Computer Science Department
Faculty of Computer & Information Sciences
Ain Shams University

A Machine Learning Approach for Web Usage Mining

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences.

By

Wael Mohamed Hamdy Mahmoud Aly Khalifa

B.Sc. in Computer and Information Sciences,
Demonstrator at Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University.

Under Supervision of

Prof. Dr. Abdel-Badeeh M. Salem
Faculty of Computer and Information Sciences,
Ain Shams University.

Dr. Shaymaa Arafat
Faculty of Computer and Information Sciences,
Ain Shams University.

Cairo 2009

Acknowledgment

First of all I would to like thank God, for his mercifulness and for giving me the energy and knowledge to finish this thesis.

I would like to thank Prof. Dr. Abdel-Badeeh Salem for his support and encouragement. I really learned a lot and he kept helping and pushing me to organize my thoughts and work to finish this thesis. I like to thank Dr. Shaymaa Arafat for her help and support in finishing this thesis.

I would also like to thank my family for all their support and love they have given my through the years. I gave them lots of hard times but they always gave me back their support. I really don't know what I could have done without them.

Last but not least my dear friends, thanks a lot, without you being there the previous years would have been even harder.

Wael

Abstract

With the huge amount of information available online, the World Wide Web is a fertile area for data mining. Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers. Web mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of users/information respectively. Machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn". Rough sets a machine learning technique; were introduced by Zdzislaw Pawlak, to provide a systemic framework for studying imprecise and insufficient knowledge

This thesis presents an approach to extract association rules from web usage data using rough sets. First we explain how to clean the data and the steps required for data cleaning. Then we describe how we convert the web usage data into a decision table, We proposed four transformation methods based on sessions, transactions, web structure mining and human experts. Then the rough set approximation is applied to the transformed decision tables to generate association rules. Finally the rules are analyzed and validated. This kind of information will help in understanding how users use the site and thus the site structure can be modified to provide an ease of use. Moreover, this could help in improving the caching techniques in the website by knowing which groups of pages are accessed together. The proposed approach is applied to four data sets, EPA, SDSC, NASA and Music website.

Publications

Abdel-Badeeh M. Salem, Shaimaa Arafat, Wael H. Khalifa, A Rough Set Approach for Web Usage Mining, 14th International Conference On Soft Comptuing MENDEL , pp 281-286, 2008

Abdel-Badeeh M. Salem, Wael H. Khalifa, Application of a Rough Set Technique in Web Log Mining, Egyptian Computer Science Journal Volume 31 Number 2,pp 38-49, January 2009

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1 Introduction

1.1 Overview

The ease and speed in business transactions that can be carried out over the Web has been a key driving force in the rapid growth of electronic commerce. Specifically, ecommerce activity that involves the end user is undergoing a significant revolution. The ability to track users' browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. It is now possible for a vendor to personalize his product message for individual customers at a massive scale, a phenomenon that is being referred to as mass customization. The scenario described above is one of many possible applications of Web Usage mining, which is the process of applying data mining techniques to the discovery of usage patterns from Web data, targeted towards various applications. Data mining efforts associated with the Web, called Web mining, can be broadly divided into three classes, i.e. content mining, usage mining, and structure mining. Web mining, much like data mining, can be said to have three operations of interests – clustering (e.g. finding natural groupings of users, pages etc.), associations (e.g. which URLs tend to be requested together), and sequential analysis (the order in which URLs tend to be accessed).

1.2 Web Mining

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the

Worldwide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining (See Figure 1.1). The necessity of web mining comes from the following problems [1][2]: low precisions due to irrelevant search results, low recall due to the inability to index all information on the web, creating new knowledge out if the information available on the web; when a large date set is returned we need to extract potentially useful knowledge out of it, personalization of the information per users profile, learning about customers behavior and specific users.

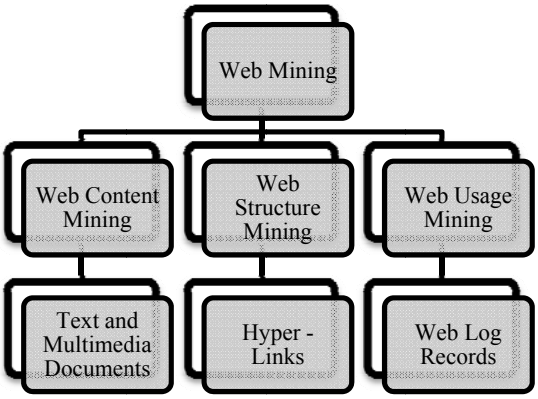

Figure 1.1 Web Mining Categories

Web Content Mining is the extraction of potentially useful patterns from the web content; the content is not only textual content, it can be multimedia data as well, for example audio, video and images. Since the content of a text document presents no machine-readable semantic, some approaches suggested restructuring the content of the document in a way that could be read by the machines. The usual approach to utilize known structure in documents is to use wrappers to map documents to some data model[1]. There are two groups of web content mining strategies: Those that

directly mine the content of documents and those that improve on the content search of other tools like search engines.

Web Structure Mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural object that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (extensible Markup Language) tags within the web page. In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as related by synonyms or ontology, they may have similar contents, and both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient. Web structure mining has a nature relation with the Web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the Web. It's quite often to combine these two mining tasks in an application. The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research.

However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining [3], which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining.

Web Usage Mining is the application that uses data mining to analyze and discover interesting patterns of user's usage data on the web [4]. The usage data stores the user's activities when the user browses or makes transactions on a web site. Web usage mining involves the automatic discovery of patterns from one or more Web servers. Organizations often generate and collect large volumes of data; most of this information is usually generated automatically by Web servers and collected in server log. Analyzing such data can help these organizations to determine the value of particular customers, cross marketing strategies across products and the effectiveness of promotional campaigns, etc. The first web analysis tools simply provided mechanisms to report user activity as recorded in the servers. Using these tools, it was possible to determine simple information such as the number of hits on the server, the times or time intervals of visits as well as the domain names and the URLs of users of the Web server. However, these tools provide little analysis of data relationships among the accessed files and directories within the Web site. More Details on web usage mining will be explained later on.

1.3 Machine Learning

As a broad subfield of artificial intelligence, machine learning is concerned with the design and development of algorithms and techniques that allow computers to "learn". The major focus of machine learning research is to extract information from data automatically, by computational and statistical methods. Hence, machine learning is closely related to data mining and statistics but also theoretical computer science [5]. Machine learning has a wide spectrum of applications including natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics and cheminformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing and robot locomotion.

Some machine learning systems attempt to eliminate the need for human intuition in the analysis of the data, while others adopt a collaborative approach between human and machine. Human intuition cannot be entirely eliminated since the designer of the system must specify how the data are to be represented and what mechanisms will be used to search for a characterization of the data. Machine learning can be viewed as an attempt to automate parts of the scientific method.

Machine learning algorithms are organized into a taxonomy, based on the desired outcome of the algorithm. There are six main learning methodologies: