Data Mining of Web Access Logs

A minor thesis submitted in partial fulfilment of the requirements
for the degree of Master of Applied Science in Information Technology

Anand S. Lalani

School of Computer Science and Information Technology
Faculty of Applied Science
Royal Melbourne Institute of Technology
Melbourne, Victoria, Australia

July, 2003

Abstract

Analysis of web visitors access patterns can lead to benefits in a wide range of areas such as decision support and website restructuring. Data mining techniques can be used to find access patterns hidden inside huge volumes of web access data. The goal of this thesis is to determine whether there are any such patterns in the web access data for the computer science website of RMIT university. In particular, this thesis investigates whether there are any differences in access patterns between: (1) Visitors from within Australia and visitors from outside Australia. (2) Visitors from within RMIT university and visitors from outside RMIT university. (3) Visitors from within RMIT university and visitors from outside RMIT university but within Australia. (4) Visitors from educational institutions other than RMIT university and visitors from non-educational institutions.

The data mining techniques of classification, association rules, clustering and attribute selection were used with four different feature sets. The entire pattern discovery process was divided into three major steps: (1) Transaction identification and feature extraction (2) Discovery of the access patterns. (3) Analysis of the discovered patterns for their interestingness.

Three major patterns were discovered: (1) Visitors from Australia generally visit the root page while visitors from outside Australia do not. The most likely reason for this is that visitors from outside Australia use search engines that direct them to specific pages. However, some (2) Visitors from outside Australia visit the root page and pages about post graduate programs (such as Master of Technology). This suggests that these visitors are mostly interested in post graduate studies. (3) Visitors from other educational institutions tend to visit pages related to staff contact information while other visitors tend to access career and industry related information.

During the course of the investigation, it was found that there were a significant number of long transactions. The long transactions were analysed manually and it was found that visitors in a significant number of transactions access information about different programs offered and towards the end of their visit they look for the information brochure of one program.

The significance of the patterns discovered during this thesis work suggests that data mining techniques with suitable feature sets can produce very interesting patterns.

Declaration

This thesis contains work that has not been submitted previously, in whole or in part, for any other academic award and is solely my original research, except where acknowledged. The work has been carried out since the beginning of my candidature on 4 March 2003.

Signed,

Anand S. Lalani
School of Computer Science and Information Technology
Royal Melbourne Institute of Technology

September 5, 2003

Acknowledgments

First of all, I would like to express my sincere gratitude to my supervisor Dr. Victor Ciesielski for his valuable guidance and advice. He has been very understanding of the problems, supportive and patient throughout the progress of my thesis. I would also like to thank Laura Thomson for providing basic scripts of preprocessing and guidance. I am thankful to Dr. Isaac Balbin for extending the thesis submission date. I also thank Alexander Rosenberg for proof-checking my thesis. And last but not least, I would like to thank my beloved father who has encouraged me throughout my career to reach new milestones.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Data are pervasive. Machines record our program schedules, preferences, events, achievements, buying and selling. Even our comings and goings are recorded today. As the volume of data increases, it becomes a more difficult task to comprehend it. Information technology has made it possible to manage a huge volume of data electronically and to be able to search for potentially very useful knowledge hidden inside this deep ocean of data. *Data Mining*, the methodology for the extraction of knowledge from data [27], seems the only solution to this ever growing problem.

With the technological advancements, businesses have gone online. The World Wide Web has, since then, been the ultimate and vast source of information. For example, people today can buy desired things by just clicking on a button in the computer. Because of the growing popularity of the World Wide Web, many websites typically experience thousands of visitors everyday. Analysis of who browsed what can give important insight into, for example, what are the buying patterns of existing customers. Interesting information extracted from the visitors browsing data help analysts to predict, for example, what will be the buying trends of potential customers. Correct and timely decisions made based on this knowledge have helped organizations in reaching new heights in the market.

The computer science website of RMIT university has information for both students studying at the university and prospective students. It has many web pages that provide information such as courses being taught in the computer science department, research topics, timetables and international programs.

Web servers store information of each page requested by web visitors in a file called the *web access log*. *Web Usage Mining* addresses the problem of extracting behavioural patterns from one or more web access logs [20]. As can be seen in Figure 1.1, the entire process can be divided into three major steps. The first step, *preprocessing*, is the task of accurately identifying pages accessed by web visitors. This is a very difficult task because of page caching and accesses by web crawlers. The second step, *pattern discovery*, involves applications of data mining algorithms to the preprocessed data to discover patterns. The last

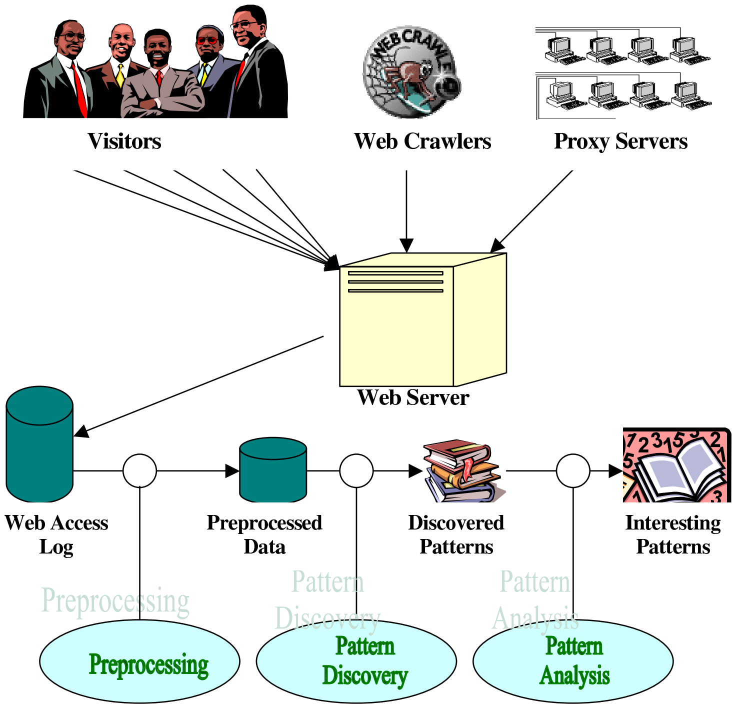


Figure 1.1: *Major Steps of Web Usage Mining Process*

step, *pattern analysis*, involves analysis of patterns discovered to judge their interestingness.

Data mining techniques such as association rules, classification, clustering and attribute selection are considered very useful in web usage mining. Association rules find the correlations among the items in large data sets [13]. Examples of association rules in accessing an educational website can be:

```
70 % of the web visitors who visited ``research'' page also visited
   ``scholarship'' page.
61 % of the web visitors who visited ``courses'' page also visited
   ``subjects'' page.
40 % of the web visitors who visited ``databases'' page also visited
   ``data-mining'' page.
```

The percentages reported in the examples above are referred to as *confidence*. Confidence is the number of instances in which all the items in a rule appeared, expressed as a proportion of total number of instances. The *support* is the percentage of the instances that contain a given pattern. Some examples of support are:

```
7.5 % of the web visitors visited both ``research'' and ``scholarship'' pages.
6.1 % of the web visitors visited both ``courses'' and ``subjects'' pages.
4.7 % of the web visitors visited both ``databases'' and ``data-mining'' pages.
```

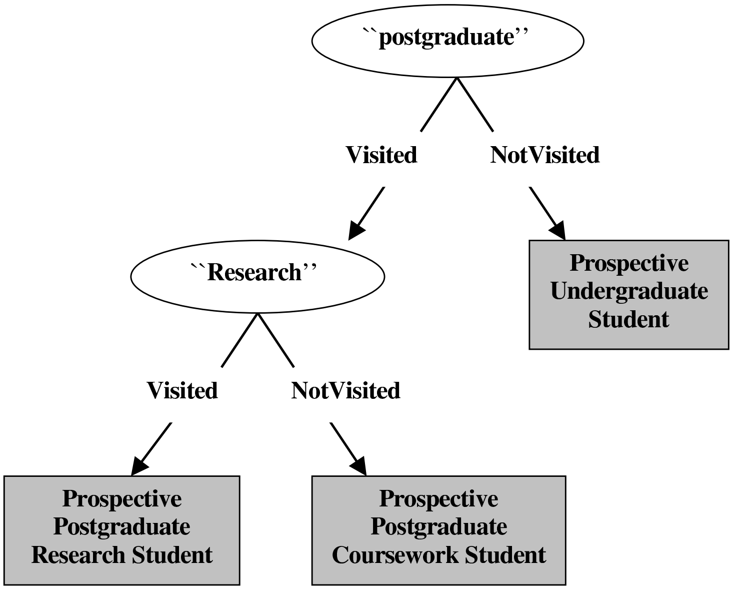

``postgraduate''
Visited
NotVisited
``Research''
Prospective Undergraduate Student
Visited
NotVisited
Prospective Postgraduate Research Student
Prospective Postgraduate Coursework Student

Figure 1.2: *An Example of a Decision Tree*

The association rules generated show the relationships among items in the data. Because a large number of such relationships can be established, *confidence* and *support* help analysts to filter out unwanted rules.

Classification is a technique to allocate instances into one of several predefined classes or categories. This requires extraction and selection of features that best describe the properties of a given class [13]. For example, in constructing a classification scheme for prospective students of a university, the interest is to accurately classify prospective students as undergraduate students or postgraduate students.

One of the most popular classification techniques is the *Decision Trees*. A decision tree is a structure that represents hierarchical rules. In a decision tree, nodes test an attribute value and leaves represent predefined classes. Figure 1.2 shows a decision tree of 3 leaves and 2 decision nodes. A part of this decision tree can be interpreted as: if visitors visit the "postgraduate" page and also visit the "research" page then they are prospective postgraduate research students.

Clustering is a technique to group together data items that have similar characteristics or properties. Clustering of web visitors, for example, tends to establish groups of visitors exhibiting similar browsing behaviours. Knowledge discovered can be used to provide preferred information to the visitors. Once clusters are formed, unknown instances may be allocated to one or more suitable clusters.

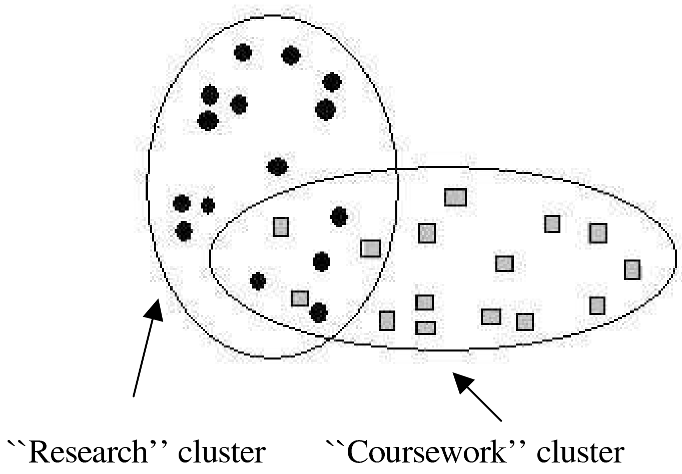


Figure 1.3: *An Example of Clustering*

Figure 1.3 shows 2 clusters formed. As can be seen from the figure, "Research" is a cluster of visitors who tend to browse information about research. Similarly, "Coursework" is a cluster of visitors who tend to browse information about coursework study. Note that some visitors belong to both the clusters and hence these clusters overlap.

Attribute Selection is a technique to select the most relevant and discriminating attributes from a data set. It has been observed that the presence of irrelevant attributes affects the performance of data mining algorithms to a considerable extent [34]. Once the most relevant attributes are identified, the data mining techniques can be applied to only these identified attributes to improve the performance as well as the results. Therefore, the process of attribute selection is often applied before application of other data mining techniques such as decision trees [34].

1.1 Goals

The goal of this thesis is to discover interesting usage patterns of the computer science website of RMIT university. In particular, we investigate whether there are any differences in access patterns between:

1. Web visitors from within Australia and visitors from outside Australia.

2. Web visitors from within RMIT university and visitors from outside RMIT university.

3. Web visitors from within RMIT university and visitors from outside RMIT university but within Australia.

4. Web visitors from educational institutions other than RMIT university and other visitors.

1.2 Scope

Due to time limitations for a minor thesis, the scope of this thesis is restricted to

- 2 supervised learning algorithms - 1R and J48.
- 1 association finding algorithm - Apriori
- 1 clustering algorithm - EM
- 1 attribute selection algorithm, correlation based feature selection subset evaluator - *CfsSubsetEval* with *Best-first* search through the attribute combinations.
- 2 web access log files of June-2001 and March-2003.

Chapter 2

Literature Survey

This chapter discusses the previous work done in this area. Section 2.1 describes the work done in different phases of data mining and the data mining algorithms used in this thesis. The research efforts made for applications of data mining techniques to the World Wide Web data are discussed in Section 2.2. Section 2.3 covers details of web access logs and work done in web usage mining. The last section describes relevant work done by researchers to this thesis work.

2.1 Data Mining

Data Mining is gaining more popularity because of its power to extract knowledge from voluminous data where it is beyond the reach of traditional techniques of knowledge discovery and human comprehension. The entire process of data mining can be divided into 4 phases: data collection, data preparation, pattern discovery and pattern analysis.

The data collection phase involves collection of data from various sources and identifying desired features for mining. Data collected from various sources may be heterogeneous. The data preparation phase involves the process of normalizing the data and representing them in a structure so that they become more manageable. Features identified in the previous phase are extracted and finally, formatting is applied to represent data in a format required by the data mining tool to be used.

2.1.1 Mining Techniques and Pattern Discovery

Once the data are prepared and formatted, the data mining techniques are applied to discover patterns. For the experiments of this thesis, a data mining tool, WEKA [11] is used. The WEKA is a collection of machine learning algorithms for solving real-world data mining problems. It is open source software issued under the GNU General Public License. Data mining tools for preprocessing, classification, clustering, attribute selection and visualization are implemented in the WEKA. The data mining techniques used for the experiments are described below.

1. **Classification:** Classification is a technique of *supervised learning* meaning that the number of categories or classes that the instances fall into are known in advance before the pattern discovery process starts. In the data set, an attribute that represents the classes is known as a *class variable.*

   In supervised learning, a model is first built using the training data. Once the model is built, it is then applied to the test data. One of the problems found in supervised learning is that the difference in the accuracies at correctly classifying instances of training and test data can be considerable. This problem is referred to as *overfitting* [34]. One of the solutions to this problem is to rerun the algorithm with different values of the parameters attempting to make the results less specific to the training data. The technique of *cross-validation*, which is based on “re-sampling” [18], is used for estimating the error rate. In a p-fold cross-validation, data are divided into p subsets of equal size. The model is trained p times, with a different subset omitted from the training set each time. The accuracy is measured only for the omitted subset. The error rate is the average of the error rates of p runs.

   The classification algorithms used for the experiments of this thesis are “1R” and “J48”. These algorithms and their parameters used in the experiments are described below.

   (a) **1R:** It is often observed that rules involving just one attribute often work astonishingly well [34]. The 1R alogorithm tests each attribute in turn and the attribute that gives the highest accuracy is chosen. As mentioned by [34], 1R often classifies instances quite accurately and finds good rules for characterizing the structure in data.

