# WEB USAGE MINING USING KUKTEM E-COMMUNITY APPLICATION

HASIMAH RAZAK

A report submitted in partial fulfilment
of the requirements for the award
of the degree of
Bachelor of Computer Technology (Software Engineering)

Faculty of Computer System & Software Engineering
University College of Engineering & Technology Malaysia

MARCH 2005

## ABSTRACT

The continuous growth of the information on the Internet enables users to get the convenient and yet accurate tools to gain the information needed. However, the explosion of unlimited information difficult the users to get the information needed and administrator having difficulties to provide the information needed by users for particular website. This is because, the administrators do not know what the users surf when they visit any particular website. Considering the problems that are faced by users and administrators, a data mining application will be developed. The aims of the application are to determine the general patterns of site usage and to determine the user's interest based on the option provided by the site. With the existence of this application, it will improve the quality of a website. To develop this data mining application, Visual Basis is used with the Rapid Software Development Life Cycle methodology. Information collected by Web servers and kept in the server logs is the main source of data for analyzing user navigation patterns. Once logs have been preprocessed and sessions have been obtained, there are several kinds of access patterns mining that can be performed depending on the needs of the analyst. Some of the navigation patterns that will be produced are the most requested pages and the most downloaded files. Beside that, support and confidence for each user's access option will be counted to know the users interest and to predict whether the visitor of a website will be more or less. If support less than confidence, there will be more visitors of a website and vice versa. In this paper, Generalized Association Rule will be used in order to optimize to content of website.

# ABSTRAK

Perkembangan maklumat didalam internet memerlukan pengguna diberi keselesaan dan alat yang tepat untuk mendapat maklumat yang diperlukan. Namun, begitu, ledakan maklumat yang melampau, menyukarkan pengguna mendapat maklumat yang mereka inginkan dan pentadbir sukar menyediakan maklumat yang diperlukan oleh pengguna bagi sesuatu laman web. Ini kerana, pentadbir tidak mengetahui apa yang pengguna lakukan apabila melayari sesuatu laman web. Justeru itu, bagi memudahkan kedua-dua belah pihak, satu aplikasi *data mining* akan dibangunkan. Ini bertujuan untuk mengetahui bentuk umum bagi penggunaan laman web tersebut dan mengetahui minat pengguna berdasarkan pada pilihan yang disediakan oleh laman web tersebut. Dengan adanya aplikasi ini, pentadbir dapat meningkatkan lagi mutu sesebuah laman web. Bagi membangunkan aplikasi ini, *Visual Basic* bersama kaedah *Rapid Software Development Life Cycle* digunakan. Maklumat akan dikumpulkan di pelayan dan disimpan dalam log pelayan sebagai sumber utama dalam menganalisis bentuk pelayaran pengguna. Apabila log telah diproses dan seisi telah dapat, akan terdapat beberapa jenis pelayaran mining yang akan dibentangkan bergantung kepada keperluan penganalisis. Antara bentuk pelayaran yang akan terhasil adalah halaman laman web yang paling banyak diakses dan fail yang paling banyak dimuat turun. Selain itu, *support* dan *confidence* bagi setiap pilihan yang diakses pengguna akan dikira untuk mengetahui minat pengguna dan membuat ramalan terhadap pelawat laman web samaada akan lebih ramai lagi orang yang akan melawat atau tidak. Jika support kurang dari confidence ini bermakna akan lebih ramai lagi pelawat laman web tersebut dan begitulah sebaliknya. Di dalam kajian ini, *Generalization Association Rules* akan digunakan untuk mengoptimakan kandungan laman web ini.

## TABLE OF CONTENT

## LIST OF FIGURES

## LIST OF TABLE

## LIST OF ABBREVIATIONS

| ABBREVIATIONS | DESCRIPTION |
|---|---|
| ASP | Active Server Pages |
| BI | Bitumen Technology |
| CASE | Computer-Assisted Software Engineering |
| CGI | Common Gateway Interface |
| DBMS | Database Management System |
| DOM | Document Object Model |
| FTP | File Transfer Protocol |
| HTML | HyperText Markup Language |
| HTTP | HyperText Transfer Protocol |
| IBM | International Business Machines Corp |
| IP | Internet Protocol |
| KUKTEM | Kolej Universiti Kejuruteraan Dan Teknologi |
| Malaysia | |
| NCSA | National Center for Supercomputing Applications |
| OLAP | Online Analytical Processing |
| RAD | Rapid Application Development |
| RSDLC | Rapid Software Development Life Cycle |
| SDLC | Software Development Life Cycle |
| SQL | Structured Query Language |
| URI | Uniform Resource Identifier |
| URL | Uniform Resource Locators |
| VB 6 | Visual Basic 6.0 |
| WUM | Web Utilization Miner |

| | |
|---|---|
| WUMUKEA | Web Usage Mining using KUKTEM E-Community Application |
| WWW | World Wide Web |
| XML | Extensible Markup Language |

# LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

This chapter will describe the Web Usage Mining Using Kuktem E-Community Application (WUMUKEA) overview at project overview and problem statement. It will briefly explain about the issues caused by the current manual system. Besides, the Project Scope and Objective will detailed the project scopes and its reason of development. Finally, a Gantt chart which attached in the Appendix A will describe Project Planning of the system.

## 1.2 Project Overview

The continuous growth of the information on the internet makes it necessary users to be provided with a convenient and yet accurate tools to capture the information needed. Information is ambiguous and possibly erroneous due to the dynamic nature of the information resource. Data mining is the search relationships and global patterns that exist in large database but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database, the objects in database and of the real world registered by the database. Data must be gathered together and organized in a consistent and useful way for learning to take place. It is primarily concerned with the discovery of knowledge and provide answers to questions that people do not know how to ask. The process extracts high quality information that can be used to draw conclusions based on relationships or pattern within the data.

According to explosion of data mining, Web Usage Mining Using Kuktem E-Community Application will developed under this term. This project will be developed for Kolej Universiti Kejuruteraan dan Teknologi Malaysia (KUKTEM). This application help the administrator of KUKTEM E-Community to determine the users' interest based on the option provided by the site.

The user of WUMUKEA is site's administrator where the administrator will be provided function to calculate the support and confidence for each options that provided by the site. This task will perform in database transaction in order to find the support and confidence for each transaction that consist of a set of item in database. Then, administrator will compare the output of support and confidence to produce the interest of site usage. If the output of support less than confidence for each option means more visitor will visit the option that provided by site.

The user will calculate the support of confidence based on National Center for Supercomputing Applications (NCSA) Common Log File format only. The WUMUKEA will contain some processes which are:-

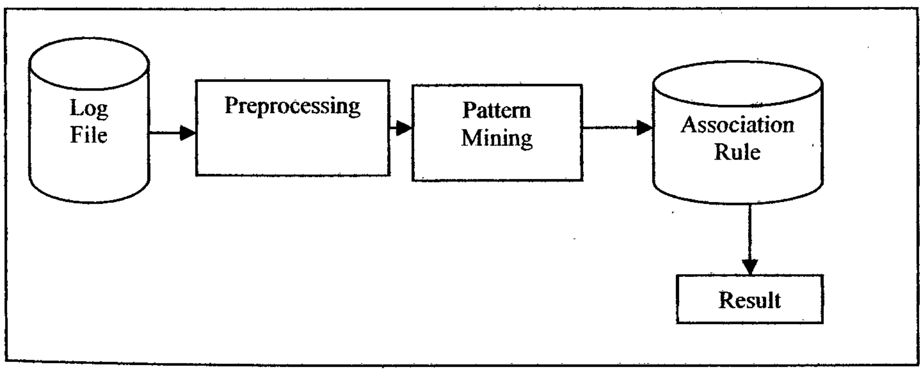

Figure 1.1: Application of WUMUKEA

## 1.3 Objective

(i) To determine the general patterns of site usage using Web Log Expert.

(ii) To determine the users' interest based on the option provided by the site using Generalized Association Rules.

