Web Mining: A Survey Paper

K.Amutha¹ Dr.M.Devapriya²
M.Phil Research Scholoar¹
PG &Research Department of Computer Science
Government Arts College (Autonomous), Coimbatore-18.

Assistant Professor²
PG &Research Department of Computer Science
Government Arts College (Autonomous), Coimbatore-18.

Abstraction:
Due to the huge amount of information available on the web, the World Wide Web has becoming one of the most important resources for extracting the information and knowledge discoveries. Web mining technology are the right solutions for knowledge finding on the Web. As the web is growing fast, the users get easily missing in the web’s rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill theirs needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed.

Key words:
Data mining, web, web content mining, web structure mining, web usage mining

I INTRODUCTION
The World Wide Web (Web) is a popular and attractive medium to distribute information today. The Web is huge, dissimilar, and dynamic and thus raises the scalability, compact disk data, and sequential issues respectively. These factor give rise to the requirement of creating server-side and client-side bright systems that can successfully mine for knowledge. The knowledge comes not only from the content of the pages themselves, but also from the exclusive rareness of the Web, such as its hyperlink organization and its diversity of content and languages.

Analysis of these characteristics often reveals interesting patterns and new knowledge. Such knowledge can be used to improve users’ good organization and effectiveness in searching for information on the Web. Thus the web mining contains two different approach were taken in initially defining Web mining. First was a ‘process-centric view’, which defined Web mining as a sequence of tasks [2]. Second was a ‘data-centric view’, which defined Web mining in terms of the types of Web data that was being used in the mining process [1].

The following are the problem encounter while retrieving in order from web:

a) Finding relevant information
People check or use the search tool when they want to discover specific information on the Web. Mostly people use some search service when they want to find exact information on the Web. A user regularly inputs a easy keyword query and a effect is a list of ranked pages. This ranking web page is similar to the query. Today’s search tools have some troubles: First one is Low accuracy and second is low evoke, mainly because of incorrect or imperfect keyword uncertainty. This leads to finding unindexed in a row.

b) Creating new knowledge from the huge information available on the Web:
This difficulty could be regard as a sub-problem of the complexity above. While the problem above is usually a query-triggered progression (rescue leaning), this crisis is a data-triggered process that presume that previously have a collection of network data and we want to extract potentially useful facts out of it (data mining oriented). up to date research [3; 4; 5] focuses on utilize the Web as a knowledge base for resolution making.
c) **Personalization of the information:**

This problem is related with the type and production of information, since client needs vary in their stuffing and presentations they prefer while interact with the Web. On the other hand, the information provider could come across these problems, among others, when trying to complete their goals on the Web.

d) **Learning about consumers or individual users:**

This problem are also expressly deal with the problem c above, which is about expressive what the customers do and want. Within this problem, there are sub-problems such as gathering customize the information to the intended consumers or even to personalize it to individual user, problems related to effective Web site design and administration, troubles related to production, etc.

**II WEB MINING OVERVIEW**

The web mining is a combination of the two singular areas of in progress research initial one is the data mining and second one is world wide web(WWW). It can be able to be mostly defined as the finding and investigation of useful information from WWW. Web mining is the make use of of data mining performance to without human intervention discover and mine information from Web documents and services. The Web mining research is a come together research area from several research community, such as database, IR, and AI research community especially from machine learning and NLP. This paper is an challenge to put the research done in a more controlled way from the machine learning end of view. However, the methods of the discover do not repeatedly use well-known machine realize algorithms. Since this is a huge, interdisciplinary, and very active examine area, there are lacking doubt some omission in this treatment.

Web mining should contain the subsequent as their subtasks:

1. **source finding:** The task of mine related information.
2. **Information collection and preprocessing:**
   - selecting and preprocessing necessary Information from related documents.
3. **Generalization:** Automatically extract general patterns at single Web sites as well as across multiple sites.
4. **Analysis:** Validation and/or interpretation of the mined patterns.

**Web Mining and Information Retrieval**

Web Mining or Web Information rescue (Web IR) is the process of retrieving useful information from among the huge set of data that make up the WWW. Web IR has a history almost as long as the web itself. Information retrieval engines may be classified into three types: directories, search engines and meta-search engines. Directories use (human generated) taxonomy to classify web contents and perform search. Search engines use automated crawlers to index documents from the web. Meta-search engines use results of existing search engines and perform further filtering. IR has the prime goal of indexing text a searching for useful documents in a group and at the present time research in IR includes modeling, script sorting and tagging, user interface, data idea, filtering, etc. [6].

**Web Mining and Information Extraction**

The World Wide Web presents a huge amount of useful in order which is usually format for its users, which makes it hard to get back relevant information from various documents. Therefore, the accessibility of robust, bendable Information Extraction (IE) systems that transform the Web documents into program-friendly structures such as a relational database will become a great necessity. Although many approaches for data extraction from Web pages have been industrial, IE has the goal.
of transform a collection of pages, usually with the help of an IR system, into useful information that is more readily digest and analyze [7]. IE aims to extract relevant facts from the documents while IR aims to select relevant credentials [8].

Web Mining and Machine Learning
Applied on the Web

Machine learning methods represent one possible approach to addressing the problem. Artificial intelligence and machine learning techniques have been applied in many important applications in both scientific and business domains, and data mining research has become a necessary subfield in this area. Machine learning techniques also have been used in information retrieval (IR) and text mining applications. Machine learning algorithms can be classified as supervised or unsupervised learning. In supervised learning, training examples consist of input/output pair patterns. The aim of the learning algorithm is to predict the output values of new examples, based on their input values. In invalid learning, training examples contain only the input patterns and no explicit target output is associated with each input. The learning algorithm needs to generalize from the input patterns to discover the output values.

III WEB MINING CATEGORIES

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely,

1. Web Content Mining (WCM)

- Describes discovery of useful information from contents, data and documents
- Two different points of view: IR view and DB view

2. Web Structure Mining (WSM)

- Model of link structures, topology of hyperlinks
- Categorizing of web pages

3. Web Usage Mining (WUM)

- Mines secondary data derived from user interactions

1. Web Content Mining (WCM)

Web content mining performed by extracting useful information from the content of a web page/site. It includes extraction of structured data/information from web pages, identification, match, and integration of semantically similar data, judgment drawing out from online sources, and concept hierarchy, ontology, or knowledge integration. Two main approaches are used in Web Content Mining: (1) Unstructured text mining approach and (2) Semi-Structured and Structured mining approach [7].

Unstructured Text Data Mining:
Web content data is much of free text data. The research around applying data mining technique to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence, one could consider text mining as an instance of Web content mining. To provide effectively usable results, preprocessing steps for any structured data is done by means of information extraction, text group, or applying NLP techniques.

Semi-Structured and Structured Data Mining:

Structured data on the Web are often very important as they represent their host pages, due to this reason it is important and popular. Structured data is also easier to extract compared to unstructured texts. Semi-structured data is a point of convergence for the Web and database communities: the former deals with documents, the latter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects like books, papers, movies, etc., without sending the application writer into contortions.
2. Web structure mining [11] tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlinks with or without the description of the links. This model can be used to categorize Web pages and is useful to generate information such as the similarity and relationship between different Web sites. Web structure mining could be used to discover authority sites for the subjects (authorities) and overview sites for the subjects that point to many authorities (hubs). Two algorithms that have been proposed to lead with those potential correlations: HITS [12] and Page Rank [10], and Web structure mining itself will be discussed in the next section. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1. Link-based Classification. Link-based classification is the most recent upgrade of a classic data mining task to linked domains [10]. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. Link-based Cluster Analysis. The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type. There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength. Links could be associated with weights.

5. Link Cardinality. The main task here is to predict the number of links between objects. There is many ways to use the link structure of the Web to create notions of authority. The main goal in developing applications for link mining is to make good use of the understanding of these intrinsic social organization of the Web.

3. Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. 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 we could classify them into the usage data that reside in the Web clients, proxy servers and servers [14]. The Web usage mining process could be classified into two commonly used approaches [15]. The first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. As is true for typical data mining applications, the issues of data quality and pre-processing are also very important here. The typical problem is distinguishing among unique users, server sessions, episodes, etc. in the presence of caching and proxy servers [15; 14]. For the details and comparison of some pre-processing methods for Web usage data we refer to [31].

• Data Collection. During this stage, data are collected either from Web servers or from clients that visit a Web site.

• Data Preprocessing. This is the stage that involves primarily data cleaning, user identification and user session identification.

• Pattern Discovery. During this stage, knowledge is extracted by applying Machine Learning techniques, such as clustering, classification, association rule discovery etc., to the data.
• **Knowledge Post-Processing.** In this last stage, the extracted knowledge is evaluated and presented in a form that is understandable to humans, e.g. by using reports, or visualization techniques. In addition to these techniques, post-processed results can also be incorporated in a Web Personalization module.

**IV Conclusion**

This survey paper research about in the area of Web mining and its types. Survey paper point out categories and then point out some of the research areas with respect to these categories. And also explore the connection between Web mining categories and the related agent paradigm. Some of the usage of the term Web mining. Paper also suggest three Web mining For the survey, focus on representation issues, on the process, and on presents new challenges to the traditional data mining algorithms that work on flat data.

**REFERENCES**


