probabilities for Previous Sequence

We could argue that the previous ordered sequence of viewed pages forms a good basis for predicting the next page(s) to be viewed. The previous ordered sequence consists of a complete sequence of pages which have been viewed in the same session. The main motivation here is the fact that most sessions of web site visitors include a small number of pages. This makes it easier to investigate the access patterns further and derive interesting predictions regarding the next page(s) to be viewed. Here paths are compressed by eliminating all repetitions of the same page in the sequence. In other words, the concentration is on the pages visited in a specific session rather than the number of times a page is visited within the same session. It is anticipated that the first few pages viewed will provide an idea about the pages expected to be visited next. Based on prediction accuracy the pattern will continue to exist as the length of the sequence increases, i.e., for longer sequences. To illustrate this, consider the following sequence of five pages from our data: 906, 185, 905, 185, 905, there is 29% (see row 8 in Table 3.3) chance that the user would visit page 905 or 30% (see row 7) chance the user would visit page 185.

It is worth noting that, for the purpose of clarity, only page numbers are reported in Table 3.3 instead of using their full names. However, the sequence in question (see row 7) should be read as 30% of the time users have visited page "images/index.html" after they visited the sequence "index.html", "images.html", "images/index.html", "images.html", "images/ index.html", "images.html". This particular sequence once analysed closely may signify that the user is browsing the catalogue of instrument images and is comparing many different instruments. It should be noted that insignificant data has been removed from the results. In other words, we will not find many very long sequences that have good prediction for the next page. This is simply because users drift widely once they get deeper into the site. It is, therefore, only practical to use this web mining method for the first couple of page views. Beyond that, it does become insignificant. web sites which provide better organisation (allowing users to find what they are looking for within their first four pages or so) would be better suited for this type of web usage mining, as they could provide some useful and informative outcome.

Table 3.4: Summary of Next Page Probabilities from Previous Unordered Set

Set	Next Page	Probability
{100, 912}	913	1
{135, 912}	913	1
{136, 912}	913	1
{176, 179, 912}	913	1
{179, 187, 912}	913	1
{179, 875, 912}	913	1
{179, 906, 912}	913	1
{185, 912}	913	1
{187, 875, 912}	913	1
{187, 888, 912}	913	1
{187, 889, 912}	913	1
{187, 891, 912}	913	1
{187, 892, 912}	913	1
{187, 904, 912}	913	1
{187, 906, 912}	913	1

Probability Based on Previous Unordered Sequence

This is similar to the above method; however, we are no longer looking at the sequence as being of higher relevance. Instead, we are looking at the first part as being a set. We consider the pages the user views as a set of pages, and from there we can calculate the probability of the next page to be viewed based upon this set. This accounts for the fact that users may be reaching the same goals, but could be approaching that in vastly different ways. In other words, if the user has viewed the given set of pages, in any order, then they are likely to view this other page next. This method of web usage mining is extremely resource intensive in computing. As such, this method may be discarded on face value because of the longer amount of time it takes to process the data. The music machines web log file contains only 643,000 (six hundred and forty three thousand) useful entries. This is fairly small size compared to some data collected by online shopping web sites. Some of the important information found is summarised in Table 3.4.

It should be noted that the probability of 100% listed in Table 3.4 is significant. It does not mean that only one particular user went from the set to the next page, these individual results were filtered out to provide only useful data. For example, row 2 in the table suggests that the user would certainly visit the page "/software.html" after visiting the page set {"/categories/software/include.html", "/search.html"}.

3.3 Sequential Association Rules

We have also applied association rules mining based analysis on the sequences. Association rules mining is an effective model for data analysis. The employed process determines the correlations between items in the analysed data. The items in our domain are the web pages and the target is to find how the visits to the web pages are correlated by considering the web log data.

As described in Chapter 2, the association rules mining model involves two steps: finding frequent itemsets and then using the frequent itemsets to derive the interesting rules. Finding frequent itemsets is driven by the support concept. The support of an itemset X is the percentage of the transactions that concurrently contain all the items in *X*. An itemset is considered frequent if its support is larger than a prespecified minimum support value; mostly specified by a domain expert. After finding the frequent itemsets, constructing the rules is straight forward. From each itemset, all possible rules are constructed such than the left hand side (antecedent) and right hand side (consequent) of a rule are disjoint. Then the confidence of each rule is determined by dividing the support of the items that appear in the consequent by the support of all the items in the rule. We keep only rules with confidence larger than a minimum confidence, normally specified by a domain expert.

For our study described in this chapter, we decided to use only closed frequent itemsets because they are more compact and serve well the target of this study. A frequent itemset is closed if and only if its support is different from all of its frequent supersets. To derive closed frequent itemsets from the *Music Machines* web log, we used the CloSpan algorithm originally proposed by Yan *et al.* [108]. As we do not have an expert to help in setting the minimum support threshold, we have two choices either to set it automatically by analyzing the existing data or run some initial tests to help us figure out what could be a reasonable minimum support threshold value that satisfies our target; we decided to proceed with the latter choice and accordingly we set the minimum support threshold value at 0.1%. Then, we identified the corresponding confidence measure for associations involving the located closed frequent sequential patterns.

Table 3.5: Sequential Association Rules

Rule	Confidence
(563, 646) ⇒(666 747)	0.92
(554, 646) ⇒(666 747)	0.91
(905, 185, 905, 185, 905, 185) ⇒(905 185)	0.67
(905, 185, 905, 185, 905, 185) ⇒(905 185)	0.67
(905, 905, 185, 185, 905) ⇒(185 905)	0.66
(185, 905, 185, 905, 185, 905) ⇒(185 905)	0.66
(905, 185, 905, 185, 905, 905) ⇒(185 905)	0.66
(905, 905, 185, 905, 185) ⇒(905 185)	0.66
(905, 185, 905, 905, 185) ⇒(905 185)	0.66
(905, 905, 905, 185, 905) ⇒(185 905)	0.66
(905, 185, 185, 905, 185) ⇒(905 185)	0.66

Table 3.5 shows the top associations based on confidence measure where the consequent of the rule has more than one page. The type of association information can give the analysts a significant insight on the dynamics of the behaviour of users. For example the first row in Table 3.5 suggests that if a user visits the page "/manufacturers/oberheim/xpander.matrix-12/patches/index.html" and then the page "/manufacturers/roland/juno/patches/index.html" in order; then he/she also would visit the page sequence "/manufacturers/roland/mc-202/patches /index.html" and "/manufacturers/sequential/pro-one/patches/index.html" in order for around 92% of the time. This type of insight may help the business organisation is strategic decision making such as site/link organisation or cleverly placing dynamic promotions, etc. In other words,

the derived rules provide a rich insight and understanding of the access patterns by various users and greatly help in predicting the behaviour of users.

3.4 Social Network Analysis

As described in Chapter 2, a social network represents relationships or ties among individuals or actors. Formally, a social network is a graph G = (V,E), where V = {v1, v2, . . . , vn} is the set of vertices representing individuals and E is the set of edges representing the relationships between vertices or individuals. This model of data analysis perfectly fits our research objective where we can represent the vertices (actors) as pages and associations between pages as edges. For our specific model, we use directional edges with weights where higher strength associations have higher weights.

3.4.1 Network Construction Methodology

To construct the social network of the hyperlinks or pages, we have used the closed sequence patterns as described in the previous section (see Section 3.3). Let us assume that the set of closed sequence patterns for a particular support threshold σ (for our analysis σ is 0.1%) is *S*. In order to calculate the relationship strength between page $a \in P$ and page $b \in P$ (where $a \neq b$, and *P* is the set of web pages constituting the analysed web site), we first identify a set of pages which have the following form - *XaYbZ*. Here, *X*, *Y*, and *Z* are different subsets of the page set *P*, any or all of which can be empty. Let us assume that this set is denoted by $S_{a,b}$.

Given the set $S_{a,b}$, the relationship strength between page *a* and page *b* can be calculated using either one of the following two formulas (Equation 3.1 or Equation 3.2)-

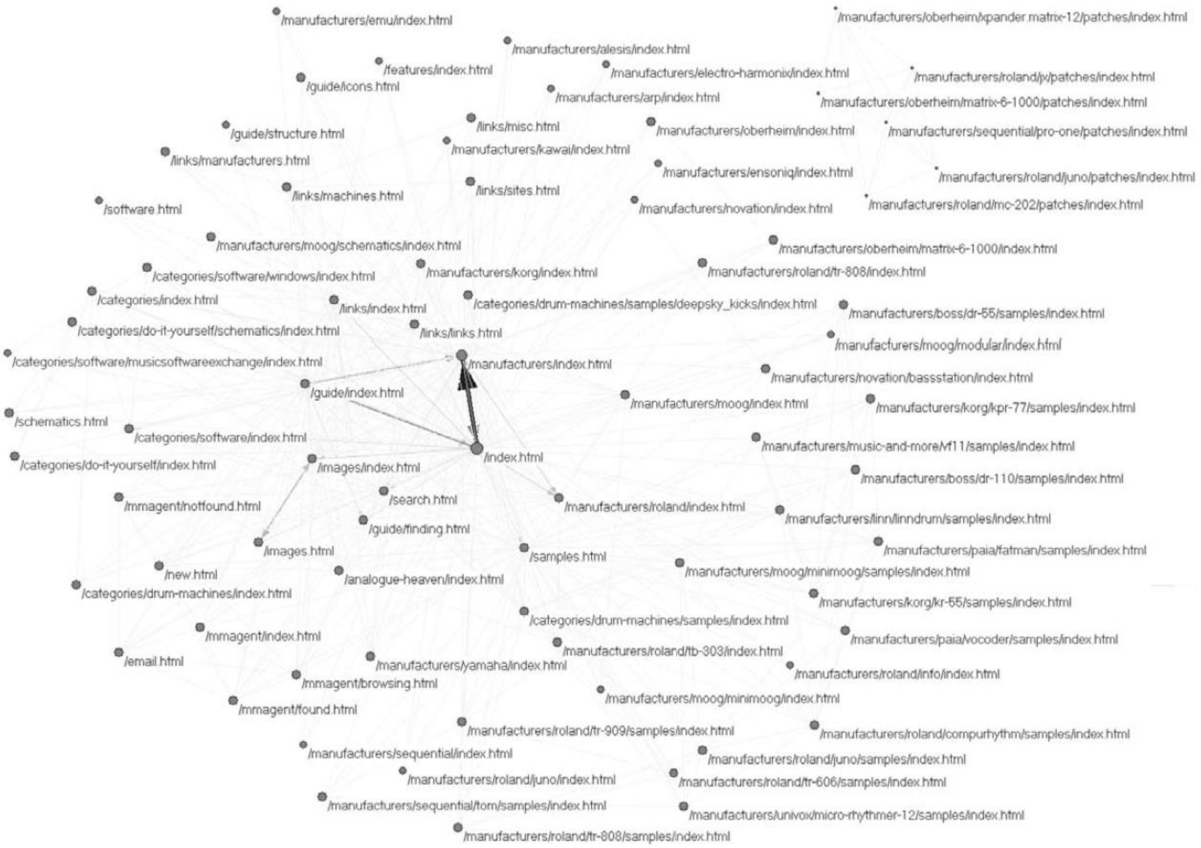

/manufacturers/emu/index.html
/manufacturers/oberheim/xpander matrix-12/patches/index.html
/manufacturers/alesis/index.html
/features/index.html
/manufacturers/electro-harmonix/index.html
/guide/icons.html
/manufacturers/roland/jx/patches/index.html
/manufacturers/arp/index.html
/manufacturers/oberheim/matrix-6-1000/patches/index.html
/links/misc.html
/guide/structure.html
/manufacturers/oberheim/index.html
/manufacturers/sequential/pro-one/patches/index.html
/manufacturers/kawai/index.html
/links/manufacturers.html
/manufacturers/ensoniq/index.html
/manufacturers/roland/juno/patches/index.html
/links/machines.html
/links/sites.html
/software.html
/manufacturers/novation/index.html
/manufacturers/roland/mc-202/patches/index.html
/manufacturers/moog/schematics/index.html
/manufacturers/oberheim/matrix-6-1000/index.html
/categories/software/windows/index.html
/manufacturers/korg/index.html
/manufacturers/roland/tr-808/index.html
/categories/index.html
/links/index.html
/categories/drum-machines/samples/deepsky_kicks/index.html
/manufacturers/boss/dr-55/samples/index.html
/categories/do-it-yourself/schematics/index.html
/links/links.html
/manufacturers/moog/modular/index.html
/categories/software/musicsoftwareexchange/index.html
/manufacturers/index.html
/manufacturers/novation/bassstation/index.html
/guide/index.html
/manufacturers/moog/index.html
/manufacturers/korg/kpr-77/samples/index.html
/schematics.html
/categories/software/index.html
/manufacturers/music-and-more/vf11/samples/index.html
/index.html
/categories/do-it-yourself/index.html
/images/index.html
/manufacturers/boss/dr-110/samples/index.html
/search.html
/mmagent/notfound.html
/manufacturers/roland/index.html
/manufacturers/linn/linndrum/samples/index.html
/guide/finding.html
/manufacturers/paia/fatman/samples/index.html
/images.html
/samples.html
/manufacturers/moog/minimoog/samples/index.html
/new.html
/analogue-heaven/index.html
/categories/drum-machines/index.html
/manufacturers/korg/kr-55/samples/index.html
/categories/drum-machines/samples/index.html
/mmagent/index.html
/manufacturers/paia/vocoder/samples/index.html
/manufacturers/roland/tb-303/index.html
/email.html
/manufacturers/yamaha/index.html
/manufacturers/roland/info/index.html
/mmagent/browsing.html
/manufacturers/moog/minimoog/index.html
/mmagent/found.html
/manufacturers/roland/tr-909/samples/index.html
/manufacturers/roland/compurhythm/samples/index.html
/manufacturers/sequential/index.html
/manufacturers/roland/juno/samples/index.html
/manufacturers/roland/juno/index.html
/manufacturers/roland/tr-606/samples/index.html
/manufacturers/sequential/tom/samples/index.html
/manufacturers/univox/micro-rhythmer-12/samples/index.html
/manufacturers/roland/tr-808/samples/index.html

Figure 3.1: Social network depicting the relations among the hyperlinks extracted from the *Music Machines* web log data.

$$rel(a,b) \ = \ |S_{a,\,b}| \tag{3.1}$$

where || denotes the cardinality of a set.

$$rel(a,b) = \sum_{s \in S_{a,b}} \sup(s) \tag{3.2}$$

where sup(s) denotes the support count for the sequence s. It may be noted that for our analysis we have found it useful to normalise the relationship strength using the maximum relation value between any two pages, i.e., max{rel(p,q)|p ≠ q, and p,q ∈ P}

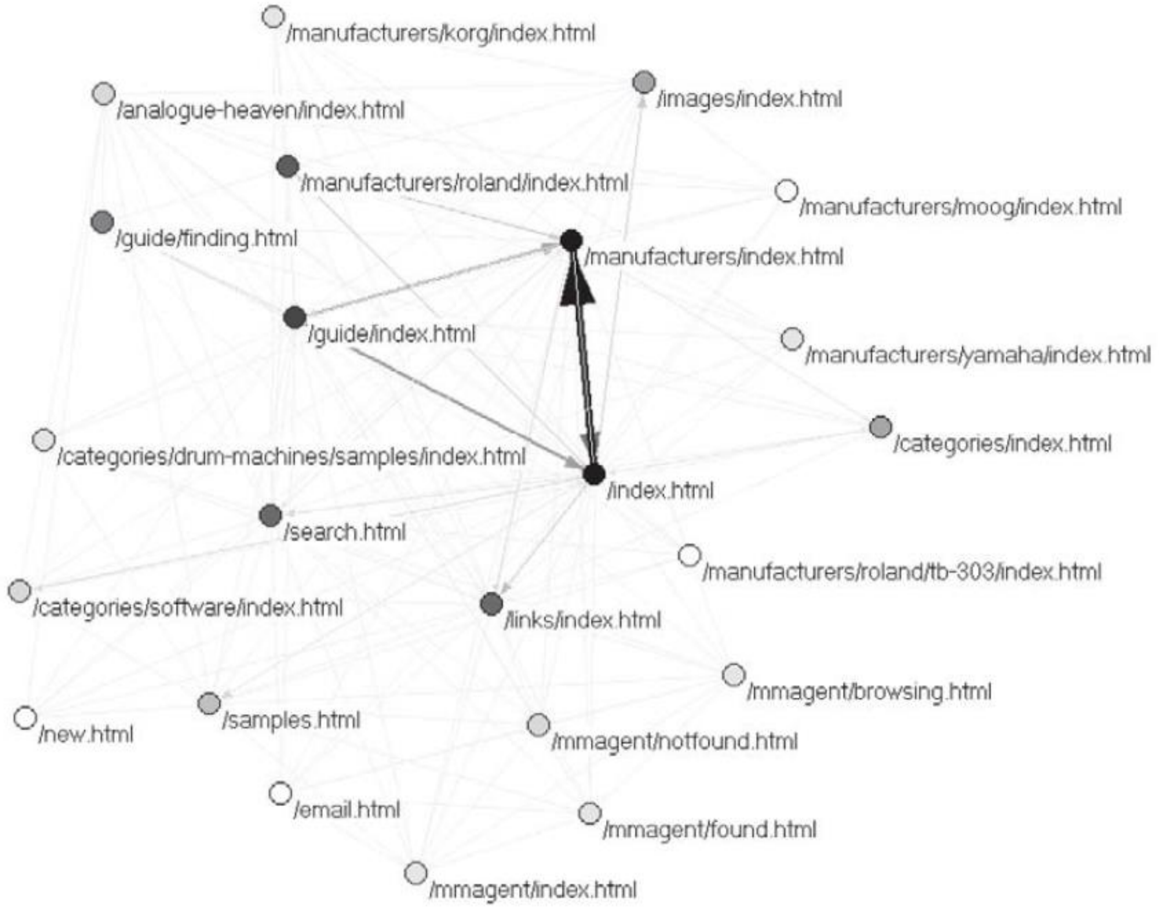

/manufacturers/korg/index.html
/images/index.html
/analogue-heaven/index.html
/manufacturers/roland/index.html
/manufacturers/moog/index.html
/guide/finding.html
/manufacturers/index.html
/guide/index.html
/manufacturers/yamaha/index.html
/categories/index.html
/categories/drum-machines/samples/index.html
/index.html
/search.html
/manufacturers/roland/tb-303/index.html
/categories/software/index.html
/links/index.html
/mmagent/browsing.html
/new.html
/samples.html
/mmagent/notfound.html
/email.html
/mmagent/found.html
/mmagent/index.html

Figure 3.2: Valued cores from the Music Machines web log data.

Figure 3.1 shows the social network extracted from the Music Machines web log using Equation 3.2 as the strength measure. We have used the tool Pajek [14] to draw the network. Figure 3.1 is actually a reduced version of the whole network. Here the network is reduced using the degree reduction technique of Pajek, where nodes having less than three outgoing or incoming edges are removed. This type of network reduction can be used to simplify the network for a better analysis. In this social network the relationships are directed and the strength of a relationship is depicted using grey-scale and thickness in such a way that the thicker and darker links signify a stronger relationship. During our analysis, we also produced a similar network using

Equation 3.1, but it is not shown here because it conveys much the same information as depicted in Figure 3.1.

3.4.2 Centrality Measures

The centrality measures of vertices within a graph identify the relative importance of a particular vertex in the graph. As described in Chapter 2, there are several centrality measures that can be found in the social network analysis literature. These include degree centrality, betweenness centrality, closeness centrality, etc. For a more comprehensive description of the various measures used for network analysis, the reader may refer to [35].

For the social network shown in Figure 3.1, the node size depicts the closeness centrality of the graph. The bigger a node the more central it is, i.e., the more likely the page is in a user's page visit path. But, we have used the valued graph concept while constructing our social network. So, a centrality measure [40, 65] that considers the importance of edge weight should be more appropriate here. The concept of valued cores [65] is used for our analysis. A core is a subnetwork of a given network where the sum of values/weights of arcs is higher than members of the same core/cluster, and it is lower than members of different cores. The target is to find highly weighted subgraphs in a weighted graph. Part of the social network presented in Figure 3.1 is shown in Figure 3.2 with the cores identified. Here, nodes of the same core have the same colour while higher sum valued cores have darker colours. These higher sum valued cores should be considered with great importance because not only they do signify the importance of a page but also a grouping of pages of the same importance level. High sum valued cores can be an ideal place to promote certain products from the business owners perspective.

