WEB MINING

Submitted by: Priya Agrawal
Supervisor: Namita Mittal
March 7, 2011
# Contents

1 INTRODUCTION .................................................. 3

2 WEB MINING ..................................................... 5

3 CHALLENGES OF WEB MINING ................................. 7

4 TAXONOMY OF WEB MINING .................................. 9
   4.1 WEB CONTENT MINING .................................... 10
      4.1.1 Agent-Based Approach ............................. 10
      4.1.2 Database Approach ................................. 11
   4.2 WEB STRUCTURE MINING ................................. 11
      4.2.1 PageRank ........................................... 12
      4.2.2 HITS .................................................. 13
   4.3 WEB USAGE MINING ...................................... 13
      4.3.1 Pre-processing ...................................... 14
      4.3.2 Pattern discovery .................................. 15
      4.3.3 Pattern Analysis .................................. 16

5 APPLICATIONS OF WEB MINING ............................. 17
   5.1 Personalized Services ................................... 17
   5.2 Improve the web site design .............................. 18
   5.3 System Improvement ..................................... 18
   5.4 Predicting trends ......................................... 18
   5.5 To carry out intelligent business ....................... 19

6 CONCLUSION .................................................. 20
List of Figures

4.1 Taxonomy of Web mining ............................. 10
4.2 Web graph structure ................................. 11
4.3 Web usage mining process ............................ 14
Chapter 1

INTRODUCTION

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and Intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular web localities[5].

At present most of the users commonly use searching engines such as www.google.com, to find their required information. Moreover, the target of the Web search engine is only to discover resource on the Web. Each searching engines having its own characteristics and employing different algorithms to index, rank, and present web documents. But because all these searching engines is build based on exact key words matching and it’s query language belongs to some artificial kind, with restricted syntax and vocabulary other than natural language, there are defects that all kind of searching engines cannot overcome:

Narrowly Searching Scope: Web pages indexed by any searching engines are only a tiny part of the whole pages on the www, and the return pages when user input and submit query are another tiny part of indexed numbers of the searching engine.
Low Precision: User cannot browse all the pages one by one, and most pages are irrelevant to the user’s meaning, they are highlighted and returned by searching engine just because these pages in possession of the key words[4].

Web mining techniques could be used to solve the information over load problems directly or indirectly. However, Web mining techniques are not the only tools. Other techniques and works from different research areas, such as DataBase (DB), Information Retrieval (IR), Natural Language Processing (NLP), and the Web document community, could also be used[2].

INFORMATION RETRIEVAL
Information retrieval is the art and science of searching for information in documents, searching for documents themselves, searching for metadata which describes documents, or searching within databases, whether relational standalone databases or hypertext networked databases such as the Internet or intranets, for text, sound, images or data[1].

NATURAL LANGUAGE PROCESSING
Natural language processing (NLP) is concerned with the interactions between computers and human (natural) languages. NLP is a form of human-to-computer interaction where the elements of human language, be it spoken or written, are formalized so that a computer can perform value-adding tasks based on that interaction[1].
Natural language understanding is sometimes referred to as an AI-complete problem, because natural-language recognition seems to require extensive knowledge about the outside world and the ability to manipulate it[2].
Chapter 2

WEB MINING

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web[1].

Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular hyper-text documents published on the Web[8]. Web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others. Web mining has new character compared with the traditional data mining. First, the objects of Web mining are a large number of Web documents which are heterogeneously distributed and each data source are heterogeneous; second, the Web document itself is semi-structured or unstructured and lack the semantics the machine can understand[2].

This area of research is so huge today due to the tremendous growth of information sources available on the Web and the recent interest in e-commerce. Web mining is used to understand customer behavior, evaluate the effectiveness of a particular Web site, and help quantify the success of a marketing campaign.

Web mining can be decomposed into the subtasks, namely:

1. **Resource finding**: the task of retrieving intended Web documents. By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the web such as
electronic newsletters, electronic newswire, the text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources.

2. **Information selection and pre-processing**: automatically selecting and pre-processing specific information from retrieved Web resources. It is a kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing that are mentioned above such as stop words, stemming, etc. or a pre-processing aimed at obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.

3. **Generalization**: automatically discovers general patterns at individual Web sites as well as across multiple sites. Machine learning or data mining techniques are typically used in the process of generalization. Humans play an important role in the information or knowledge discovery process on the Web since the Web is an interactive medium.

4. **Analysis**: validating and/or interpretation of the mined patterns[6].
Chapter 3

CHALLENGES OF WEB MINING

1. Today World Wide Web is flooded with billions of static and dynamic web pages created with programming languages such as HTML, PHP and ASP. It is significant challenge to search useful and relevant information on the web.

2. Creating knowledge from available information.

3. As the coverage of information is very wide and diverse, personalization of the information is a tedious process.

4. Learning customer and individual user patterns.


6. Much of the information present on web is redundant, as the same piece of information or its variant appears in many pages.

7. The web is noisy i.e. a page typically contains a mixture of many kinds of information like, main content, advertisements, copyright notice, navigation panels.

