“Web Usage Mining using ART Neural Network”

Ms. Parminder Kaur,
Lecturer CSE Department
MGM’s Jawaharlal Nehru College of Engineering,
N-1, CIDCO, Aurangabad – 431003

&

Ms. Ruhi M. Oberoi,
Lecturer CSE Department
MGM’s Jawaharlal Nehru College of Engineering,
N-1, CIDCO, Aurangabad – 431003

Abstract

As the web and its usage continues to grow, so too grows the need to analyze web data and extract all types of useful knowledge from it. It has become increasingly necessary for users to utilize automated tools to find the desired information resources, and to track and analyze their usage patterns. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine knowledge. In this paper we will study web mining & types of web mining. We further illustrate about web usage mining, how ART neural network can be used for web usage mining and some applications where web mining can be useful.

Keywords: Web mining, web usage mining, ART neural network
1. Introduction

The explosive growth of World Wide Web has resulted in a large amount of data that is now in general freely available for user access. Now need is that different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently.

Currently, users can choose from three major approaches when accessing information stored on the Web:

- **Keyword-based search or topic-directory browsing** with search engines such as Google or Yahoo, which use keyword indices or manually built directories to find documents with specified keywords or topics;
- **querying deep Web sources**—where information, such as amazon.com’s book data and realtor.com’s real-estate data, hides behind searchable database query forms—that, unlike the surface Web, cannot be accessed through static URL links; and
- **random surfing** that follows Web linkage pointers.

Although keyword-, address-, and topic-based Web search engines already support information searches, data mining will play an important role in Web intelligence because the Web’s current incarnation still cannot provide high-quality, intelligent services. Several data mining methods are used to discover the hidden information in the web.

The aim of this paper is to provide an overview on types of web mining generally and web usage mining in particular. The paper is organized as follows: Section 2 introduces web mining. In Section 3 web mining tasks are described. In Section 4 web mining methodologies are explained and web usage mining is explained in detail in Section 5. Section 6 concludes the paper with some typical application areas.

2. What is Web Mining?

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

3. Web Mining Tasks

The following tasks embody research problems that must be solved if we are to use data mining effectively in developing Web intelligence.
3.1 Mining Web search-engine data

An index-based Web search engine crawls the Web, indexes Web pages, and builds and stores huge keyword-based indices that help locate sets of Web pages that contain specific keywords. By using a set of tightly constrained keywords and phrases, an experienced user can quickly locate relevant documents. However, current keyword-based search engines suffer from several deficiencies. First, a topic of any breadth can easily contain hundreds of thousands of documents. This can lead to a search engine returning a huge number of document entries, many of which are only marginally relevant to the topic or contain only poor-quality materials. Second, many highly relevant documents may not contain keywords that explicitly define the topic, a phenomenon known as the polysemy problem. For example, the keyword data mining may turn up many Web pages related to other mining industries, yet fail to identify relevant papers on knowledge discovery, statistical analysis, or machine learning because they did not contain the data mining keyword. Based on these observations, we believe data mining should be integrated with the Web search engine service to enhance the quality of Web searches. To do so, we can start by enlarging the set of search keywords to include a set of keyword synonyms. For example, a search for the keyword data mining can include a few synonyms so that an index-based Web search engine can perform a parallel search that will obtain a larger set of documents than the search for the keywords alone would return. The search engine then can search the set of relevant Web documents obtained so far to select a smaller set of highly relevant and authoritative documents to present to the user. Web-linkage and Web-dynamics analysis thus provide the basis for discovering high-quality documents.

3.2 Analyzing the Web’s link structures

Given a keyword or topic, such as investment, we assume a user would like to find pages that are not only highly relevant, but authoritative and of high quality. Automatically identifying authoritative Web pages for a certain topic will enhance a Web search’s quality. The secret of authority hides in Web page linkages. These hyperlinks contain an enormous amount of latent human annotation that can help automatically infer the notion of authority. When a Web page’s author creates a hyperlink pointing to another Web page, this action can be considered as an endorsement of that page. The collective endorsement of a given page by different authors on the Web can indicate the importance of the page and lead naturally to the discovery of authoritative Web pages.

Web-page authors create some links for other purposes, such as navigation or to serve as paid advertisements. Overall, though, if most hyperlinks function as endorsements, the collective opinion will still dominate. Second, an authority belonging to a commercial or competitive interest will seldom have its Web page point to rival authorities’ pages. For example, Coca-Cola will likely avoid endorsing Pepsi by ensuring that no links to Pepsi’s Web pages appear on Coca-Cola’s sites. Third, authoritative pages seldom provide illuminating descriptions. For example, Yahoo’s main Web page may not contain the explicit self-description “Web search engine.”
3.3 Classifying Web documents automatically

Although Yahoo and similar Web directory service systems use human readers to classify Web documents, reduced cost and increased speed make automatic classification highly desirable. Typical classification methods use positive and negative examples as training sets, and then assign each document a class label from a set of predefined topic categories based on pre-classified document examples. For example, developers can use Yahoo’s taxonomy and its associated documents as training and test sets to derive a Web document classification scheme. This scheme classifies new Web documents by assigning categories from the same taxonomy. Developers can obtain good results using typical keyword-based document classification methods—such as Bayesian classification, support vector machine, decision-tree induction, and keyword-based association analysis—to classify Web documents. Since hyperlinks contain high quality semantic clues to a page’s topic, such semantic information can help achieve even better accuracy than that possible with pure keyword-based classification.

Thus, ideally, a Web document classification scheme should not require explicitly labeled negative examples. Using positive examples alone can be especially useful in Web document classification, prompting some researchers to propose a classification method based on a refined support-vector-machine scheme.

3.4 Mining Web page semantic structures and page contents

Fully automatic extraction of Web page structures and semantic contents can be difficult given the current limitations on automated natural-language parsing. However, semiautomatic methods can recognize a large portion of such structures. Experts may still need to specify what kinds of structures and semantic contents a particular page type can have. Then a page-structure-extraction system can analyze the Web page to see whether and how a segment’s content fits into one of the structures.

Developers also can test user feedback to enhance the training and test processes and improve the quality of extracted Web page structures and contents. Detailed analysis of Web page mining mechanisms reveals that different kinds of pages have different semantic structures. For example, a department’s homepage, a professor’s homepage, and a job advertisement page can all have different structures. First, to identify the relevant and interesting structure to extract, either an expert manually specifies this structure for a given Web page class, or we develop techniques to automatically induce. Such a structure from a set of pre-labeled Web page examples. Second, developers can use Web page structure and content extraction methods for automatic extraction based on Web page classes, possible semantic structures, and other semantic information. Page class recognition helps to extract semantic structures and contents, while extracting such structures helps to confirm which class the extracted pages belong to. Such an interaction mutually enhances both processes. Third, semantic page structure and content recognition will greatly enhance the in-depth analysis of Web page contents and the building of a multilayered Web information base.
4. Web Mining Methodologies

4.1 Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video or structured records such as lists and tables. Web content mining can also be defined as an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested, restructuring the document content in a representation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. 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.

4.2 Web Structure Mining

Here, we examine only the relationships between web documents by utilizing the information conveyed by each document's hyperlinks. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages. it is the process of discovering structure information from the web. This can be divided into two kinds based on the kind of structure information used.

- **Hyperlinks:** It 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:**
  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.

World Wide Web can reveal more information than just the information contained in documents. For example, links 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 document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. A graph representation can be user for web page structure. Here nodes in the graphs are web pages and edges indicate hyperlinks between pages. By examining these "web graphs" it is possible to find documents or areas of interest through the use of certain graph-theoretical measures or procedures. Structures such as web rings, portals, or affiliated sites can be identified by matching the characteristics of these structures (e.g. we can identify portal pages because they have an unusually high out-degree).
4.3 Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. In web usage mining the goal is to examine web page usage patterns in order to learn about a web system's users or the relationships between the documents. For example, a tool creates association rules from web access logs, which store the identity of pages accessed by users along with other information such as when the pages were accessed and by whom; these logs are the focus of the data mining effort, rather than the actual web pages themselves. Rules created by their method could include, for example, "70% of the users that visited page A also visited page B."

5. Web Usage Mining

This is defined as the application of data mining techniques to discover usage patterns from web data in order to understand and better serve the needs of web based applications. It is also called as Web Log Mining. It is a process of extracting interesting patterns from the web access logs. Analyzing the web access logs of websites can help to understand the user behavior and the web site graph.

The process of WUM is basically a combination of two processes:

- Domain Dependent Processes,
- Domain Independent Processes

Domain dependent processes transform the web data into suitable form. This process includes preprocessing, data integration and transaction identification.

Domain independent processes involve the domain independent application of generic data mining and techniques involving formatting of data and query mechanism.

Fig 1. Web Usage Mining Architecture

The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration.
components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine.

Data cleaning is the first step performed in the Web usage mining process. Currently, the WEBMINER system uses the simplistic method of checking filename suffixes. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc.

After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The clean server log can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. A transaction identification module can be defined as either a merge or a divide module. Both types of modules take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the module in the same format as the input. The requirement that the input and output transaction format match allows any number of modules to be combined in any order, as the data analyst sees fit. The WEBMINER system currently has reference length, maximal forward reference, and time window divide modules, and a time window merge module.

Access log data may not be the only source of data for the Web mining process. User registration data, for example, is playing an increasingly important role, particularly as more security and privacy conscious client-side applications restrict server access to a variety of information, such as the client user IDs. The data collected through user registration must then be integrated with the access log data. There are also known or discovered attributes of references pages that could be integrated into a higher level database schema. Such attributes could include page types, classification, usage frequency, page meta information, and link structures. While WEBMINER currently does not incorporate user registration data, various data integration issues are being explored in the context of Web usage mining. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns.

Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints. The emerging data mining tools and systems lead naturally to the demand for a powerful data mining query language, on top of which many interactive and flexible graphical user interfaces can be developed. Such a query mechanism can provide user control over the data mining process and allow the user to extract only relevant and useful rules. In WEBMINER, a simple Query mechanism has been implemented by adding some primitives to an SQL-like language. This allows the user to provide guidance to the mining engine by specifying the patterns of interest.
5.1 Web usage mining using ART neural Network

To derive URL clusters ART neural network can be used:

How can a learning system be designed to remain plastic, or adaptive, in response to significant events and yet remain stable in response to irrelevant events? How does the system know how to switch between its stable and its plastic modes to achieve stability without rigidity and plasticity without chaos? In particular, how can it preserve its previously learned knowledge while continuing to learn new things?

And, what prevents the new learning to wash away the memories of prior learning?

ART Algorithm

1. Initialise the number N of categories. Initialize every Prototype Vector (Category) Pi, \( i \in [1, N] \) to the unitary vector. Initialize the vigilance parameter \( \rho \in [0,1] \).
2. Apply Input Get the next input vector \( X_j \). Enable all Output.
3. Compute Activation for every enabled category:

   \[
   T = \frac{\| P_i \cap X_j \|}{\alpha + \| P_i \|}
   \]

4. Select Category \( P_i \) with Max \( T \), if set of enabled category is empty goto 2
5. Check for resonance:

   \[
   \frac{\| P_i \cap X_j \|}{\| X_j \|} \geq \rho
   \]

6. if resonance equation is false disable output \( P_i \) and goto 4
7. Adjust winning Category.

We have used this approach to dynamically group web users based on their web access patterns using Adaptive Resonance Theory Neural Network. Web usage mining works on user profiles, user access patterns and navigation paths that are being heavily used by e-commerce companies, for tracking customer behavior on their sites.
Group users based on their web interests and then organize the structure of site according to the needs of different groups. This approach adapts to changes in user access patterns over time without losing earlier information. Each cluster is represented as prototype vector by generalizing the URLs most frequently accessed by all cluster members.

6. Conclusion

Analysis of how users are accessing a site is critical for determining effective marketing strategies and optimizing the logical structure of the Web site.

Areas where web mining can be useful:

Business intelligence systems have traditionally relied on internally developed repositories. With the advent of the Web and WM, the opportunity exists to extend the business intelligence repository external to the organization. For example, systems have been built to analyze securities and Exchange Commission electronic filings, looking for potential items of interest to investors or business partnerships. The larger information base afforded by the Web, while less structured, provides the opportunity to become proactive by using WM tools to deliver business intelligence before it is requested – i.e., push technology. The integration of WM and traditional data warehouse - based business intelligence systems therefore becomes critical to achieving user acceptance.

Also, applications that make use of textual information on the Web to assist in predicting stock movements. The Web could be considered one vast digital library. The function of a librarian is to make it easier for people to find information. WM could be viewed as the digital library’s librarian, or at least a powerful tool to those acting as their own librarian.

References
1. http://www.w3.org/2001/sw/
2. Yan Wang “Web mining and knowledge discovery of usage patterns”.
4. [August 2006], CSI communications Article by Dr. Shyam Kumar Garg on “Web mining “ pp. 30-34.
5. http://www.galeas.de/webmining
7. Artificial Neural Network by B. Yegnanarayana, Ch 7, pp 258-264.

***