3.4.3 Hubs and Authorities

Hubs and authorities [63] of a web site tells us about where the information is typically sought and where the navigation originated. If a page is very popular, many other pages will be directed to this page where the user will typically find the anticipated important piece of information. These pages are identified as authorities. On the other hand, hubs are pages that point to a large number of authorities. Actually, this is the basic concept of the HITS algorithm developed by Kleinberg [63] as mentioned in Chapter 2.

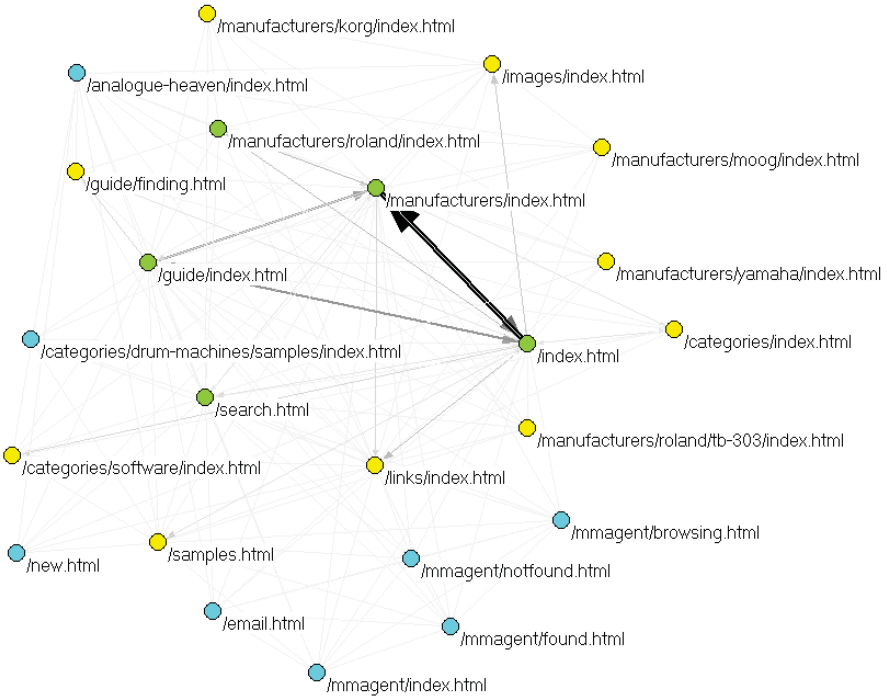

/manufacturers/korg/index.html
/images/index.html
/analogue-heaven/index.html
/manufacturers/roland/index.html
/manufacturers/moog/index.html
/guide/finding.html
/manufacturers/index.html
/manufacturers/yamaha/index.html
/guide/index.html
/categories/index.html
/categories/drum-machines/samples/index.html
/index.html
/search.html
/manufacturers/roland/tb-303/index.html
/categories/software/index.html
/links/index.html
/mmagent/browsing.html
/new.html
/samples.html
/mmagent/notfound.html
/email.html
/mmagent/found.html
/mmagent/index.html

Figure 3.3: Hubs and Authorities from the *Music Machines* web log data.

For the case of the Music Machines web log, we also performed similar analysis of finding hubs and authorities using the tool Pajek (the reader should refer to Figure 3.3). The top five hubs are shown as green coloured nodes, and the top 15 authorities are shown as yellow coloured nodes. It is worth noting that for this dataset the five hubs are also identified as authorities. Also note that pages like /guide/index.html, /manufacturers/index.html, and /samples.html are identified as authorities. Intuitively speaking, these are the pages a user should look for in order to find useful information. From a business perspective, these authoritative pages are the most useful spaces to place an important promotional notice. In other words, these pages may get the most expensive promotions because they are anticipated to be the most visited and browsed by potential customers.

Table 3.6: The top 10 pages based on page-rank measure

Rank	Page
1	*/index.html*
2	*/manufacturers/index.html*
3	*/guide/index.html*
4	*/samples.html*
5	*/links/index.html*
6	*/images/index.html*
7	*/manufacturers/roland/index.html*
8	*/search.html*
9	*/images.html*
10	*/guide/finding.html*

3.4.4 Applying PageRank

PageRank [26] is a link analysis algorithm that is used by the Google search engine to determine the relative importance of a page. The PageRank mechanism assigns to each page a global importance ranking which helps users in finding the important information quickly. The PageRank mechanism exploits the link structure of web pages to determine the relative significance of a page. In general, pages that have high back links are more important and will receive a higher rank according to PageRank [26]. This process is used recursively and it is dependent on the PageRank measures of the back links. For the case of the Music Machines web log, we have calculated the PageRank of the web pages of the network shown in Figure 3.1 using the statistical tool R [86]. Table 3.6 shows the top 10 pages from the Music Machines web site. Like authoritative pages, these highly ranked pages can be identified as good to place promotional notices by business organisations.

3.5 Groups and Sub-Networks

Clustering outcome is usually discovered by applying machine learning or statistical techniques. While statistical techniques have been used traditionally for clustering, machine learning techniques have been recently developed as attractive alternatives which scale well for large data and provide flexibility. As described in Chapter 2, clustering is unsupervised learning and there are a wide variety of clustering algorithms that can be used for grouping objects based on a specific similarity measure that can be used to compare the objects in hand. For the domain tackled in this thesis, the grouping is done based on comparing the similarity of users by considering their behaviour and interests.

Statistical and machine learning techniques may be employed to group a given set of objects. However, machine learning techniques are more attractive as they provide for better scalability. Clustering is preferred as a grouping technique because it is not feasible to label the input data for supervised learning techniques, e.g., classification. As clustering groups objects such that similarity is high within the same group and low across the group, it is possible to derive several community models from the outcome from the clustering outcome. Such models represent usage patterns of the analysed web site. The analysis of these community models can provide a deeper understanding of the users who can then be provided with a more suitable and customised experience.

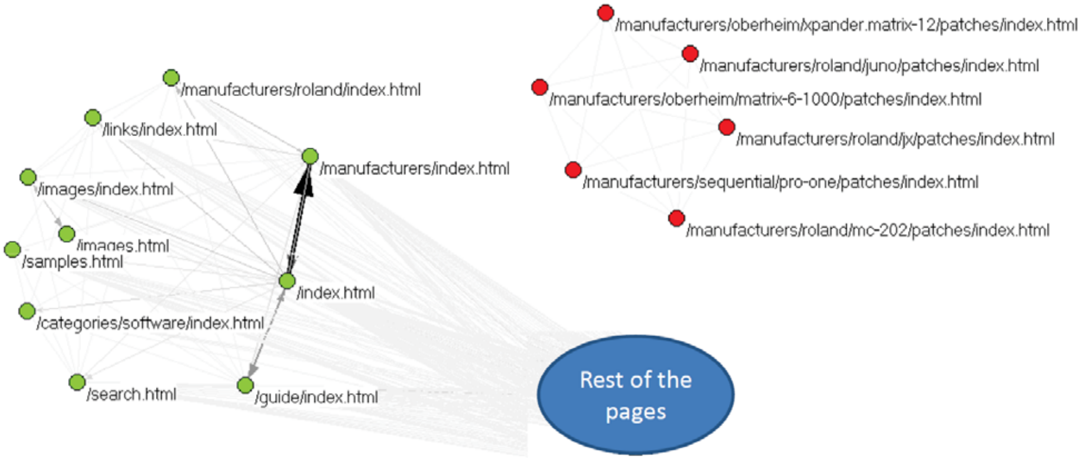

/manufacturers/oberheim/xpander.matrix-12/patches/index.html
/manufacturers/roland/juno/patches/index.html
/manufacturers/oberheim/matrix-6-1000/patches/index.html
/manufacturers/roland/jx/patches/index.html
/manufacturers/sequential/pro-one/patches/index.html
/manufacturers/roland/mc-202/patches/index.html
/manufacturers/roland/index.html
/links/index.html
/manufacturers/index.html
/images/index.html
/images.html
/samples.html
/index.html
/categories/software/index.html
/search.html
/guide/index.html
Rest of the pages

Figure 3.4: Islands from the *Music Machines* web log data.

In Section 3.4.2 we have discussed the sum valued cores. Each core actually represents a cluster or a sub-network which exhibits a stronger connection strength among the nodes inside the core. Following a similar concept Batagelj [15] has developed the island graph partitioning algorithm. Here an island is defined as a connected small subnetwork of size in the interval [*k,K*], where the links within the same subnetwork have stronger weights than the links among different subnetworks. In Figure 3.4, we have shown the islands extracted for the size interval [5,10]. Two main disconnected islands have been identified. The green and red coloured nodes represent these islands; while the white coloured nodes are considered as submerged and represented as the blue ellipse. It is interesting to note that the pages related to software patches are identified as a different group (red coloured nodes). This means if the user visits one of the software patches related pages, then he/she tends to visit other software patches related pages. This kind of insight into user behaviour certainly can help in the decision making process of the business organisation/web site manager. As a proof of concept and to validate this result, we applied k-means to produce the same number of islands (i.e., two) and we used the centroid of each island as the initial cluster seed. The outcome from k-means was similar to the outcome obtained from Pajek.

As a user browses the site, a window of the last *n* pages is maintained. The number of transactions to keep in a user's window can be determined based on the average user transaction length. Every time a user visits a new page on the site, the partial session is matched against the web page clusters/subnetworks. This can be used to classify the users into web usage community model(s). Note that they might fall into more than one community model because not all users will have exactly the same browsing habits. The community model information can be used to dynamically provide the user with customised content or targeted advertising.

3.6 conclusions

The study described in this chapter demonstrates the power of combining various techniques into a robust method. Indeed, the utilisation of multiple perspectives has been widely used in various applications from health informatics, bioinformatics, business, etc. In here, we once again emphasised the importance of employing multiple perspectives by utilising three techniques from different fields, namely statistics, data mining and network analysis. The results reported in this chapter showed that each of these three techniques reveals some unique aspects of the analysed data. Though there is overlap in the discoveries by the three techniques, however, once combined the results turn into more solid outcome that increases the confidence in the reported recommendations. Explicitly speaking, the statistical techniques concentrated on the probability of visiting a certain page given the sequence of previously visited one or more pages. On the other hand, the association rules mining approach studied the actual links between pages by analysing the existing hyper-links to determine the degree of support and satisfaction such links afford to users who are willing to navigate the given web site. Further, the log is processed using frequent pattern mining to determine pages which are mostly visited together and hence should be linked as a sequence. The linkage is realised by employing network construction and analysis. The three results complement each other and provide excellent guidance for restructuring the analysed web site in a way that maximises the overlap between the results from the three techniques. In other words, a better web design would lead to more overlap between the outcome from the three techniques employed in this chapter and hence will lead to higher confidence in the applied methodology.

Chapter 4

How Network Construction and Analysis Serves web Structure Optimisation

The World Wide web is growing continuously and rapidly; it is quickly facilitating the migration of daily life tasks into web-based ones. This trend shows time will come when everyone is forced to use the web for daily activities; indeed this does not look far ahead. Naive users are the major concern of such a shift. So, it is necessary to have the web ready to serve them. We argue having well optimised web sites will increase the trust of the visitors who will easily browse to their target pages to find the required information. The importance of having easy to access web sites is increasing due to the widespread reliance on the many services available on the Internet nowadays. It is true that search engines may be used to directly find the required information. However, search engines will never replace, but do complement the optimization of a web site's internal structure based on previously recorded user behaviour. Personalised search is developing fast and may dominate in the near future. While in Chapter 3 we utilized web usage mining to highlight important pages, in this chapter, we present a novel approach for analysing the structure of a given web Site to identify problematic connections and pages and hence suggest some

The content of this chapter is mainly based on the following published papers:

M. Nagi, A. M. ElSheikh, I. Suleiman, P. Peng, M. Rifaie, K. Kianmehr, P. Karampelas, M. Ridley, J. Rokne and R. Alhajj, "Association rules mining based approach for web usage mining", *Proceedings of IEEE International Conference on Information Reuse and Integration*, Las Vegas, Aug. 2011.

M. Nagi, A. Elhajj, O. Addam, A. Qabaja, O. Zarour, T. Jarada, S. Gao, J. Jida, A. Murshed, I. Suleiman, T. Özyer, M. Ridley, R. Alhajj, Robust Framework for Recommending Restructuring of web sites by Analysing web Usage and web Structure Data, Journal of Business Intelligence and Data Mining, 7(1/2): 4-20, 2012.

recommendations based on web structure and web usage mining. This method consists of two phases. The first phase compares user behaviour, derived by employing web usage mining, to a combined analysis of the web site's link structure obtained by applying three methods leading to more robust framework and hence strong and consistent outcome: (1) constructing and analysing a social network of the pages constituting the web site by considering both the structure and the usage information; (2) applying a Weighted PageRank algorithm; and (3) applying the Hypertext Induced Topic Selection (HITS) method. In the second phase, we use the term frequency/inverse document frequency (TFIDF) measure to investigate further the correlation between the page that contains the link and the linked-to pages in order to further support the findings of the first phase of the described approach. The obtained results will be utilised to identify problematic web site structures which are anticipated to be a major concern of web site owners. This chapter is organised as follows. Section 4.1 is an overview of the problem tackled in this chapter; we highlight its importance and briefly describe the proposed solution. Section 4.2 describes the proposed approach; we first discuss how to benefit from web structure and web usage mining for web site optimisation; then we discuss how web content mining helps in increasing the confidence in the recommendation; and finally, we discuss how the overall recommendation is delivered to the users/owner of the analysed web site. Section 4.3 reports test results that demonstrate the applicability and effectiveness of the proposed integrated approach.

4.1 Introduction

The rapid development in technology has motivated a shift towards electronic documentation and communication. Almost everything is turning into web-based service and more sophisticated web sites are being developed in a very competitive environment. However, the increased accessibility and availability of e-commerce sites even on mobile phones necessitate the existence of flexible web sites with easy to reach pages. Actually, it is becoming harder for visitors to locate the required pages and information due to the increased complexity of web sites. Thus, it is becoming more important to restructure web sites in order to increase their visitors' group by targeting visitors who are not professional enough to find their way through a web site. Analysing the visitor's behaviour along with the internal web site structure and content will provide insight on how to optimise the web site's structure in order to improve its usability. This is possible and feasible by a combined approach that integrates web mining techniques with social network analysis. We are mainly motivated by the successful application of the social network methodology in various domains, e.g., [35, 52, 101]. Social network analysis is intended to combine web usage information into the process in order to find a network structure alternative to the one produced by web structure mining. Analysing both networks could lead to stronger and more consistent results. This is true because the combined approach investigates the web site structure from two perspectives, the actual structure and the structure induced from web usage.

The approach proposed in this chapter involves two phases. The first phase uses web structure and web usage mining combined with social network construction and analysis techniques in order to find an initial recommendation for restructuring the analysed web site. The second phase relies on web content mining to determine a more appropriate linking between pages by considering the context around each link. The combined result from both phases leads to a higher confidence in the final recommendation; we use the term frequency/inverse document frequency measure (TFIDF) [96, 47, 94], which is the most common measure used in analysing documents for information retrieval.

For the first phase, we applied two trends. The first trend investigates how to use both the Hypertext Induced Topic Selection (HITS) method [63] and the Weighted PageRank algorithm [8, 16, 107] for web-structure mining; this leads to structure based analysis of the hyperlink structure of a web site. We derive another structure which is realised as a social network constructed by considering web log content. Actors in the social network are the pages constituting the web site and links are derived by applying association rules mining technique on the web log content. Two pages are linked together if they, respectively, appear in the antecedent and consequent of the same association rule. The weight of the link increases to reflect the number of supporting association rules.

We further illustrate how to employ web usage mining to obtain data on the site user's specific navigational behaviour. We then describe a scheme to interpret and compare the intermediate results in order to measure the web site's efficiency in terms of usability. Based on this, we provide some recommendations to web site owners in order to assist them in improving their site's usability. It is obvious that such an approach is needed to make web sites more attractive to end users who navigate web sites looking for particular information. Having the information buried deep within the web site discourages users from continuing their link navigation process. They may stop visiting such web sites and move to the competitors; a situation not appreciated by web site owners especially in such a highly competitive environment. Therefore, the result from the approach described in this chapter is intended to make web sites more attractive to most users by considering users' behaviour and web site structure.

Pages are clustered based on web structure data by considering each page as an object and its adjacent pages as its features. The result reports in the same cluster, those pages which are highly connected. Frequent pattern mining is applied on web structure data by considering pages and their adjacent pages as transactions and items, respectively. Each maximal closed frequent set of pages includes pages which are highly connected. Finally, the adjacency matrix that summarises the web structure data and its transpose are different because the graph is directed. Multiplying the two matrices will lead to a rich network where each link is characterised by a weight which reflects how strong the link is by considering other pages adjacent to the two pages connected by the link. Finally, the knowledge produced by each of the three techniques is reflected on to the main network which has already been enriched by the knowledge discovered from the web log data.

Explicitly speaking, motivated by the need of visitors to have flexibility in reaching their target pages and by the need of web site owners to increase the number of satisfied visitors, this chapter describes a robust framework which is capable of analysing web log data and web site structure data to recommend a restructuring of the web site. The target is a win-win situation when visitors and owners are considered. Each data source is treated as a two dimensional table. For the web log rows are users or IP addresses and columns are pages of the web site. We get one row per user session and an entry (*i, j*) indicates the amount of time user *i* spent at web page *j*. The table is analysed using three independent techniques, namely clustering, frequent pattern mining and network analysis. For clustering, each column or web page is one instance and rows are features. The clustering technique distributes pages into groups such that pages that have larger number of visitors in common are in the same group. The result from the clustering suggests pages in the same group should be linked together more than other pages because users mostly access them together and providing smooth transition between these pages will increase the number of satisfied users.

For frequent pattern mining, we consider each IP address or user as a transaction and each page as item. The result produces frequent sets of pages. Each set includes pages which are frequently accessed together. This is mainly a huge collection of frequent sets. We reduce the number of frequent sets to be considered by concentrating on maximal closed frequent sets. This way, we eliminate small sets which are mostly subsumed by larger sets. A frequent set *A* is closed if one of two conditions is satisfied. It is not subsumed by a larger frequent set *B* or the two sets *A* and *B* do not have the same frequency. A closed frequent set is maximal if it is not subsumed by any other frequent set. As in clustering, each maximal closed frequent set of pages deserves special attention to recommend linking such pages together as they are frequently accessed together.

Network analysis considers the two dimensional table as a bipartite graph connecting users or IP addresses to pages. In other words, the two dimensional table corresponds to the adjacency matrix. The network may be folded to get a one mode network of web pages which is the concentration of the study described in this chapter. This is possible by multiplying the transpose of the adjacency matrix and the matrix itself to produce a new table where rows and columns are web pages. Folding the other way around will lead to one mode network of users or IP addresses. Though the latter network may lead to valuable discoveries it is out of the scope of our study described in this chapter. We mainly concentrate on the former one mode network to find communities of pages. Actually the one mode network of web pages may be a good recommendation for actual linking of web pages because it reflects the linkage between web pages based on patterns of their actual access by users. This network is enriched further by reflecting on to it the knowledge discovered by the other two techniques. The weight of every link in the network is adjusted up to incorporate the number of maximal closed frequent sets which contain the two pages connected by the link. The weight is also increased if the two pages connected by the link exist in the same cluster. This network is also enriched by considering web structure data as described next.

web structure data reflects how pages are linked in the current design. This data may be represented as a two dimensional table reflecting the adjacency between the pages. We can incorporate this directly as another piece of information to enrich the network described above. We benefit from the web structure data by applying the three techniques (i.e., clustering, frequent pattern mining and network analysis) to determine more informative knowledge to enrich the main network.

The reported test results demonstrate the effectiveness and applicability of the proposed approach. Finally, it is worth emphasising that use of search engines will work in parallel, not as an alternative, to the web site optimisation process described in this chapter. Rather search engines could complement this process when a user is not willing to navigate through the pages of different web sites. The user in such a case may prefer to use a search engine as a first step to land at a particular page and from there navigate to other pages. To sum up, the contributions of the work described in this chapter could be summarised as follows.

- Benefiting from web structure, web log and web content information in order to analyse the structure of a given web site.
- Utilising the social network analysis methodology to derive an alternative structure by analysing the web log.
- Clustering of web pages based on the information derived from the structure of the web site to identify similar pages based on neighbourhood.
- Combining all the results into an integrated robust approach that leads to more consistent and stronger recommendations for altering the current structure of the analysed web site.

As a result, the weight of each link reflects the necessity to have it considered in maintaining the web site. Links with lower weight can be eliminated in case there is an alternative convenient short path which consists of links with higher weight. The final enriched network is reported to web site owner(s) as the recommended alternative design.

4.2 The Proposed web site Analysis Approach

Hyperlinks within a given web site encode what developers anticipate as the desired connections between pages of the web site as to be navigated by end users. Specifically, the existence of a link from page p to page q in the web represents a concrete indication that the two pages p and q contain some related information; such pages are mostly visited together, which is mostly not true given the diversity of web users who may have different interests. Consequently, it is more appropriate to revisit, evaluate, and adjust a web site design periodically by considering users' behaviour as reflected in the web log and the web site structure which is recognised as a directed graph. The other directed graph considered and analysed in this chapter is a social network of the pages with the links derived by mining the web log data.

In general, a set *V* of hyperlinked pages can be viewed as a directed graph *G* = (*V,E*), where the nodes correspond to the pages, and a directed edge (*p,q*) $\in E$ indicates the presence of an actual link from page *p* to page *q*. Each web page presented in the graph has an outdegree and in-degree; each of the two degrees may be any positive integer greater than or equal to zero. The out-degree of a page is the number of links present in the page; such links lead to other mostly related pages. The in-degree of a page *p* is the number of pages which include a hyperlink to page *p*. So, our target is to check the links and validate them by applying a systematic approach that incorporates certain criteria to evaluate the overall structure of a web site.

To achieve our target of recommending modifications to the link structure of a web site, we decided on two main subproblems that should be initially tackled. The first subproblem involves determining important pages by investigating the structure of the web site; two sources are used for this information: the actual hyperlink structure and the social network structure. The actual hyperlinks structure is analysed using two methods, namely HITS and weighted PageRank. The other structure is derived as a social network by using the web log. The current web site structure is said to satisfy the visitors if the social network based structure is very similar to the actual hyperlinks structure. In other words, the social network reflects the trends in users' navigational patterns and the target is to adjust the current web site structure to best fit users' expectations by matching the actual structure of the web site to the deduced social network structure. The outcome from this first subproblem complements the investigation of the second subproblem where the main target is to conclude which pages the users of the studied web site consider as important, based on the information harvested from the web log.

From the above description of the two subproblems, it is obvious that the web log serves two purposes for solving the two subproblems. By solving the two subproblems, there are two methods which are separately capable of ranking the same web pages. After the two rankings are produced, the next step is to implement a scheme which utilises the results and makes meaningful recommendations. These tasks and the related algorithms are described in the next subsections. The outcome from this first phase is supported by a second phase that emphasizes web content mining to study the correlation around each link in a page and the pages directly and indirectly pointed to by the current page.

In the rest of this section, we describe the different components of the first phase of the proposed approach, namely web structure and web log mining; we then analyse the combined result from the first phase. Finally, we concentrate on the second phase which completes the overall analysis process.

4.2.1 From web Structure Mining to Page Ranking

A web site generally consists of a set of web pages mostly interrelated and well connected. To study the structure of a web site it is necessary to first crawl through its pages to locate and extract relevant information to be used in the analysis. Then, each page is given a rank based on the analysis. The crawling process proceeds as follows. First, the root of a set of web pages is provided. Second, an application called a crawler is employed to traverse the set of pages by starting at the given root page. The crawler extracts the required information from all the visited pages. The particular information relevant for the work described in this chapter are the hyperlinks that exist within the visited pages. This is the basic information needed for producing a graph summarising the web site structure.

As alternative to crawling, it is possible to analyse pages of a web site and extract the required information using regular expressions. However, there are some risks associated with using regular expressions. For instance, regular expressions assume that the code used within the web pages follows certain standards. Further, simple errors, such as having some HTML tags not closed or improperly formatted and non-HTML code, such as CSS or JavaScript, may negatively affect the parsing of the page and this may lead to inaccurate results.

The system administrator decide on the portion of the web site to be crawled and analysed, whether the full set of pages or a subset of pages constituting a web site. Starting with the first web page specified, the crawler locates, retrieves and adds to a queue all links in the page. Each link leads to another web page which is added to the queue of pages to be crawled next. The crawler recursively crawls all pages in the queue to locate, retrieve and add to a queue all links in each page. Of course duplicates and already visited pages are ignored. After all pages in the queue are processed, for each page p_i the standard page rank value, denoted $PR(p_i)$ is computed as follows.

$$PR(pi) = \frac{1-d}{N} + \mathrm{d} \sum_{p_j \in M(p_i)} \frac{PR(pj)}{L(pj)} \tag{4.1}$$

where N is the number of crawled pages, M(pi) is the number of pages that include the links connecting directly to p_i, $L(p_j)$ is the number of number of links that exist in page p_j, and d is a damping factor, usually chosen around 0.85, refer to [12] for more details and justification of this choice. The damping factor refers to the probability that a visitors of a given page will follow the links that exist in the page. It is necessary to consider this factor to differentiate between two cases: (1) visiting a given page, say p_l , directly by entering its link in the header bar of a browser or by using a search engine which will lead directly to the target page, or (2) reaching page p_l using a link from another page. The latter case should be when computing the page rank of p_l . Therefore, the fraction $\frac{1-d}{N}$ the above formula may be interpreted as the influence of a random jump to page p_i based on the page rank $PR(p_i)$. As a result, the above formula is comprehensive enough to cover all possible cases from navigating through pages of the web site to landing directly at a particular page.

Xing and Ghorbani [107] proposed what they called the *weighted page rank algorithm* (*WPR*); it is an improved version of the standard page rank algorithm described above. Their basic goal is to differentiate between popular and unpopular pages. Their argument is that the rank of a popular page should have a higher weight than the rank of an unpopular page. The *WPR* value is computed as follows [107].

$$WPR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} PR(p_j)\, W^{in}_{(p_j,p_i)}\, W^{out}_{(p_j,p_i)} \qquad (4.2)$$

Here, $W^{in}_{(p_j,p_i)}$ and $W^{out}_{(p_j,p_i)}$ are the weights of the link between pages p_j and p_i. These weights are computed as follows [107].

$$W^{in}_{(p_j,p_i)} = \frac{I_{p_i}}{\sum_{p \in R(p_j)} I_p} \qquad (4.3)$$

$$W^{out}_{(p_j,p_i)} = \frac{O_{p_i}}{\sum_{p \in R(p_j)} O_p} \qquad (4.4)$$

where I_x is the number of links leading to page *x*, O_x is the number of links in page *x* and *R*(*x*) is the set of pages that are linked to/from page *x*. Each component of the sum of page ranks is multiplied by its respective weight. This is the main difference between the standard page rank and the weighted page rank. While the former splits the rank of a given page p_k evenly between pages reachable directly from page p_k, the latter emphasises the fact that more important pages should be assigned a higher rank. Indeed incorporating weight in the page rank computation would lead to more consistent, realistic and accurate outcome as already pointed out by Xing and Ghorbani [107] .

After thorough analysis of the possible techniques already reported in [55], we decided to use *WPR* instead of the standard page rank; the weighted page rank algorithm produces better results as demonstrated in the experiments conducted by Xing and Ghorbani [107]. Using this concept also fills the gap between the structure based analysis and the social network based analysis because the latter method derives a weight per link by considering association rules. In other words, the outcome from the two methods will be comparable without any of the two methods taking advantage over the other; both are deriving links between pages of the web site and both assign weight to each derived link. However, the structure derived by the social network methodology is more practical because it reflects the actual links by analysing the web log. On the other hand, the web structure mining method derives all the available links regardless of whether they have been used on not; every hyperlink leads to an edge in the derived structure and this edge may not have a corresponding edge in the structure derived based on the social network methodology. We decided to assign zero weight to all such links unless they are highly favoured by weight from the weighted PageRank algorithm. By assigning zero weight, we drop from further consideration every link not favoured by the web log and hence the integration of the results becomes smoother.

The output of the weighted PageRank based processing stage is a list where each entry contains a web page and its weight computed by the weighted PageRank, denoted [$p_i, WPR(p_i)$]. Entries in the list are sorted in descending order by the weight values, *WPR*. However, the list is revised by reordering the entries based on the outcome from applying another web page ranking algorithm, namely HITS [63] as well as the social network methodology.

HITS is another algorithm for ranking web pages based on two values for each page, namely `authority` and `hub`. Here it is important to mention that hubs and authorities have been used in Chapter 3 to highlight important pages in order to guide owners to those pages as the best target for posting important advertisements. Here the two measures, hubs and authorities are used to help in restructuring a given web site.

Authority refers to pages that provide important and trustworthy information on a given topic; and hubs are pages that contain links to authorities. These two values are defined in terms of one another in a mutually reinforcing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs. In other words, authority of a page *p* is computed as the sum of the scaled hub values that point to page *p*; and hub of page *p* is computed as the sum of the scaled authority values of the pages pointed to by page *p*. Relevance of the linked pages is also considered in some implementations. In-degree of a page measures in a sense its authoritativeness; and the out-degree measures its value as a hub. Hubs and authorities together form a bipartite graph.

Classifying pages into authority and hubs directly influences the result from the weighted PageRank algorithm by favouring more pages classified as authorities.

4.2.2 Clustering based method

From the variety of clustering algorithms described in the literature, we decided to use a genetic algorithm (GA) based clustering approach which has been developed by our research group [75, 79]. This would have the benefit of local technical support. We modified the approach to better fit our needs. The GA based process involves a number of steps including initialisation

of the individuals (interchangeably called chromosomes), cross-over and mutation to produce new individuals with better convergence towards the final solution, and selection based on fitness which is achieved as a combination of the three objectives enumerated above. We apply a variation of arithmetic cross-over in order to produce the number of clusters as a by-product of the process without requiring it as an input parameter. This leads to more natural clustering with fewer input parameters.

The length of every individual's vector is equal to the number of users or IP addresses in the given dataset. The initialisation process starts by distributing the given users into clusters by considering the time spent at each page to produce 50 initial individuals. We produce alleles of an individual by taking the time in seconds modulus *m*, where *m* is the number of closed frequent sets of pages (produced by the frequent pattern mining technique) which is considered as the upper limit for the number of clusters. This number of clusters will be tuned up by the iterations of the GA to end up reporting the appropriate number of clusters.

Cross-over is an operation that leads to the evolution of the population towards the final stable state. It is supported by mutation to help in quick convergence. There are several crossover operators in use [99]. They mostly take two individuals as input and produce one or two new individuals that may have better fitness than their parents. The cross-over operator used in our work takes two of the existing individuals and processes them to produce a new individual. Corresponding alleles within the two input individuals *p* and *q* are used to produced alleles of the new individual *r* as follows: $r_i = ((p_i \times q_i) mod m)+1$ (where *i* ranges from 1 to the number of users in the input dataset) to produce integer values in the range [1,*m*]. If some values from the range [1,*m*] do not appear in the new individual then values of alleles within the new individual are mapped to guarantee consecutive cluster numbers starting with 1. Mutation is applied to guarantee faster convergence. After the cross-over operation is completed, the number of individuals increases to 75. Then the fitness of each individual is measured as the sum of the average homogeneity of its clusters, average separateness between the clusters and average size of the clusters. Homogeneity and separateness are measured using Euclidean distance between the pages within the same cluster and between the centroids of each two clusters, respectively. Finally, the 75 individuals are ranked based on their fitness in descending order and the best 50 individuals are kept for the next iteration. The process continues up to 500 iterations and the best individuals are considered as the final clustering solution. However, to validate the latter selection process, we apply cluster validity indexes on the top 10 individuals from the final solution produced by the GA process. The outcome from the validity analysis will confirm the appropriateness of the selected solution, which is the solution favoured by the majority of the validity indexes [75]. The solution returned as the most appropriate for the given set of pages gives an idea about pages which should be connected together. The normalised distance between each two pages in a cluster will be added to the value of their entry in the adjacency matrix of the social network to be discussed in the next section.

4.2.3 Social Network Based Ranking

The social network is built by analysing the web log data. Actors in the social networks are the pages. We construct the adjacency matrix by mining association rules from the transactional database obtained after preprocessing the web log data; each transaction is a set of pages accessed together in one session. Frequent patterns of pages are determined by the same techniques applied in Chapter 3. From the identified frequent pattern, rules are derived and lead to the adjacency matrix as outlined next.

First, all entries in the adjacency matrix are set to zero. Then, Apriori is applied on the derived transactional data and association rules are derived. Each rule is reflected in the adjacency matrix by incrementing every entry (*i, j*) such that pages *i* and *j* exist in the antecedent and consequent of the rule, respectively. Further, for each two pages that ended up in the same cluster from the previous section their each entry in the matrix is adjusted by adding their normalised distance. Finally, entries in the adjacency matrix are normalised by dividing each value by the overall average of the values that exist in the matrix. The outcome is used as input to Pajek (a social network construction and analysis tool) [14]. The social network is analysed to rank the pages by considering their in-degrees, out-degrees, and betweenness centralities. Pages with high betweenness centrality are considered as important to link pages from different communities. Each page gets its rank value as the average of the three values: normalised indegree, normalised outdegree and normalised betweenness measure.

4.2.4 Ranking Pages based on web Log Mining

Seeking user opinion is always a target but mostly hard to achieve thoroughly. However, we discovered that the web log could be the best available source that may be used to pull out users' opinions without hurting any party and even without any precautions that might lead to bias. Users normally access web sites with certain targets in mind; they navigate between pages at their own choice; they decide how long to spend on each page; and they may decide to leave and come back later. All this valuable trace is recorded and summarised in the web log. Therefore, we can analyse the web log and extract a model of users' "opinions" which could be another metric for ranking the web pages. We base the latter ranking on two parameters: (1) frequency of access expressed as the number of visits, and (2) time computed as the total time spent by all visitors browsing to a web page. These are important factors which could be computed by analysing the web log. This illustrates how rich the web log is because in Chapter 3 we analysed the web log from another perspective and to satisfy another target, i.e., to decide on important pages for posting advertisements.

A page may be visited as a target or as an intermediate "hop" in the way to a target page. Thus, the time spent at a visited page is more important than having the page visited. To illustrate this, consider three pages *A*, *B* and *C*, which are linked as follows: *A* → *B* →*C*,. This means a viewer of page "A" should pass through page "B" to reach page "C". Therefore, the three pages will be seen as equally important because they have the same number of visits by considering the number of accesses to pages as recorded in the web log file. However, by taking the time spent at each page into consideration it becomes obvious that page "B" serves only as an intermediate "hopping" page and hence should not be considered important based on this particular case. As a result, it may be a better choice to add a direct link from page "A" to page "C" to avoid the need to pass through page "B". This will be more attractive solution for visitors interested in the two pages "A" and "C". The importance of adding a direct link increases as the number of "hopping" pages between "A" and "C" increases. To better recognise and differentiate target pages from "hopping" pages in the set of visited pages, it is important to give in the ranking scheme higher weight to the time spent viewing a page and less weight to just having the page visited.

Consider a page *p* and let let v_p be the number of visits to page *p* and let t_p be the total time spent by all visitors at page *p*, the *log rank* value l_p is computed as follows.

$$l_p = \alpha v_p + \beta t_p \tag{4.5}$$

where α and β are two parameters to decide on the importance of each of v_p and t_p; their values are specified by a domain expert or the investigator involved in the analysis of a web site based on the specific case under investigation. We arbitrarily set the values of α and β to 0.2 and 0.8, respectively.

Here it is important to note that the number of visitors to a given page *p* is computed by summing all occurrences of page *p* in the log file. In other words, in the case where a user visited page *p* say three times then v_i of page *p* is incremented by 3, not by 1. This way, we are able to include in the log rank a measure of how important page *p* is to every user. Such measure will be lost when the number of visits by each user is normalised down to 1 for visited pages and to 0 for unvisited pages. Actually, this ranking is directly reflected on to the social network of the web pages because the number of times a web page is visited is taken as its frequency each time a sequence of pages constituting a session are processed by the frequent pattern mining algorithm.

Computing a weighted sum by considering the number of visitors and the total time spent viewing the page will lead to a value which reflects the importance of each page. Ranking the pages by their computed weights will highlight the importance of each page with respect to the other pages. This way, pages which are more frequently visited and viewed longer will rank on the top while pages which are only visited with short view time will rank lower. To emphasise the importance of view time as compared to number of visits, we decided to give higher contribution to the time in the computation and hence we have chosen to split the share into 20% for the number of visits and 80% for the time spent; both values are normalised to avoid dominance and bias of any of the two over the others. In other words, each of the number of visits and the time spent viewing a page is normalised to a value in the interval [0,1]. Here it is worth mentioning that the share of the number of visits and the time spent viewing a page may be adjusted depending on the web site to be analysed because each web site has its own characteristics. Some web sites are text intensive and hence require more time to grasp the content compared to other less text intensive web sites. Keeping in mind the balance between the two parameters per web site will lead to a more accurate weight function that better reflects the particular case being analysed.

4.2.5 Combining the Ranks into a Recommendation

The measures obtained from the HITS, weighted PageRank, web log mining, and the social network methodology are combined to produce a final ranking for the pages; the process is mostly important to break ties and to validate the ranking of pages by recomputing the rank for each page as the average rank from the five approaches. This turns the page ranking process into a more stable task with high confidence in the final ranked list.

The results from this process are reported directly to web site owners in the form of recommendations to guide them regarding the possible alternatives for changing their web sites. The input required for this stage includes:

- A sorted list of web pages where each page appears with its rank value pi ∈ R+. This list should be sorted by the page rank value, which is the combined rank produced by the three approaches weighted PageRank, HITS and social network.
- A sorted list of web pages where each page appears with its ranking value lp ∈ R+ derived from the web log mining task (the list is sorted by the ranking values).

These two criteria are utilised by the three steps of the analysis as described in the remainder of this section.

Preprocessing:

Initially, it is required to preprocess the page rank and log rank values. This is realised by normalising the two values by mapping them to a common index of integers. First, the largest page rank value is determined.. Then each page rank value is divided by the maximum page rank value. This way all page rank values are mapped into the interval [0,1]. Recall that one of the main targets of developing the integrated approach that employs the weighted PageRank, the social network methodology, and the HITS algorithm in the process has been to get a more robust framework that produces more consistent results. In case of a tie in the rank mapping (two rank values map to the same value in the interval [0,1]), then the page that has higher rank according to the social network methodology is given preference. We decided on giving higher priority to the social network result because it combines both the web structure and web log features of the analysed web site. In case two pages get a tie on the social network analysis based rank then the tie is broken by considering the values produced from the weighted PageRank. Finally, the same normalisation process is repeated for the log rank values, i.e., each log rank value is mapped into the interval [0,1] by dividing each log rank value by the largest log rank value.

This step transforms and maps the page rank and log ranking values from two different distributions into a simple linear distribution in the interval [0,1]. This provides for more realistic and consistent comparison of values from the two domains based on their corresponding values from the [0,1] interval. The outcome from this step consists of two lists, namely the page rank and the log rank indexes in the interval [0,1] along with their respective pages. At this stage of the process, the two lists are still independent of each other.

Guideline to Recommendations The outcome from the preprocessing step can be used to compute the following value d_p for each page p:

$$d_p = index(l_p) - index(p) \tag{4.6}$$

Equation (4.6) computes the difference in rank between the two rank values. Ideally, we will find $d_p = 0$, because little deviation is expected in the final ranking of the page rank and log rank values. This point has been indirectly raised above when we discussed the analysis of the social network of web pages. Here, it is worth highlighting that d_p will be zero when the page gets the same rank from weighted PageRank, HITS and the social network process because the social network combines both the web structure and web log information; it is actually another way of quantifying the value of d_p. Finally, pages are sorted in ascending order by considering their d_p values. The following two cases may be witnessed at the top and the bottom of the list, respectively:

Case 1 Pages characterised by a large page rank index and a small log rank index (d_p very low).

Case 2 Pages characterised by a small page rank index and a large log rank index (d_p very high).

In case 1, the outcome from the analysis may be articulated as a recommendation to the web site owner to move the page into a location where it is not directly reachable. This may favour pages which may need to be reachable more easily. Some of the legitimate action which may be taken by the web site owner may include:

- Removing links which lead to pages classified under case 1, i.e., links that appear in pages with high page rank.
- Linking to pages classified under case 1 from other pages characterised by low page rank value.

In case 2, the recommendation would be to modify the link structure in a way that makes a popular page easier to reach. This may be realised by adding inside some appropriate pages links which will lead to popular pages. Pages with high page rank are considered the most appropriate pages to which the mentioned links could be added. One alternative may be to direct jump onto popular pages from pages far away towards the main page of the web site. In other words, it is recommended to shorten the path that should be traversed in order to reach pages highlighted in case 2. Here, it is worth emphasising that these are only recommendations which may be ignored or partially considered by owners of web sites depending on their expectations and needs.

The page rank value is a good indicator about the difficulty of accessing a given web page. For instance, a high page rank value means an easily reachable page while a low page rank is mostly associated with pages which are hard to reach. The former pages are generally reachable from important pages, it is like the rich get richer. On the other hand, the latter pages are reachable from less popular pages and hence could not be easily noticed. Similar to the page rank value, a high log rank value marks a popular page, whereas a low log rank value indicates unpopular page. Therefore, by combining the two ranks to compute d_p and by considering the two cases described above, it can be easily realised that in case 1 a page may be easy to reach but only may have been viewed by a small number of visitors; such page has a very low d_p value. On the other hand, pages classified in case 2, where the value of d_p is very high, are distinguished as pages that contain some information which may be of interest to many visitors of the containing web site, however these pages are mostly hard to reach due to their high d_p value. As a result, the best case is depicted when $d_p \approx 0$; this characterizes ideally positioned pages.

4.2.6 Ranking Pages by Employing web Content Mining

For this phase of the process, we decided on using the term frequency-inverse document frequency (TFIDF) measure, which has been successfully applied and produced promising results in information retrieval and text mining. Given a document which could be treated as a collection of words, TFIDF is a statistical measure which depicts the importance of a given word in a document from a collection or corpus. The importance of a word increases proportional to the number of times it appears in a given document; however, this value depends also on the frequency of the word in the corpus. In other words, given a document *D* from a corpus, a term/word which appears frequently in document *D* and in the whole corpus will get a lower score, e.g., very frequent words like “the”, “and”, “it”, etc.

The web content mining problem may be classified under text mining because we want to find the correlation between the term that appears in the link and other documents directly or indirectly pointed to by the page. As TFIDF is used to evaluate how important a word is to a document in a collection or corpus, we could smoothly use the measure by considering the corpus in our case as the set of documents pointed to directly or indirectly by following a particular link within the given page. For the term constituting the link, its frequency is measured within the linked-to page(s). Frequency of a term is measured as the number of times the term appears in a document; it is normalised to avoid potential bias towards longer documents, which might have more occurrences of the given term as compared to shorter documents. Term frequency, denoted *TF*, of a term t_i in document p_j is measured by:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{4.7}$$

where $n_{i,j}$ is the number of occurrences of term t_i in document p_j, and the denominator is the size of document p_j, i.e., the number of occurrences of all terms in document p_j. The inverse document frequency, denoted *IDF*, measures the importance of term t_i. It is computed as the logarithm of the value determined after dividing the number of all documents by the number of documents that contain term t_i.

$$IDF_i = \log(\frac{|D|}{|\{p_j : t_i \in p_j\}|}) \tag{4.8}$$

where $|D|$ is the total number of documents in the corpus, $|\{p_j : t_i \in p_j\}|$ is the number of documents in which term t_i appears (*i.e.*, $n_{i,j}$_= 0).

Combining both values computed above, we get *TFIDF* as:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i$$

A high term frequency in the given document leads to a high weight value of *TFIDF*.

We applied this process to find the *TFIDF* value for each term that appears in a hyperlink inside a page. The returned result guides the optimisation process. It is a step to confirm or tune up the outcome from the first phase. It is considered an important step because the first phase considers only the structure and usage of the web pages. On the other hand, by computing the *TFIDF* values for the hyperlinks, we dive deeper inside the web pages, analyse their content around the link(s) and comment on the importance of the link to the web page it points to as compared to the other web pages indirectly reached from the current page by following the hyperlinks.

4.2.7 The Relinking Process

Although we have a systematic approach that involves two phases to analyse a web site seeking how its links could be optimised, our approach is semi-automated in the sense that user (expert) validation of the outcome is needed before the optimisation is physically applied to the web site. In particular, the proposed approach will highlight the restructuring decisions to be taken for a better optimised web site. These decisions are to be investigated by domain experts and guide them for better restructuring the web site. So, the "relinking" process described above should be completed manually by the web site owner (i.e., the assigned Webmaster), who is considered the domain expert since he/she is assumed the most familiar with the content and structure of the pages constituting his/her web site; thus the suggestions of the proposed process could be articulated as follows. Consider two web pages, say "A" and "B" such that there is no direct link from "A" to "B", it will be highly recommended to add a direct link from "A" to "B" in case the analysis reveals that most visitors navigate to "B" right after viewing "A". On the other hand, in case a direct link to "B" already exists in "A" and the outcome from the analysis shows that visitors of page "A" rarely navigate to "B" then the suggestion would be to remove the link from "A" to "B". In other words, the outlined process determines any flows in the current structure of a web site, including misplaced pages or links, and to remedy the problem suggests some appropriate restructuring by adding, removing or moving links in a way that will optimise the navigation between pages.

The outcome from the analysis should be presented as recommendations to the web site owner. All recommendations are based on the values of 'd_p and *TFIDF*. The information to be presented to the web site owner is determined based on two threshold values (ε_1 and ε_2), which are mostly specified by a domain expert.

1. *UNLINK* list, which is a sorted list of pages characterised by having $d_p < 0$ and $d_p^* > \varepsilon_1$.
2. *LINK-TO* list, which is a sorted list of pages characterised by having $d_p > 0$ and $d_i^* > \varepsilon_2$.
3. For each web page *p* that exists in any of the two lists derived in (1) and (2) above, produce *INCOMING-LINKS*(*p*), which includes the set of pages that include links to *p*.
4. For *each* web page *p* that exists in any of the two lists derived in (1) and (2) above, produce *OUTGOING-LINKS*(*p*), which includes the set of pages directly reachable from *p*, i.e., there is a link from *p* to each of them.
5. The importance of each link by considering the *TFIDF* values computed for the terms that appear in the hyperlinks within the pages: this is presented as pairs of hyperlink and related pages, where the pages are sorted by their *TFIDF* values.

To sum up, it is important to emphasise the fact that the *TFIDF* value combined with the page rank and log rank values give better insight to the web site owner to help and encourage him/her to take action with higher confidence.

4.3 Evaluation of the Proposed Approach

We run some experiments to demonstrate the applicability and effectiveness of the proposed framework. For this purpose, we used the Music Machines web site (≈ 915 pages) [98], which provides references for HiFi devices. Its structure is more generally wide than deep (it provides breadth instead of depth in terms of links between pages) when it lists the manufacturers of documented devices.

The first set of experiments presented in Section 4.3.1 demonstrate the various aspects presented in this chapter leading to a recommendation suggesting web site restructuring. We also employ the ARM process to recommend access patterns based on the produced association rules as reported in Section 4.3.2. The second set of experiments extend the study beyond what has been described above to report the effectiveness of combining the mining of maximal closed frequent sets of pages with clustering and network construction. In a sense the second set of experiments combined some aspects from Chapter 3 with others from this chapter and studied the effectiveness of this integration on deriving important pages as well as on suggesting web site restructuring.

4.3.1 Experiments Conducted to Recommend web site Restructuring

The Music Machines web site already comes with a log file that had been parsed into sessions. This saved us the time and effort required to derive the sessions from the raw web log. In general a log file is not ready for direct use as input to the analysis process. It is a sequence of entries that summarise all access information to a given web site. Before a log file is used it is necessary to group its entries into sessions where a session could be seen as a sequence of log file entries which altogether last for almost 30 minutes.

We analysed the sessions and produced the transactions which are lists of pages accessed per session. These transactions were used as the input to Apriori which produced the association rules; the minimum support and minimum confidence thresholds were arbitrarily set to 20% and 75%, respectively. The association rules were then reflected into the adjacency matrix where entry (*i, j*) measures the number of rules that include the two pages *i* and *j* in the antecedent and consequent, respectively. The derived social network is shown in Figure 4.1. The social network of the top ranked pages is shown in Figure 4.2.

The employed social network analysis technique has been used to rank the web pages according to their importance; this could be used to validate the results outputted by the weighted

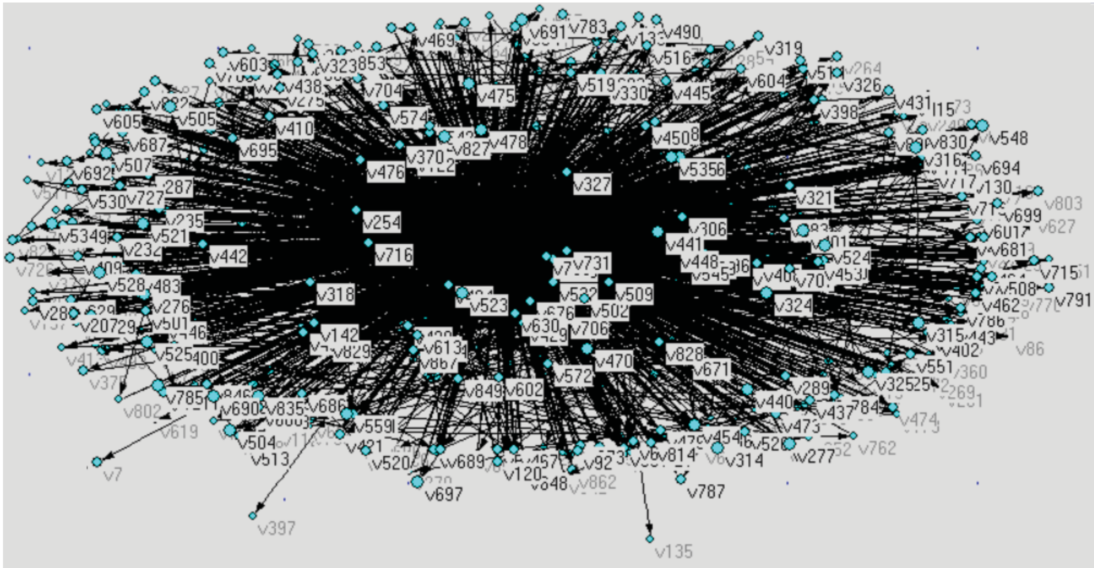


Figure 4.1: The complete social network

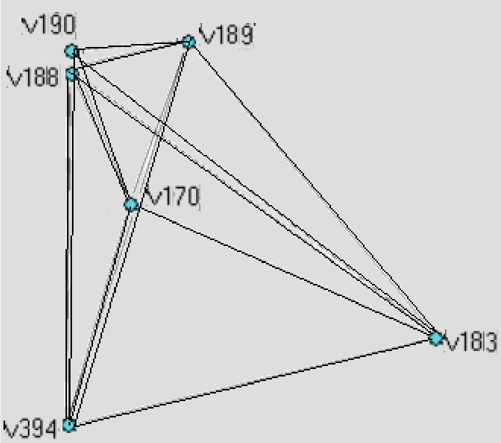
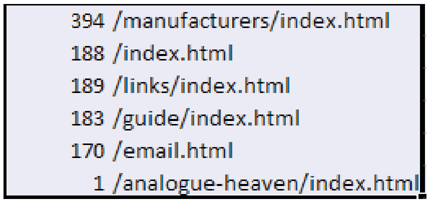


Figure 4.2: The most central pages in the social network



394	/manufacturers/index.html
188	/index.html
189	/links/index.html
183	/guide/index.html
170	/email.html
1	/analogue-heaven/index.html

Figure 4.3: The top six pages as ranked by the first phase

PageRank and HITS. A centrality analysis has been performed on the social network which has been constructed from the web site's pages and links between these pages. After comparing the results produced by the web site structure based techniques (PageRank and HITS) with the results produced by the social network analysis based method we have found that both agree almost totally on ranking the most important pages. Therefore, the most important pages discovered by both approaches are almost the same. The top ranked pages are listed in Figure 4.3. For example: /manufacturers/index.html -394-, /guide/index.html -183- and /email.html -170- have been ranked exactly in the same place by both approaches and they are in the top rank in both lists. The other most important pages have close ranking but not exactly the same. The exception is the following page /analogue-heaven/index.html. It has been ranked as the sixth most important page by the web site structure based techniques but it has not been considered as a central point according to the social network centrality measure. After closer analysis of this particular page, the explanation of this phenomenon could be articulated as follows: this specific page has two incoming edges from two of the very powerful pages. Therefore, the web site structure based techniques have considered this page as very important because this is how the web site structure based techniques work. On the other hand, this page was not considered as a central page by the social network analysis based method because it has very few incoming and outgoing edges.

The combined approach of the first phase analysis on the site used in the testing yielded interesting distribution of deviation values d_p. These results from the first phase of the proposed approach have been supported by consistent results from the second phase of the proposed approach. In other words, the computed d_p values and *TFIDF* values are consistent for this particular web site.

4.3.2 Experiments Conducted to Recommend web Access Patterns

We conducted some experiments to demonstrate the applicability and effectiveness of the proposed framework for recommending some access patterns based on web log analysis obtained from the Music Machines web site. The web log used in the testing has been anonymised. We downloaded a zip file of all of the logs for October of 1997. Each file corresponds to one day of server transactions by different users. Every transaction in the log file is separated by a user identifier.

After identifying all maximal frequent sets of pages, we found a weight for each page in each maximal frequent itemset. Of course the weight is zero if the page does not belong to the maximal frequent itemset. Otherwise, the weight for a page, say *p*, in itemset *X* is computed as average support of *X* divided by the sum of the average support of all itemsets *Y* (including set *X*) which contain *p*. The average support of an itemset *X*, denoted *AvgSpt*(*X*) is determined by dividing the support *X* by the number of pages in *X*. Formally, $AvgSpt(X) = \frac{support(X)}{|X|}$ and $Weight(p) = \frac{AvgSpt(X)}{\Sigma(AvgSpt(Y))}$.

By considering the maximal closed frequent sets of pages, we derive all possible association rules and retain only frequent rules. These rules constitute the recommendation system to be employed for guiding visitors to the web site. A rule is fired if all pages in its antecedent are visited and it is the rule that has the highest overall weight for the mentioned pages. Firing the latter rule means recommending to the visitor pages that appear in the consequent of the rule as the pages to be visited next because they have been visited by most of the previous visitors who visited pages appearing in the antecedent.

Some of the rules derived by analysing the input web log are shown in Table 4.1 where the weight of each item is shown between the parentheses. Notice how different pages have different weights in different rules. This means the rules have been derived from different maximal closed frequent sets of pages. We arbitrarily set minimum support and minimum confidence to 0.1 and 0.7, respectively.

Table 4.1: Sample rules from the “Music Machines” web log

P18 (0.3) P12 (0.125) ⇒P10 (0.02) P5 (0.14)
P9 (0.2) P5 (0.105) ⇒P0 (0.05)
P12 (0.1) P18 (0.024) P45 (0.01) ⇒P5 (0.25) P2 (0.1)
P3 (0.08) P37 (0.19) P1 (0.12) ⇒P2 (0.23) P6 (0.05)

The analysis of the web structure data produces some rules similar to the sample rules enumerated in Table 4.1. The two sets of rules are matched in order to enrich the set of rules produced from the web usage data. The enrichment process may increase the confidence in a rule. Once the latter set of rules is finalised it is kept as the core of the web usage recommendation system.

We also benefited from the association rules discovered from the web usage mining to find pages that are in most cases accessed directly after accessing some other pages. These pages may not have direct links and hence, it is recommended to consider them for adding some direct links to the satisfaction of accessing users.

Given the set of rules from web usage mining like the rules shown in Table 4.1, we rank the rules based on the weight characterising individual pages appearing in each rule. Recall that the same page may appear in different rules with different weights assigned to it as reported in Table 4.1. Further, the confidence in the rule is also used in the ranking process. So, for each rule, we find a new measure computed by adding its confidence and the average of the weights of the pages appearing in the rule. We consider the average in order to avoid any bias from rules involving more pages.

For each rule starting from the rule with the highest rank we check if page(s) appearing in the antecedent have direct link to pages appearing in the consequent. If the link does not exist, then we add such a link to the recommendation set with confidence computed as the average of the rule confidence and the average weight of the pages that appear in the rule. This process produces a list of recommendations which includes all possible links to be added to the analysed pages. The list is presented to web site owner for final approval. All approved links may be later on added to the pages.

4.3.3 Testing the Effectiveness of Frequent Pattern Mining Combined with Clustering

In this section, we combine the outcome from frequent pattern mining with clustering. In particular, we concentrate on maximal closed frequent sets and enrich the model by the result from the clustering approach. The target is to report some recommendations regarding important pages and web site restructuring. For this purpose, we used the Music Machines web site which contains 254413 different values for IP addresses available in the data set. The data is available for years 1997-1999. Since this web site has been provided for experiments with data mining techniques, it already came with a log file that had been parsed into sessions. This saved us the time and effort required to derive the sessions from the raw web log data.

We analysed web log data by concentrating on sessions to derive the table of users versus accessed pages. Each session is assumed to correspond to a user. The produced table is used in the analysis for the first stage of the experiments, namely to find clusters of pages by applying the clustering techniques described in Section 4.2.2, to find maximal closed frequent sets of pages by employing the frequent pattern mining technique described in Section 4.2.4, and to find the network of pages based on access patterns by applying the folding technique described in section 4.2.3.

We run Apriori by arbitrarily setting the minimum support to 20%. Then we concentrated on only maximal closed frequent sets of pages. The result reports in the same set pages which are mostly accessed together. A total of 18 maximal closed sets of pages were reported. Based on the result, the connections between the pages in each of the 18 sets are enriched in the network of pages produced from the technique described in Section 4.2.3.

We run the genetic algorithm based clustering technique which returned 15 clusters as the most appropriate solution. To further validate this result which is somewhat different from the result of 18 sets obtained from the frequent pattern mining approach, we applied cluster validity analysis techniques on the top 10 clustering solutions reported by the last generation of the genetic algorithm. The latter check confirmed 15 as the most appropriate number of clusters. We manually compared the fifteen clusters and the eighteen sets. We realised that three of the eighteen sets have major overlap with three other sets and this was the main reason for reducing the number of clusters to fifteen. We reflected the information related to pages in each cluster on to the network. For each two pages *i* and *j* which fall in the same cluster, we increment the entry (*i, j*) in the adjacency matrix by the normalised distance between pages *i* and *j*.

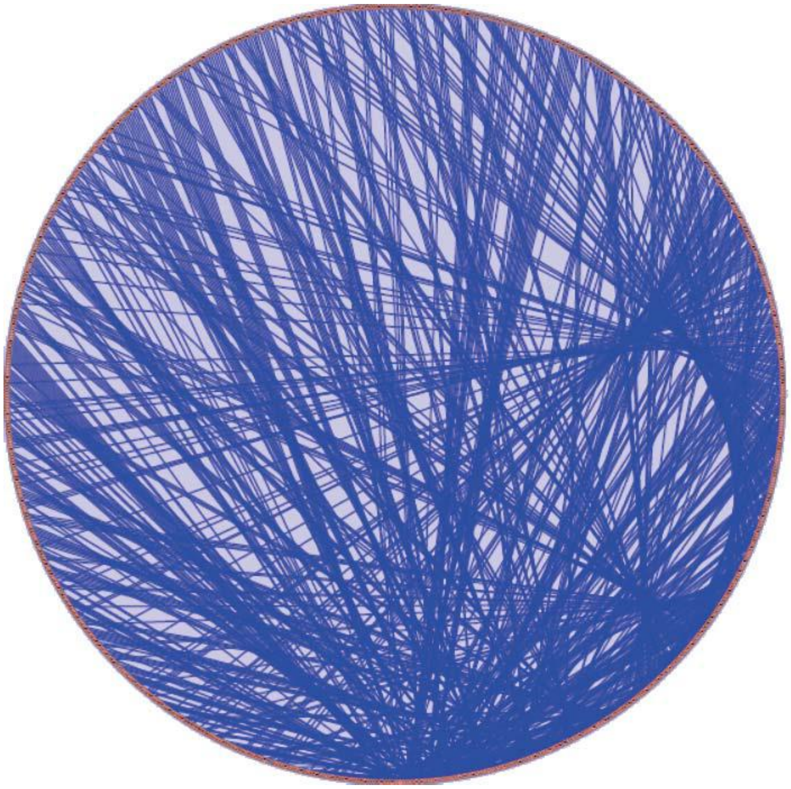


Figure 4.4: The complete social network

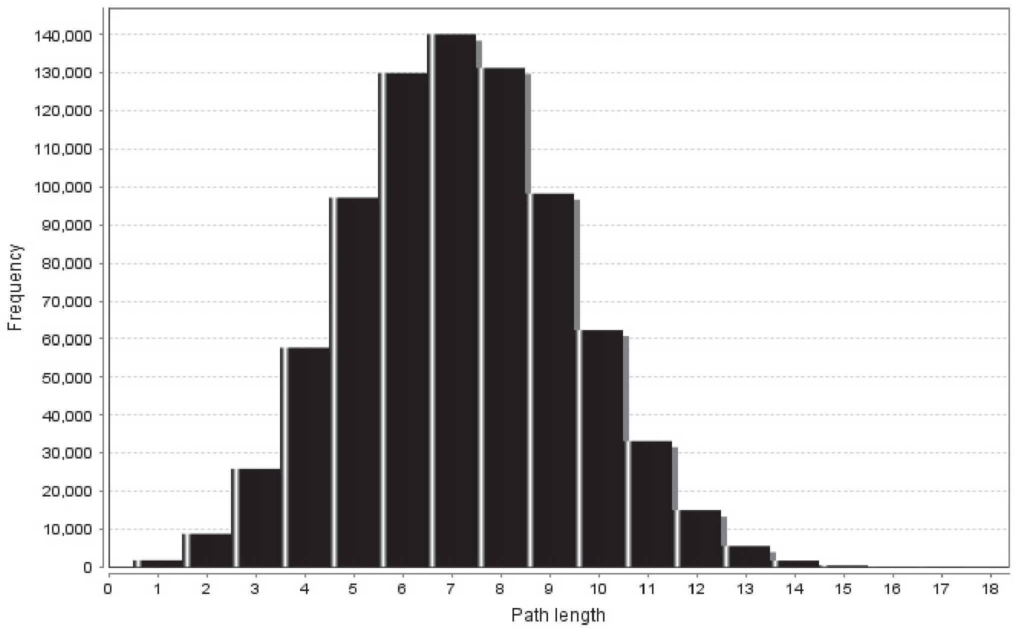


Figure 4.5: Frequency of shortest paths

After completing web log data analysis, we want to enrich the network of pages further by incorporating knowledge extracted from web structure data. This has been made possible by first producing a table of 831 rows and 831 columns, i.e., one row per page and one column per page. Each entry (*i, j*) in the table is set to 1 if and only if page *i* contains a link to page *j*. The produced table is used to serve three purposes as was the case with the table of users versus pages which was produced from web log data. For the first purpose, we consider each row as a transaction leading to 831 transactions with 831 items corresponding to columns. We run the frequent pattern mining technique to find all maximal closed frequent sets of pages. A frequent set contains pages which have in common a large number of referencing pages. These pages are anticipated to be related in some way. This information is reflected on to the major network that was derived from web log data. Entry (*i, j*) in the network is incremented if pages *i* and *j* are in the same maximal closed frequent set. The derived network is shown in Figure 4.4. The frequency of shortest paths is shown in Figure 4.5. The closeness and betweenness centrality measures are plotted in Figure 4.6 and Figure 4.7, respectively.

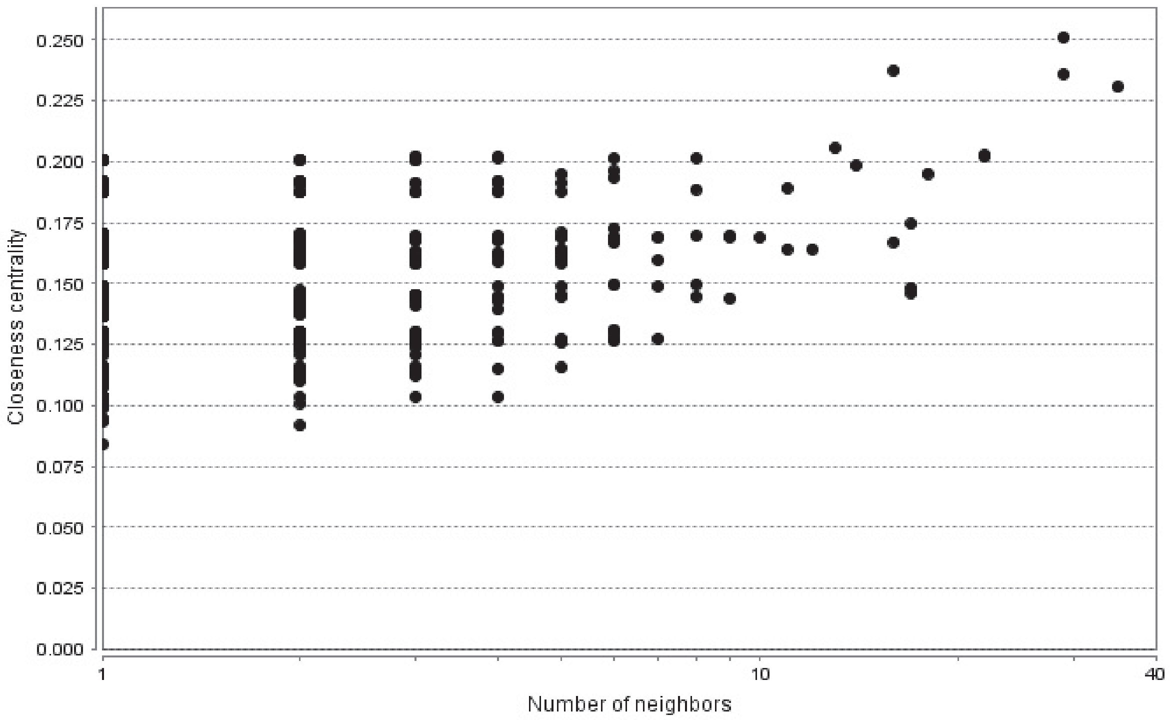


Figure 4.6: Plot of the closeness centrality measure showing number of neighbours per node

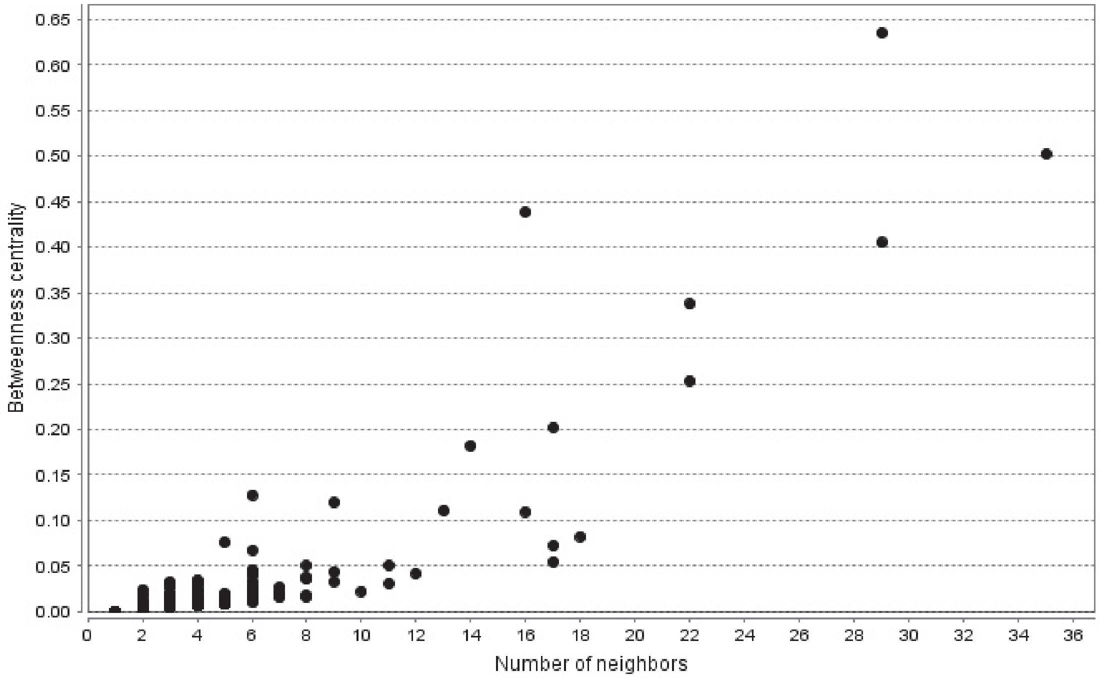


Figure 4.7: Plot of the betweenness centrality measure showing number of neighbours per node

For the second purpose, we consider the 831 rows as objects and the 831 columns as features. We applied the clustering technique described in Section 4.2.2 and produced 10 clusters which is the same as the number of maximal frequent sets of pages produced from the frequent pattern mining technique. We reflected the information from each of the 10 clusters on to the major network of pages by applying the same strategy used to reflect the information from the clusters produced from web log data. In other words, the normalised distance between each two pages in the same cluster is reflected on to the network. For the third purpose, we multiply the adjacency matrix by its transpose to produce a new adjacency matrix where each entry emphasises the strength of the link between two pages by considering the number of their common neighbour pages. At the end, we get an adjacency matrix of the same size as the adjacency matrix of the major network. We add corresponding entries to produce a richer adjacency matrix for the major network. This final adjacency matrix combines all the information from the six sources, three sources are based on web log data and three sources are based on web structure data. This matrix is normalised further by setting to zero any entry which is considered as an outlier compared to the other entries in the matrix. An entry is said to be an outlier if its value is less than 10% of the average value of all the entries in the adjacency matrix.

The information from the final normalised major network is reported to the web site owner as the recommendation produced by the framework. Links are classified based on their weights to give better idea to the web site owner who may decide to take into consideration all or part of the recommendations. For better visualisation, a number of sub networks are extracted from the major network and displayed to the user. Each subnetwork contains only links which share the same weight. The owner may decide to set a threshold and neglect all recommendations with weight below the set threshold.

The outcome from the process applied in this section will reflect the most appropriate setup for the given web site based on the access pattern reported in the web log and the structure. This may lead to rapid increase in the number of visitors who may like the new look of the web site with the links provided to meet their preference. Shall the user group preferences change, this will be reflected on to the new web log data to be generated after the web site has been redesigned. The new data will be used to readjust the web site in the next maintenance phase.

These are the most important pages reported by the proposed framework:

/manufacturers/index.html -394-,
/guide/index.html -183-,
/email.html -170-, and
/analogue-heaven/index.html.

4.4 Conclusion

As users access web sites they leave a trace in the log file. Such a trace form a valuable source for further analysis to produce informative recommendations which might be beneficial for web site owners. In this chapter we described an effective approach for web data analysis leading to recommendations. To sum up, the integrated robust framework described and analysed in this chapter is powerful enough to produce consistent and legitimate recommendations. It is a rich framework because it integrates three main pieces of information, namely web structure, web content and web log data. The integration of the network analysis methodology into the framework is a major gain as reflected by the consistent results produced. The influence of the network produced is high because it integrates both the structural and usage information together. In other words, it smoothly bridges the outcome from web structure and web log mining. As a result, the proposed framework has succeeded in identifying problematic locations in the web site's structure.

Chapter 5

Integrating Data Mining and Network Analysis for Feature Selection and Its Application to Gene Expression Data Analysis

Data analysis techniques developed for a specific domain may turn out to be effective for several other domains. The only challenge to realise such applicability is figuring out the mapping between a new domain for which a solution is sought and another existing domain for which a solution has been materialised. Once the mapping is resolved, application of the techniques becomes trivial. This has been demonstrated in this chapter by describing how the methods described and utilised in Chapter 3 and Chapter 4 are powerful enough to tackle various applications. One such application is the analysis of gene expression data to identify a reduced set of genes as potential disease biomarkers. This is an essential problem with both scientific and social impacts. The mapping process between the two domains tackled in this thesis could be generalised to cover other domains. For instance, a direct mapping could be realised between web log data and gene expression data where on one side a web log file contains sessions which reflect access patterns to pages constituting the corresponding web site and on the other side gene expression data contains values reflecting the expression levels of genes in samples. The mapping between the two domains is straight forward; each session or IP address could be seen as equivalent to a sample and each gene is equivalent to a web page.

The content of this chapter is mainly based on the following published paper:

S. Gao, O. Addam, M. Nagi, A. Qabaja, A. ElSheikh, P. Peng, O. Sair, W. Almansoori,T. Özyer, J. Zeng, J. Rokne, R. Alhajj, "Robust Integrated Framework for Effective Feature Selection and Sample Classification and Its Application to Gene Expression Data Analysis", *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, May 2012

While the analysis of web log data reveals pages that are frequently accessed together, the analysis of gene expression data would highlight genes which demonstrate similar expression profiles in most of the processed samples.

In general, genes are encoding regions that form essential building blocks within the cell and lead to proteins which are achieving various functions. However, some genes may be mutated due to internal or external factors and this is a main cause for various diseases. The mutation could be discovered by closely examining samples taken from patients to identify faulty genes. In other words, it is important to identify mutated genes as disease biomarkers. We can then build a classifier model based on certain normal and infected samples capable of successfully classifying new samples as infected or normal. The work described in this chapter addresses this problem by describing a framework that incorporates the two stages of the process, namely feature selection and sample classification. In fact, gene expression data is distinguished by high dimensionality in terms of the number of genes and small number of samples. Reducing the dimensionality is essential to efficiently analyse the samples for effective knowledge discovery. There is a trade-off between feature selection and maintaining acceptable accuracy. The target is to find the reduction level or compact set of features which once used for knowledge discovery will lead to improved performance and acceptable accuracy. For the first stage, we concentrate on three feature selection techniques which have been inspired from the methods presented in the previous three chapters and effectively used for web mining, namely frequent pattern mining and clustering from data mining, and community detection from network analysis. The effectiveness of the feature reduction techniques is demonstrated in the second stage by coupling them with classification techniques, namely associative classification, support vector machine and naive Bayesian classifier. Majority voting is applied for both stages. The results reported for four cancer datasets demonstrate the applicability and effectiveness of the proposed framework.

The rest of the chapter is organised as follows. A general introduction to feature reduction within the scope of gene expression data is presented in Section 5.1. The related work is summarised in Section 5.2. In Section 5.3, we describe the framework which consists of two stages, feature reduction and classification. Section 5.3.1 presents the various methods that constitute the feature reduction stage and the employed classification stage is described in Section 5.3.2. The test results are reported in Section 5.4.

5.1 Introduction

Real world applications like gene expression data analysis and image processing are characterised by high dimensionality which is a main bottleneck when the data is to be analysed for effective knowledge discovery. Realising this essential problem facing various applications, dimensionality reduction is one of the main research areas addressed in the literature, e.g., [17, 18, 34, 42, 62, 90, 100, 103, 104, 111, 43]. The target is to analyse a given set of features to decide on the ones most effective for a given task. For instance, it is possible to reduce a large set of genes into a small set of genes capable of distinguishing between infected and normal samples [93, 58, 68, 84]. The development of microarray technology allowed researchers to simultaneously study the expression levels of thousands of genes leading to huge amounts of gene expression data. Gene expression data analysis is an important research area that has attracted the attention of a large number of research groups who share a common goal to identify discriminating genes or biomarkers of various diseases. A biomarker is a molecule (mainly gene or protein) that exists in the cell and does not function normally because its chemical structure has been mutated due to internal or external factors. Minimising the number of biomarkers for a given disease is an essential process that contributes to huge reduction in the cost and feasibility of sample classification. The whole reduction process is based on the expression level of genes which is one of the important aspects that guide researchers in their effort to reveal various phenomena related to diseases threatening human life.

To tackle this essential problem, we present an integrated framework for feature reduction and sample classification. The framework incorporates two stages. Feature reduction is the main stage used to produce a reduced set of features which are intended to be informative. The second stage is classification which validates the outcome from the other stage by using the reported features to build a set of classifiers which should report high accuracy for the reduced set of features to be acceptable. We describe some techniques for feature selection and then study how their outcome positively affect classification techniques when applied to gene expression data analysis. In other words, the developed framework incorporates feature selection methods to reduce the number of genes and classification methods to classify samples based on the reduced set of features. The techniques employed for feature reduction are frequent pattern mining and community discovery by employing network analysis. We intentionally utilised techniques from various domains to avoid biased result. The outcomes from the feature selection methods are integrated into a more informative set of biomarkers. Then three classifiers are employed to build a robust classifier model. Majority voting is applied at both stages of the process, i.e., biomarkers favoured by most of the feature selection methods are chosen to the final set and the class favoured by at least two of the classifiers is decided as the target class for a given sample.

For frequent pattern mining, we concentrate on closed frequent sets of features which constitute a reduced and rich source of information. Clustering based feature selection distributes the features into groups based on their expression levels in samples. Then representatives are selected from each cluster to constitute the actual reduced set of features. Finally, a two-mode network is built between genes and samples based on the expression levels of genes in samples. Then the network is folded to produce a one-mode network of genes. The network of genes is then analysed to find communities of genes. From each community, the most influential gene is selected as the representative of the community in the reduced set of features. As a result, we get three different reduced sets of features, one from each of the three methods employed in the framework. Majority voting is applied to decide on the final set of potential biomarkers.

We used the outcome from the feature reduction process to compare the methods in order to decide on the most effective method which is the method that favours most of the selected features [95, 85]. The test results from the four cancer datasets favour the three techniques, namely frequent pattern mining, clustering and the network based approach as effective and efficient methods for feature reduction.

5.2 Related work

Monitoring gene expressions of the whole genome is one of the most challenging tasks in experimental molecular biology. Microarray technology has provided an enormous progress in cancer prognosis and drug discovery processes by enabling the detection of differentially expressed genes between two biological states, e.g., normal versus cancer cells, or untreated versus drug treated cells. However, it is well known that only a specific set of genes (differentially expressed between two different states) are biologically important. Hence, there is a need to filter out some irrelevant genes by using feature selection methods [46, 42].

Many feature selection methods have been proposed, e.g., entropy based method [39], correlation based [53], Singular value decomposition (SVD), and unsupervised feature selection method (UFF) [100]. The entropy based method works by filtering out features whose expression distributions are random by minimising the heuristic entropy. This is simply achieved by identifying the expression levels for all the features or patterns, and then identifying the frequency of these patterns and finally emerging patterns whose frequency can significantly change across different subsets of samples (cancer versus normal cells).

A correlation based feature selection method scores subsets of features rather than scoring individual features [53]. This method takes the importance of specific features into account for predicting a class. Another correlation based approach to define the features with the highest discrimination value is the *T*-statistic. In this case, every sample must be labelled with +1 if it belongs to a specific class or −1 if it does not. Then for each feature the *T* score is calculated by dividing the mean difference of that gene between the +1 labelled and the −1 labelled samples. This value is then divided by the standard deviation difference for the gene between the +1 labelled and the −1 labelled samples. Higher score implies the gene is more functionally important. Finally, we need to mention the commonly used SVD method which finds singular values for the purpose of feature selection. Using singular value's entropy based approach, the UFF method has been proposed to filter out biological data [100]. UFF defines the entropy contribution of every single feature and uses features with positive entropy contribution to be the best discriminate genes.

5.3 The Feature Reduction Framework

The feature reduction framework consists of two stages which are described in this section. The first stage produces the reduced set of features and the second stage validates the outcome by building a set of classifiers.

5.3.1 The Feature Reduction Stage

In this section, we concentrate on the three methods that produce reduced feature-set by applying global analysis on the whole feature set leading to more informative result. Frequent pattern mining based method As described in Chapter 2, frequent pattern mining is a powerful technique capable of identifying in a set of objects (called items) those which demonstrate similar behaviour. For instance, in a supermarket, customers purchase patterns are kept as transactions each includes a set of items purchased together. Analysing the set of transactions may lead to items that are frequently purchased together. This information may be valuable in planning for promotions on items and for the layout of items on the shelves. The same methodology applies to genes expression data by considering genes as items and samples as transactions to determine sets of genes that show similar behaviour.

To determine closed frequent itemsets we proceed as follows. First we discretise the gene expression data into binary values based on their expression level. This will lead to more reasonable knowledge discovery. Second, we determine frequent itemsets by applying the Apriori algorithm. Any other algorithm may be applied and will produce the same result. Third, we add to the set of closed frequent itemsets the largest frequent itemsets which are frequent itemsets that are not subsumed by other frequent itemsets. Forth, we concentrate on frequent itemsets not added to the closed group yet. From the latter collection, we add to the closed group any itemset that satisfies the following three criteria. It is large compared to the frequent itemsets not yet added to the closed group; it is not subsumed by other frequent itemsets not yet added to the closed group; and it has larger frequency than all the closed itemsets that subsume it.

After all closed frequent itemsets are identified, we proceed to identify representatives from the itemsets. We sort items in descending order based on the number of closed itemsets that contain them. We choose to include in the reduced set of features items that satisfy the following criteria. The item should rank above average in the list and the number of closed itemsets where the item co-exists with already selected items is less than the average. At the end of this process, we get the set of features (items) that form the reduced set. Of course only features favoured by the other techniques will be used in building the classifier models described next in Section 5.3.2.

Clustering based method

From the variety of clustering algorithms described in the literature, we decided to use a genetic algorithm (GA) based clustering approach which has been developed by our research group [75, 79], As with the web site analysis, we applied the same genetic algorithm which we had found successful in that domain. Also the algorithm produces the nature clustering without expecting more input from the user as it is the case with the other algorithms like k-means. While k-means requires the number of clusters as input, the genetic algorithm based approach employed in this study produces the appropriate number of clusters in an automated process. We modified the approach to better fit our needs. The GA based process involves a number of steps including initialisation of the individuals (interchangeably called chromosomes), crossover and mutation to produce new individuals, and selection based on fitness which is achieved as a combination of the three objectives enumerated above. We apply a variation of arithmetic cross-over in order to produce the number of clusters as a by-product of the process without requiring it as an input parameter. This leads to more natural clustering with fewer input parameters.

The length of every individual is equal to the number of genes in the given dataset. The initialisation process starts by distributing the given genes into clusters by considering their expression levels in the given samples to produce 50 initial individuals. The expression levels within each sample are scaled up to guarantee that the smallest is greater than zero; then we produce alleles of an individual by taking the scaled expression levels modulus *m*, where *m* is the number of closed frequent itemsets which is considered as the upper limit for the number of clusters. This number of clusters will be tuned up by the iterations of the GA to end up reporting the appropriate number of clusters.

As already mentioned in Chapter 4, the arithmetic cross-over operator employed in this study takes two of the existing individuals and processes them to produce a new individual. Corresponding alleles within the two input individuals *p* and *q* are used to produced alleles of the new individual *r* as follows: $ri = ((pi \times qi)\ modm) + 1$ (where *i* ranges from 1 to the number of genes in the input dataset) to produce integer values in the range [1,*m*]. If some values from the range [1,*m*] do not appear in the new individual then values of alleles within the new individual are mapped to guarantee consecutive cluster numbers starting with 1. Mutation is applied to guarantee faster convergence. After the cross-over operation is completed, the number of individuals increases to 75. Then the fitness of each individual is measured as the sum of the average homogeneity of its clusters, average separateness between the clusters and average size of the clusters. Homogeneity and separateness are measured using Euclidean distance between the genes within the same cluster and between the centroids of each two clusters, respectively. Finally, the 75 individuals are ranked based on their fitness in descending order and the best 50 individuals are kept for the next iteration. The process continues up to 500 iterations and the best individuals are considered as the final clustering solution. However, to validate the latter selection process, we apply cluster validity indexes on the top 10 individuals from the final solution produced by the GA process. The outcome from the validity analysis will confirm the appropriateness of the selected solution, which is the solution favoured by the majority of the validity indexes [75].

The clustering solution returned as the most appropriate for the given data is further analysed to decide on the number of features that should represent each cluster. For compact and homogeneous clusters, the feature closest to the centroid is selected as representative. As a result, there will be *n* clusters leading to *n* features as the reduced set of features produced by the clustering approach.

Network analysis based method

The relationship between samples and genes is a natural outcome from the microarray analysis. The expression levels are considered as a two dimensional matrix that represents a bipartite graph which corresponds to a two mode network between samples and genes. Researchers have used bipartite graphs to tackle other problems in bioinformatics, e.g., [13, 19]. In other words, for the gene expression data utilised in this chapter we produce a two mode network where in the corresponding adjacency matrix a row corresponds to a feature or gene and a column corresponds to a sample. An entry (*i, j*) reflects the expression level of gene *i* in sample *j*. Folding is applied on the two mode network to produce a one mode network of the genes. As described in Chapter 2, the folding process works as follows. We get the transpose say *T* of the original adjacency matrix say *A* which has one row corresponding to each sample and one column corresponding to each gene. We multiply the two matrixes *T* ×*A* to get a matrix *Y* where rows and columns are genes and hence it is the adjacency matrix of the one mode network between genes. Each entry (*i, j*) in matrix *Y* reflects the correlation between the two genes g_i and g_j across all the samples.

The one-mode network of genes is preprocessed by discretising the weights of all edges such that zero weight (edge is eliminated) is assigned to each edge which has weight lower than the average weight of the edges in the graph; weight one is assigned to all other edges. The discretised graph is processed further to find communities of genes. A community of genes includes a set of genes which are more connected to each other than to other genes outside the community. We obtain communities by eliminating edges (links) from the graph by considering the betweenness centrality. The higher the betweenness centrality the more the edge becomes a candidate to be removed. Edges are removed from a network based on two criteria, they should have high betweenness centrality and their removal should lead to more communities in the network. Our strategy for the network of genes is to satisfy either of the following two constraints, the one that could be achieved first: (1) to have the number of communities equal to the number of clusters produced by the method presented in Section 5.3.1; (2) not to remove any edge whose betweenness centrality is below the average betweenness centrality of all the edges in the network. This will lead to reasonable number of communities where each community is somehow homogeneous. We select from each community a representative which has the highest average of two centrality measures, namely degree centrality and eigen-vector centrality. Interestingly, the conducted experiments reported close to 90% overlap (at least in functionality) between the reduced features produced by the closed frequent sets of genes, clustering and the network based methods.

5.3.2 Classification Stage

Clustering could be used as a preprocessing step for classification. First instances are clustering used a clustering approach capable of finding the number of clusters like the one described in Section 5.3.1. Then representative instances from the various clusters are used to build a classifier model. Other instances from the clusters could be used to test the model. The classifier will report better accuracy if the training set is a good representative of the classes hence it should be selected by considering the distribution of the objects in the clusters; the same should be valid for selecting the test set. The predictor is subsequently used to classify unknown objects and hence determine the accuracy of the classifier.

The accuracy is determined as the percentage of the correctly classified instances from the test set. In other words, classification [58] centres around exploring through data objects (training set) to find a set of rules, a formula or a function, etc. which determine the class of each object based on the values of its attributes. The discovered rules are later used to build a classifier to predict the class or missing attribute value of unseen objects whose class might not be known. Support vector machine (SVM) [83], associative classification, Bayesian network, *k* nearest neighbours, decision tree, and neural network are well established techniques for classification. In this study, we use the first three classifiers.

Support Vector Machine Classifier

Support vector machine (SVM) is a powerful classifier with a strong mathematical foundation. It represents a particular instance of a large class of learning algorithms known as kernel machines. In general, a two class SVM projects data into higher dimensional space where the two classes are linearly separable. For a given set of instances represented as data points in the space, SVM tries to find a hyper-plane that separates the two classes of the data, and maximises the width of a separating band between the data points and the hyper-plane. The model is defined by considering support vectors which are formed using points nearest to the margin.

The objective of SVM is to select the optimal hyper-plane which can separate the two classes because there may exist many hyper-planes that can separate the two sets of points. The optimal hyper-plane is defined as the one which can separate the classes with the largest margin. A hyper-plane equation can be determined based on two parameters: *w* and *b*, where *w* is a weight vector perpendicular to the hyper-plane, and *b* is a bias that moves the hyper-plane parallel to itself. The equation of a hyper-plane can be written as:

$$\vec{w}^T . \vec{x}_i + b = 0 \tag{5.1}$$

Recall that a classification task usually involves training and test sets which consist of data instances, and each instance in the training set has several features and a target value called the class label. The goal of SVM is to produce a model which can predict the target value of data instances in the test set that have values for the features used in building the model. Given a training set of instance-label pairs $(x_i, y_i), i = 1, \ldots, l$, where $X_i \in R_n$ and $y \in \{1, -1\}^l$, SVM requires solving the following optimisation problem:

$$min_{w,b,\varepsilon} \frac{1}{2} ||\mathrm{w}||^2 + C \sum_{i=1}^{l} \varepsilon_i \tag{5.2}$$

subject to

$$y_i(w \bullet x_i + b) \geq 1 - \varepsilon_i, \varepsilon_i > 0, i = 1, \ldots, l \tag{5.3}$$

where x_i is an input vector, *l* is the number of instances in the training set, *C* is a cost factor for misclassification, and *ε* is a slack variable for misclassification points.

There are various fundamental advantages of SVM compared to other methods. First, SVM produces a unique solution because it is basically a linear method and does not have such a pitfall as multiple local minima. Second, SVM is inherently able to deal with very large amounts of dissimilar information. Third, the discriminant function is characterised by only a comparatively small subset of the entire training data set, thus making the computations noticeable faster.

Recent comparative study among all existing classification methods have shown the outperforming nature of SVM [83]. Therefore, SVM is an attractive classifier model. It is even an ideal model for the problem tackled in this chapter because we have two classes infected and normal samples. Also, the reduced set of genes lead to the construction of a powerful SVM model.

Bayesian Network Classifier

A Bayesian network [77] is a graphical probabilistic model that consists of a directed acyclic graph and a set of conditional probability tables. Nodes in the network represent features or variables and links encode conditional independence between the variables. The probability distribution is unconditional for a node without any parents. If a node has one or more parents, the probability distribution is a conditional distribution, where the probability of each node value depends on the values of the parents. This requires the probability distribution for each node be defined by a conditional probability table and by considering its parent nodes.

The learning process in a Bayesian network consists of two stages. First the network structure is built (structure learning) and then probability distribution estimations are calculated in the form of conditional probability tables. Structure learning often has high computational complexity as the number of possible structures is huge. To solve the computational complexity, heuristic and approximate learning algorithms have been proposed [37, 50]. There are many combinations of structure learning and search technique that can be used to create Bayesian networks. Consequently, the application of feature reduction before building the classifier is an ideal step to produce a manageable space for deriving conditional probability tables. In other words, constructing a Bayesian network model by considering all the given genes is not feasible. However, the reduced set of genes form an excellent input for the Bayesian network construction process.

Associative Classifier

Associative classification is a simple yet powerful classifier model which derives a set of association rules where the antecedent of each rule is a subset of the given features and the consequent of the rule is the class variable. To derive the rules constituting the associative classifier we utilised the approach described in Section 5.3.1 to find a special type of frequent itemsets; each special frequent itemset contains genes and a class variable.

The input to the process consists of the given samples as transactions. However, for each transaction we consider as items only the reduced set of genes and a class variable which indicates whether the sample is normal or infected. The process consists of two stages. In the first stage, we apply the frequent pattern mining process to produce only frequent itemsets of size at least two and which contain the class variable. The identified frequent itemsets are processed further to keep only the concise sets by eliminating any frequent set which has at least one of its subsets frequent. The second stage is applied on the concise frequent sets to produce one rule from each set. It is the rule with only the class variable as consequent. Only rules which have high confidence and which are interesting are maintained as classification rules to constitute altogether the associative classifier.

5.4 Experiments

We conducted a set of experiments to demonstrate the applicability and effectiveness of the proposed framework. For this purpose, we have used different existing software tools for realizing the feature reduction models and the classification tasks. We used four cancer datasets in the testing in order to illustrate further the power of the proposed framework. The environment, the datasets and the results are described next in this section.

5.4.1 Testing Environment

We used a number of existing software packages because our target is to demonstrate the applicability and effectiveness of the proposed framework. Pajek was used to realise the network based feature selection method, our implementation of the GA-based clustering method [75, 79] and WEKA were used for the other feature selection methods. The reduced set of features is extracted from the ranked list. Explicitly, we decided on the final reduced set of features after analysing the effect of using different sizes of the reduced feature set (starting with the top ranked feature as set by itself) and by considering the rank of each feature. Finally, our GA-based clustering algorithm was used to find the clustering outcome. We did not consider the best solution, we rather concentrated on the top ten alternative solutions produced by the GA. We then applied five clustering validity indexes, namely Dunn, Davies-Bouldin, Silhouette, Jaccard and Rand. Majority voting is used to find the best clustering solution. For the cases considered in the testing, the solution which ranked on the top was confirmed as the best solution.

To classify the reduced feature sets two tools were used, namely MATLAB and WEKA. MATLAB was used to run the SVM classifier and for the naive Bayesian classifier. On the other hand, WEKA was used to realise the associative classifier. For each of the classification algorithms the reduced feature set was used to train the classifier. Then the accuracy was determined based on the result from classifying the samples in the test set.

5.4.2 Datasets

Four data sets have been used in the experiments conducted in this study. The essential information related to the datasets is summarised next.

Leukaemia (AML/ALL) [46]: The ALL/AML data set was produced from Affymetrix microarray with 6,817 genes. The data has 73 ALL/AML samples, 38 (27AML/11 ALL) samples for training and 35 (23AML/ 12 ALL) samples for testing.

Colon caner [7]: The Colon data set contains 62 samples collected from cancer patients. Among them 40 tumour biopsies and 22 normal biopsies are from healthy parts of the colons of the same patients. Two thousands out of around 6,500 genes were selected based on the confidence in the measured expression levels. The Colon data set was downloaded from the University of Texas, Human Genetics Centre http://www.sph.uth.tmc.edu/hgc/default.aspx?id=2775. Samples were split as: 15 normal samples were used for training and 7 normal samples for testing; in addition, 23 tumour samples were used for training and the other 17 tumour samples were used for testing.

Prostate cancer [92]: The expression profile of 12,600 genes were derived from 136 samples: 102 samples for training (52 prostate & 50 normal) versus 34 samples for testing (25 prostate & 9 normal).

Lung caner [44]: 918 genes are profiled with total 73 samples. The adenocarcinoma is classified to 7 classes: AC-group-1 (21 samples), AC-group-2 (7 samples), AC-group-3 (13 samples), Normal (6 samples), Squamous (16 samples), Small-cell (5 samples) and Large-cell (5 samples).

Table 5.1: Top 10 genes (features) extracted from the four cancer datasets using the three feature reduction methods: these are the top agreed upon genes by majority voting

Leukemia	Colon	Prostate	Lung
CST3	*CSRP1*	*PHF16*	*EGFR*
MYB	*IL8*	*VGLL4*	*GSTM1*
CD33	*CKS2*	*HNRNPM*	*MYCN*
CFD	*DES*	*LAMP2*	*MAP3K8*
LEPR	*HNRNPA1*	*GRSF1*	*KRAS*
PPBP	*DARS*	*CYTSA*	*ERCC6*
CEBPD	*CLNS1A*	*CALM2*	*XRCC1*
ELA2	*FBL*	*CCR2*	*PIK3CA*
CSTA	*CXCL2*	*ZNF148*	*CYP2A6*
SPTAN1	*GUCA2B*	*HTATIP2*	*MET*

5.4.3 Gene Selection

We first applied the feature reduction methods described in Sections 5.3.1, 5.3.1, and 5.3.1. Our goal is to derive a gene set (i.e., features or biomarkers) that can be used as input to the classifiers.

For the frequently pattern mining method, we set the threshold arbitrarily to 75%; in other words, genes that occur in less than 25% of the samples are discarded in order to maintain high confidence in the outcome. For clustering, we rank genes based on their co-expression strength (“distances”) to centroids. For the network analysis feature selection, we select hubs with degree greater than 5 to represent most influential genes in the network. We then applied majority voting by considering the function of the genes because various genes reported by the employed methods may have the same function. The latter genes are considered equivalent in the voting scheme. As a result, we extract from each dataset the most influential genes as disease biomarkers. Listed in Table 5.1 are the top ten biomarkers in descending order of their functional importance in the network. These are the genes which were the outcome of the majority voting scheme from the employed feature reduction methods.

The interactions between genes in the lung cancer data set are visualised in Figure 5.1 with size of nodes corresponding to the expression levels of genes. It is interesting to see the extracted genes from feature reduction methods densely interact with each other. This suggests that the feature selection step is effective in the network analysis.

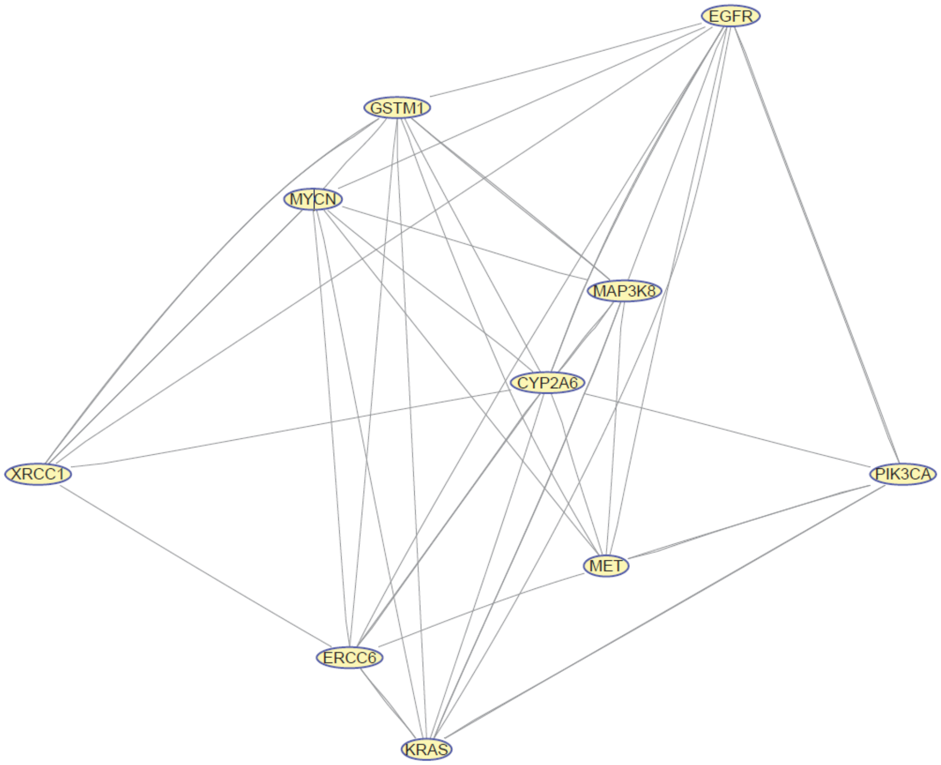


Figure 5.1: Interaction between genes selected for the lung cancer data set [44]

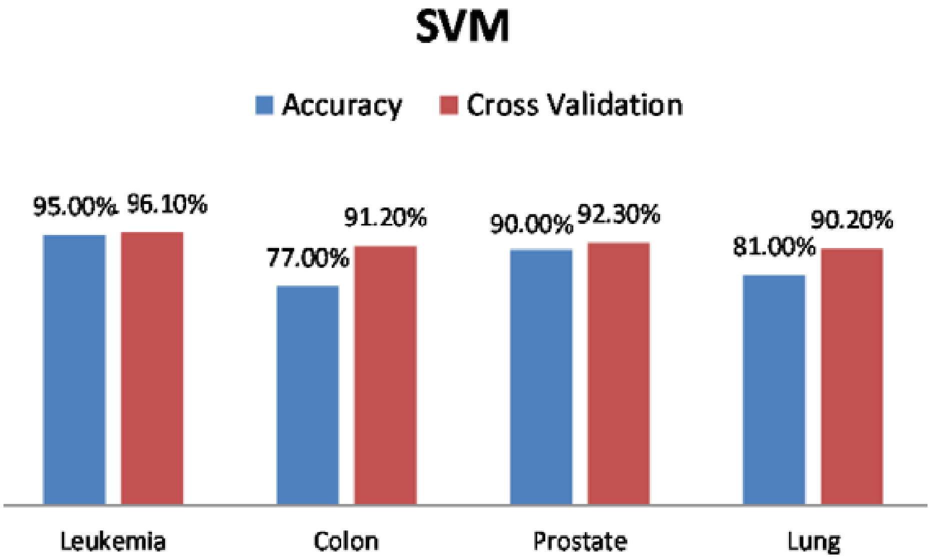


Figure 5.2: SVM Classification Statistics

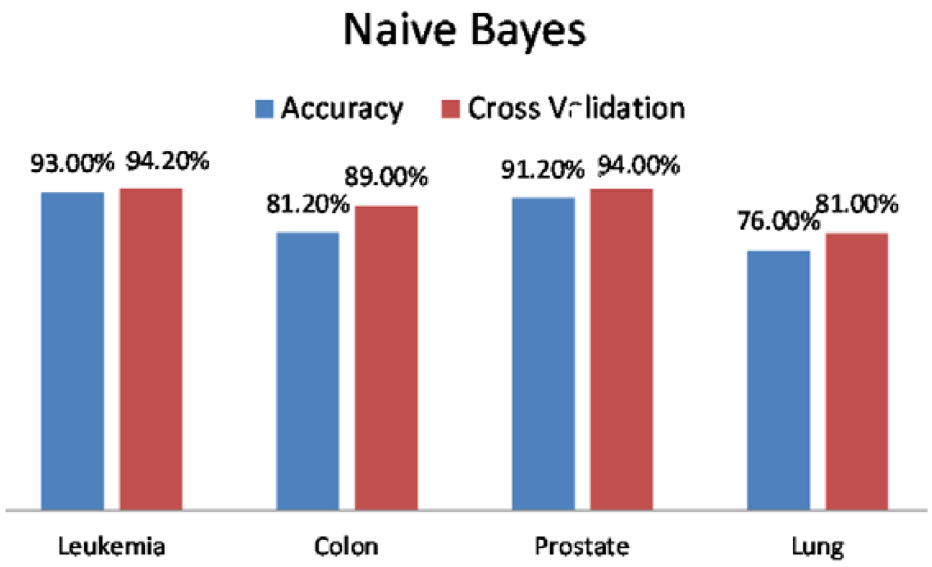


Figure 5.3: Naive Bayes Classification Statistics

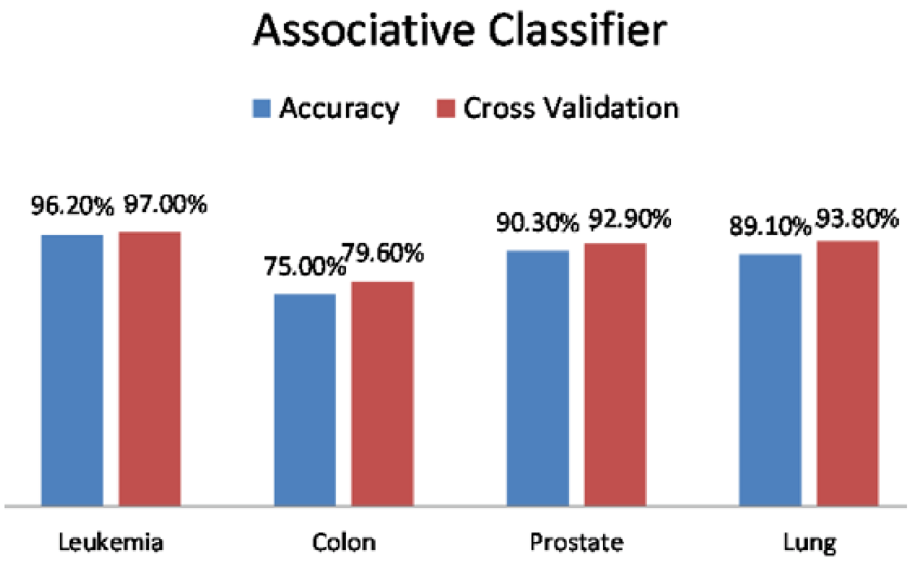


Figure 5.4: Associative Classifier Classification Statistics

5.4.4 Classification of Samples

As in most gene expression analysis cases, the number of samples is very small ($\ll$) compared to the number of genes in the original data sets. However, with the efficacy of the features listed in Table 5.1, the ill-defined problem is resolved and we are able to input the compact gene set to the classifiers. For each classifier, we use 10-fold cross validation, and we take the average over the results from individual runs. The results are reported in Figures 5.2-5.4 for the three classifiers, namely SVM, naive Bayesian network and associative classifiers, respectively.

Figures 5.2-5.4 demonstrate that the accuracy generally varies for each cancer data set. This suggests that the extracted gene sets have different sensitivity to different classifiers, and further indicate the need to use the ensemble approach so as to provide comprehensive analytical results. An ensemble approach involves employing a number of classifiers to analyse a give data and then taking a (may be weighted) vote of their predictions. The results also show that cancer as a genetic disease at the cellular level is different from type to type. For example, the accuracy of the colon gene set is lower than others. Therefore, deriving more efficient gene sets remains a challenging problem, owing not only to data mining techniques but rather the relatively small number of samples available for the analysis.

Chapter 6

Summary, Conclusions and Future Work

People are changing their life style, they are moving from traditional to a virtual life where the availability and accessibility of almost everything is possible by anyone anywhere with an Internet connection and at any time. People do not anymore need to travel for achieving certain targets. While sitting on a desk at home with an Internet connection, it is possible to surf the web to shop online, to socialise, to read the news, to enrol in online courses, to do online banking, etc. This uncontrollable diversity in the user group makes it unfortunately hard if at all possible to predict the skills and qualifications of end-users who surf the web. They could range from very naive to very professional. Satisfying naive users needs extra care compared to satisfying professionals; this is a major requirement to achieve the main target of web site owners, i.e., to increase the number of visitors. In fact there is a small group of professionals and everyone else falls in the range from professionals down to unskilled users. The latter group forms the majority of visitors to web sites whether for curiosity, fun, or to complete something serious, like online shopping.

This dissertation contributed a novel approach for web mining for identifying important pages and recommending web site restructuring. This has been achieved by integrating various data mining and social network analysis techniques. The developed framework is general enough to serve a wide range of applications as illustrated in Chapter 6 where the framework has been utilised to identify disease biomarkers by analysing gene expression data. To support this claim further, currently we are trying apply the framework in the business domain to analyse stock market data.

6.1 Dissertation Overview

Web site owners are capable of keeping track of visitors to the various web pages constituting the web site. This is affordable by the server which keeps a log file where it records traces including IP of the computer from which the web site was accessed, the time spent on each page, etc. Some popular pages may be buried deep in the web site such that a visitor must pass by several other pages before landing at the target page. This process may be frustrating to the visitors who may change their minds and decide not to continue the surfing. As a result, the web site owner will lose visitors who may shift to web sites of competitors. This situation is unavoidable because it is impossible to predict ahead of time the best possible linking of pages which satisfies a larger group of visitors. However, by depending on the web log kept over time, behaviour of visitors could be analysed and the result could be used to guide future maintenance of the web site.

In this thesis, we presented an integrated framework that combines the power of social network methodology and web mining techniques in order to produce a comprehensive and robust approach capable of effectively optimising web sites for better navigation. We did not restrict the process to only web structure mining. We rather integrated the three web mining techniques, namely web structure, usage and content mining. We also utilised the web log in order to construct the social network of the pages constituting the web site under investigation. For constructing the social network, we demonstrated the power of association rules mining by considering user sessions as transactions and the pages themselves as the items. The result is another perspective to view the web structure. We were pleased to find that the results obtained from the social network based methodology are consistent with the results reported by the web structure based techniques, namely PageRank and HITS. We also used *TFIDF* in order to analyse the correlation between the term in each hyperlink and the pages it points to directly or indirectly. We used the "Music Machines" dataset to demonstrate the effectiveness of combining these values for measuring a web site's usability. The outcome from the testing showed that the proposed approach is simple but viable to solve the given problem. As a result, we would argue that the proposed approach could be integrated into a more global framework for managing and optimising the usability of web sites.

6.2 Discussion and Conclusion

Daily life is shifting from a traditional style that mainly depends on face to face communication for completing various activities into a virtual style which totally depends on the web as the main medium for communication. In fact web based applications impacted all aspects of our life and hence the number of users in increasing rapidly. It is very common to have people moving around with hand-held devices which have the capability to browse web sites. This new trend in technology made web sites available and accessible anywhere and at any time. However, it is not possible to have every web site satisfy every visitor because web sites are mainly designed and deployed based on general requirement analysis. Fortunately, the servers hosting web sites are keeping track of all accesses by users without any need to notify the users. The maintained log forms a valuable source to guide web site owners how the web site could be adjusted to satisfy users better. The framework described in this thesis heavily depends on web log data to investigate the usage patterns leading to recommendation of the most appropriate way to link pages of a web site. Finding pages that are frequently accessed together and pages which are clustered together based on access patterns is necessary to link these pages in order to allow for faster access to the related information accessed together by certain group of users. This will satisfy the latter users and will mostly increase the number of visitors to the web site especially by users who share the same expectations with a given user group. Further, we are able to suggest restructuring the web site into a better accessible site with better reachable pages by recommending adding direct links between pages most frequently accessed by the visitors.

The techniques used in this research show that there is a great amount of information about groups of online customers based on which a web site can be customised in order to improve business policies; the outcome enriches and maximises the benefit of users visits to the web site. The typical web site owner will often be provided with aggregate statistics which are either too confusing or are of no use to them. The information is of no use because there have been no relevant conclusions determined from the data. In this research, it has been demonstrated that with the right data analysis, conclusions can be achieved which provide a more valid and easy to understand solution in order to develop strategies for good promotional policy and better web site usage. One example of this type of conclusion comes from calculating the backtrack pages from the dataset to determine pages which may contain little relevant information or of no use to most users; this information would alert the vendors about the need to review these pages in order to ensure that they are valid and relevant for users of the web site. The probabilities and sequential association rules also provide some useful information with regards to site navigation and setup of promotions.

The information on how a user has viewed their pages up until the current point of time can be analysed to determine what their next step is likely to be. Without any doubt, this is useful information to have – knowing what the user will go to next can be used to help them get there, or even divert them away. Beyond guiding a user's interaction, we can also use the knowledge for product bundling and promotional offers. For instance, if we are viewing a certain product and we know within a certain percentage of the user viewing another product, we can offer to sell the user both products in a bundle, with some incentive attached. This information is extremely valuable for the development and successful deployment of promotional offers and discounts. This type of grouping and association helps to increase the value of the sale and allows customers to find the products they are looking for with more ease.

While association rules capture relationships among different web pages, based on the same navigational patterns of users, we can extract the social network of pages and thereafter the sub-communities/clusters of pages. For a large web site, analysing these clusters is not only a daunting task, but often will not provide the higher level of knowledge that could be useful. Using the sum valued core technique or the island technique, we can identify most useful and notable groups/clusters of pages which is relatively easily manageable. This approach is very useful in classifying users according to their patterns of usage of the web site. The resulting clusters of web pages can then be used in order to cluster a transaction by looking at the cluster that a transaction belongs to. This information should then be used to specifically target advertisements to a user or to provide them with a custom interface dynamically. Clearly, the aforementioned resources allow a great deal of information to be found with regards to user behaviour. By adding the data calculated, some creative marketing and product placement to the design of the web site, business owners can lead to large scale economic benefits. The ability to better control the movements on the web site can be likened to arrows guiding a customer in a store to products with higher sale margins. Until recently, such analysis has been seen to be too complex for the average online shopping owner, but now new solutions to the problem are being actively created. As realised from the outcome of the work done in this thesis, going the extra step to extract this information from web site logs is without any doubt an extremely beneficial application of log mining.

The measures proposed in Chapter 3 can guide web site owners in *what to promote*. If for example the outcome from the sequential association rules mining process leads to the prediction that the user is likely to visit product *p* and product *q* in his/her next few visits and these two products fall in the same user groups of pages then the organisation may choose to promote these products dynamically to the user. This dynamic advertising policy actually answers the question *when to promote* a product as opposed to static advertising where web site owners have little or no clue on what could be interesting to a user and when the web site owners should change the promotion. On the other hand, in order to answer the *where to promote* question we have proposed the centrality, PageRank, hubs and authority measures. As too much promotion can negatively impact the visitors, we propose to place the promotional notices in only a selected few pages. The above measures can guide the organisation in selecting those few important pages. These grouping and users' visit information may also guide the organisation to restructure their site topology in order to make the user visit experience smoother and hence the user will be satisfied. Such a strategy will help in maintaining existing visitors and would probably attract new visitors who may belong to the social networks of the current visitors and hence may get affected by the opinions of the current visitors.

Feature reduction is an essential research area that has attracted considerable attention in the research community. It has a large number of applications from various domains, including gene expression data analysis as demonstrated by our study. The target is to identify small set of discriminative genes as biomarkers for specific diseases. Human cells are the home for genes which encode the genetic code that produces proteins which accomplish the major tasks within the body. Genes are expected to follow a certain trend within each cell. Failing to do so leads to malfunctioning genes which are the main cause for diseases. Identifying such malicious genes is not feasible without employing some computational methods; otherwise it is like seeking a needle in a haystack. In other words, pure wet-lab based analysis is hard to achieve the target if at all feasible due to the high cost and tremendous effort required to undertake the experiments. Fortunately computational techniques have provided attractive methods to help in the analysis. This brings up the beauty of computation whether discrete or applied, including statistical, machine learning and data mining techniques. However, applied computational techniques do suffer and do not produce satisfactory results when the dataset to be analysed is characterized by high dimensionality. Here comes the role of dimensionality reduction approaches which are capable of finding an informative reduced set of features that could be used to discriminate between infected and uninfected samples.

6.3 Future Work

Although the proposed approach is powerful enough to handle web log and web structure mining, there is still much work that can be done in the following aspects–

- There is a need to investigate how it would be possible to minimise the user involvement in the process by trying to automate the discovery of the parameters and possibly the final analysis of the results. An alternative immediate solution could be realised by presenting the results to domain experts in fuzzy terms as they are more understandable by humans. All these issues need to be investigated for their applicability and feasibility.
- To move this project forward, it is possible to work on enriching the developed framework with a variety of new functionalities. Further research may turn the web usage mining process into an incremental approach that periodically realises changes in access and reflects them onto the recommendation system. It is also possible to zoom into certain periods of the year, month, week, days, etc. This will provide the ability to say visitors who accessed these page(s) at noon on Friday also accessed the following page(s). It is important to track how the interests of certain people change over time and what drives the change.
- Content of the web site may be analysed to find out if the content of each page is coherent and whether it is possible to recommend partitioning and moving content around for better accessibility and readability. Avoiding dense pages is always more attractive. So, it is important to suggest splitting a web page into sub-pages where each sub-page contains coherent information which is easier to track and read.
- The work may be further expanded by incorporating web content mining to investigate how the content of various pages are related and suggest restructuring of a web site based on the degree of similarity between the content of the various pages constituting

the web site. Pages with higher degree of similar based on content should be linked directly and the distance between any two pages should be proportional to the degree of similarity between them. Once combined with web log access patterns, the recommendation system will lead to a more robust web site restructuring facility.

- The three feature reduction methods employed in this study have different characteristics and capabilities. The three methods (closed frequent pattern mining, GA-based clustering and network analysis) consider the features collectively and hence produce more robust results. The conducted testing demonstrates that the reported disease biomarkers are related in some sense and considering them in isolation could lead to information loss and hence could negatively affect the final outcome. Our study demonstrated further that the frequent pattern, clustering and network based approaches work well and produce a good discriminative set of features. Not having common genes between the set of genes reported as discriminative features for classification is another interesting phenomenon to comment on. Closer analysis of the genes revealed the fact that (when different) the genes reported by the employed methods have close functionalities.

Bibliography

[1] S. Abiteboul, M. Preda, and G. Cobena. “Adaptive on-line page importance computation”. *Proceedings of the International Conference on World Wide Web*, pp.280-290, 2003.

[2] M. Adnan and R. Alhajj, “DRFP-Tree: Disk-Resident Frequent Pattern Tree,” Applied *Intelligence,* 30(2):84-97, 2009.

[3] M. Adnan, R. Alhajj, J. G. Rokne. “Identifying Social Communities by Frequent Pattern Mining”. *Proceedings of the International Conference on Information Visualisation*, pp. 413-418, 2009.

[4] [4] M. Adnan and R. Alhajj, “A Bounded and Adaptive Memory-Based Approach to Mine requent Patterns from Very Large Databases”. *IEEE Transactions on Systesms, Man, and Cybernetics, Part B: Cybernetics*, 41(1):154-172, 2011.

[5] B. B. Agarwal, D. M. Khan and S. Dhall, “Web Mining: Information and Pattern Discovery on theWorldWideWeb”. *International Journal of Science, Technology and Management*, 2010.

[6] R. Agarwal, T. Imielinski and A. Swami, “Mining Association Rules between Sets of Items in Large Databases”. *Proceedings of ACM SIGMOD international conference on Management of data*, 1993.

[7] U. Alon, et al., “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *PNAS,* 96:6745-6750, 1999.

[8] A. Altman and M. Tennenholtz. “Ranking systems: the pagerank axioms”. *Proceedings of ACM Conference International on Electronic commerce*, pp.1-8, 2005.

[9] A. Amir, R. Feldman, and R. Kashi. "A New and Versatile Method for Association Generation". *Information Systems*, 22(6-7):333–347, 1997.

[10] L. Ardissono and P. Torasso, "Dynamic User Modeling in a Web Store Shell", *In Proceedings of the European Conference on Artificial Intelligence,* Berlin, Germany, pp.621- 625, 2000.

[11] R. Armstrong, D. Freitag, T. Joachims and T. Mitchell, "WebWatcher: A Learning Apprentice for the World Wide Web". *Spring Symposium on Information Gathering from Distributed, Heterogeneous Environments*. Stanford, CA, USA, 1995.

[12] K. Avrachenkov, N. Litvak and K. S. Pham, "A singular perturbation approach for choosing PageRank damping factor", *arXiv:math/0612079v1 [math.PR].*

[13] J. Ballesteros and K.G. Palczewski, "protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin," *Current Opinion in Drug Discovery and Development,* 4(5):561-74, 2001.

[14] V. Batagelj, A. Mrvar, *Pajek – Program for Large Network Analysis.* http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (Last accessed on 30/10/2014))

[15] V. Batagelj. "Analysis of large networks - Islands", *Presented at Dagstuhl seminar 03361: Algorithmic Aspects of Large and Complex Networks Dagstuhl*, 2003.

[16] M. Bianchini, M. Gori, and F. Scarselli. "Inside pagerank". *ACM Transactions on Internet Technology*, 5(1):92–128, 2005.

[17] S. Bicciato, M Pandin., G. Didon and C. Di Bello, "Pattern identification and classification in gene expression data using an autoassociative neural network model." *Biotechnology and Bioengineering,* 81:594-606, 2002.

[18] R. Bijlani, et al., “Prediction of biologically significant components from microarray data: independently consistent expression discriminator(iced),” *Bioinformatics,* 19:62-70, 2003.

[19] K. Bleakley and Y. Yamanishi,“Supervised prediction of drug-target interactions using bipartite local models,” *Bioinformatics,* 25(18):2397, 2009.

[20] P. Boldi, M. Santini, and S. Vigna. “Pagerank as a function of the damping factor”. *Proceedings of the International Conference on World Wide Web*, pp.557-566, 2005.

[21] C. Borgelt. “Efficient implementations of apriori and eclat”, *Proceedings of theWorkshop of Frequent Item Set Mining Implementations,* Melbourne, FL, 2003.

[22] J. Borges and M. Levene, “Data mining of user navigation patterns”, *Proceedings of Workshop on Web Usage Analysis and User Profiling (WEBKDD), in conjunction with ACMSIGKDD International Conference on Knowledge Discovery and Data Mining,* San Diego, CA., pp.31-36, 1999.

[23] P. Berkhin, “A Survey of Clustering Data Mining Techniques”, *Grouping Multidimensional Data,* pp 25-71, 2006.

[24] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. “Link analysis ranking: algorithms, theory, and experiments”. *ACM Transactions on Internet Technology*, 5(1):231– 297, 2005.

[25] J. T. Bradley, D. V. de Jager, W. J. Knottenbelt, and A. Trifunovic. “Hypergraph partitioning for faster parallel pagerank computation”. *Proceedings of Formal Techniques for Computer Systems and Business Processes, European Performance Engineering Workshop*, pp.155–171, 2005.

[26] S. Brin, R. Motwani, L. Page, and T. Winograd. “What can you do with a Web in your pocket”. *In Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 1998.

[27] S. Brin and L. Page. “ The anatomy of a large-scale hypertextual Web search engine”, *In Proceedings of the Seventh International World Wide web Conference*, 21(2): 37-47, 1998.

[28] L. D. Catledge and J. E.Pitkow. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. *Computer Networks*, 30(1-7): 107-117, 1998.

[29] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. “Automatic resource compilation by analyzing hyperlink structure and associated text”. *Proceedings of the International Conference on World Wide Web*, 1998.

[30] Z. Chen, A. W.-C. Fu, and F. C.-H. Tong. “Optimal Algorithms for Finding User Access Sessions from Very Large Web Logs”, *Journal of World Wide Web*, 7, pp. 65-74, 2004.

[31] Y.-Y. Chen, Q. Gan, and T. Suel. “I/o-efficient techniques for computing pagerank”. *Proceedings of ACM International Conference on Information and knowledge management*, pp.549–557, 2002.

[32] P.-A. Chirita, J. Diederich, and W. Nejdl. “Mailrank: using ranking for spam detection”. *Proceedings of ACM International Conference on Information and knowledge management*, pp.373–380, 2005.

[33] J. Cho, S. Roy, and R. E. Adams. “Page quality: in search of an unbiased Web ranking”. *Proceedings of ACM SIGMOD*, pp.551–562, 2005.

[34] F. Chu and L. Wang, “Cancer classification with microarray data using support vector machines.” *Bioinformatics*, 176:167-189, 2005.

[35] L. da F. Costa, F. A. Rodrigues, G. Travieso and P. R. Villas Boas. “Characterization of complex networks: a survey of measurements”, *Advanced Physics*, Vol. 56, pp 167–242, 2007.

[36] R. Cooley, B. Mobasher and J. Srivastavaa, “Data preparation for mining World Wide Web browsing patterns”, *Journal of Knowledge and Information Systems,* 1(1), pp.55-32, 1999.

[37] G. Cooper and E. Herskovits, “A Bayesian method for the induction of probabilistic networks from data”. *Machine Learning,* 9, 309-347, 1992.

[38] J. Dean and M. Henzinger. “Finding related pages in the world wide Web”. *Proceedings of the Eighth International Conference on World Wide Web*, pp.1467-1479, 1999.

[39] U. Fayyad, K. Irani, “Multi-interval discretization of continuous-valued attributes for classification learning,” *, Proc. of the International Joint Conference on Artificial Intelligence,* 1022-1029, 1993.

[40] L. C. Freeman, S. P. Borgatti, D. R. White. “Centrality in valued graphs: A measure of betweenness based on network flow” *Social Networks*, 13(2), pp. 141–154, 1991.

[41] Y. Fu, K. Sandhu and M.Y. Shih, “Clustering of WebUsers Based on Access Patterns”, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1999.

[42] T. S. Furey, et al., “Support vector machine classification and validation of cancer tissue samples using microarray expression data,” *Bioinformatics,* 16:906-914, 2000.

[43] S. Gao, et al., “A Closer Look at ‘Social’ Boundary Genes Reveals Knowledge to Gene Expression Profiles,” *Curr Protein Pept Sci,* 12:602-613, 2011.

[44] M. E. Garber, O. G, Troyanskaya et al., “Diversity of gene expression in adenocarcinoma of the lung,” *PNAS,* 98:13784-13789, 2001.

[45] Geetika. “Article: A Survey of Classification Methods and its Applications”. *International Journal of Computer Applications*, 53(17):14-16, September 2012.

[46] T. R. Golub, et al., “Molecular Classification of Cancer: class discovery and class prediction by gene expression monitoring,” *Science*, 286, 531-537, 1999.

[47] G. Guo, et al., “An kNN Model-Based Approach and Its Application in Text Categorization”. *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pp.559–570, 2004.

[48] J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”. *Proceedings of ACM SIGMOD international conference on Management of data*, pp.1-12, 2000.

[49] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. “Clustering based on association rule hypergraphs”, *Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery,* pp.9-13, Tucson, Arizona, 1997.

[50] D. Heckerman, D. Geiger and D. M. Chickering, “Learning Bayesian networks: The combination of knowledge and statistical data”. *Machine Learning,* 20, 197-243, 1995.

[51] J. Hou and Y. Zhang. “Effectively finding relevantWeb pages from linkage information”. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):940–951, 2003.

[52] W. H. Hsu, A. King, M. S. Paradesi, T. Pydimarri, and T. Weninger. “Collaborative and structural recommendation of friends usingWeb log-based social network analysis”. *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs (CAAW)*, volume SS-06-03, pages 55–60, Menlo Park, CA, 2006.123

[53] L. Huiqing, L. Jinyan and W. Limsoon, “A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns,” *Genome Informatics,* 13:51-60, 2002.

[54] R. Ivancsy and I. Vajk, “Frequent Pattern Mining in Web Log Data”, *Acta Polytechnica Hungarica, Journal of Applied Science at Budapest Tech Hungary, Special Issue on Computational Intelligence* , pp.77-90, 2006.

[55] J. Jeffrey, P. Karski, B. Lohrmann, K. Kianmehr and R. Alhajj, “Optimizing Web Structures Using Web Mining Techniques”, *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning,* Springer-Verlag LNCS, Brimingham, UK, pp.653-662 , 2007.

[56] X.-M. Jiang, G.-R. Xue, W.-G. Song, H.-J. Zeng, Z. Chen, and W.-Y. Ma. “Exploiting pagerank at different block level”. *Proceedings of the International Conference on Web Information Systems Engineering*, pp.241–252, 2004.

[57] A. Joshi, K. Joshi and R. Krishnapuram, “On MiningWeb Access Logs”. *Proceedings of ACM SIGMODWorkshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 63-69, 2000.

[58] M. Kantardzic, *Data Mining - Comcepts, Models, Methods, and Algorithms.* Picastaway: Wiley-Interscience, 2003.

[59] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An Efficient k-Means Clustering Algorithm: Analysis and Implementation”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881-892, 2002.

[60] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. “Multilevel hypergraph partitioning: Applications in VLSI domain”. *Proceedings of ACM/IEEE Design Automation Conference,* pp.526- 529, 1997.

[61] M.J. Keiser, et al., “Predicting new molecular targets for known drugs,” *Nature,* 462(7270):175-181, 2009.

[62] O. Kent and J. Mendell, “A small piece in the cancer puzzle:microRNAs as tumor suppressors and oncogenes,” *Oncogene,* 25:6188-6196, 2006.

[63] J. M. Kleinberg. “Authoritative sources in a hyperlinked environment”. *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pp.668–677, 1998.

[64] A. Kobsa, J. Koenemann and W. Pohl, “Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships”, *The Knowledge Engineering Review,* 16(2), pp.111-155, 2001.

[65] T. V. Le, C. A. Kulikowski and I. B. Muchnik. “Coring method for clustering a graph.” *Proceedings of the Internatonal Conference on Pattern Recognition*, pp.1–4, 2008.

[66] S. Lee, H. S. Yong. “Web Personalization: My Own Web Based on Open Content Platform”. *Web Information Systems Engineering WISE 2005*, 3806, pp.731-739 2005.

[67] C.H. Li and C.K. Chui. “Web structure mining for usability analysis”. *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pp.309-312, 2005.

[68] J. Li and L. Wong, “Identifying good diagnosis gene group from gene expression profile using the concept of emerging patterns,” *Bioinformatics*, 18:725-734, 2002.

[69] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations”. *Proceedings of the Fifth Symposium on Math, Statistics and Probability*, pp. 287-297. Berkeley, CA, United States, 1967.

[70] P. Massa and C. Hayes. “Page-rerank: Using trusted links to re-rank authority”. *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pp.614–617, 2005.

[71] E. Moler, M. Chow, and I. Mian, “Analysis of molecular profile data using generative and discriminative methods,” *Physiol genomics,* 4:109-126, 2000.

[72] A. Moore, "Lecture on Bayesian Networks".http://www.autonlab.org/ tutorials/bayesnet09.pdf. (Last accessed 2011.)

[73] S. Orlando, P. Palmerini, and R. Perego. "Enhancing the apriori algorithm for frequent set counting". *Proceedings of the International Conference on Data Warehousing and Knowledge Discovery*, pp.71–82, London, UK, 2001.

[74] S. Orlando, P. Palmerini, R. Perego, and F. Silvestri. "Adaptive and resource-aware mining of frequent sets". *Proceedings of IEEE International Conference on Data Mining*, page 338, Washington, DC, USA, 2002.

[75] Özyer T. and Alhajj R., "Achieving natural clustering by validating results of iterative evolutionary clustering approach," *Proceedings of IEEE International Conference on Intelligent Systems,* pp.488-493, 2006.

[76] J. S. Park, M.-S. Chen, and P. S. Yu. "Using a hash-based method with transaction trimming for mining association rules". *IEEE Transactions on Knowledge and Data Engineering*, 09(5):813–825, 1997.

[77] J. Pearl, "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning". *Proceedings of the Conference of the Cognitive Science Society,* Irvine, CA, pp.329-334, 1985.

[78] H. Peng, "Discovery of Interesting Association Rules Based on Web Usage Mining". *Proceedings of the International Conference on Multimedia Communications*, pp. 272-275, 2010.

[79] P. Peng, M. Nagi, et al., "From Alternative Clustering to Robust Clustering and Its Application to Gene Expression Data", *Proc. of IDEAL*, Springer-Verlag LNCS, Norwich, UK, Sep. 2011.

[80] M. Perkowitz and O. Etzioni, “Log Files for Music Machines at Hyperreal”, 1997. http://www.cs.washington.edu/research/adaptive/download.html (Last accessed 25/9/2013)

[81] M. Perkowitz and O. Etzioni, “Adaptive Web Sites: Automatically Synthesizing Web Pages”. *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, Madison, Wisconsin, United States, pp. 727-732, 1998.

[82] P. Pirolli, J. Pitkow and R. Rao. “Silk from a sow’s ear: extracting usable structures from the Web”, *CHI ’96: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp.118–125, 1996.

[83] M. Pirooznia, J. Yang, M. Yang and Y. Deng, “A comparative study of different machine learning methods on microarray gene expression data,” *BMC Genomics*, 2007.

[84] A. Qabaja, M. Alshalalfa, R. Alhajj and J. Rokne, “Multiagent Approach for Identifying Cancer Biomarkers,” *Proceedings of IEEE BIBM,* pp.228-233, Nov. 2009.

[85] J. Quackenbush, “Microarray data normalization and transformation.” *Nature Genetics Supplement,* December 2002: 496-501.

[86] R Development Core Team. R: A language and environment for statistical computing, reference index version 2.10.1. R Foundation for Statistical Computing, 2009.

[87] I. V. Ren´ata Iv´ancsy. “Frequent pattern mining in Web log data”,. *Journal of Applied Sciences at Budapest Tech*, 3(1):77–90, 2006.

[88] B. Schafer, J.A. Konstan and J. Riedl, “E-commerce Recommendation Applications”, *Data Mining and Knowledge Discovery,* 5(1-2),115-152, Kluwer Academic Publishers, 2001.

[89] M. W. Seeger, "Cross-Validation Optimization for Large Scale Structured Classification Kernel Methods" *Journal of Machine Learning Research*, 9, 1147-1178, 2008.

[90] S. R. Setlur, et al., "Integrative Microarray Analysis of Pathways Dysregulated in Metastatic Prostate Cancer," *Cancer Research,* 67:10296, 2007.

[91] C. Shahabi, A. M. Zarkesh, J. Abidi and V. Shah, "Knowledge discovery from user's Webpage navigation", *Proceedings of the IEEE International Workshop on Research Issues in Data Engineering (RIDE),* pp.20-29, 1997.

[92] D. Singh, P. G. Febbo, et al., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, 1, 203-209, 2002.

[93] R. Soumya , P. D. Sutphin, J. T. Chang and R. B. Altman. "Basic Microarray Analysis: Grouping and Feature Reduction." *Trends in Biotechnology,* pp.189-193, 2001.

[94] P. Soucy and G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model". *Proceedings of the International Joint Conference on Artificial Intelligence*, pp.1130–1135, 2005.

[95] D. Stekel, *Microarray Bioinformatics.* New York: Cambridge University Press, 2003.

[96] R. Steinberger, B. Pouliquen and J. Hagman, "Cross-Lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC". *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pp.415–424, 2002.

[97] D. Tanasa, "Web Usage Mining: Contributions to Intersites Logs Preprocessing and Sequential Pattern Extract with Low Support, *PhD thesis, University of Nice*, June 2005.

[98] U. of Washington Artificial Intelligence Research. Music machines Web site. http://www.cs.washington.edu/ai/adaptive-data/. (Last accessed 10/3/2012)

[99] K. Vekaria and C. Clack, “Selective Crossover in Genetic Algorithms: An Empirical Study”, A.E. Eiben et al. (Eds.): PPSN V, LNCS 1498, pp. 438-447, 1998

[100] R. Varshavsky, et al., “Novel unsupervised feature filtering of biological data,” *Bioinformatics*, 22:507-513, 2006.

[101] X.Wan, E. Milios, N. Kalyaniwalla, and J. Janssen, “Link-based event detection in email communication networks,” in *Proceedings of ACM symposium on Applied Computing*. New York, NY, USA: ACM, pp.1506-1510, 2009.

[102] E. Wall et al., “Singular value decomposition and principal component analysis,” *In P. Berrar et al. (eds.) A Practical Approach to Microarray Data Analysis,* Kluwer Academic Publishers, pp.91-109, 2003.

[103] J.Wang, et al., “Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data,” *BMC Bioinformatics*, 4:60-70, 2003.

[104] L. Wang, F. Chu, and W. Xie, “Accurate cancer classification using expressions of very few genes.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 4:40-53, 2007.

[105] Warner, D., Richter, J. N., Durbin, S. D., and Banerjee, B. “Mining user session data to facilitate user interaction with a customer service knowledge base in RightNow Web”. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2001.

[106] Z. Xia, L.-Y. Wu, X. Zhou and S.T.C. Wong, “Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces,” *BMC Systems Biology*, 4(Suppl 2):S6, 2010.

[107] W. Xing and A. A. Ghorbani. “Weighted pagerank algorithm”. *Proceedings of Annual Conference on Communication Networks and Services Research*, pp.305–314, 2004.

[108] X. Yan, J. Han, and R. Afshar. “CloSpan: Mining Closed Sequential Patterns in Large Datasets”, *Proceedings of the SIAM Int. Conf. Data Mining: SDM*, 2003.

[109] X. Yan, J. Han, and R. Afshar. “CloSpan: Mining Closed Sequential Patterns in Large Datasets”. *Proceedings of the SIAM Int. Conf. on Data Mining*, 2003.

[110] J. X. Yu, Y. Ou, C. Zhang, and S. Zhang. “Identifying interesting customers through Web log classification.” *IEEE Intelligent Systems*, 20(3):55–59, 2005.

[111] X. Zhang and H. Ke, “ALL/AML cancer classification by gene expression data using SVM and CSVM,” *Genomics informatics,* 11:237-239, 2000.

[112] Zhang, F., and Chang, H.-Y.. “Research and development in Web usage mining systemkey issues and proposed solutions: a survey”. *Proceedings of the International Conference on Machine learning and Cybernetics*, pp.986-990, 2002.

Appendix

List of Publications

Various parts of this thesis have been already published in reputable venues and approved by the research community. Here is a list of publications to which I have contributed since 2010.

1. A. Guerbas, O. Addam, O. Zarour, M. Nagi, A. Elhajj, M. Ridley and R. Alhajj, "Effective Web Log Mining and Online Navigational Pattern Prediction", *Knowledge-Based Systems*, Volume 49, September 2013, Pages 50-62.

2. M. Nagi, A. Elhajj, O. Addam, A. Qabaja, O. Zarour, T. Jarada, S. Gao, J. Jida, A. Murshed, I. Suleiman, T. Özyer, M. Ridley, R. Alhajj, Robust Framework for "Recommending Restructuring ofWeb sites by AnalyzingWeb Usage andWeb Structure Data", *Journal of Business Intelligence and Data Mining*, 7(1/2): 4-20, 2012.

3. M. Adnan, M. Nagi, K. Kianmehr, R. Tahboub, M. Ridley and J. Rokne, "Promoting where, when and what?: An analysis of Web logs by integrating data mining and social network techniques to guide eCommerce business promotions", *Social Networks Analysis and Mining*, 1(3):173-185, 2011.

4. Naji G., Nagi M., A. M. Elsheikh, Gao S., Kianmehr K., Özyer T., Demetrick D., Alhajj, R., Rokne, J., and Ridley, M., "Effectiveness of Social Networks for Studying Biological Agents and Identifying Cancer Biomarkers". In U. K. Wiil (Ed.), *Counterterrorism and Open Source Intelligence*. Springer, 2011.

5. M. Nagi, A. Guerbas, K. Kianmehr, P. Karampelas, M. Ridley, R. Alhajj, and J. Rokne, "Employing Social Network Construction and Analysis in Web Structure Optimization", in *Social Networks Analysis and Mining: foundations and applications*, by Springer, pp.13-34, 2010.

6. S. Gao, O. Addam, M. Nagi, A. Qabaja, A. ElSheikh, P. Peng, O. Sair, W. Almansoori,T. Özyer, J. Zeng, J. Rokne, R. Alhajj, “Robust Integrated Framework for Effective Feature Selection and Sample Classification and Its Application to Gene Expression Data Analysis”, *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, May 2012.

7. P. Peng, M. Nagi, O. Sair, I. Suleiman, A. Qabaja, A. M. ElSheikh, S. Gao, T. Özyer, K. Kianmehr, G. Naji, M. J. Ridley, J. G. Rokne, R. Alhajj,, “From Alternative Clustering to Robust Clustering and Its Application to Gene Expression Data”, *Proceedings of the International Conference on Intelligent Data Engineering and Automated Learning*, Springer-Verlag LNCS, Norwich, UK, Sep. 2011.

8. M. Nagi, A. M. ElSheikh, I. Suleiman, P. Peng, M. Rifaie, K. Kianmehr, P. Karampelas, M. Ridley, J. Rokne and R. Alhajj, “Association rules mining based approach forWeb usage mining”, *Proceedings of IEEE International Conference on Information Reuse and Integration*, Las Vegas, Aug. 2011.

9. M. Rifaie, M. Nagi, A. Rahmani, K. Kianmehr, M. Ridley and R. Alhajj, “Improving Database Performance by Building and Analyzing Social Network of Tables from Query Access Patterns”, *Proceedings of the International Conference on Applications of Social Network Analysis*, Zurich, Switzerland, Sep. 2011.

10. A. M. ElSheikh, T. N. Jarada, M, Nagi, G. Naji, P. Karampelas, O. Sair, P. Peng, K. Kianmehr, T. Özyer, M. Ridley, J. Rokne and R. Alhajj, “Effectiveness of Feature Selection and Classification Techniques for Gene Expression Data Analysis”, *Proceedings of the International Conference on Information Technology*, May 2011.

11. A. Rahmani, M. Nagi, M. Rifaie, K. Kianmehr, M. Ridley, R. Alhajj and J. Rokne, “Employing Frequent Pattern Mining for Finding Correlations between Tables in Relational Databases,” *Proceedings of International Conference on Information Technology: New Generations*, April 2011.