   Table 2.1 shows a sample of fictitious “weather” data set that supposedly concerns the conditions that are suitable for playing some unspecified game [34]. As can be seen from Table 2.1, the attributes in this data set are *outlook*, *temperature*, *humidity* and *windy*. All the attributes have nominal values. *Play* is the class variable that corresponds to the condition for playing and whose value can be either “yes” or “no”.

   Figure 2.1 shows partial output of the WEKA 1R program using “weather” data set. As can be seen from the Figure, the first clause of the rule can be interpreted as: if the *outlook* is “sunny” then *play* = “no”. Similarly the third clause of the rule can be interpreted as: if the *outlook* is “rainy” then *play* = “yes”.

   (b) **J48:** J48 is an implementation of an improved version of the C4.5 decision tree algorithm in the WEKA, as mentioned in [11]. C4.5 is a landmark machine

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

Table 2.1: *A Sample of the Weather Data*

```
   === 1R Classifier model (full training set) ===

outlook:
        sunny     -> no
        overcast  -> yes
        rainy     -> yes
```

Figure 2.1: *Partial Output Produced by the WEKA 1R Program Using the Weather Data*

learning method that is most widely used [34]. The J48 algorithm works with nominal as well as numeric attributes. An important step in building the decision tree is *pruning*, which removes the least reliable branches, generally resulting in faster classification and an improvement in the accuracy of correctly classifying independent test data [13]. The default value of *confidence* in J48 is 0.25. The lower the value of this parameter, the more the pruning. The second important parameter is *Minimum number of objects* whose effect is to eliminate tests for which the number of instances are less than the value set for this parameter. The complete description of C4.5 is available in [26].

Table 2.2 shows a sample of the iris data set [10]. As can be seen from this table, the four attributes in the data set are *sepal length*, *sepal width*, *petal length* and *petal width*. All the attributes have numeric values. *Class* is the class variable that corresponds to the types of plants.

The WEKA J48 program produced the decision tree shown in Figure 2.2 using the "iris" data. The decision tree classifies a plant based on the values of the four attribute values. It can be seen from Figure 2.2, for example, if the *petal-width* is greater than 0.6 and less than or equal to 1.7 and *petal-length* is less than or equal to 4.9 then the plant can be classified as "Iris-versicolor".

2. **Association Rules:** An association rule is a simple probabilistic statement about

SepalLength	SepalWidth	PetalLength	PetalWidth	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor
6.3	3.3	6.0	2.5	Iris-virginica
7.1	3.0	5.9	2.1	Iris-virginica
6.3	2.9	5.6	1.8	Iris-virginica

Table 2.2: *A Sample of the Iris Data*

the co-occurence of certain events in a database, and is particularly applicable to sparse transaction data sets [14].

The Apriori algorithm was proposed in [1] for mining frequent item sets. An item is considered frequent if its frequency is more than the *confidence* value. The algorithm uses an iterative approach to generate the next level of frequent item sets. i.e. from m level frequent item sets, it generates $m + 1$ level frequent item sets. The Apriori algorithm works efficiently due to the "apriori property" which states that, if an item is not frequent then all its supersets are also not frequent. Hence, they can be discarded. As described earlier on page 3, *confidence* and *support* parameters help to get association rules of the required interestingness. These parameters are supplied by the user. The Apriori algorithm does not work with numeric attributes.

Figure 2.3 shows some association rules generated by the WEKA Apriori program using the "weather" data set. The first rule can be interpreted as: if the *humidity* is "normal" and *windy* is "false" then *play* = "yes"(play the game). The confidence of this rule is 1.

3. **Clustering:** Clustering is a technique to group together items having similar characteristics. It is a technique of *unsupervised learning* because there is no knowledge about classes in advance. Clustering allocates instances into groups or clusters based on the principle of maximizing the similarity within groups and minimizing the similarity between groups [13]. The clusters formed at the end of the process may be probabilistic which means an instance belongs to a cluster with a certain probability, as described by [34].

   The EM (Expectation-Maximization) algorithm was proposed by [9] to maximize the overall probabilities of cluster memberships. In the first step, the EM algorithm calculates probabilities of cluster memberships and in the second step, it maximizes

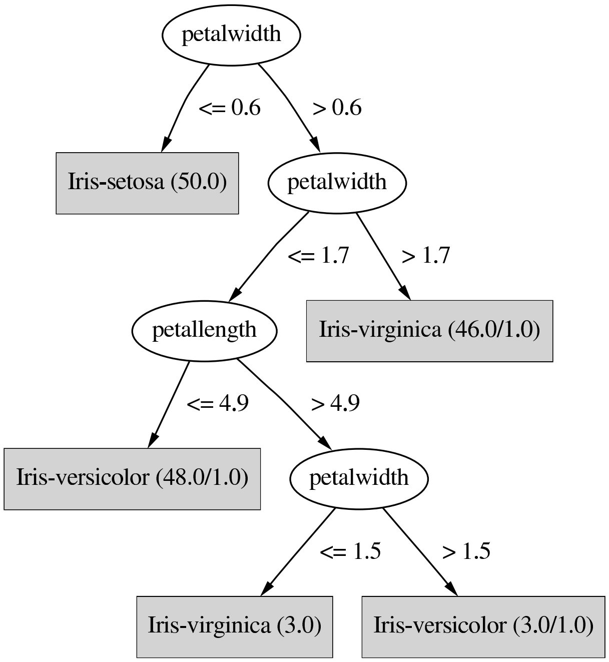


Figure 2.2: *Decision Tree Produced by the WEKA J48 Program Using the Iris Data*

these probabilities, as mentioned by [34]. The EM algorithm works well with missing data and when the probability distribution of the data is analytically intractable, as mentioned by [3].

Figure 2.4 shows annotated partial output of a cluster formed by the EM algorithm using "weather" data. The EM algorithm generates output by calculating *counts* for each nominal attribute. The *count* is the number of each distinct attribute values normalized into probabilities [34]. A part of output of the *outlook* attribute from Figure 2.4 can be interpreted as: instances which belong to this cluster have 6% (out of a total of 17%) probability that their *outlook* attribute value is *sunny*. Similarly, instances which belong to this cluster have 5% (out of a total of 17%) probability that their *outlook* attribute value is *overcast*. The output of other attributes can be interpreted similarly. Hence, this is a cluster of instances in which the probability of each value of each attribute is as shown in Figure 2.4.

4. **Attribute Selection:** The process of attribute selection attempts to identify the more relevant and discriminating attributes. As mentioned by [34], the performance of data mining algorithms degrades because of irrelevant attributes. Attribute selection attempts to reduce the dimensionality of data by deleting unsuitable attributes and hence improves the performance of data mining algorithms [34].

```
Minimum support: 0.4
Minimum confidence: 0.9

Best rules found:

 1. humidity=normal windy=false ==> play=yes conf:(1)
 2. temperature=cool ==> humidity=normal conf:(1)
 3. outlook=overcast ==> play=yes conf:(1)
 4. temperature=cool play=yes ==> humidity=normal conf:(1)
```

Figure 2.3: *Partial Output of the WEKA Apriori Program Using the Weather Data*

```
Attribute: outlook          (sunny) (overcast) (rainy)
Discrete Estimator. Counts =  6        5          6  (Total = 17)
Attribute: temperature      (hot)   (mild)     (cool)
Discrete Estimator. Counts =  5        7          5  (Total = 17)
Attribute: humidity         (high)  (normal)   (low)
Discrete Estimator. Counts =  1        8          8  (Total = 17)
Attribute: windy            (false) (true)
Discrete Estimator. Counts =  9        7             (Total = 16)
Attribute: play             (no)    (yes)
Discrete Estimator. Counts =  6       10             (Total = 16)
```

Figure 2.4: *Annotated Partial Output of the WEKA EM Program Using the Weather Data*

The *CfsSubsetEval* attribute evaluator [11] evaluates subsets of attributes by the correlation among them, as described by [24]. It selects attributes that correlate with the class variable while having less inter-correlation among them, as mentioned in [11]. In the attribute selection process, a search method is required to search through the possible subsets of attributes that best predict the class. *Best-first* is a search method that keeps a list of subsets evaluated in the order of their performance, as described by [34]. Hence, when the performance starts to drop for a subset, the search can revisit an earlier subset.

Figure 2.5 shows the sample output produced by CfsSubsetEval attribute evaluator with Best-first search method using the iris data. As can be seen from the Figure, *petal length* and *petal width* are the two attributes selected as the most relevant and discriminating.

```
Search Method:
        Best first.
        Total number of subsets evaluated: 12

Attribute Subset Evaluator (supervised, Class (nominal): 5 class):
        CFS Subset Evaluator

Selected attributes: 3,4 : 2
                     petallength
                     petalwidth
```

Figure 2.5: *Partial Output Produced by the WEKA for Attribute Selection Using the Iris Data*

2.1.2 Pattern Analysis

Pattern Analysis is the last phase of the knowledge discovery process. In this phase, the discovered patterns are judged for their interestingness. The exact methodology depends upon the application for which mining is done [31]. The analysis can be done using the content and structure of the website. Visualization techniques that represent the values at different positions using different colours are often useful in the pattern analysis [31].

2.2 Web Mining

As a result of the tremendous growth of the World Wide Web, the raw web data has become a vast source of information. Consequently, this has turned researchers attention towards the use of data mining techniques to this data. *Web Mining* is referred to as the application of data mining technologies to the web data [4].

Web mining can be categorized into *content mining*, *structure mining* or *usage mining* depending upon which part of the web to mine [19].

2.2.1 Content Mining

The web content is the *real* data the web page was designed to convey to the users [31]. It consists of several types of data such as unstructured text, graphics, sound, video and semi-structured hypertext. Content mining can be referred to as the application of data mining algorithms to the content of the web [31]. A *conceptual schema* can be created [32] that can describe the semantics of a large volume of unstructured web data to manage them. [19] discussed various categories of the web content mining such as *text mining* which is mining of unstructured texts and *multimedia data mining* which is mining of multiple types of data such as unstructured and image data.

2.2.2 Structure Mining

The purpose of web structure mining is to classify web pages based on their organization. The strucuture mining can be used to categorize web pages. The potential use of this category of web mining is to generate information such as similarity between different websites [19].

2.2.3 Usage Mining

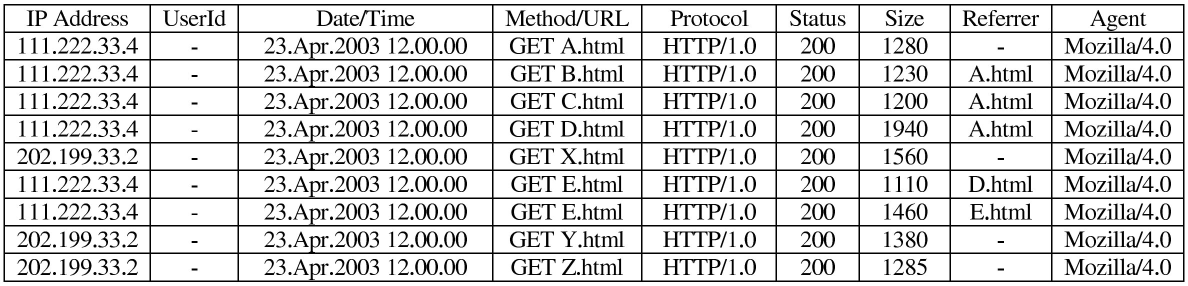
Analysis of web visitors usage habits can give important clues about current market trends and help organizations to predict the future trends of potential customers. Analysis of long visit-paths of users may indicate the need of restructuring of the website to help visitors reach desired information quickly. Also, the mined knowledge can be used to offer preferred web content to visitors.

A lot of research work has been conducted on web personalization. *Adaptive Sites* use information about user access patterns to improve their organization and presentation for different types of users [24]. A technique for capturing common visitor profiles using association rule discovery and usage-based clustering of URLs is proposed by [21]. A technique for web personalization proposed in [22] is based on association rule discovery from usage data. [29] developed a methodology to assess the quality of a web site based on the discovery and comparison of navigation patterns of customers and non-customers and also proposed a technique of dynamically adapting the site. An algorithm to reorganize a web site using page access frequency and classification of pages is proposed by [12]. A tool is developed and described in [20] for customizing the website dynamically.

[17] performed association rule mining to discover interesting behavioural patterns of mobile device users. Finding page locations that are different than where visitors expected them to be during their visit also helps to restructure website organization [30].

2.3 Web Access Logs and Web Usage Mining

The web access log is the main resource for web usage mining because it stores data pertaining to accesses of the website. The usage data can be stored in Common Log Format (CLF) or Extended Log Format (ELF). The web access log in CLF format has information of the *IP address* of a visitor's machine, the *userid* of visitor if available and *date/time* of the page request. The *method* is a means of page request. It can be GET, PUT, POST or HEAD. The *URL* is the page that is requested. The *protocol* is the means of communication used, HTTP/1.0 for example. The *status* is the completion code. For example, 200 is the code for success. The *size* field shows the bytes transferred as a result of a page request. The Extended Log Format, in addition to these information, stores *referrer*, which is the page this request has come from and *agent* is the web browser used. Figure 2.6 shows the web access log in Extended Log Format.



IP Address	UserId	Date/Time	Method/URL	Protocol	Status	Size	Referrer	Agent
111.222.33.4	-	23.Apr.2003 12.00.00	GET A.html	HTTP/1.0	200	1280	-	Mozilla/4.0
111.222.33.4	-	23.Apr.2003 12.00.00	GET B.html	HTTP/1.0	200	1230	A.html	Mozilla/4.0
111.222.33.4	-	23.Apr.2003 12.00.00	GET C.html	HTTP/1.0	200	1200	A.html	Mozilla/4.0
111.222.33.4	-	23.Apr.2003 12.00.00	GET D.html	HTTP/1.0	200	1940	A.html	Mozilla/4.0
202.199.33.2	-	23.Apr.2003 12.00.00	GET X.html	HTTP/1.0	200	1560	-	Mozilla/4.0
111.222.33.4	-	23.Apr.2003 12.00.00	GET E.html	HTTP/1.0	200	1110	D.html	Mozilla/4.0
111.222.33.4	-	23.Apr.2003 12.00.00	GET E.html	HTTP/1.0	200	1460	E.html	Mozilla/4.0
202.199.33.2	-	23.Apr.2003 12.00.00	GET Y.html	HTTP/1.0	200	1380	-	Mozilla/4.0
202.199.33.2	-	23.Apr.2003 12.00.00	GET Z.html	HTTP/1.0	200	1285	-	Mozilla/4.0

Figure 2.6: *Web Access Log in Extended Log Format*

The entire process of web usage mining can be divided into three steps as shown in Figure 1.1 on page 2. Pattern discovery involves the application of various techniques and algorithms to preprocessed data. Pattern analysis is the identification of useful patterns out of the patterns discovered.

2.3.1 Web Usage Preprocessing

Web usage preprocessing is probably the most difficult phase in the entire web usage mining process because of the complexity and amount of time involved. This phase covers tasks of removal of unwanted data, identification of visitors, visitor transactions and feature extraction. This process can be divided into four distinct steps:

1. **Removal of web robot accesses:** Web access logs, in general, contain entries by web robots like crawlers, spiders, indexers and other robots. The decision to remove entries by these robots depends upon the goal of web usage mining. The web robots in general access the "robots.txt" file created by the web site administrator for access permissions. Also, some robots can be identified by observing the Host name of the IP Addresses.

2. **Filtering image and noisy data:** The web pages also carry image, sound or video files along with the data. Consequently, the web server records entries of the content page that was requested and also records any image, sound or video files that were sent. Generally entries of image, sound and video files are discarded for web access log analysis. However, access patterns of these files can give interesting clues regarding web site structure, traffic performance and user motivation [35].

   Sometimes, visitors request pages that do not exist or no longer exist. At times, the server fails or visitors fail to authenticate themselves. In all of these scenarios, the web server makes a log entry with the appropriate code. Depending upon the purpose of the analysis, some or all of these entries may be included or discarded.

The web server also records accesses to file "proxy.pac" which is an auto-configuration file to configure all web browser clients. Typically, these entries are filtered out to avoid their effect on discovered patterns.

3. **Extracting Transactions:** Once the irrelevant entries are removed from the web access log, the next step is to extract *transactions* pertaining to individual users. There is no natural definition of a transaction in the site navigation scenario [33]. A transaction can be considered as a single entry in the log or a set of entries accessed by a visitor from the same machine in a defined time window. The desired transaction can be thought of as a set of log entries of a visitor in a single visit.

   The problems involved in extracting transactions for each user have been discussed in [25, 31]. Use of proxy servers, setting up a session limit that is the maximum time period set for browsing the web in a single visit, hyperlink structure and use of multiple IP addresses by a single user make it difficult to identify unique visitors.

   Several heuristics like *referrer log* can be applied to identify single user accesses [7]. Referrer log is a file that stores an entry of the document when it is requested along with an entry of the *referring* page through which the request has come. For example, if the visitor is currently viewing the page "/courses" and requests the page "/courses/postgraduate" then an entry in the referrer log can be made with "/courses" as the referring page. Use of referrer log along with the web site structure can help to identify unique visitors.

   The concept of *reference length* proposed in [8] can also be used to identify transactions. The reference length module is based on the assumption that visitors spend less time on *navigational* pages and more time on *content* pages. Transactions can also be extracted using *maximal forward references*, proposed by [6]. A maximal forward reference is a page that is accessed before a page is revisited. For example, in a single visit, if a visitor has visited pages I-J-K-I-L-J then the maximal forward references are K and L.

4. **Feature Extraction and Formatting:** The final step of preprocessing is to extract features from the transactions available. Feature extraction involves identifying the relevant attributes and reducing the dimensionality of the data by excluding irrelevant attributes. The task is to convert variable length transactions into fixed-length feature vectors.

   The web visitors usually spend more time on the pages that interest them compared to less interesting pages. A feature called "interestingness" is extracted by [33] to model user preferences in accessing different pages of a website. The size of the pages and the network speed are taken into consideration in defining this feature as these two factors affect the time spent in visiting the pages.

Finally, fixed length feature vectors are converted into the format required by the data mining tool to be used. Several researchers have made use of OLAP (Online Analytical Processing) and data cube technology for web usage mining. The relational database and a multi-dimensional data cube are used by [35] to represent preprocessed data. A tree structure is suggested by [23] to efficiently store access sequences to web pages. A cube structure can be used along with the web access log, marketing and site topology to discover internet marketing intelligence [5]. The web log data can be represented in a statistically and structurally aggregated tree form [28] to reduce the size and improve the performance of mining process. [16] suggested a relational OLAP approach to build a warehouse of web accesses and presented a tool to execute analytic queries on the warehouse.

2.4 Related Work on Mining of Web Access Logs

Many data preparation techniques are presented by [7, 15] in order to perform web usage mining on web access logs. However, there is not much literature available that describes preprocessing in detail for web usage mining [2].

The problems involved in web usage mining have been discussed in [7, 15]. One of the problems in getting an accurate picture of the website access is caused by web browsers and proxy servers. Web browsers store pages that have been visited and if the same page is requested, the web browser displays the page rather then sending another request to the web server. Proxy servers cache frequently visited pages locally to reduce network traffic and improve server performance. This problem caused by web browsers and proxy servers can be solved by use of *Cookies* and *Remote Agents*, as described in [7].

The entries of web robots can be deleted by heuristically finding them in web access logs, as mentioned by [15]. The web robots can also be identified by the host names. Hence, by preparing a list of the web robots, the entries can be removed from web access logs. As mentioned in Section 2.3.1, web robots in general, access “robots.txt” file for website permissions. The entries by web robots can be filtered by preparing a list of the IP addresses that have accessed “robots.txt” file but brand-new robots and page scanning tools remain in the web access log [15].

[7] has presented several data preparation techniques to identify visitors and their transactions. Visitors can be identified by applying several heuristics. [7] defined the term *path completion* as the process of identifying missing page accesses that are not recorded in the access log and suggested the use of site topology and referrer log to accomplish this task.

It is quite common that web visitors visit a website more than once. Also, it is possible that one user ends his or her visit and another user starts a visit to the same website from the same computer. A simple way to identify individual transactions is to use a timeout. If the time period between two consecutive page visits is more than a fixed time period then it should be considered as the end of a transaction and beginning of another. Thirty minutes is often used as a timeout period, as mentioned in [7, 15, 31].

Chapter 3

Data Mining Experiments

This chapter describes the preprocessing tasks done and the experiments conducted. The first section provides an overview of the structure of the experiments performed. Section 3.2 covers the data preparation tasks performed. This is followed by details of the experiments conducted.

3.1 Overview

This section briefly describes the organization of experiments conducted and the pattern discovery tasks performed for each experiment. Figure 3.1 shows a hierarchical view of the pattern discovery tasks performed in the experiments of this thesis and is intended to show the organization of the experiments.

As mentioned in Section 1.1, there are four goals of this thesis. The access log files of the June-2001 and the March-2003 were used for the experiments. One experiment addresses one goal on one access log file. Hence, a total of 8 experiments were conducted. For example, as can be seen from Figure 3.1, experiment 4 was conducted for the goal RMITVsNotRMIT using the March-2003 log file.

As can be seen from Figure 3.1, for each experiment, 4 data mining techniques were used. These techniques were classification, association rules, clustering and attribute selection. A pattern discovery task for one of these data mining techniques was performed using one of four different feature sets extracted from the preprocessed data. The four different feature sets used for the experiments are First3-Last2, First5-Last5, 20-Most-Frequent-TF and 20-Most-Frequent-Time. As an example, a pattern discovery task can be performed to find association rules using the First3-Last2 feature set. The feature sets are described in detail in Section 3.2.3.

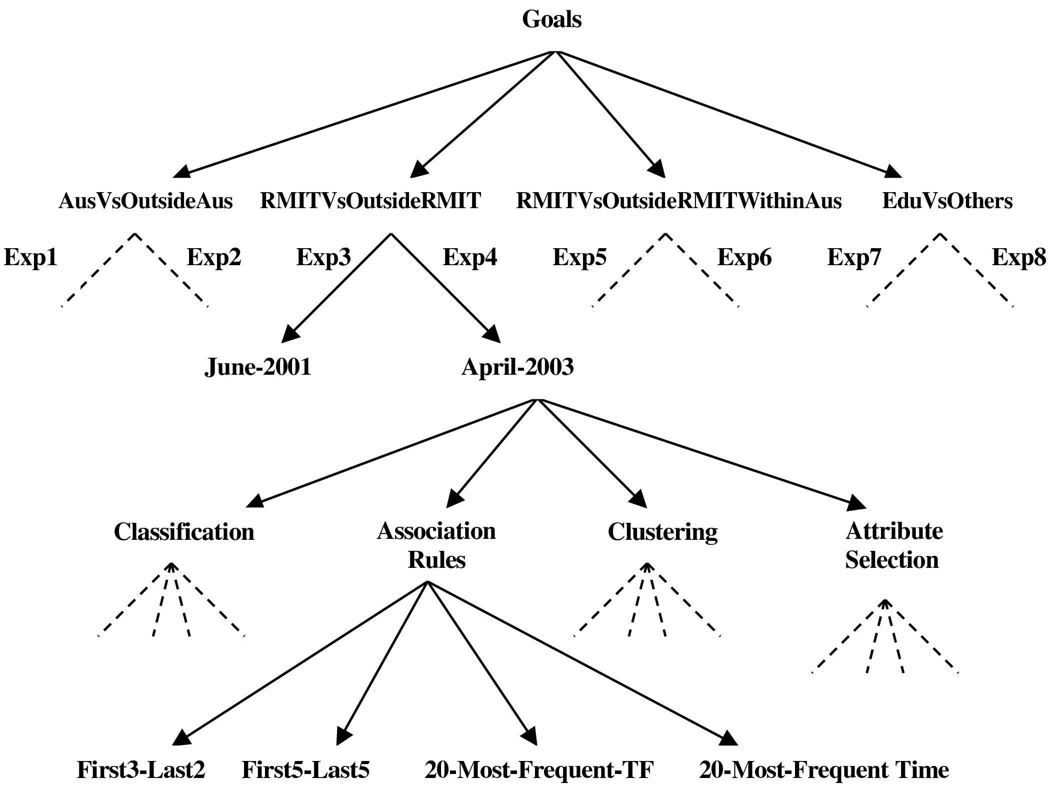


Figure 3.1: *A Hierarchical View of the Pattern Discovery Tasks Performed*

3.2 Data Preparation

3.2.1 Web Log Data

Details of web log data used for the experiments are shown in Table 3.1. The web crawler entries were not useful for the experiments. As mentioned in Section 2.3.1, crawlers generally access "robots.txt" file for the website access permissions. Therefore, these entries were removed by preparing a list of crawlers that accessed the "robots.txt" file. However, this technique does not remove all crawler entries as mentioned in Section 2.4.

The log entries of IP addresses that no longer existed were also not useful for the experiments. Hence, those entries were removed.

The image entries, entries by proxy servers and entries of bad requests, as described in Section 2.3.1, were not useful for the experiments of this thesis. Therefore, they were removed.

Web Access Log File	Number of Entries	Time Period
access2001	1000000	29/05/2001 - 03/06/2001
access2003	11390257	04/02/2003 - 23/04/2003

Table 3.1: *Details of Web Access Log Files*

3.2.2 Transaction Identification

An *IP-Day* consists of all entries of the pages visited from one IP address in a day. Once the irrelevant entries were removed from the web log data, IP-Days were extracted based on the IP addresses. The transactions were identified from the IP-Days by the time duration between two consecutive visits. If the time duration between accesses of two pages "X" and "Y" in an IP-Day is more than 30 minutes then "X" was considered as the last page accessed in one transaction and "Y" was considered as the first page accessed in another transaction. A time period of 30 minutes was considered appropriate to distinguish two transactions as discussed in Section 2.4 on page 17. It may be possible that a visitor starts a visit at 11:55 pm and ends at 00:10 am in which case there will be two transactions generated instead of one.

It was conjectured that transactions that have atleast 5 pages visited would be useful for this data mining task. Transactions with fewer than 5 visited pages were not used. Table 3.2 shows the number of transactions used for the experiments from each log file.

Web Access Log File	Number of Transactions
access2001	4591
access2003	55602

Table 3.2: *Number of Transactions Used from Web Access Log Files*

3.2.3 Feature Sets

Once the transactions are identified, the next step is to extract features from the transactions, as mentioned in Section 2.3.1 on page 15. For the experiments, four feature sets were extracted for the pattern discovery process:

1. **First3-Last2:** In this feature set, an instance consists of the first 3 and the last 2 pages accessed by a visitor in a transaction. It is possible that there may be less than five pages visited in a transaction. In this case, an instance will have one or more missing attribute values. This feature set is based on the conjecture that the first 3 and last 2 pages visited will contain information critical to distinguish classes of visitors.

A sample of instances represented using this feature set is shown in Figure 3.2. The instances are represented in "comma separated" format. An attribute name in this feature set indicates the order of the pages visited. For example, the "link0" attribute corresponds to the first page visited in a transaction. Similarly, "link2last" attribute corresponds to the second last page visited in a transaction. The last attribute was used to act as the class variable. This attribute represents the different classes that the instances belong to.

```
Host,Link0,Link1,Link2,Link2last,Linklast,Location

i-gate.abz.nl,/,/employment,/,/,/students,NotRMIT
vail.cs.ucsb.edu,/,/timetable,/timetable/city,/timetable/city/01,/course,NotRMIT
knu.cs.rmit.edu.au,/,/course,/course/pgraduate,/course/pgraduate/mit,/,RMIT
csse.monash.edu.au,/,/staff,/general/contact/phone.shtml,/,/course,NotRMIT
```

Figure 3.2: *A Sample of Instances Using the First3-Last2 Feature Set*

2. **First5-Last5:** This feature set is the same as the First3-Last2 feature set except that in this feature set, an instance consists of the first 5 and the last 5 pages visited by the visitor in a transaction. It may be possible that there may be less than ten pages visited in a transaction. In this case, an instance will have one or more missing values. This feature set is chosen to mine more information compared to First5-Last2 feature set.

3. **20-Most-Frequent-TF:** The web access log files were analysed to find the most frequently visited pages. The 20 most frequently visited pages were selected as attributes in this feature set. An attribute value in an instance is "T" if that particular page was visited in the transaction and "F" otherwise. A sample of instances using the 6-Most-Frequent-TF feature set is shown in Figure 3.3.

   It was thought that it would be possible to categorize visitors browsing behaviours based on whether they visited a frequently visited page or not. Hence, this feature set was selected for the experiments.

4. **20-Most-Frequent-Time:** This feature set has the same attributes as those in 20-Most-Frequent-TF. In this feature set, an attribute value is the amount of time the visitor spent on that particular frequently visited page, i.e. the attribute value is the time spent in seconds instead of either "T" or "F" as used in 20-Most-Frequent-TF feature set. The duration can be calculated by taking the time difference between two consecutive page requests. The attribute value "0" indicates that this particular page was not visited by the visitor in that transaction. A sample of instances using 6-Most-Frequent-Time feature set is shown in Figure 3.4. It may be possible that one

```
Host,/,/course,/student,/timetable,/course/pgraduate,/staff,Location

i-gate.abz.nl,F,F,T,F,F,F,NotAus
vail.cs.ucsb.edu,F,T,F,F,F,T,NotAus
knu.cs.rmit.edu.au,T,F,F,F,T,T,Aus
csse.monash.edu.au,T,T,T,F,T,F,Aus
```

Figure 3.3: *A Sample of Instances Using the 6-Most-Frequent-TF Feature Set*

of the most frequently visited pages is visited last in a transaction. In this case, visit duration for this page can not be calculated. Hence, in this case, the corresponding attribute will have a missing value.

This feature set is based on the conjecture that time spent on frequently visited pages might be a very distinguishing factor to categorize visitors.

```
Host,/,/course,/student,/timetable,/course/pgraduate,/staff,Location

i-gate.abz.nl,55,0,8,0,127,0,NotAus
vail.cs.ucsb.edu,0,9,210,0,0,0,NotAus
knu.cs.rmit.edu.au,67,0,0,0,56,0,Aus
csse.monash.edu.au,44,345,43,0,89,0,Aus
```

Figure 3.4: *A Sample of Instances Using the 6-Most-Frequent-Time Feature Set*

3.2.4 Formatting

The instances represented using a feature set must be converted into a specific format required by the data mining tool to be used for the pattern discovery process as described in Section 2.3.1. The WEKA data mining tool, as described in Section 2.1.3, was used for the experiments. The WEKA tool requires that the input data set should be represented in the ARFF format. Figure 3.5 shows a sample of instances represented in ARFF format using the 20-Most-Frequent-Time feature set.

3.3 Experiment1: AusVsOutsideAus2001

The first experiment was conducted to compare access patterns between visitors from within Australia and visitors from outside Australia using the web access log file “access2001” as shown in Table 3.1. It was conjectured that there would be some differences in patterns of usage between visitors to the computer science website from within and from outside

Australia. This experiment was designed to determine whether this is in fact the case and what the differences are.

The data mining techniques of classification, clustering, association rules and attribute selection as described in Section 2.1.3 were used for the pattern discovery process. The value of the class variable for each instance for this experiment was either "Aus" or "NotAus". This value was determined from the host name of the instance. For example, if the host name ended in ".au" substring then that particular visitor was considered from within Australia and hence the value was set as "Aus". If the host name did not contain ".au" substring at the end then this visitor was considered to be from outside Australia and the value of the class variable was set as "NotAus".

3.4 Experiment2: AusVsOutsideAus2003

The second experiment was conducted to compare access patterns between visitors from within Australia and visitors from outside Australia using the web access log file "access2003" as shown in Table 3.1. This experiment was the same as the first experiment except that the web access log file used was different.

The number of transactions generated from the web access log file "access2003" was very large and beyond the capacity of the WEKA EM clustering program to produce the output in a reasonable time. For example, it takes 19 minutes to produce the output for 1000 instances, 1 hour and 26 minutes for 2000 instances and 6 hours for 5000 instances. Given the limited time period for the minor thesis, it was decided to take sample transaction sets from different parts of the "access2003" file and perform the experiments using these subsets. For the experiments, the first 5000 transactions, the last 5000 transactions and central 5000 transactions were taken from the "access2003" log file. A discovered pattern was analysed only if that pattern was discovered in the output of all the subsets.

3.5 Experiment3: RMITVsOutsideRMIT2001

The third experiment was conducted to compare access patterns between visitors from within RMIT university and from outside RMIT university using the web access log file "access2001" as shown in Table 3.1. This experiment was designed based on a presumption that the visitors within RMIT university might be accessing the computer science website differently compared to visitors from outside RMIT university. The aim of this experiment was to find the differences if there are any.

The data mining techniques used for the pattern discovery process were the same as those used in Experiment1. The value of the class variable for each instance for this experiment was either "RMIT" or "NotRMIT". This value was determined from the host name of the instance. For example, if the host name had ".rmit." substring in it then that particular visitor was considered to be from within RMIT university and hence the value was set as "RMIT". If the host name did not contain ".rmit." substring then this visitor was considered to be from outside RMIT university and the value of the class variable was set as "NotRMIT".

3.6 Experiment4: RMITVsOutsideRMIT2003

The fourth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university using the web access log file "access2003" as shown in Table 3.1. This experiment was the same as the third experiment except that the web access log file used was different.

As the number of transactions generated from the web access log file "access2003" was very large, the WEKA EM clustering program was run on subsets of the "access2003" file as explained in experiment 2 on page 23.

3.7 Experiment5: RMITVsOutsideRMITWithinAus2001

The fifth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university but from within Australia using the web access log file "access2001" as shown in Table 3.1. This experiment was based on the conjecture that there might be some differences in patterns of access to the computer science website between visitors from within RMIT university and outside RMIT university but within Australia.

The data mining techniques used for the pattern discovery process were same as those used in Experiment1. The value of the class variable for each instance was either "WithinRMIT" or "OutsideRMIT". This was determined from the host name of the instance. If the host name had ".rmit." substring in it then the visitor was considered to be from within RMIT university. If the host name did not contain ".rmit." but ended in ".au" substring then the visitor was considered to be from within Australia but outside RMIT university. Hence, the value was set as "OutsideRMIT". The instances in which the host name did not contain ".rmit." or did not end in ".au" were discarded. The number of transactions used for this experiment is shown in Table 3.3.

3.8 Experiment6: RMITVsOutsideRMITWithinAus2003

The sixth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university but from within Australia using the web access log file "access2003" as shown in Table 3.1. This experiment

Experiment	Number of Transactions
Experiment5: RMITVsOutsideRMITWithinAus2001	3054
Experiment6: RMITVsOutsideRMITWithinAus2003	48333

Table 3.3: *Number of Transactions Used for Experiments 5 and 6*

was the same as the fifth experiment except that the web access log file used was different. The number of transactions used for this experiment is shown in Table 3.3.

As the number of transactions generated from the web access log file “access2003” was very large, the WEKA EM clustering program was run on subsets of the “access2003” file as explained in experiment 2 on page 23.

3.9 Experiment7: EduVsOthers2001

The seventh experiment was conducted to compare access patterns between visitors from educational and non-educational organizations using the web access log file “access2001” as shown in Table 3.1. It was hypothesized that visitors from educational institutions might be accessing the computer science website differently compared to other visitors. This experiment was designed to look for differences in their patterns of usage.

The data mining techniques used for the pattern discovery process were same as those used in Experiment1. The value of the class variable for each instance for this experiment was either “Edu” or “NotEdu”. This value was determined from the host name of the instance. For example, if the host name ended in “.edu” or “.ac.uk” or “.ac.nz” then that particular visitor was considered to be from an educational organization and hence the value was set as “Edu”. If the host name did not contain any of these substrings in it then this visitor was considered from a non-educational organization and the value of the class variable was set as “NotEdu”. It was thought that accesses of the visitors from RMIT university will dominate the patterns of visitors from educational institutions to be discovered. Therefore, entries of visitors from RMIT university were discarded to avoid this effect. The number of transactions used for this experiment is shown in Table 3.4.

Experiment	Number of Transactions
Experiment7: EduVsOthers2001	2022
Experiment8: EduVsOthers2003	35482

Table 3.4: *Number of Transactions Used for Experiments 7 and 8*

3.10 Experiment8: EduVsOthers2003

The eighth experiment was conducted to compare access patterns between visitors from educational and non-educational organizations using the web access log file “access2003” as shown in Table 3.1. This experiment was the same as the seventh experiment except that the web access log file used was different. The number of transactions used for this experiment is shown in Table 3.4.

As the number of transactions generated from the web access log file “access2003” was very large, the WEKA EM clustering program was run on subsets of the “access2003” file as explained in experiment 2 on page 23.

```
@relation weblogdata

@attribute Host {’i-gate.abz.nl’,’vail.ucsb.edu’,’knu.edu.au’,’csse.edu.au’}
@attribute / real
@attribute /courses/ real
@attribute /students/ real
@attribute /timetables/ real
@attribute /courses/postgraduate real
@attribute /staff/ real
@attribute /employment/ real
@attribute /general/contact/phone.shtml/ real
@attribute /timetables/city/2001s1/index.htm real
@attribute /timetables/city/2001s1/list.htm real
@attribute /timetables/city/2001s1/help.htm real
@attribute /timetables/city/2001s1/cover.htm real
@attribute /conf/doa/2001/ real
@attribute /˜pmcd real
@attribute /˜pmcd/teaching/cs208/ real
@attribute /˜winikoff/palm/dev.html real
@attribute /˜caspar/turbo-mazda-323.htm real
@attribute /˜shyam/cs492meta.html real
@attribute /timetables/city/2001s2/help.htm real
@attribute /˜linpa/SE/design.html real
@attribute Location {’Aus’,’NotAus’}

@data
’i-gate.abz.nl’,55,0,8,0,0,0,0,0,17,0,0,0,0,0,55,0,0,0,0,0,’NotAus’
’vail.ucsb.edu’,0,9,0,0,210,0,0,303,0,0,41,0,0,6,0,0,111,0,0,300,’NotAus’
’knu.edu.au’,67,0,0,0,0,56,78,0,0,91,0,23,441,0,0,1700,0,43,112,0,’Aus’
’csse.edu.au’,44,345,43,0,89,0,8,4,88,0,777,0,5,8,111,0,0,0,0,0,’Aus’
```

Figure 3.5: *A Sample of Data in ARFF Format Using the 20-Most-Frequent-Time Feature Set*

Chapter 4

Experimental Results

The experiments described in Chapter 3 were conducted and the discovered patterns were analysed for their interestingness. This chapter provides details of the results obtained for each experiment. The results of the first experiment are discussed in detail. For the rest of the experiments, only the significant results are discussed. A table is presented for each experiment to show the significance of the results obtained.

4.1 Experiment1: AusVsOutsideAus2001

As mentioned in Section 3.3, the first experiment was conducted to compare access patterns between visitors from within Australia and visitors from outside Australia using the web access log file "access2001" as shown in Table 3.1.

Table 4.1 summarizes the experimental work done and results obtained for this experiment. The last column shows whether the results obtained through the pattern discovery process are significant or not. All the results obtained are discussed in detail in their respective sections.

4.1.1 Classification

It was decided that the classification accuracy needs to exceed some threshold before the results can be considered to be significant. An accuracy of 70 % was chosen as this threshold as high accuracy can not be expected because of the known inaccuracies involved in the preprocessing work as discussed in Section 3.2. An accuracy of 50 % on a two class problem can be achieved by guessing. An accuracy of 70 % is a considerable improvement on guessing. The problem of *overfitting* was solved by rerunning the algorithm with different parameter values as described in the literature survey of classification technique in Section 2.1.1.

1. **First3-Last2**
   The classification accuracy on the First3-Last2 feature set did not reach 70 % and

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2001	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Attribute Selection	access2001	First5-Last5	YES
		First3-Last2	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.1: *Experiment1: AusVsOutsideAus2001 - Summary of results*

hence pattern discovery using this feature set did not provide any meaningful patterns of usage.

2. **First5-Last5**
   Pattern discovery using the First5-Last5 feature set did not provide any meaningful patterns of usage for the same reason as the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   Pattern discovery using this feature set discovered two interesting patterns of usage. The root page of the RMIT computer science website was selected as the most significant page by both the 1R and the J48 algorithms. Sample outputs of the 1R and the J48 algorithms are shown in Figure 4.1 and Figure 4.2 respectively. The outputs can be interpreted as explained in Section 2.1.3. The output of the 1R algorithm shown in Figure 4.1 can be interpreted as: if visitors visit the root page then they are from within Australia and if they do not then they are from outside Australia.

   People from within Australia mostly visit the root page first. This result is probably dominated by students of RMIT university who know the home page.

   Results obtained through the J48 algorithm as shown in Figure 4.2 were also in accordance with the results produced by the 1R algorithm. The decision tree produced by the J48 algorithm also revealed one more pattern. As can be seen from the decision

```
/:
   T        -- Aus
   F        -- NotAus
(3226/4591 instances correct)
Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances    3226  70.2755 %
Incorrectly Classified Instances  1365  29.7245 %
Total Number of Instances         4591
```

Figure 4.1: *A Sample Output of the WEKA 1R Program in Experiment1*

```
Number of Leaves : 9
Size of the tree : 17
Time taken to build model: 4.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances     3274     71.3134 %
Incorrectly Classified Instances   1317     28.6866 %
Total Number of Instances          4591
```

Figure 4.2: *A Sample Output of the WEKA J48 Program in experiment1*

tree in Figure 4.3 that, if the visitors visit the root page and also visit pages related to postgraduate study then they are mostly from outside Australia.

The decision tree also shows a pattern of access to a specific course page. This result is obtained because there may be some activities running for this course during the time period of this log file and hence, many students have accessed this page. This may be a temporary situation for this log file and this pattern is not expected in log file for another period of time.

4. **20-Most-Frequent-Time**
The results produced by both the 1R and the J48 algorithms selected root page as the most significant attribute. This result is consistent with the results obtained in Section 4.1.1 using the 20-Most-Frequent-TF feature set. The decision tree shown in Figure 4.4 indicates that if visitors visit the root page and also visit the pages related to post graduate study then they are mostly from outside Australia. This result is

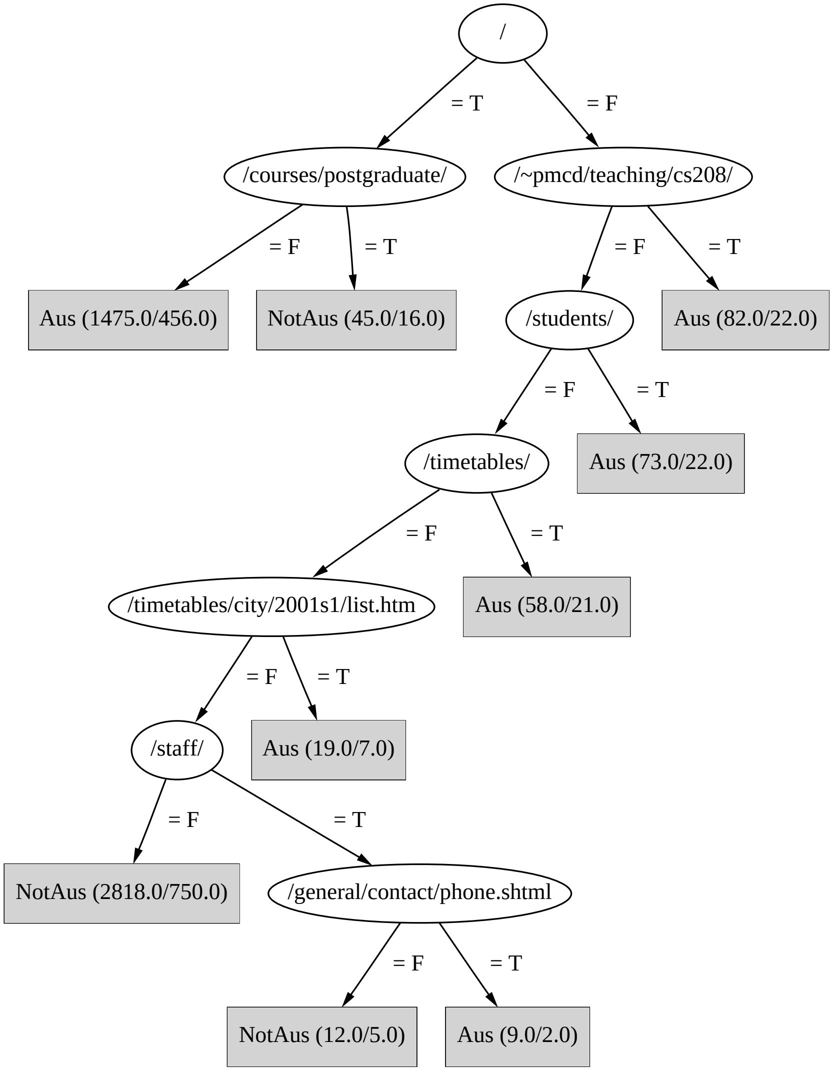


Figure 4.3: *Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment1*

consistent with the result obtained using the 20-Most-Frequent-TF feature set.

It can also be seen from Figure 4.4 that if visitors do not visit the root page and they visit the "/students" page then they are from within Australia. This result is probably dominated by the students of RMIT university who know "/students" page and hence, visit this page directly.

4.1.2 Association Rules

The association rules generated by the Apriori algorithm were analysed to find their interestingness.

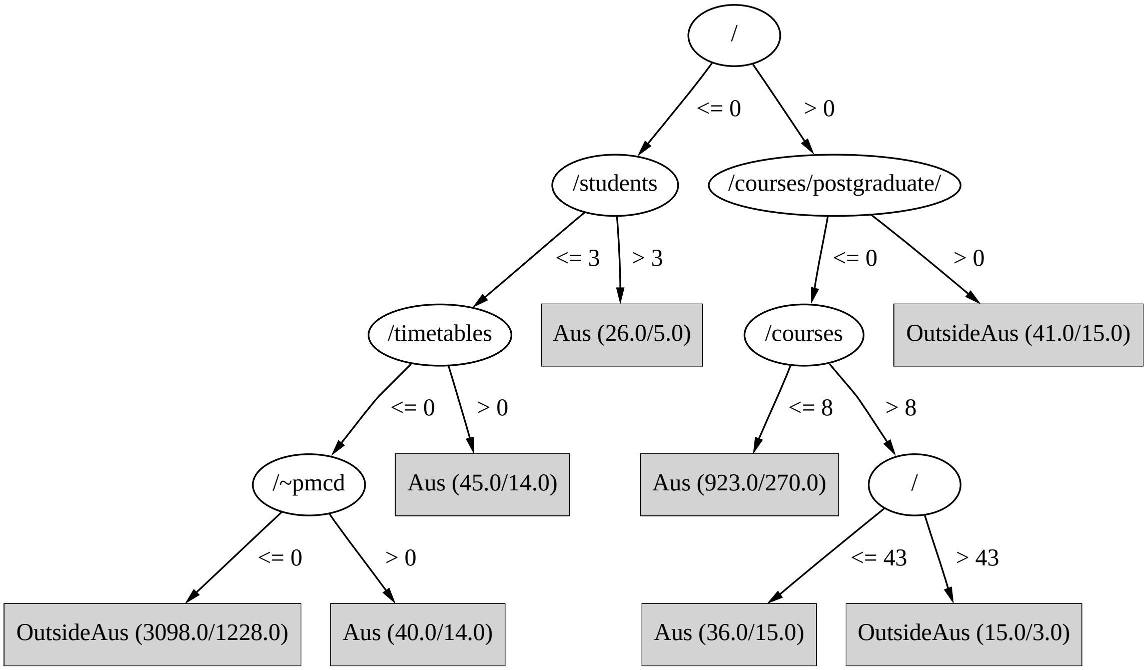


Figure 4.4: *Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-Time Feature Set in Experiment1*

1. **First3-Last2**
   The Apriori algorithm found only 2 rules. These are shown in Figure 4.5. The association rules can be interpreted as explained in Section 2.1.3. The first rule can be interpreted as: if the first page visited is the root page then the visitor is from inside Australia. The second rule can be interpreted as: if visitors are from inside Australia then the first page they visited was the root page. Both of these rules indicate that if visitors are from within Australia then they visited the root page. This result is consistent with the classification result observed in Section 4.1.1.

2. **First5-Last5**
   The Apriori algorithm found one interesting rule which was the same as that using the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   The results produced were not interesting due to the fact that the number of visited pages was very much less than number of pages which were not visited in a transaction. Hence, the Apriori algorithm generated rules in which all the attribute values had "F" value. A sample output of the association rules obtained is shown in Figure 4.6.

   Hence, it was decided to consider only the instances in which at least one attribute value is "T". However, there was no significant difference in the rules generated. It was also attempted to consider only those instances in which 2 or more attributes had the "T" value. But there were not many instances that satisfied this condition. Hence, it was concluded that the Apriori algorithm with this feature set is not likely

```
Apriori
=======

Minimum support: 0.05
Minimum metric (confidence): 0.4
Number of cycles performed: 17

Generated sets of large itemsets:
Size of set of large itemsets L(1): 3
Size of set of large itemsets L(2): 1

Best rules found:

1. link0=/ 339 ==> Location=Aus 301    conf:(0.89)
2. Location=Aus 531 ==> link0=/ 301    conf:(0.57)
```

Figure 4.5: *Partial Output of the WEKA Apriori Program Using the First3-Last2 Feature Set in Experiment1*

```
/~caspar/turbo_mazda_323.htm=F 911 ==> /~shyam/cs492meta.html=F
/~caspar/turbo_mazda_323.htm=F 911 ==> /~winikoff/palm/dev.html=F
/conf/doa/2001/=F 907 ==> /~winikoff/palm/dev.html=F /~shyam/cs492meta.html=F
/~winikoff/palm/dev.html=F /conf/doa/2001/=F 907 ==> /~shyam/cs492meta.html=F
/~shyam/cs492meta.html=F /conf/doa/2001/=F 907 ==> /~winikoff/palm/dev.html=F
```

Figure 4.6: *Partial Output of the WEKA Apriori Program Using the 20-Most-Frequent-TF Feature Set in Experiment1*

to produce interesting association rules. Therefore, this feature set was not used for association finding in further experiments.

4. **20-Most-Frequent-Time**
The Apriori algorithm does not work with numeric attributes. Hence, this feature set was not used for association rule mining.

4.1.3 Clustering

The clusters formed by the EM algorithm were analysed to find whether there are any groups of visitors who exhibit similar web browsing behaviour.

1. **First3-Last2**
The clusters formed did not provide any interesting information about access patterns.

The clusters generated by the EM algorithm had very small numbers of associated instances and allocation of instances to clusters appeared to be arbitrary. This is due to the fact that each attribute has a very large number of values. Hence, this feature set was not found suitable for the use of the clustering technique and was not used for further experiments.

2. **First5-Last5**
   The clusters formed did not provide any interesting information about access patterns. This feature set was not used for the clustering technique in further experiments for the same reason as the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   The output of the EM algorithm showed two interesting clusters. The output of the EM algorithm can be interpreted as explained in Section 2.1.3. One cluster was formed of visitors who browsed pages related to staff members and their contact information. A sample output of this cluster is shown in Figure 4.7. As can be seen from this figure, this cluster of visitors tend to visit both pages "/staff/" and "/general/contact/phone.shtml" pages since their *Counts* for the "T" value is much higher than their *Counts* for "F" value. The output also shows that these 2 are the only pages that this cluster of visitors have mostly visited. Hence, it can be deduced that some visitors visit the RMIT computer science website to access information about staff members and their contact details.

   The other interesting cluster formed was of visitors who browsed information about postgraduate courses. An annotated partial output of this cluster is shown in Figure 4.8. The output can be interpreted in the same manner as the output of Figure 4.7. This cluster is consistent with the results obtained using the 20-Most-Frequent-TF feature set in Section 4.1.1.

4. **20-Most-Frequent-Time**
   The WEKA EM program using this feature set did not produce any interesting cluster. As described in Section 3.2.3, the attribute values in this feature set represent the time period a visitor has spent during the visit, the majority of the attribute values in an instance were "0". This presented problems for the probability density estimation component of the EM algorithm and no meaningful clusters were produced. Therefore, this feature set was not found suitable for the use with the clustering technique and further pattern discovery tasks using the clustering technique with this feature set were not performed.

4.1.4 Attribute Selection

The attributes selected by the process of Attribute Selection using "cfssetEval" attribute evaluator together with "BestFirst" search method were analysed for their interestingness.

1. **First3-Last2**
   The results produced using this feature set showed "link0" as the most significant

```
Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: log
Instances: 4591
                              (F)   (T)
Attribute: /
Discrete Estimator. Counts =  73.29 12.71 (Total = 86)
Attribute: /students/
Discrete Estimator. Counts =  83.51 2.49  (Total = 86)
Attribute: /staff/
Discrete Estimator. Counts =  7.91 78.1   (Total = 86)
Attribute: /courses/
Discrete Estimator. Counts =  78.53 7.47  (Total = 86)
Attribute: /timetables/
Discrete Estimator. Counts =  79.28 6.72  (Total = 86)
Attribute: /employment/
Discrete Estimator. Counts =  84.08 1.92  (Total = 86)
Attribute: /general/contact/phone.shtml
Discrete Estimator. Counts =  5.97 80.03  (Total = 86)
Attribute: /courses/postgraduate/
Discrete Estimator. Counts =  83.72 2.28  (Total = 86)
```

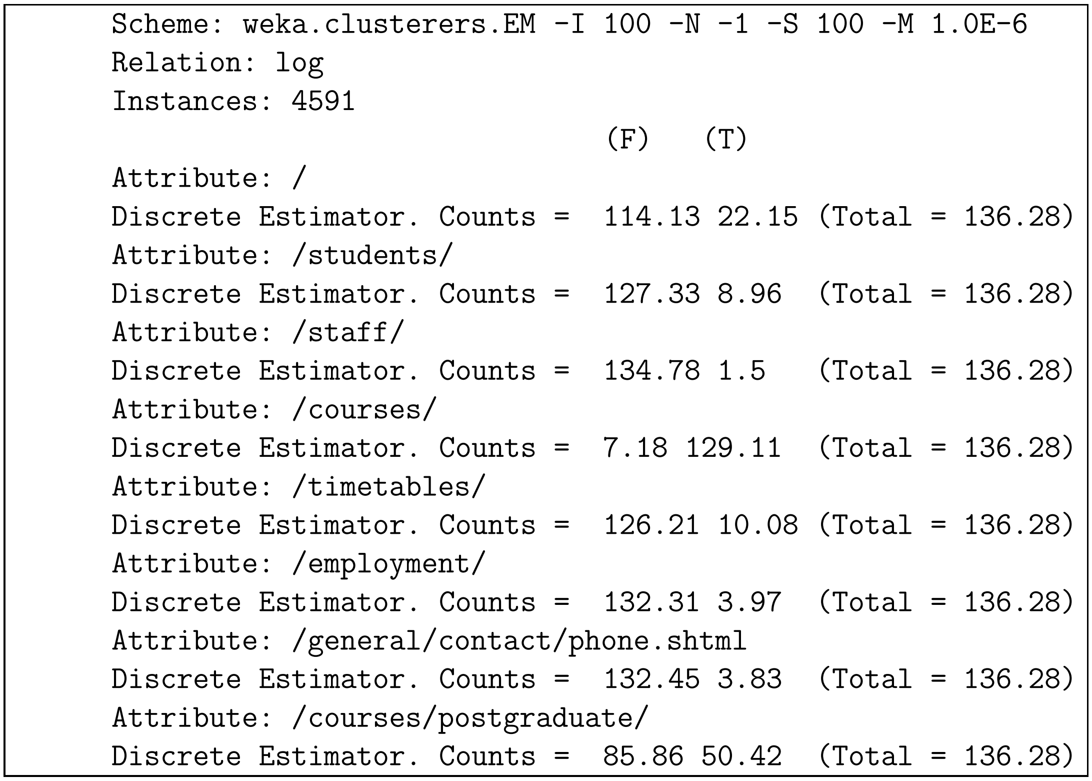
Figure 4.7: *Annotated Partial Output of Cluster-1 by the WEKA EM Program Using the 20-Most-Frequent-TF Feature Set in Experiment1*

attribute. "link0" was the only attribute selected. As explained in Section 3.2.3, "link0" corresponds to the first page visited in a transaction. This result indicates that the first page a visitor visits is the most relevant and discriminating.

2. **First5-Last5**
   Attribute selection using this feature set selected "link0" as the most significant attribute. The second and third most significant attributes selected were "link2last" and "linklast" respectively. "link2last" is the second last page and "linklast" is the last page visited in a transaction.

3. **20-Most-Frequent-TF**
   Attribute selection using this feature set indicated the root as the most singificant attribute. The root page was the only attribute selected. This result is in accordance with the results obtained in Section 4.1.1 using the 20-Most-Frequent-TF feature set.

4. **20-Most-Frequent-Time**
   The output using this feature set indicated the root as the most singificant attribute. The root page was the only attribute selected. This result is the same as that using the 20-Most-Frequent-TF feature set.



```
Scheme: weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
Relation: log
Instances: 4591
                              (F)   (T)
Attribute: /
Discrete Estimator. Counts =  114.13 22.15 (Total = 136.28)
Attribute: /students/
Discrete Estimator. Counts =  127.33 8.96  (Total = 136.28)
Attribute: /staff/
Discrete Estimator. Counts =  134.78 1.5   (Total = 136.28)
Attribute: /courses/
Discrete Estimator. Counts =  7.18 129.11  (Total = 136.28)
Attribute: /timetables/
Discrete Estimator. Counts =  126.21 10.08 (Total = 136.28)
Attribute: /employment/
Discrete Estimator. Counts =  132.31 3.97  (Total = 136.28)
Attribute: /general/contact/phone.shtml
Discrete Estimator. Counts =  132.45 3.83  (Total = 136.28)
Attribute: /courses/postgraduate/
Discrete Estimator. Counts =  85.86 50.42  (Total = 136.28)
```

Figure 4.8: *Annotated Partial Output of Cluster-2 by the WEKA EM Program Using the 20-Most-Frequent-TF Feature Set in Experiment1*

4.2 Experiment2: AusVsOutsideAus2003

As mentioned in Section 3.7, the second experiment was conducted to compare access patterns between visitors from within Australia and visitors from outside Australia using the web access log file “access2003” as shown in Table 3.1.

Table 4.2 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

4.2.1 Classification

As stated in Section 4.1.1, the results were analysed only if the classification accuracies were above 70 %.

1. **20-Most-Frequent-TF**
   The output of the J48 and the 1R algorithms indicated the root as the most significant attribute. The partial decision tree in Figure 4.9 shows that if visitors visit the root page at all then they are from within Australia. This result is consistent with the result obtained in Section 4.1.1 using the 20-Most-Frequent-TF feature set.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2003	First3-Last2	NO
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO
Attribute Selection	access2003	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.2: *Experiment2: AusVsOutsideAus2003 - Summary of results*

The decision tree also shows that if visitors do not visit the root page and visit the "/timetables" page then they are from within Australia. This may be because visitors from within Australia tend to look at the timetables of various subjects of their interest. It may be possible that this result is dominated by students of RMIT university who know the URL of the "/timetables" page and hence visit this page directly. This conjecture is tested in experiments 3 and 4.

The decision tree also shows that if visitors do not visit the root page and they also do not visit the "/timetables" page but visit "/students" page then they are from within Australia. This result is obtained may be because the students of RMIT university visit this page frequently for various activities such as checking the emails.

2. **20-Most-Frequent-Time**
The output of the 1R algorithm indicated "/courses" as the most significant attribute. The partial decision tree in Figure 4.10 shows the root as the most significant attribute. The decision tree also shows that if visitors do not visit the root page and visit the "/timetables" page then they are from within Australia. The decision tree also showed that if visitors do not visit the root page and they also do not visit the "/timetables" page but they visit "/students" page then they are from within Australia. This outcome is consistent with the result obtained using the 20-Most-Frequent-TF feature set.

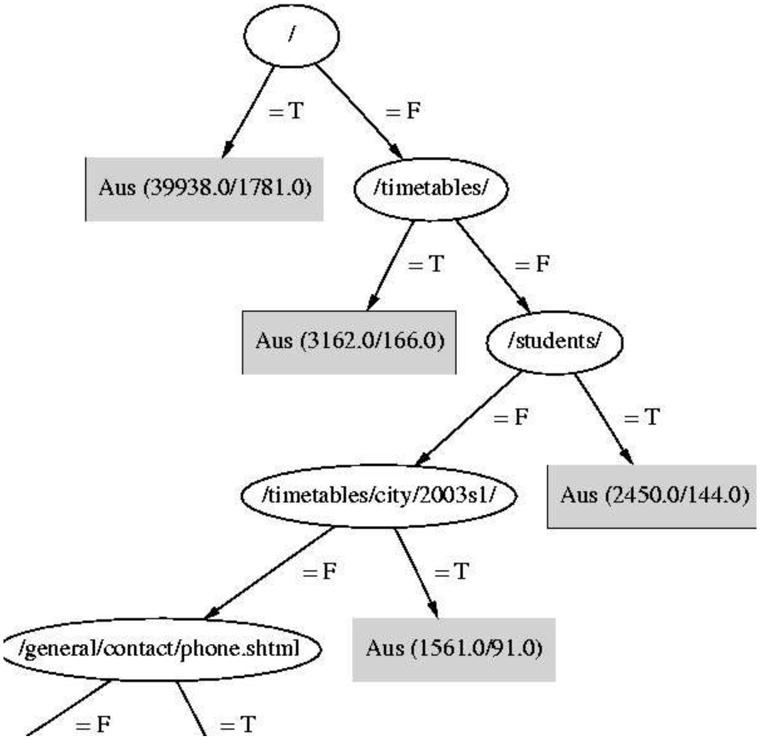


Figure 4.9: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment2*

4.2.2 Association Rules

The association rules generated by the Apriori algorithm were analysed to find their interestingness.

1. **First5-Last5**
   The Apriori algorithm found one interesting rule which can be interpreted as if visitors visit the "/timetable" page then they are from within Australia. This result is consistent with the results obtained using the 20-Most-Frequent-TF feature set in Section 4.2.1.

4.2.3 Attribute Selection

The attributes selected by the process of Attribute Selection using "cfssetEval" attribute evaluator together with "BestFirst" search method were analysed for their interestingness.

1. **First3-Last2**
   Attribute selection using this feature set indicated "link0", which is the first page

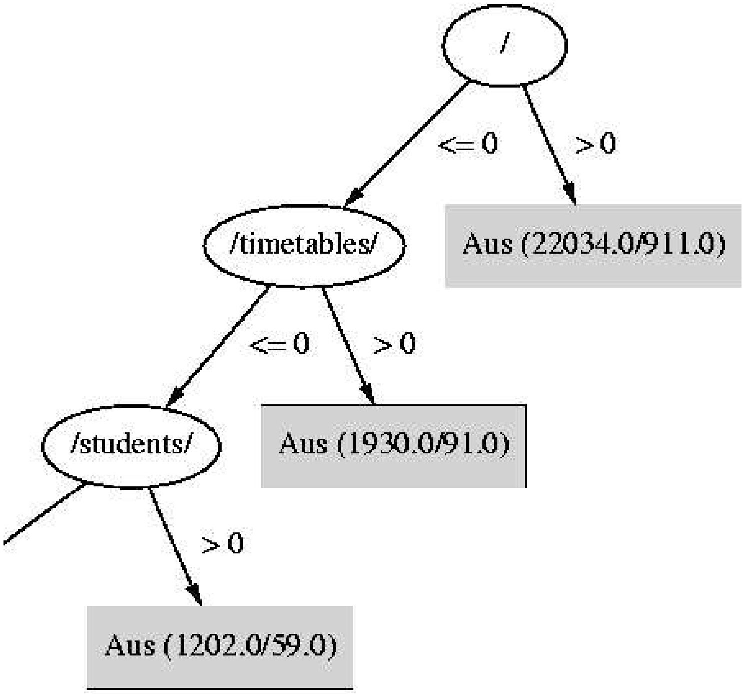


Figure 4.10: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-Time Feature Set in Experiment2*

visited in a transaction, as the most significant attribute. This result is consistent with the results obtained in Section 4.1.4 using the First3-Last2 feature set.

2. **First5-Last5**
   Attribute selection using this feature set indicated "link0" as the most significant attribute. This result is consistent with the result obtained using the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   Attribute selection using this feature set indicated the root page as the most significant attribute. This result is consistent with the result obtained in Section 4.1.4 using the 20-Most-Frequent-TF feature set.

4. **20-Most-Frequent-Time**
   Attribute selection using this feature set indicated the root page,"/students" and "/timetables" as the most significant attributes in order. This result is consistent with the result shown by decision trees using the 20-Most-Frequent-TF and the 20-Most-Frequent-Time feature sets in Section 4.2.1.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2001	First3-Last2	NO
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Attribute Selection	access2001	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.3: *Experiment3: RMITVsOutsideRMIT2001 - Summary of results*

4.3 Experiment3: RMITVsOutsideRMIT2001

As mentioned in Section 3.4, the third experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university using the web access log file “access2001” as shown in Table 3.1.

Table 4.3 summarizes the experimental work done and results obtained for this experiment. Only significant results obtained are discussed in their respective sections.

4.3.1 Classification

1. **20-Most-Frequent-TF**
   The results obtained using this feature set are consistent with the results obtained in Section 4.1.1 using the 20-Most-Frequent-TF feature set. The partial outputs of the 1R and the J48 algorithms are shown in Figure 4.12 and Figre 4.11 respectively. The decision tree generated by the J48 algorithm as shown in Figure 4.13 indicated that if visitors visit the root page and they also visit pages about postgraduate studies then they are from outside RMIT. The root attribute was selected by both the 1R and the J48 algorithms as the most significant attribute.

2. **20-Most-Frequent-Time**
   The results obtained through this feature set indicate that if visitors do not visit the root page then they are from outside RMIT university. This result is consistent with

```
Number of Leaves : 5
Size of the tree : 9
Time taken to build model: 2.95 seconds

     === Stratified cross-validation ===
     === Summary ===
Correctly Classified Instances   3830  83.4241 %
Incorrectly Classified Instances  761  16.5759 %
```

Figure 4.11: *Partial Output of the WEKA J48 Program Using the 20-Most-Frequent-TF feature set in Experiment3*

```
Test mode:    user supplied test set: 4591 instances
=== Classifier model (full training set) ===
/ :
  T  -- NotRMIT
  F  -- NotRMIT

(3844/4591 instances correct)
Time taken to build model: 0.2 seconds
=== Evaluation on test set ===
=== Summary ===
Correctly Classified Instances        3844   83.729 %
Incorrectly Classified Instances       747   16.271 %
```

Figure 4.12: *Partial Output of the WEKA 1R Program Using the 20-Most-Frequent-Time feature set in Experiment3*

the result obtained in Section 4.1.1 using the 20-Most-Frequent-TF feature set.

4.3.2 Association Rules

1. **First5-Last5**
   The Apriori algorithm found one interesting rule. The rule indicated that if the first page the visitors visit is the root page then they are from outside RMIT univeristy. This rule is consistent with the results obtained in Section 4.2.1 using the 20-Most-Frequent-TF feature set.

4.3.3 Clustering

1. **20-Most-Frequent-TF**
   The output of the EM algorithm showed two interesting clusters. These clusters were

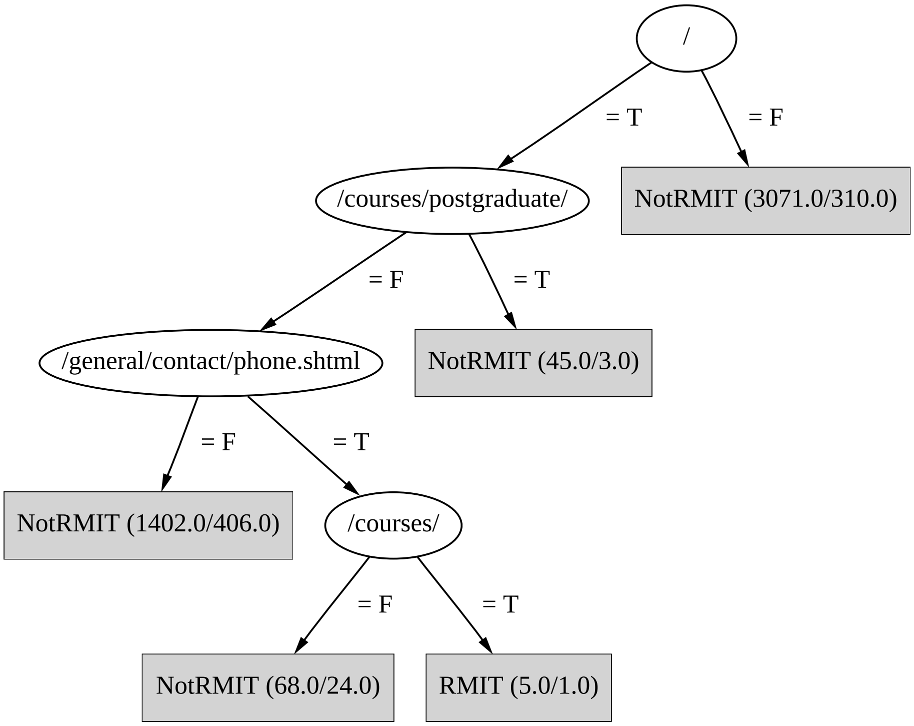


Figure 4.13: *Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment3*

same as those found in Section 4.1.3 using the 20-Most-Frequent-TF feature set. A further pattern discovery experiment was conducted including the class variable whose value in this experiment can be either "RMIT" or "NotRMIT". The output showed that the two clusters were formed mostly of visitors from outside RMIT university.

4.3.4 Attribute Selection

The "cfssetEval" attribute evaluator together with "BestFirst" search method produced the same results as those described in Section 4.1.4 for all feature sets.

4.4 Experiment4: RMITVsOutsideRMIT2003

As mentioned in Section 3.8, the fourth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university using the web access log file "access2003" as shown in Table 3.1.

Table 4.4 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Association Rules	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
Clustering	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Attribute Selection	access2003	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.4: *Experiment4: RMITVsOutsideRMIT2003 - Summary of results*

4.4.1 Classification

1. **20-Most-Frequent-TF**
   The results of the J48 and the 1R algorithms indicated the root as the most significant attribute. The partial decision tree shown in Figure 4.14 indicates that if visitors visit the root page and also visit "/industrycareers" then they are from outside RMIT university. This result suggests that some visitors tend to look at the industry collaborations of RMIT university and career related information. These visitors may be prospective students. The decision tree also shows that if the visitors visit the pages related to contact information and also visit the "/students" page then they are from outside RMIT university.

4.4.2 Clustering

1. **20-Most-Frequent-TF**
   The output of the EM algorithm showed one interesting cluster. The visitors who belong to this cluster tend to access staff contact information. This cluster is the same as a cluster discovered in experiment1 using the 20-Most-Frequent-TF feature set in Section 4.1.3.

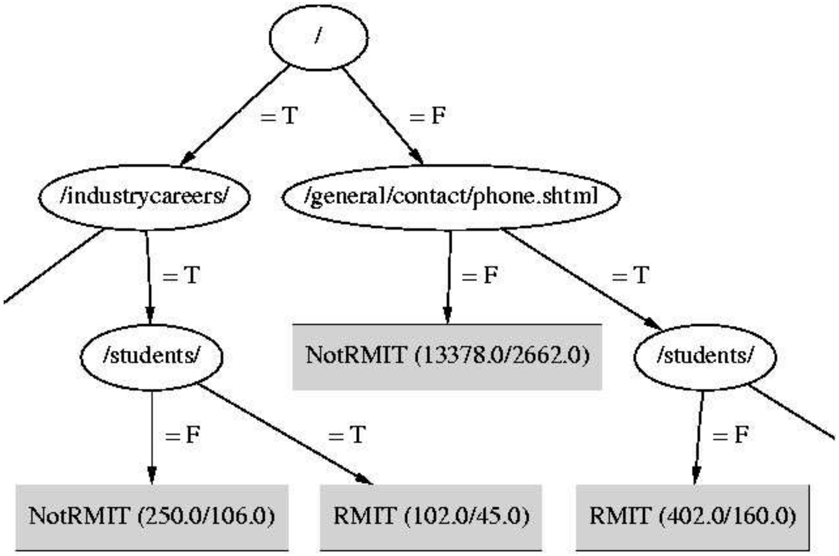


Figure 4.14: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-Time Feature Set in Experiment4*

4.4.3 Attribute Selection

1. **First3-Last2**
   Attribute selection using this feature set indicated "link0" as the most significant attribute. This result is consistent with the results obtained in Section 4.1.4 using the First3-Last2 feature set.

2. **First5-Last5**
   Attribute selection using this feature set indicated "link0" as the most significant attribute. This result is consistent with that obtained using the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   Attribute selection using this feature set indicated the root page as the most significant attribute. This result is consistent with the results obtained in Section 4.1.4 using the 20-Most-Frequent-TF feature set.

4. **20-Most-Frequent-Time**
   Attribute selection using this feature set indicated the root page as the most significant attribute. This result is consistent with that obtained using the 20-Most-Frequent-TF feature set.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
Clustering	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO
Attribute Selection	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.5: *Experiment5: RMITVsOutsideRMITWithinAus2001 - Summary of results*

4.5 Experiment5: RMITVsOutsideRMITWithinAus2001

As mentioned in Section 3.7, the fifth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university but within Australia using the web access log file "access2001" as shown in Table 3.1.

Table 4.5 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

4.5.1 Classification

1. **20-Most-Frequent-TF**
   Results obtained indicated that if the visitors do not visit the root page then they are from outside RMIT university. Further experiment was conducted to determine whether there is any difference in the access patterns if the root page is the first page visited in a transaction. The difference is that if a visitor does visit the root page but the root is not the first page that was visited then the visitor is from within RMIT university. This is mostly due to the fact that students within RMIT university know the URLs of specific subjects and lecturers. Hence, they do not visit the root page; rather they directly visit more specific pages.

2. **20-Most-Frequent-Time**
   The results obtained indicate that if the visitors do not visit the root page then they

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
Clustering	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Attribute Selection	access2003	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES

Table 4.6: *Experiment6: RMITVsOutsideRMITWithinAus2003 - Summary of results*

are from outside RMIT university but from within Australia. This result is consistent with the result obtained in Section 4.3.1 using the 20-Most-Frequent-Time feature set.

4.5.2 Attribute Selection

1. **20-Most-Frequent-TF**
   The output of the pattern discovery process using this feature set showed the root attribute and the "/courses/postgraduate" as the two most significant attributes.

2. **20-Most-Frequent-Time**
   The output of pattern discovery process using this feature set showed "/courses/postgraduate" as the most significant attribute.

4.6 Experiment6: RMITVsOutsideRMITWithinAus2003

As mentioned in Section 3.8, the sixth experiment was conducted to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university but within Australia using the web access log file "access2003" as shown in Table 3.1.

Table 4.6 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

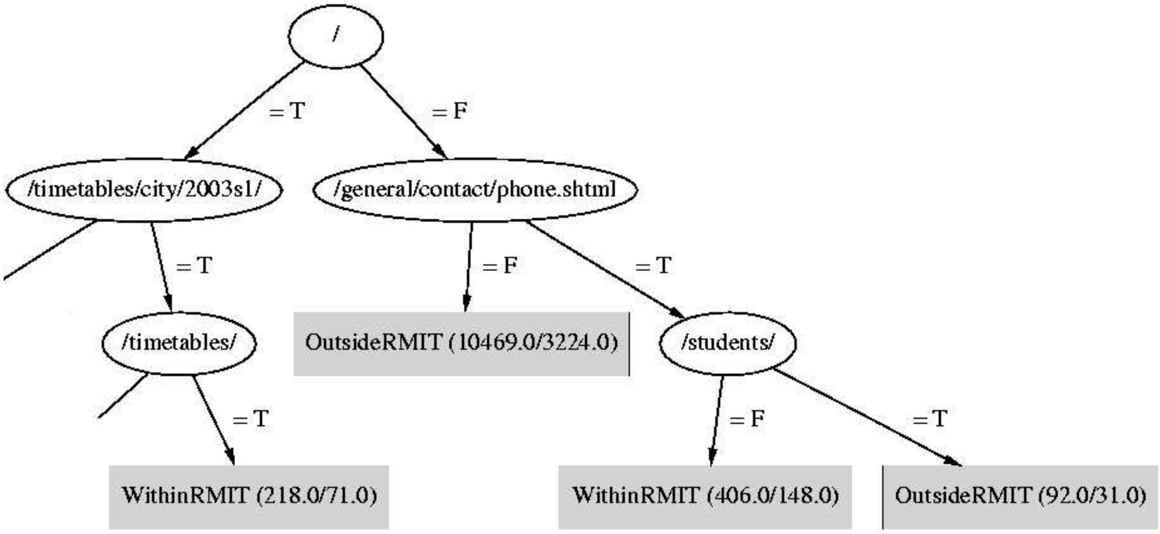


Figure 4.15: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment6*

4.6.1 Classification

1. **20-Most-Frequent-TF**
   The results of the J48 and the 1R algorithms indicated the root as the most significant attribute. The partial decision tree in Figure 4.15 shows that if the visitors visit the pages related to contact information and also visit the "/students" page then they are from outside RMIT university but within Australia. This result is consistent with the result described in Section 4.4.1 for the 20-Most-Frequent-TF feature set.

2. **20-Most-Frequent-Time**
   The partial decision tree of the J48 algorithm in Figure 4.16 shows that if the visitors do not visit the root page but if they either visit "/courses" page or "/timetables" page and spend some time at either of these pages then they are from outside RMIT university but within Australia. The decision tree also shows that visitors from within RMIT university spend less time compared to visitors from outside RMIT university but within Australia. It can be conjectured that visitors from within RMIT university know the structure of the website very well and hence navigate quickly.

4.6.2 Clustering

1. **20-Most-Frequent-TF**
   The output of the EM algorithm showed one interesting cluster. Figure 4.17 shows an annotated partial output of this cluster produced by the WEKA EM program.

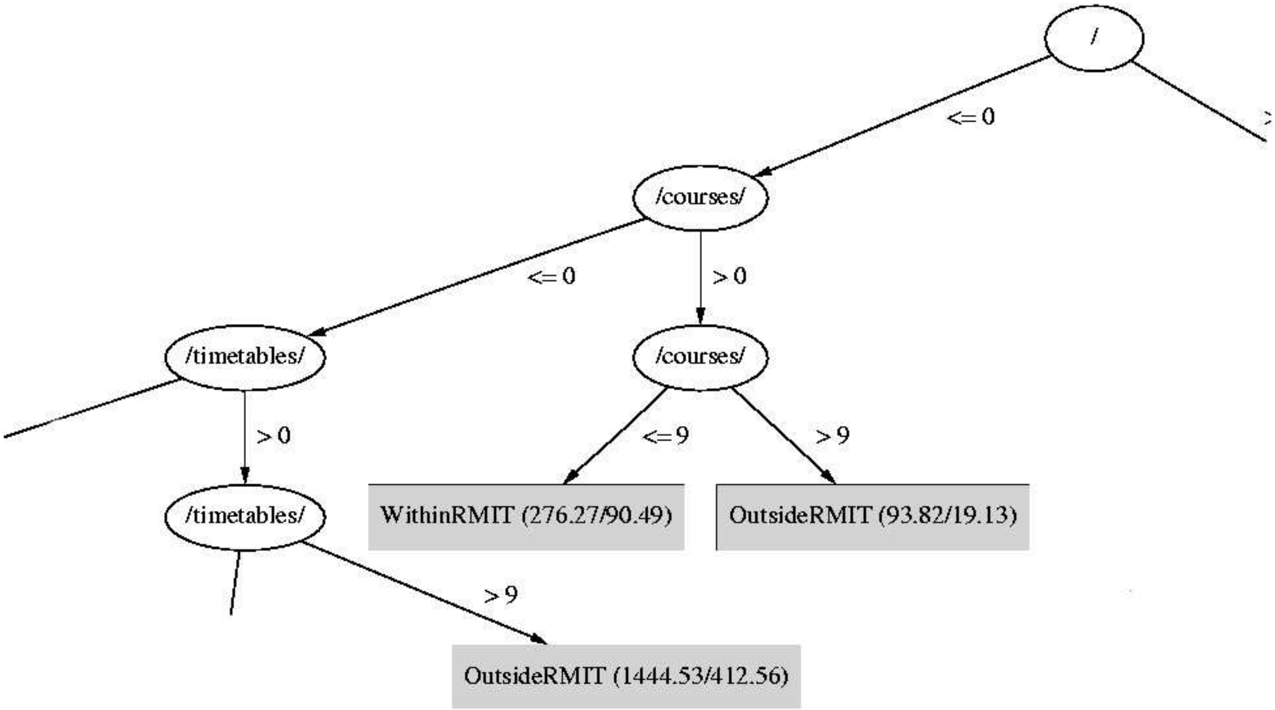

/
<=0
/courses/
<=0
>0
/timetables/
/courses/
>0
<=9
>9
/timetables/
WithinRMIT (276.27/90.49)
OutsideRMIT (93.82/19.13)
>9
OutsideRMIT (1444.53/412.56)

Figure 4.16: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-Time Feature Set in Experiment6*

This cluster of visitors tend to access industry and career related information and employment prospects. These visitors are mostly from outside RMIT university but within Australia. This result is consistent with the one obtained using the 20-Most-Frequent-TF feature set in Section 4.4.1.

4.6.3 Attribute Selection

1. **First3-Last2**
   Attribute selection using this feature set indicated "link0" as the most significant attribute. This result is consistent with the results obtained in Section 4.1.4 using the First3-Last2 feature set.

2. **First5-Last5**
   Attribute selection using this feature set indicated "link0" and "linklast" as the most significant attributes.

3. **20-Most-Frequent-TF**
   Attribute selection using this feature set indicated the root page as the most significant attribute. The root was the only attribute selected. This result is consistent with the results obtained in Section 4.1.4 using the 20-Most-Frequent-TF feature set.

```
Scheme:       weka.clusterers.EM -I 100 -N -1 -S 100 -M 1.0E-6
                        (WithinRMIT) (OutsideRMIT-WithinAus)
Attribute: /
Discrete Estimator. Counts =   10.58      268.06 (Total = 278.64)
Attribute: /employment/
Discrete Estimator. Counts =    4.30      274.34 (Total = 278.64)
Attribute: /general/contact/phone.shtml
Discrete Estimator. Counts =  276.92        1.71 (Total = 278.64)
Attribute: /search/
Discrete Estimator. Counts =  277.63        1.01 (Total = 278.64)
Attribute: /general/
Discrete Estimator. Counts =  276.54        2.10 (Total = 278.64)
Attribute: /industrycareers/
Discrete Estimator. Counts =    1.01      277.63 (Total = 278.64)
Attribute: Location
Discrete Estimator. Counts =   56.29      222.34 (Total = 278.64)
```

Figure 4.17: *Annotated Partial Output of a Cluster Produced by the WEKA EM Program Using the 20-Most-Frequent-TF Feature Set in Experiment6*

4. **20-Most-Frequent-Time**
   Attribute selection using this feature set indicated the root page as the most significant attribute. The root was the only attribute selected.

4.7 Experiment7: EduVsOthers2001

As mentioned in Section 3.9, the seventh experiment was conducted to compare access patterns between visitors from educational organizations and non-educational institutions using the web access log file "access2001" as shown in Table 3.1.

Table 4.7 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

4.7.1 Association Rules

1. **First3-Last2**
   One interesting rule was found from the output of the Apriori algorithm. The rule was interpreted as: if the first page that visitors visit is the root page then they are from a place other than an educational institution.

2. **First5-Last5**
   The output of the Apriori algorithm showed one interesting rule, which was the same as that using the First3-Last2 feature set.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO
Association Rules	access2001	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO
Attribute Selection	access2001	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	NO

Table 4.7: *Experiment7: EduVsOthers2001 - Summary of results*

4.7.2 Clustering

1. **20-Most-Frequent-TF**
   The analysis of the output revealed one interesting cluster. This cluster of visitors tend to browse information about the staff members of the RMIT university and their contact information and are mostly from educational institutions.

4.7.3 Attribute Selection

1. **20-Most-Frequent-TF**
   The output of the attribute selection process showed root and "/staff" as the two most significant attributes. The selection of "/staff" as the second most significant attribute provides more evidence to the result obtained in Section 4.7.1 using the 20-Most-Frequent-TF feature set.

4.8 Experiment8: EduVsOthers2003

As mentioned in Section 3.10, the eighth experiment was conducted to compare access patterns between visitors from educational organizations and other visitors using the web access log file "access2003" as shown in Table 3.1.

Table 4.8 summarizes the experimental work done and results obtained for this experiment. The significant results obtained are discussed in their respective sections.

Data Mining Technique	Web Access Log File	Feature Set Used	Significant Patterns Discovered
Classification	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	YES
		20-Most-Frequent-Time	YES
Association Rules	access2003	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	NO
Clustering	access2003	First3-Last2	NO
		First5-Last5	NO
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO
Attribute Selection	access2003	First3-Last2	YES
		First5-Last5	YES
		20-Most-Frequent-TF	NO
		20-Most-Frequent-Time	NO

Table 4.8: *Experiment8: EduVsOthers2003 - Summary of results*

4.8.1 Classification

1. **20-Most-Frequent-TF**
   The partial decision tree in Figure 4.18 shows that visitors from educational institutions look for pages related to contact information. At the same time they also visit the root page and the "/students" page. The decision tree also shows that if the visitors do not visit "/students" and if they visit "/timetables" page then they are from a place other than an educational institution. It can also be seen from the decision tree that other visitors tend to access career and industry related information.

2. **20-Most-Frequent-Time**
   The partial decision tree in Figure 4.19 shows that if visitors spend some time on the pages related to the contact information then they are from an educational institution. This result is in accordance with the result obtained using the 20-Most-Frequent-TF feature set. It can also be seen from the decision tree that other visitors tend to access career and industry ralated information. This result is the same as that obtained using the 20-Most-Frequent-TF feature set.

4.8.2 Association Rules

1. **First3-Last2**
   One interesting rule was found from the output of the Apriori algorithm. The rule was interpreted as: if the first page that the visitors visit is the root page and second page

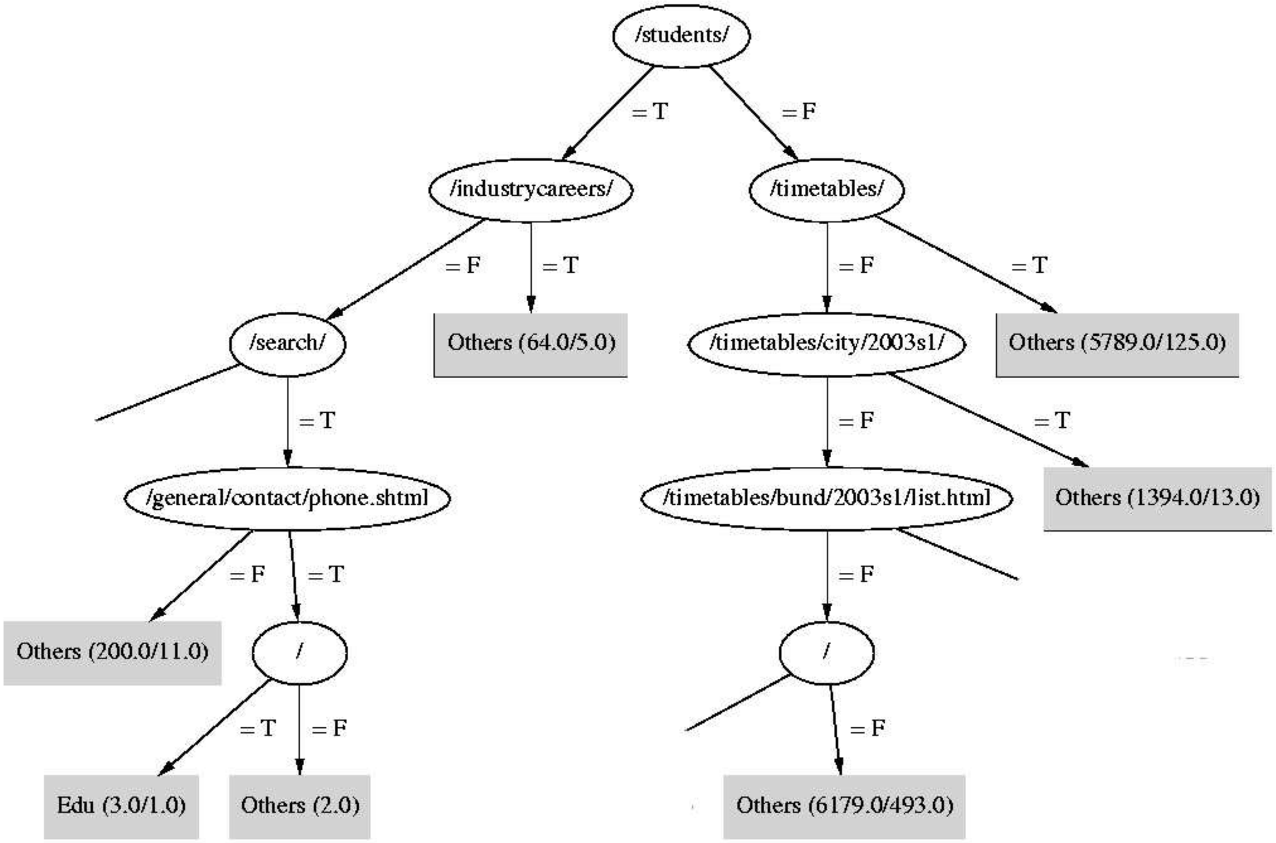


Figure 4.18: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-TF Feature Set in Experiment8*

they visit is "/timetables" page then they are from a place other than an educational institution. This result is consistent with the result obtained in Section 4.8.1 using the 20-Most-Frequent-TF feature set.

2. **First5-Last5**
   The output using this feature set found one interesting rule which was the same as that using the First3-Last2 feature set.

4.8.3 Attribute Selection

1. **First3-Last2**
   The output of attribute selection showed "link0" and "link1", the first and the second page visited respectively, as the most significant attributes.

2. **First5-Last5**

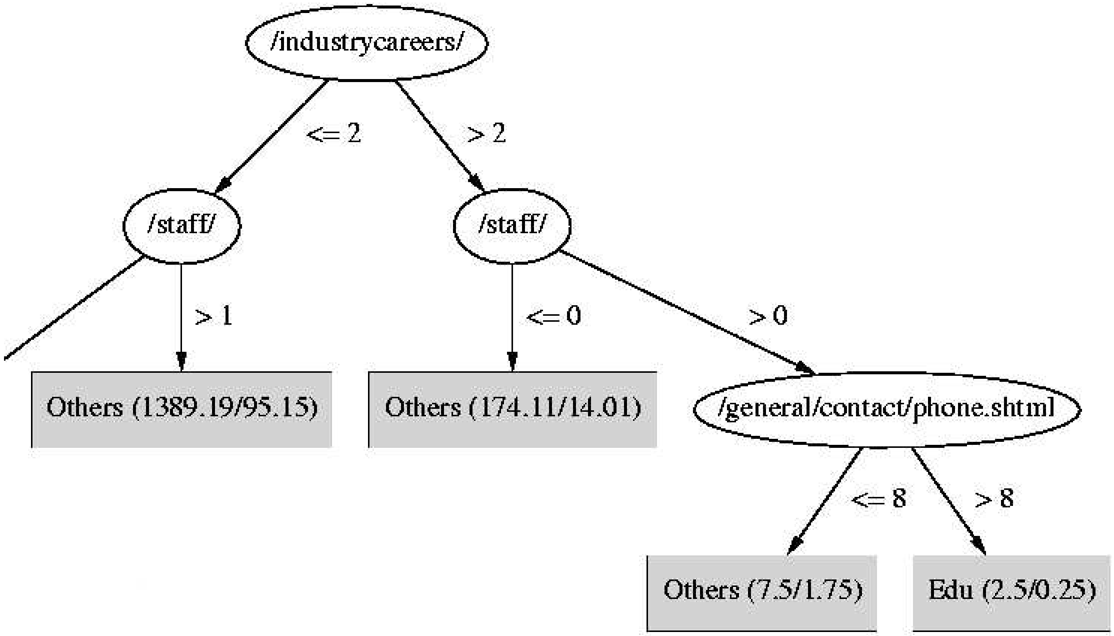


Figure 4.19: *Partial Decision Tree of the WEKA J48 Program Using the 20-Most-Frequent-Time Feature Set in Experiment8*

The attributes selected using this feature set were the same as those using the First3-Last2 feature set.

3. **20-Most-Frequent-TF**
   The attributes selected using this feature set were the root, "/students","/staff" and "/timetables". These attributes were also observed at the top in the decision trees described in Section 4.8.1 which shows that the results are consistent.

4. **20-Most-Frequent-Time**
   The first four attributes selected using this feature set are the same as those using the 20-Most-Frequent-TF feature set. The other two attributes selected were "/general/contact/phone.shtml" and "/industrycareers".

4.9 Discussion of Experimental Results

The results of experiments 1 and 2 showed the differences in acccess patterns between visitors from within Australia and visitors from outside Australia. The major difference is that if the web visitors visit the root page at any time during a transaction then they are from Australia and if they do not then they are from outside Australia. This result is most likely due to the fact that visitors from outside Australia use search engines to find relevant information. Search engines provide links to specific pages based on the query entered by the visitors. By clicking on these links, visitors go directly to the specific pages. It was also observed that visitors from within Australia start their visit with the root page. This result is probably dominated by students of RMIT university who know the home page or have it set as the default in the browser. Results of pattern discovery using different data mining techniques also indicated that visitors from outside Australia mostly access information about postgraduate courses. Hence, it can be hypothesized that they are mostly interested in postgraduate studies. It was also found that some visitors browse information about staff members and tend to find contact information probably for further information or queries.

Experiments 3 and 4 were conducted to compare access patterns between visitors from within RMIT university and visitors of outside RMIT university. The results using different data mining techniques indicated that if visitors visit the root page then they are from within RMIT university and if they do not they are from outside RMIT university. This difference in patterns of usage is consistent with that observed in experiments 1 and 2. It was observed that visitors from outside RMIT university mostly access information about postgraduate courses. This access pattern was also discovered in experiments 1 and 2. One interesting difference is that if the visitors visit the root page but if the root is not the first page visited then they are from within RMIT university. This pattern is observed probably because students of RMIT university know specific URLs of lecturers and courses and hence visit these pages directly.

One interesting pattern that differentiates access patterns is that some visitors from outside RMIT university tend to access information about the employment prospects and career and industry collaborations of the computer science department of RMIT university. It can be conjectured that prospective students are usually concerned about their career and employment prospects before doing a course at a university.

The results of experiments 5 and 6 show the differences in access patterns between visitors from within RMIT university and visitors from outside RMIT university but within Australia. The experimental results show that pages about the timetables are mostly accessed by visitors from within Australia. It can be assumed, since people from within Australia can work full time and study part time, that they tend to check the timetables for the suitability of timings of the subjects they want to study, whereas most people from outside Australia do not bother about the timings as they have to study full time. It was also observed that visitors from outside RMIT university generally spend more time on the pages compared to visitors from within RMIT university. This is mostly because students at RMIT university know the website structure very well and hence quickly navigate through the website.

Experiments 7 and 8 were conducted to compare access patterns between visitors from educational institutions and other visitors. The major difference is that some visitors from educational institutions access staff details and their contact information, whereas other visitors access career and information related information. It would appear that academics use the website for getting contact information.

In most of the pattern discovery tasks using the 20-Most-Frequent-TF and the 20-Most-Frequent-Time feature sets the root page appeared as the most significant attribute. Also, several pattern discovery tasks using the First3-Last2 and the First5-Last5 feature sets showed that "link0", which is the first page accessed in a visit, is the most significant attribute. It was observed that in the majority of the transactions, the root page was the first page visited in a transaction. All these results indicate that the first page accessed in a visit is the most significant page that differentiates the access patterns of visitors. The classification accuracies of the significant patterns discovered ranged from 70 % to 85 %.

It was thought that analysis of long transactions could give interesting insights into the patterns of access. Hence, the long transactions from the transaction set were manually analysed. An interesting pattern discovered was that some visitors tend to look for different programs available and towards the end of their visit they try to find pages related to the information brochure for the courses they want to pursue. Other discovered pattern revealed that some visitors try to find contact information of staff members. Some visitors also try to find the timetables of specific subjects of interest. The last two patterns are consistent with two of the patterns discovered using data mining techniques.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The goal of the experiments 1 and 2 was to compare access patterns between visitors from within Australia and visitors from outside Australia. Visitors from within Australia mostly visit the root page, whereas visitors from outside Australia mostly visit specific pages directly. This is probably because they use search engines. Several results indicated that overseas visitors mostly access pages related to postgraduate courses. Therefore, the website administrator should ensure that the information in these pages is accurate. Also, once a visitor is identified as an overseas visitor, dynamically providing more links to pages related to post graduate courses would help him or her quickly access desired information.

The goal of experiments 3 and 4 was to compare access patterns between visitors from within RMIT university and visitors from outside RMIT university. The results show that visitors from within RMIT university mostly visit the root page, whereas visitors from outside RMIT university visit specific pages directly. This pattern is consistent with a pattern discovered in experiments 1 and 2. An additional pattern discovered reveals that visitors from outside RMIT university tend to access information about career and employment prospects. This could be because prospective students tend to look at information about the employment prospects and industry collaborations of the university before commencing their study.

The goal of experiments 5 and 6 was to compare access patterns between visitors from inside RMIT university and outside RMIT university but within Australia. Some of the patterns discovered were consistent with those discovered in experiment 1 and 2. A discovered pattern revealed that visitors outside RMIT university generally spend more time on pages compared to visitors from within RMIT university. It can be concluded that this difference in time spent may be due to the fact that visitors from within RMIT university know the strucuture of the website very well and hence quickly browse through the website. Also, visitors from outside RMIT but within Australia might be experiencing delays because of some kind of network difficulty.

The goal of experiments 7 and 8 was to compare access patterns between visitors from educational institutions other than RMIT university and other visitors. A group of visitors from educational institutions tend to access staff contact information, whereas a group of visitors from places other than educational institutions tend to browse career and industry related information. This suggests that a group of visitors from educational institutions contact academic colleagues.

It was observed that the First3-Last2 and the First5-Last5 feature sets are suited well for association rules, whereas the 20-Most-Frequent-TF and the 20-Most-Frequent-Time feature sets are suitable for the classification technique.

The experiments of this thesis discovered a number of interesting access patterns. Evidence for these patterns comes from a number of pattern discovery tasks. For example, results of pattern discovery using both classification and clustering techniques showed that some visitors from outside Australia tend to access information related to post graduate programs. However, it should be noted that there were some inaccuracies involved in the preprocessing work. In experiments of this thesis, the classification results were analysed for their interestingness only if their accuracies exceeded a threshold of 70 %. The pattern discovery tasks could be performed with a higher accuracy threshold to obtain results that are potentially more significant. This may produce some different patterns, but the overall results are not expected to change very much as evidence for the patterns discovered comes from a number of pattern discovery tasks.

The patterns discovered show that some visitors look for specific information. For example, one pattern shows that visitors from outside Australia look for information about postgraduate courses. Hence, the web master should ensure that this information is accurate.

5.2 Future Work

While this thesis work has revealed a number of interesting access patterns, it would be worth conducting the experiments with broader scope compared to the scope of this thesis as outlined below.

1. Small Transactions
   Data Mining experiments with inclusion of transactions that have less than 5 pages visited in them may discover interesting patterns of web usage.

2. Larger Data Sets:
   Using large sized web access log files might allow the data mining algorithms to discover more access patterns.

3. Visitor Identification:
   In the experiments of AusVsNotAus, it is considered that if the ".au" suffix is found in the host name of the machine of the visitor then he or she is from within Australia. However, many websites within Australia may not have ".au" suffix in their host names. Hence, more sophisticated techniques to identify visitor locations will provide more accurate information. Therefore, the end results may be improved.

4. Domain Knowledge:
   Incorporation of knowledge about the web site structure, cookies and path completion can provide more accurate visit information. Use of cookies will enable management of state of visitors and servers. Hence, visitors can be more reliably identified. The path completion task helps in filling in the missing page accesses. Incorporation of path completion with knowledge of website structure will provide more accurate information about user visits.

5. Removal of the web robots:
   The results obtained are to some extent inaccurate due to entries of web robots that did not access the "robots.txt" file. Using better approaches like identifying web robots through the web browser information for their removal will improve results.

6. Infrequently Visited Pages:
   Feature sets having infrequently visited pages may discover interesting patterns of web access.

Bibliography

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. Very Large Databases (VLDB)*, 1994.

[2] C. R. Anderson. A machine learning approach to web personalization. *PHD Thesis, University of Washington*, 2002.

[3] J. Bilmes. A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *Technical Report, University of Berkeley, ICSI-TR-97-021*, 1997.

[4] J. Borges and M. Levene. Data mining of user navigation patterns. In *Proc. of the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99)*, pages 92–111, 2001.

[5] A.G. Buchner and M.D. Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. *SIGMOD Record*, 27(4):54–61, 1998.

[6] M.S. Chen, J.S. Park, and P.S. Yu. Data mining for path traversal patterns in a web environment. pages 385–392, 1996.

[7] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems V1(1).*

[8] R. Cooley, B. Mobasher, and J. Srivastava. Grouping web page references into transactions for mining world wide web browsing patterns. *IEEE Knowledge and Data Engineering Exchange Workshop (KDEX '97)*, 1997.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, pages 39, PP. 1–38, 1977.

[10] R. A. Fisher. The use of multiple measurements in taxanomic problems. *Annual Eugenics 7 (part II): 179-188. Reprinted in Contributions to Mathematical Statistics,1950. New York, John Wiley*, 1936.

[11] WEKA: Waikato Environment for Knowledge Analysis. www.cs.waikato.ac.nz/ml/weka.

[12] Y. Fu, M. Creado, and C. Ju. Reorganizing websites based on user access patterns. In *Proc. of the ACM CIKM International Conference on Information and Knowledge Management, Atlanta, Georgia, USA*, pages 583–585, 2001.

[13] J. Han and M. Kamber. Data mining: Concepts and techniques. *Morgan Kauffmann Publishers*, 2001.

[14] D. Hand, H. Mannila, and P. Smyth. Principles of data mining. *Massachusetts: The MIT Press*, 2001.

[15] H. Ishikawa, M. Ohta, S. Yokoyama, J. Nakayama, and K. Katayama. On the effectiveness of web usage mining for page recommendation and restructuring. *Web, Web-Services, and Database Systems.*

[16] K.P. Joshi, A. Joshi, Y. Yesha, and R. Krishnapuram. Warehousing and mining web logs. In *Proc. of ACM CIKM Workshop on Web Information and Data Management (WIDM'99)*, 1999.

[17] M. Kitsuregawa, T. Shintani, and I. Pramudiono. Web mining and its SQL based parallel execution. *IEEE Workshop on Information Technology for Virtual Enterprises*, 2001.

[18] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Max-Plank-Institute Proceedings*, pages 1137–1145, 1995.

[19] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD*, 2(1):1–15, 2000.

[20] F. Masseglia, P. Poncelet, and M. Teisseire. Using data mining techniques on web access logs to dynamically improve hypertext structure. *In ACM SigWeb Letters, 8(3): 13-19*, 1999.

[21] B. Mobasher, R. Cooley, and J. Srivastava. Creating adaptive web sites through usage-based clustering of urls. *In IEEE Knowledge and Data Engineering Workshop (KDEX'99)*, 1999.

[22] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Proc. of the 3rd ACM Workshop on Web Information and Data Management (WIDM01),At-lanta*, 2001.

[23] J. Pei, J. Han, B. Mortazavi-asl, and H. Zhu. Mining access patterns efficiently from web logs. In *Pacific Asia Conference on Knowledge Discovery and Data Mining PAKDD 2000*, pages 396–407, 2000.

[24] M. Perkowitz and O. Etzioni. Adaptive sites: Automatically learning from user access patterns. In *Proc. of the Sixth International WWW Conference, Santa Clara, CA.*, 1997.

[25] James Pitkow. In search of reliable usage data on the www. In *Proc. of the Sixth International WWW Conference*, pages 1–13, 1997.

[26] J. R. Quinlan. C4.5: Programs for machine learning. *San Francisco: Morgan Kaufmann*, 1993.

[27] M. Spiliopoulou. Web usage mining for web site evaluation. *Communications of the ACM*, 43(8), 2001.

[28] M. Spiliopoulou, L.C. Faulstich, and K. Winkler. A data miner analyzing the navigational behaviour of web users. In *Proc. of workshop on Machine Learning in User Modelling of the ACAI'99, Creta, Greece*, 1999.

[29] M. Spiliopoulou, C. Pohle, and L.C. Faulstich. Improving the effectiveness of a website with web usage mining. *In Advances in Web Usage Analysis and User Profiling, Berlin, Springer, pp. 14162*, 2000.

[30] R. Srikant and Y. Yang. Mining web logs to improve website organization. In *Proc. of the tenth International World Wide Web Conference, Hong Kong*, 2001.

[31] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

[32] K.W. Tan, H. Han, and R. Elmasri. Web data cleansing and preparation for ontology extraction using WordNet. In *First International Conference on Web Information Systems Engineering (WISE'00)*, volume 2.

[33] Feng Tao and Fionn Murtagh. Towards knowledge discovery from www log data. In *Proc. of the International Conference on Information Technology: Coding and Computing*, 2000.

[34] I.H. Witten and E. Frank. Data mining - Practical machine learning tools and techniques using JAVA implementations. *Morgan Kauffmann Publishers*, 2000.

[35] O. R. Zaiane, M. Xin, and J. Han. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. *In Advances in Digital Libraries*, pages 19–29, 1998.