8. The web is dynamic, information keeps on changing constantly. Keeping up with the changes and monitoring them are very important.

9. The Web is not only disseminating information but it also about services.
Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.

10. The most important challenge faced is Invasion of Privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, when it occurs without their knowledge or consent [7].
Chapter 4

TAXONOMY OF WEB MINING

In general, Web mining tasks can be classified into three categories:

1. Web content mining,

2. Web structure mining and

3. Web usage mining.

However, there are two other different approaches to categorize Web mining. In both, the categories are reduced from three to two: Web content mining and Web usage mining. In one, Web structure is treated as part of Web Content while in the other Web usage is treated as part of Web Structure. All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web[2].
4.1 WEB CONTENT MINING

Web Content Mining deals with discovering useful information or knowledge from web page contents. Web content mining analyzes the content of Web resources. Content data is the collection of facts that are contained in a web page. It consists of unstructured data such as free texts, images, audio, video, semi-structured data such as HTML documents, and a more structured data such as data in tables or database generated HTML pages[1]. The primary Web resources that are mined in Web content mining are individual pages. They can be used to group, categorize, analyze, and retrieve documents. Web content mining could be differentiated from two points of view:

4.1.1 Agent-Based Approach

This approach aims to assist or to improve the information finding and filtering the information to the users. This could be placed into the following three categories:

a. Intelligent Search Agents: These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.

b. Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.

c. Personalized Web Agents: These agents learn user preferences and
discover Web information based on these preferences, and preferences of other
users with similar interest[5].

4.1.2 Database Approach

Database approach aims on modeling the data on the Web into more struc-
tured form in order to apply standard database querying mechanism and
data mining applications to analyze it. The two main categories are:

a. Multilevel databases: The main idea behind this approach is that the
lowest level of the database contains semi-structured information stored in
various Web sources, such as hypertext documents. At the higher level(s)
meta data or generalizations are extracted from lower levels and organized
in structured collections, i.e. relational or object-oriented databases.

b. Web query systems: Many Web-based query systems and languages
utilize standard database query languages such as SQL, structural informa-
tion about Web documents, and even natural language processing for the
queries that are used in World Wide Web searches[5].

4.2 WEB STRUCTURE MINING

Web structure mining is the process of discovering structure information from
the web[1]. The structure of a typical web graph consists of web pages as
nodes, and hyperlinks as edges connecting related pages. This can be further
divided into two kinds based on the kind of structure information used.

![Web graph structure](image)

**Figure 4.2**: Web graph structure
Hyperlinks
A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an Intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

Document Structure
In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents[7].

Web structure mining focuses on the hyperlink structure within the Web itself. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models.

Two algorithms that have been proposed to lead with those potential correlations are:
1. HITS and
2. PageRank.

4.2.1 PageRank
Page Rank is a metric for ranking hypertext documents that determines the quality of these documents. The key idea is that a page has high rank if it is pointed to by many highly ranked pages. So the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined[4].

The rank of a page \( p \) can thus be written as:

\[
PR(p) = \frac{d}{n} + (1 - d) \sum_{(q,p) \in G} \left( \frac{PR(q)}{Outdegree(q)} \right)
\]  

(4.1)

Here, \( n \) is the number of nodes in the graph, \( OutDegree(q) \) is the number of hyperlinks on page \( q \) and \( d \) damping factor is the probability at each page the random surfer will get bored and request another random page.
4.2.2 HITS

Hyperlink-induced topic search (HITS) is an iterative algorithm for mining the Web graph to identify topic hubs and authorities. Authorities are the pages with good sources of content that are referred by many other pages or highly ranked pages for a given topic; hubs are pages with good sources of links. The algorithm takes as input, search results returned by traditional text indexing techniques, and filters these results to identify hubs and authorities[1]. The number and weight of hubs pointing to a page determine the page’s authority. The algorithm assigns weight to a hub based on the authoritativeness of the pages it points to. If many good hubs point to a page $p$, then authority of that page $p$ increases. Similarly if a page $p$ points to many good authorities, then hub of page $p$ increases[4].

After the computation, HITS outputs the pages with the largest hub weight and the pages with the largest authority weights, which is the search result of a given topic.

4.3 WEB USAGE MINING

Web usage mining is a process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on the Internet[1].

Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. It collects the data from Web log records to discover user access patterns of Web pages. Usage data captures the identity or origin of web users along with their browsing behavior at a web site.

In the using and mining of Web data, the most direct source of data are Web log files on the Web server. Web log files records of the visitor’s browsing behavior very clearly. Web log files include the server log, agent log and client log (IP address, URL, page reference, access time, cookies etc.))[3].

There are several available research projects and commercial products that analyze those patterns for different purposes. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization[6].
The Web Usage Mining can be decomposed into the following three main sub tasks:

4.3.1 Pre-processing

It is necessary to perform a data preparation to convert the raw data for further process. The actual data collected generally have the features that incomplete, redundancy and ambiguity[7]. In order to mine the knowledge more effectively, pre-processing the data collected is essential. Preprocessing can provide accurate, concise data for data mining. Data preprocessing, includes data cleaning, user identification, user sessions identification, access path supplement and transaction identification.

- The main task of data cleaning is to remove the Web log redundant data which is not associated with the useful data, narrowing the scope of data objects.

- Determining the single user must be done after data cleaning. The purpose of user identification is to identify the users uniqueness. It can be complete by means of cookie technology, user registration techniques and investigative rules.

- User session identification should be done on the basis of the user identification. The purpose is to divide each user’s access information into several separate session processes. The simplest way is to use time-out estimation approach, that is, when the time interval between the page requests exceeds the given value, namely, that the user has started a new session.
Because the widespread use of the page caching technology and the proxy servers, the access path recorded by the Web server access logs may not be the complete access path of users. Incomplete access log does not accurately reflect the user’s access patterns, so it is necessary to add access path. Path supplement can be achieved using the Web site topology to make the page analysis.

The transaction identification is based on the user’s session recognition, and its purpose is to divide or combine transactions according to the demand of data mining tasks in order to make it appropriate for demand of data mining analysis[3].

4.3.2 Pattern discovery

Pattern discovery mines effective, novel, potentially useful and ultimately understandable information and knowledge using mining algorithm. Its methods include statistical analysis, classification analysis, association rule discovery, sequential pattern discovery, clustering analysis, and dependency modeling.

• **Statistical Analysis:** Statistical analysts may perform different kinds of descriptive statistical analyses (frequency, mean, median, etc.) based on different variables such as page views, viewing time and length of a navigational path when analyzing the session file. By analyzing the statistical information contained in the periodic web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions.

• **Association Rules:** In the web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions.

• **Clustering analysis:** Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies.

• **Classification analysis:** Classification is the technique to map a data item into one of several predefined classes. The classification can be
done by using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc.

- **Sequential Pattern:** This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis.

- **Dependency Modeling:** The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the web domain. The modeling technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future web resource consumption[3].

### 4.3.3 Pattern Analysis

Pattern Analysis is a final stage of the whole web usage mining. The goal of this process is to eliminate the irrelevant rules or patterns and to understand, visualize and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. There are two most common approaches for the pattern analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations[3].
Chapter 5

APPLICATIONS OF WEB MINING

Web mining techniques can be applied to understand and analyze such data, and turned into actionable information, that can support a web enabled electronic business to improve its marketing, sales and customer support operations. Based on the patterns found and the original cache and log data, many applications can be developed. Some of them are:

5.1 Personalized Services

The so-called personalized service, that is, when the user browses Web sites, as far as possible to meet each user’s browsing interest and constantly adjusts to adapt to the users browsing interests change, so that make each user feel he/she is a unique user of this Web site.

In order to achieve personalized service, it first has to obtain and collect information on clients to grasp customer’s spending habits, hobbies, consumer psychology, etc., and then can be targeted to provide personalized service. To obtain consumer spending behavior patterns, the traditional marketing approach is very difficult, but it can be done using Web mining techniques[8].

Early on in the life of Amazon.com, its visionary CEO Jeff Bezos observed, In a traditional (brick-and mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase - since the cost of going to another store is high and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer experience budget (which
keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store. This fundamental observation has been the driving force behind Amazons comprehensive approach to personalized customer experience, based on the mantra a personalized store for every customer. A host of Web mining techniques, e.g. associations between pages visited, click-path analysis, etc., are used to improve the customers experience during a store visit. Knowledge gained from Web mining is the key intelligence behind Amazons features such as instant recommendations, purchase circles, wish-lists, etc[3].

5.2 Improve the web site design

Attractiveness of the site depends on its reasonable design of content and organizational structure. Web mining can provide details of user behavior, providing web site designers basis of decision making to improve the design of the site.

5.3 System Improvement

Performance and other service quality attributes are crucial to user satisfaction from services such as databases, net-works, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission, load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc[3].

5.4 Predicting trends

Web mining can predict trend within the retrieved information to indicate future values. For example, an electronic auction company provides information about items to auction, previous auction details, etc. Predictive modeling can be utilized to analyze the existing information, and to estimate the values for auctioneer items or number of people participating in future auctions.
The predicting capability of the mining application can also benefits the society by identifying criminal activities[1].

5.5 **To carry out intelligent business**

A visit cycle of customer network marketing activities can be divided into four steps: Being attracted, presence, purchase and left. Web mining technology can dig out the customers’ motivation by analyzing the customer click-stream information in order to help sales make reasonable strategies, custom personalized pages for customers, carry out targeted information feedback and advertising. In short, in e-commerce network marketing, Using Web mining techniques to analyze large amounts of data can dig out the laws of the consumption of goods and the customer’s access patterns, help businesses develop effective marketing strategies, enhance enterprise competitiveness[8].

The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer[1].
Chapter 6

CONCLUSION

As the Web and its usage continue to grow, so does the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past few years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing. The key component of web mining is the mining process itself. Here we have described the key computer science contributions made in this field, including the overview of web mining, taxonomy of web mining, the prominent successful applications, and outlined some promising areas of future research.
Bibliography


[8] Web mining: applications and techniques By Anthony Scime