## 1.4 Problem Statement

The amount of information available on World Wide Web (WWW) and databases has increased and is still rapidly increasing (Mohamadian, 2001)., For a particular website, normally hundred thousand of users will be accessing a particular site. Currently, KUKTEM do not have its own system to mine the users' interest based on the option provided by the site. It is important to know the user of the site and what the purpose of the site. At the same time, the system administrator can know whether the option of the site popular or never been visited. The administrator of a system has an access to the server log. However, the pattern of site usage cannot be analyzed without the use of a tool. Therefore, Data Mining method would ease the System Administrator to mine the usage patterns of a particular site.

## 1.5 Scope

Most of applications in web mining have been developed in business and finance, so this project focuses on the application of Web Mining in education. The numbers of successful web mining applications in education are currently small and not widely applied. Domain area that has been use is std-community.kuktem.edu.my and the techniques will be used is generalized association rules. This project will focus on:-

(i) In cleaning process, only certain attributes will clean up which are .gif and .jpg.

(ii) The option that will be calculated in generalization stage is forum. add subject, memo and email.

## 1.6 Project Planning

The project planning of WUMUKEA is shown in the Gantt chart. It will divide to five phase which are:-

(i) Planning Phase

(ii) Analysis Phase

(iii) Design Phase

(iv) Development Phase

(v) Operation and Support Phase

(refer APPENDIX A).

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

This chapter presents the literature review which discusses the issues and challenges in web mining that cover the problem solved by previous researchers, the definition of term web mining and techniques used by the previous research and the application areas where web mining have been successfully applied.

## 2.2 Previous issue and challenges

The challenge to discover valuable knowledge from chaotic WWW has become a challenging task to research community. Explosive growth in size and usage of the World Wide Web has made it necessary for web site administrators to track and analyze the navigation patterns of web site visitors. The scale of the web data exceeds any conventional databases, and therefore there is a need to analyze the data available on the web. There are also needs from the users of the web and business built around the web to benefit more from the web data. For example many users still complain from the poor performance of the websites and the difficulty to obtain their goal in the current websites because of the poor site structure or mismatches between site design and user needs (Pramudiono, 2004).

However, data mining techniques are not easily applicable to web data due to problems both related with the technology underlying the web and the lack of standards in the design and implementation of web pages. Information collected by web servers and kept in the server log is the main source of data for analyzing user navigation patterns. Once logs have been preprocessed and sessions have been obtained, there are several kinds of access pattern mining that can be performed depending on the needs of the analyst. One way to overcome this problem is by using web mining techniques in order to make use of the web data.

As a result, various web mining techniques have been developed to assist web administrator in understanding the users' patterns of site usage. Analyzing such data can help business organizations in particular to determine the life time value of customers, cross-marketing strategies across products, and effectiveness of promotional campaigns, among other things. Analysis of server access logs and user registration data can also provide valuable information on how to better structure a website in order to create a more effective presence for the organization

## 2.3 Data Mining term and definition.

Data mining can best be described as a Bitumen Technology (BI) technology that has various techniques to extract comprehensible, hidden and useful information from a population of data. Data mining makes it possible to discover hidden trends and patterns in large amounts of data. The output of a data mining exercise can take the form of patterns, trends or rules that are implicit in the data. . Web access pattern, which is the sequence of accesses pursued by users frequently, is a kind of interesting and useful knowledge in practice (Pei, 2000). Web mining is now a popular term, of techniques to analyze the data from World Wide Web (Pramudiono, 2004). A widely accepted definition of the web mining is the application of data mining techniques to web data.

As an important extension of data mining, Web mining is an integrated technology of various research fields including computational linguistics, statistics, informatics, artificial intelligence (AI) and knowledge discovery (Fayyad et al., 1996; Lee and Liu, 2000). Pal (2002) classified Web Mining into three categories: Web content mining, Web structure mining and Web usage mining (see Figure 2.1).

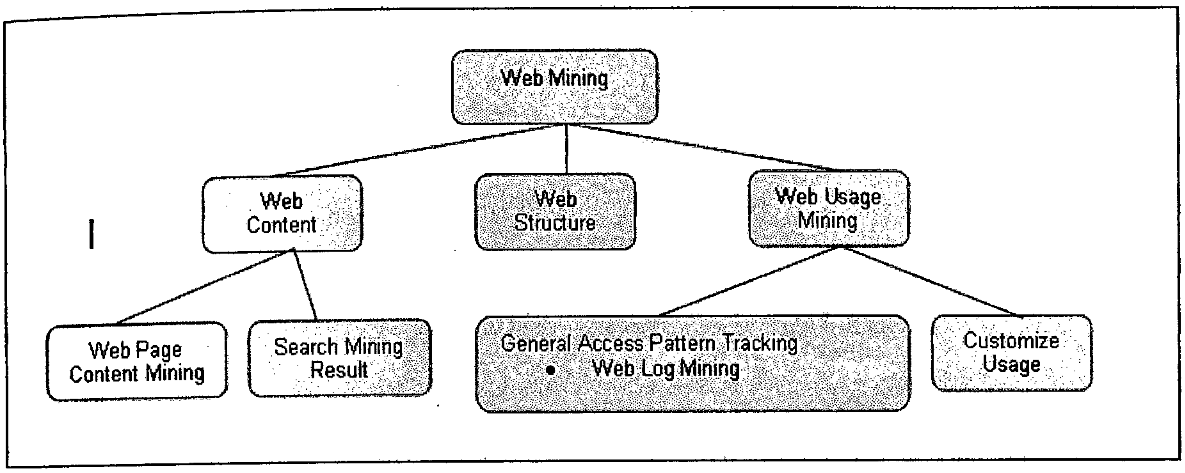

Figure 2.1: Taxonomy of Web Mining.

Content Mining involves mining web data contents (Madria, 1999) ranging from the HTML based document and XML-based documents found in the web servers to the mining of data and knowledge from the data source. Content Mining consists of two domain areas: Web Page Content Content Mining and Search Mining Result. Content data corresponds to the collection of facts from a Web page that was designed to convey information to the users (Srivastava et al., 2000). It may consist of text, images, audio, video, or structured records such as lists and tables.

Text Mining and its applications to Web Content has been the most widely researched (Srivastava et al., 2002). Some of the research issues addressed in text mining are, topic discovery, extracting associations patterns, clustering of web documents and classification of web pages while Search Mining Result can be thought of as extending the work performed by basic search engines. Most search engines are keyword-based and search mining goes beyond through the basic Information Retrieval

(IR) technology. It can improve traditional search engines through such techniques as concept hierarchies and synonyms, user profiles and analyzing the links between pages.

Structural Mining emphasizes on knowledge discovery for the structure of the Web system, including the mining of the user preferences on web browsing, the usage of the different URLs in a particular websites, external structure mining and internal structure mining (Lee and Liu, 2001). It is used in order to discover users' patterns and understand their behavior. This type of mining can be further divided into two types based on the structural data used.

(i) Hyperlinks: A Hyperlinks is a structural unit that connects a Web page to different location, either within the same page or to a different Web page. A Hyperlink that connects to a different part of the same Web Page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called Inter-Document Hyperlink. There has been a significant body of work on hyperlink analysis of which provides an up-to-date survey (Desikan et al., 2002)

(ii) Document Structure: In addition, the content within a Web page cal also be organized in a tree structure format, based on various HTML and XML tags within the page. Mining efforts have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Lui, 1998;Moh et al., 2000)

In this project, Server logs from KUKTEM E-Community were used to demonstrate the application of Web Structure mining. Figure 2.2 shows sample link structure and levels of the site std-comm.kuktem.edu.my. Structure Mining reveals more information than just the information contained in documents. For example, link pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the documents (Pal et al., 2002)

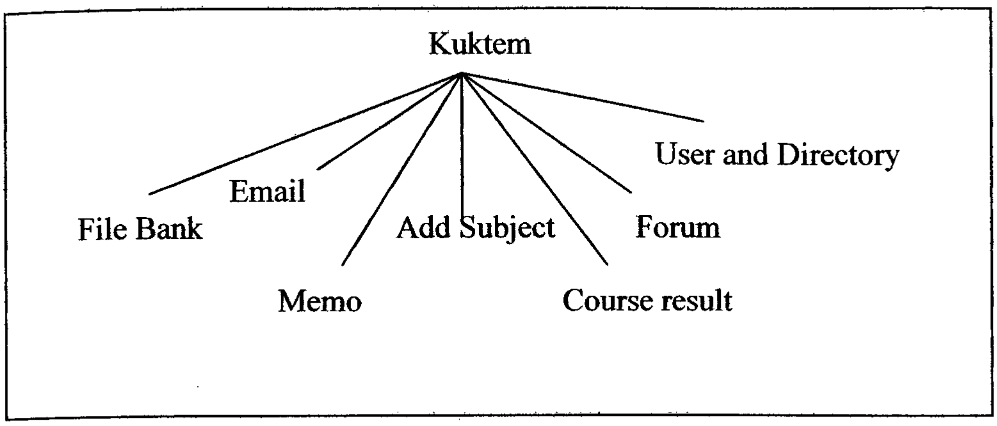

Figure 2.2: Web Structure Mining of Kuktem Education Portal.

While content mining and structure mining utilize the real or primary data on the Web, Web Usage Mining focuses on discovery of meaningful patterns from data generated from client-server transactions on one or more web server in order to study the navigation behavior and access patterns of website visitors (Mobasher *et al.*, 1999). Cooley *et al.* (1997) suggested that web usage mining includes data related to the usage of the pages of a website such as IP address, page references and the date and time access.

Web Usage data includes data from web server access logs, proxy server logs, browser logs, user profiles, registration files, user sessions or transactions, user queries, bookmark folders, mouse clicks and scrolls and any other data generated by the interaction of users and the web (Pal et al., 2002). Cooley et al. (1997) suggested that web usage mining includes data related to the usage of the pages of a website such as IP address, page references and the date and time access. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally the data will be classified into the usage data that resides in the Web clients, proxy server and servers (Srivastava et al., 2000)

Web usage mining consists of three phases, namely preprocessing, pattern discovery and pattern analysis (Srivasta,J., Cooley,R., 2000). This section presents an overview of each phase.-

### 2.3.1 Preprocessing

The preprocessing task is arguably the most difficult task in the web usage mining process due to the incompleteness of the available data (Srivastava et al., 2000). For instance, there are might be more than a user using the same client in a different period of time. Difficulties in determining a single user browsing patterns will arise as a single IP address is actually presenting multi user. It is one of the subtasks in preprocessing phase to identify users besides that cleaning and filtering the data, transaction identification and user session identification.

### 2.3.2 Pattern Discovery

Once the user transactions have been identified, several access patterns mining can be performed such as association rule mining, clustering and sequential mining (Cooley,R., Mobasher, B. and Srivastava,J. 1997). Association rule mining (Agrawal, R. and Srikant,R. 1994) discovery produces rules that described the relationships among items, which is in the case of web usage mining is the URLs, based on the certain threshold. Clustering (Kohonen,T. 1982) eventually group users or transactions according to their similarity in usage patterns while sequential pattern mining (Srikant,R and Agrawal, R. 1996) discover the sequence of items in a time ordered set.

### 2.3.3 Pattern Analysis

Pattern analysis is the important phase to filter out uninteresting rules or patterns from the set found in the pattern discovery phase (Srivastava et al., 2000). Lots of tools use variety of form to analyze the pattern such as SQL and OLAP. Graphical visualization tools that assigning colors to different values bring more advantages as it often highlight overall patterns or trends in the data.

## 2.4 Web Mining definition and research area

Research in web mining is at cross road of several research communities such as database, information retrieval, and within artificial intelligence (AI), especially in sub areas of machine learning, natural language processing (Kosala and Blockeel, 2000) and in business and e-commerce domain areas (Mobasher et al., 1996). Web mining can be broadly defined as the discovery and analysis of useful information from the WWW (Mobasher, 1996). In general, web mining can be classified into web structure mining, web content mining and web usage mining.

This study focuses on web usage mining, which analyzes the history of user behavior in the form of access patterns recorded in web access logs of web server. Organizations often generate and collect large volume of data in their daily operations. Most of this information is usually greeted automatically by web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts