A Study on Different Aspects of Web Mining and Research Issues

Santosh Kumar¹, Ravi Kumar²

Department of Computer Science and Engineering
ABES Engineering College, Ghaziabad, Uttar Pradesh, India-201009
¹santoshg25@gmail.com, ²ravikumarabes@gmail.com

Corresponding author email: santoshg25@gmail.com,

Abstract. Web is huge and dynamically increasing day by day with exponential rate. If query is little bit complex then often or sometimes it becomes difficult to get relevant information in response. Several search engines try to optimize the users query and/or search results to provide the relevant information. Also many websites are ecommerce sites which are also competing to attract their customers and some are recommendation based. Now almost all good sites are using web mining techniques to improve their performance. In this paper, an attempt has been made to cover the key aspects used in web mining research. Initially problems faced by users and/or owners have been highlighted. Further it discusses types of data and techniques used by different categories of web mining are discussed. Finally research issues and application areas have been discussed.

Keywords: Web usage mining, content mining, structure mining, link mining, precision, recall, web personalization, information retrieval, information extraction, opinion mining.

1. Introduction
WWW is used in its content size and it is growing very rapidly continuously. Information about almost anything is available in the Web. So, Today, WWW is the huge source of information in world. The size of web has grown so huge that finding relevant information is becoming a tedious task. Although, the search engines (like google, yahoo, lycos etc) are contributing a lot in finding the required information. For a simple query they are returning thousands and even lacs of documents in the result. Most of these documents are irrelevant. So again we have problem of finding desired document from result of query by the search engine. Although today’s search engines are using several advanced techniques to assist their users in finding the required information for example recommender system [1], outcome of web mining, is playing a vital role while searching in the web. Due to its vast size and dynamically changing information, the problems like scalability and temporal issues occurring. This is also raising the important problem of information overload. Precision is percentage of relevant results returned in response to user’s query. Recall is percentage of total relevant results with respect to a query. If search engine returns x results, y are relevant results while search engine fails to return z relevant results. Precision is y/x and...
recall is $y/(y+z)$. Precision shows the usefulness of result while recall shows the completeness of result. Additionally, the web users/owners face the some of the following problems:

1. **Low precision:** Most of the search engines use keyword based searching. User input the query in the form of keyword. Result of query is the list of pages ranked on the basis of similarity of keywords. Most of these pages are irrelevant. So there is difficulty in finding the relevant and required pages.

2. **Low recall:** Most of the search engines maintain the indices of web documents. The indices are used in searching while answering the user’s query. All the available web documents cannot be indexed. So there is difficulty in finding the required information from these un-indexed web pages.

3. **Find new knowledge from web data:** The aim in web system is to create potentially useful information from the vast data available in the web and process is called web mining. Authors in [2] tried to utilize the web as a knowledge base for decision making activities.

4. **Web Personalization:** for the success of any e-commerce website, it is necessary to take care of individual user’s preferences. Owners of the websites should understand the needs of users and accordingly represent the content i.e. they should create the adaptive websites. Research shows that there should be individual store on the web for or each customer [3]. Most of the internet users prefer to receive the personalized information. So, website owners can place the advertisement and offers based on customer’s preferences. How web personalization works is shown in Fig. 1.

![Fig. 1. Web Personalization](image)

5. **Learning about customers:** knowledge about customer’s history and patterns plays an important role in effective website design, development and management.
Web mining techniques have been proven to be very useful to address all above problems directly or indirectly [4]. Some other techniques are also available in the literature to address the above problems like Information retrieval, database management system, machine learning, natural language processing, web document community etc. In direct approach, the web mining tools or techniques directly address the above problems like newsgroup agent categorizes the relevant and irrelevant news, recommender system is used to suggest the content or items to users. In indirect approach, web mining approaches are used as a part of other applications. For example, spam email detection market basket analysis, credit scoring, Information retrieval, fraud detection, data visualization.

2. What is Web Mining?
Web mining and uncovers the hidden patterns in the large amount of data. It finds the unknown, relevant and useful information contained in the web documents [5, 17]. Web mining techniques are inspired from data mining techniques. It does not directly uses the data mining techniques due to diverse nature of web data which is available in the form of unstructured, semi-structured, and structured data. For analysis of web documents, there exist several mining tasks and algorithms in the literature. Unlike data warehousing, web has mixed type of data e.g. content data (text, audio, video, and graphics), structure data (hyperlinks, web graph), and usage data (web log data). On the basis of types of data used, web mining can be categorized as web content mining, web structure or link analysis mining, and web usage mining [17].

2.1 Web content mining:
Web content mining discovers the useful and relevant information from the content of webpage which could be unstructured text, XML data, structured tables, graphical information, images, videos etc [17]. For example, classification of webpage according to their content, mining product reviews, user’s sentiments in blog data.

2.2 Web structure mining:
It specially deals with intra documents and inter documents structure i.e. link structure of the content within a webpage and interconnectivity of webpage across the websites. Structure of the webpage affects its ranking. Web structure mining can be classified as hyperlink structure and document structure [16]. Link structure connects the content at different locations in this same webpage or it can be used to interconnect the different webpage of the same or different website while document structure organizes the page content in the form of the structure due to various tags of HTML and XML.

2.3 Web usage mining:
Web usage mining discovers user’s traversal patterns from web logs which record clickstreams by the user. Many data mining algorithms are also applicable in web usage mining. Web usage mining uses several data mining algorithms. The main problem with web usage mining is unprocessed clickstream data in web usage log file. The web mining inherits the process used in data mining. Both of these differ in their data collection techniques. Data in the data warehouse is collected form different heterogeneous sources like databases flat files. This process involves data cleaning, integration and transformation. Data for mining in the data warehouse is already collected. While for web mining data collection task is tedious but somehow web crawlers are useful in this activity. After data collection is done, this needs preprocessing, integration, transformation, and selection of data required for web mining. Finally generalization and analysis is done.

3. Web Mining Subtasks
Web mining includes four subtasks are shown in Fig. 2:
Resource gathering: This phase retrieves intended documents and is done by web search engines or web crawlers [7].

Information selection/pre-processing: after the resource finding the relevant web documents are selected and transform into standard form. Most of the methods used papers to select the data and they represent the data in tabular form [8].

Generalization: It tries to find out the general user access pattern within and across the websites. This determines the user’s interest and behavior. The web mining techniques like classification, clustering association rule techniques etc. are used.

Analysis/Validation: This step analyzes, interprets and validates the potential information against the information patterns. The aim of this task is the dredging of knowledge from the information obtained by the previous steps. Several models to simulate and validate the web data for mining. All the steps are shown in Fig. 1.

Web Mining inherits the techniques of data mining to automatically extract and you will wait the information for getting the knowledge from web content. Pattern evaluation involves generalization, classification clustering and analysis.

4. Web mining vs. other related techniques

4.1 Web Mining: Not IR

- Information Retrieval automatically retrieves the relevant documents as well as some unimportant documents.
- It uses classification step of web mining for indexing the retrieved documents so that searching becomes efficient.

4.2 Web Mining: Not IE

- Information Extraction extracts the useful facts from the web documents.
- Information extraction is not feasible for general web
- It mainly focuses on particular web page or web content

4.3 Web Mining vs. Machine Learning

- Machine learning enables the computers to learn from experiences. For this, focus is given to develop the algorithms and techniques so that machine is able to automatically learn.
- Web mining does not automatically learn from web.
- There exist some machine learning applications in the web which is not web mining for example automatic suggestions in Google search.
- Web mining techniques are not restricted to machine learning.

5. Opportunity and challenges in web mining

The heterogeneous information available in the web and characteristics provide great opportunity as well as challenges for the researchers for data mining. A few characteristics are as below:
i. The information in the web is very huge and growing rapidly. Also this huge information is available to all easily.

ii. The information about anything is available in different formats and presentation. Also information about any topic is diverse.

iii. Sample data is freely available for training and predictions in different file formats like .xls, .xlsx, .json, .csv, .arff etc.

iv. Web servers log data is also available to find out users navigational pattern, recommendations, and target marketing.

v. The data is available in different formats like structured, unstructured, semi-structured, usage data (server log, user’s click), hyperlink data, multimedia data, news, audio, video, images, advertisement data, encrypted data etc.

vi. Information about any topic is available totally in different formats and presentations due to different authorship. Presenting this information to users of the web in single format is very big challenge for the search engines.

vii. Information is interlinked with the help of hyperlinks.

viii. Most of the information in the web is noisy. Noise is to be filtered before mining the data which is a big challenge.

ix. Internet also provides services in different forms like purchasing items through online marketing, paying bills, registering on websites, purchasing domain names, playing online games, watching movies, listening songs and many more. Each of these services provides an opportunity of mining e.g. prediction, recommendations, service mining.

x. Information in the web is changing dynamically. Accounting the change is definitely helpful for many applications.

xi. Except data, information and services, web also maintains a virtual society. There is interaction among people, computers, and companies. This interaction data is available in the form of transaction logs, reviews, blogs which is very useful to improve and strengthen the interaction.

6. Taxonomy of Web Mining

Taxonomy of Web Mining is shown in Fig. 3. and is described next.

Fig. 3. Web Mining Taxonomy
6.1 Web content mining

It is used to extract useful and relevant information from content data which includes structured, semi-structured and unstructured data [6]. Rapid growth of the web is leading to several problems like difficulty in finding relevant information, doing statistical analysis, learning about customers and their behavior. For this unstructured and semi-structured data can be transformed to structured form to make analysis easier. Conversion of unstructured and semi-structured data to structured form is future area to work with. Web content mining uses two approaches, First, to mine the content of web documents and seconds to mine or improve the search results of web search engines.

A few of the challenges of web content mining are given below:

i. Data/information extraction: There are many techniques to extract structured data from the web also called wrapper generation [9]. The first is to write dedicated program to extract data from any particular website. In this approach a lot of efforts are required and also this is time consuming process. So it is not feasible. The second is wrapper induction or wrapper learning. In this, programmer first labels some training pages which are used to make a learning system and also some rules are generated from these. These rules are used to extract content data form other websites. This technique is also called supervised learning technique. Some examples of second technique are WIEN, Stalker [10], BWI [11] etc. The third technique automatically extracts structured data from the websites. This technique finds patterns/grammars from the web pages and then use these patterns to extract data from other websites e.g IEPAD [12], MDR [13], RoadRunner EXALG [14], [15], are some examples of third approach.

ii. Web Information Integration: Web has very huge amount of data. Different websites represent the same information in their own format. Even also in the same website presentation of information at two distinct places may be different. Usually the mining techniques require the data in some standard format. This information is to be integrated at one place so that it could be used for mining purpose. There are two problems related to web integration. First, web query interface integration to query multiple deep web databases and second is schema matching in which we match the concept hierarchies e.g. integrating directories of two search engines Yahoo and Google to match concept hierarchies [18]. Extraction of information from many deep web databases is difficult because this vast data can not be indexed by traditional search engines[19].

iii. Building Concept Hierarchies: Information in the web is so huge that it should be organized. Organized information becomes easier to manage and use. But due to large size it is difficult to organize the whole web. Although we can organized the search result of a query. Web pages resulted based on ranking in response to user’s query are not sufficient for many applications so another way to organize the information is concept hierarchies so sometimes categorization [17]. The mostly used technique for hierarchy creation is clustering the results and is used by researchers in [20].

iv. Automatic Web Page Segmentation and Noise Removal: generally a web page consists of several parts e.g. main content areas, menus, ads, etc. which do not contribute in knowledge extraction and performance of website can be improved after removing this noise. Also it reduces the performance of web mining process [21].

v. Mining Web Opinion Sources: In the growing age of market, competition is increasing. In such situations, companies require feedback in the form of consumer opinions in the form of surveys or use manual methods to get feedback about their products and services. And this information is publically available in the websites itself or in the form of blogs in some bloggers sites. This information is now frequently used to improve the performance of website and its reorganization.

vi. Deep Web: It is invisible or hidden web that is not indexed and not coded in HTML. So it becomes difficult for search engines to extract information from this source of world wide web.
6.2 Web structure mining
It analyses the link structure of the website. This link structure can be organized in the form of topology and can be used to find similarity and relationships between websites. This link information can also be used for website reorganization and web page ranking. Popular link based algorithm of page ranking are HITS (Hypertext Induced Topic Search) and PageRank [22]. After seeing the outcome of analysis of web structure mining a new area of research called Link Mining is also becoming popular. Some of the possible tasks of link mining are as follows:
   i.  Link-based Classification: Web pages are represented as nodes of the web graph. This labels or classifies the nodes or objects in the graph based on the characteristics of nodes or neighboring nodes.
   ii. Link-based Cluster Analysis: Link of the web page includes enough information for clustering the websites. This uses unsupervised learning approach.
   iii. Link Type: This is used to predict the type or purpose of link between two web pages.
   iv. Link Strength: tells the importance of a link by association weight to it. Weight is assigned based on the degree of closeness between two nodes/pages in the web graph.
   v. Link Cardinality: tells the numbers of links existing between two nodes in the web graph.

6.3 Web Usage Mining
Every activity of user is recorded in the server log known as web usage data. Understanding this usage log can help to the owners of websites to improve the website design, interests of the user and web personalization. This data contains user profiles, sessions, cookies, bookmarks, click streams etc. and hides potentially useful information. Before using this data needs preprocessing then pattern discovery analysis can be performed.
   i. Preprocessing: It performs data cleaning, user identification, session identification. Data cleaning filters the noise like unsuccessful requests, or some unwanted data.
   ii. Pattern Discovery: Many popular data mining techniques like as statistical analysis, association rules, clustering, and classification can be used to find interesting patterns.
   iii. Pattern analysis: In this discovered patterns are evaluated based on intuitive knowledge and unwanted patterns are discarded. Besides this there are some pattern analysis tools which tell how users are using website and how to present the information to its users, reordering the structure of content placement in the site. From analysis frequency of users, users profiles, and which links are important can be identified.

7. Web Mining Applications
   i. It is used in personalizing the web portal such as Yahoo is the first portal to do so and introduced the people with My Yahoo.
   ii. eBay is using web mining for understanding the behavior of auction. It makes the concept of bidding to come to online and facilitates user to buy or sell products.
   iii. America online (AOL) makes use of mining technique in order to understand user behavior associated with community and use this information to target advertisements.
   iv. DART performs web-wide tracking that is it follows the user on each and every web site he visits.
   v. Web search is mostly performed by most popular search engine Google. It makes use of Page Rank in order to prioritize its search result.
   vi. It makes the business to come online. Amazon is the first web site which personalizes the experience of customer in daily business.
   vii. Web mining is used in detection of fraud in e-commerce by maintaining the record of each user participating in purchasing of products.
   viii. Advertisements vulnerable area is identified by using web mining such as in Lycas and yahoo etc.
ix. Opinions are provided online about any product which helps in identifying the advantages and Disadvantages of any product before purchasing.

x. Web mining acts as a communication medium in the form of virtual society which helps people to share their views with each other.

8. Conclusion

This paper covered how precision and recall are useful parameters in information extraction. As most of the business sites are personalizing the content to the user on need basis. So we included Web personalization which is a technique used to increase the performance and business of websites. Different tasks of web mining help in extracting knowledge from the huge web. Opportunities and challenges will help the researchers to uncover other aspects of the web which are still hidden. Next, how web content, usage and structure data can be used to rank the web pages, find the similarities and dissimilarities among websites and is the basic content for applying any mining techniques.

References

[1] Cho, Y. H., Kim, J. K. and Kim, S. H. A personalized recommender system based on web usage mining and decision tree induction, in Expert Systems with Applications, 23(3): 329--42, September 2002.

[2] Mitchell, T.. Machine Learning. McGraw Hill, 1997.

[3] Schaefer, J.B., Konstan, J.A., Riedl, J., E-commerce recommendation applications, Data Mining and Knowledge Discovery 5 (1/2)(2001) 115–153.

[4] Raymond, K., Hendrik, B.. Web Mining Research: A Survey. July 2000

[5] Srivastava J., Cooley R., Deshpande M, Tan, P-N. Discovery and Applications of Usage Patterns from Web Data ACM SIGKDD Explorations Newsletter, 2000, 1(2) 12-23

[6] Mendez-Torreblanca, A. and Monte, M. A Trend Discovery for Dynamic Web Content Mining, IEEE, Inteligence System, Vol 14, pages.20-22, 2002.

[7] Crimmins, F., Smeaton, A. F., Dkaki, T. and Mothe, J. TetraFusion: information discovery on the Internet. Journal of IEEEexpert, pp 55-62, July 1999.

[8] Kushmerick, N. Gleaning Answers from the Web. IEEE Intelligent Systems. Vol. 14, No. 2, pp. 20-22,1999.

[9] Bing Liu, Kevin Chen-Chuan Chang, Editorial Issue on Web Content Mining, issue2, 2004.

[10] Muslea, I., Minton, S. and Knoblock, C. A Hierarchical Approach to Wrapper Induction. Agents-99, 1999.

[11] Freitag, D and McCallum, A. Information Extraction with HMM Structures Learned by Stochastic Optimization. AAAI-00, 2000.

[12] Chang, C-H., Lüi, S-L. IEPAD: Information Extraction Based on Pattern Discovery. WWW-10, 2001.

[13] Liu, B., Grossman, R. and Zhai, Y. Mining Data Records in Web Pages. KDD-03, 2003.

[14] Arasu, A. and Garcia-Molina, H. Extracting Structured Data from Web Pages. SIGMOD-03, 2003.

[15] Lerman, K., Getoor L., Minton, S. and Knoblock, C. Using the Structure of Web Sites for Automatic Segmentation of Tables. SIGMOD-04, 2004.

[16] Tyagi N., Gupta S.K. (2018) Web Structure Mining Algorithms: A Survey. In: Aggarwal V., Bhatnagar V., Mishra D. (eds) Big Data Analytics. Advances in Intelligent Systems and Computing. vol 654. Springer, Singapore.

[17] Johnson, F., Gupta, S. K., Web Content Mining Techniques: A Survey, International journal of computer applications (0975-888), vol. 47, no. 11, June 2012.

[18] Agrawal, R., Srikant, R. On Integrating Catalogs. WWW-01, 2001.
[19] Bergman, M. K. The Deep Web: Surfacing Hidden Value. Technical report, BrightPlanet LLC, Dec. 2000

[20] Chuang, S.-L. and Chien, L.-F., A Practical Web-based Approach to Generating Topic Hierarchy for Text Segments. CIKM-04, 2004.

[21] Yi, L., and Liu, B. Web Page Cleaning for Web Mining through Feature Weighting IJCAI-03, 2003.

[22] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.