International Journal of Science and Research (IJSR)
ISSN: 2319-7064
ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426

Web Mining Techniques

Neha Doomra[1], Rashmi Verma[2]

[1]Student of Master of Engineering in (CSE), DPGITM, Gurugram, India

[2]H.O.D at department of (CSE), DPGITM, Gurugram, India

Abstract: ***The Internet is a worldwide, publicly accessible series of interconnected computer networks that transmit data. Internet today is called as "Information Super Highway", the web purports to provide access to vast quantities of information..The wide adoption of the Internet has fundamentally altered the ways in which we communicate, gather information, conduct businesses and make purchases. As the use of the World Wide Web and email skyrocketed, computer scientists and physicists rushed to characterize this new phenomenon. While initially they were surprised by the tremendous variety the Internet demonstrated in the size of its features, they soon discovered a widespread pattern in their measurements: there are many small elements contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others. The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. The long-term success of the WWW depends upon fast response time. People use the web to access information from remote sites, but do not like to wait long for their results. A widely cited study from Zonal Research provide evidence for the "eight second rule" – in electronic commerce which states that, if a web site takes more than eight seconds to load, the user is much more likely to become frustrated and leave the site. The latency of retrieving a web document depends upon several factors, such as network bandwidth, propagation time, and the speed of the server and client computers. For reducing user perceived latency, many researchers have been working constantly***

Keywords: Data mining, Web mining, Web Content Mining, Web Structure Mining, Web Usage Mining

1. Introduction

The WWW has become a huge, diverse, and dynamic information reservoir accessed by people with different backgrounds and interests. On the Web, access information is generally collected by Web servers and recorded in the access logs. Web mining and user modeling are the techniques that make use of these access data, discover the surfer's browsing patterns, and improve the efficiency of Web surfing .Web mining can be generally defined as the use of data mining techniques to automatically discover useful knowledge from the Web. It is a converging area from several research communities such as Information Retrieval, Database, Machine Learning, and Natural Language Processing. Web mining is the application of data mining techniques to extract useful knowledge from web data that includes web documents, hyperlinks between documents, usage logs of web sites, etc. This technique enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc . As there is large amount of data present in web pages, the World Wide Web Data Mining may include content mining, hyperlink structure mining, and usage mining. Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. Depending on the location of the source, the type of collected data differs .It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are Unlabeled, Distributed, Heterogeneous (mixed media) ,Semi structured, Time varying, High dimensional. Therefore, web mining basically deals with mining large and hyper-linked information base having the aforesaid characteristics. Also, being an interactive medium, human interface is a key component of most web applications. Some of the issues which have come to light, as a result, concern

1) Need for handling context sensitive and imprecise queries;
2) Need for summarization and deduction;
3) Need for personalization and learning.

Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research, mainly because of the aforesaid characteristics of the data and human related issue.

2. Web Mining Component and Methodology

Web mining can be viewed as consisting of four tasks, shownin **Fig. 1** each task is described belowalong with a survey of the existing methodologies/tools for thetask.

1) Information Retrieval (IR) (Resource Discovery): Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non relevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

An index is, basically, a collection of terms with pointers to places where the information about documents can be found. However, indexing of web pages to facilitate retrieval is quite a complex and challenging problem as compared to the corresponding one associated with classical databases where straightforward techniques suffice. The enormous number of pages on the web, their dynamism, and frequent updating make the indexing techniques seemingly impossible. At

present, four approaches to index documents on the web are human or manual indexing, automatic indexing, intelligent or agent-based indexing, and metadata-based indexing.

Search engines are programs written to query and retrieve information stored in databases (fully structured), HTML pages (semi structured), and free text (unstructured) on the web. The most popular indexes have been created by web robots such as AltaVista and WebCrawler which scan millions of web documents and store an index of the words in the documents. There are over a dozen different indexes currently in active use, each with a unique interface and a database covering a different fraction of the web.

2) Information Selection/Extraction and Preprocessing: Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content. Until now, the major methods of IE involve writing *wrappers* (hand coding) which map the documents to some data model. Information integration systems operate by interpreting the various sites as knowledge sources and extracting information from them.

3) Generalization: In this phase, pattern recognition and machine learning techniques are usually used on the extracted information. Most of the machine learning systems, deployed on the web, learn more about the user's interest than the web itself. A major obstacle when learning about the web is the labeling problem: data is abundant on the web but it is unlabeled. Many data mining techniques require inputs labeled as positive (yes) or negative (no) examples with respect to some concept. For example, if we are given a large set of web pages labeled as positive and negative examples of the concept *homepage*, then it is easy to design a classifier that predicts whether any unknown web page is a home page or not; unfortunately web pages are unlabeled. Techniques such as uncertainty sampling reduce the amount of unlabeled data needed, but do not eliminate the labeling problem. An approach to solve this problem is based on the fact that the web is much more than just a linked collection of documents, it is an interactive medium.

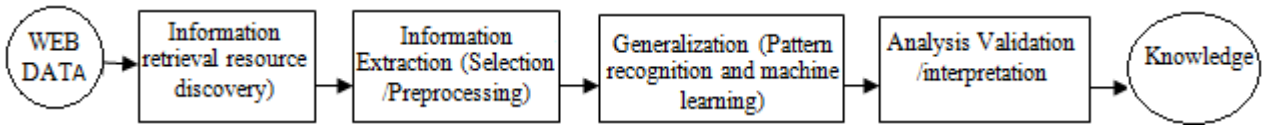

Figure 1: Web mining subtask

3. Overview of Mining

A. Data mining

Data mining is the process of analyzing large amounts of data in order to discover patterns and other information. It is typically performed on databases, which store data in a structured format. By "mining" large amounts of data, hidden information can be discovered and used for other purposes. Data mining is used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. The data is integrated and cleaned so that the relevant data is retrieved..

B. Web Mining

Web is collection of inter – related files on one or more web servers. "*Mining*" literally means the operations involved in digging for hidden treasures. Similarly data mining is used for the operations involved in digging out critical information from within an organization stored data for better decision support. It is a nontrivial process of extracting implicit, previously unknown and potentially useful patterns from large database. Web mining can be generally defined as the application of data mining techniques to extract useful knowledge from the Web Data. Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services . Web mining is the application of data mining techniques to extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. Web mining can be further categorized as web content that includes text, images, record etc, web structure which includes hyperlinks, tags etc, and web usage including http logs, app server logs etc. **Figure 2** shows the taxonomy of the Web mining

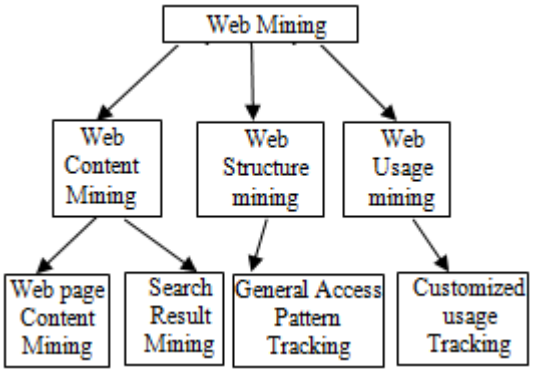

Figure 2: The taxonomy of the Web mining

1) Web Content Mining

Web content mining describes the discovery of useful information from the web contents/data /documents. However, what consist of web content could encompass a very broad range of data. Previously the internet consists of different type of services and the data sources such as Gopher, FTP and Usenet. Now most of those data are either ported to or accessible from the web. Basically the web content consists of several types of data such as textual image, audio, video, meta data as well as hyperlinks. The web content data consist of unstructured data such as free text, semi-structured data such as html documents, and a more structured data such as data in the tables or database generated HTML pages. However much of the web content data is unstructured text data. The research around applying data mining techniques to unstructured text is termed

Knowledge discovery in text, or text data mining, or text mining

2) Web Structure Mining

Web structure mining is the process of discovering structure information from the web. It tries to discover the model underlying the link structure of the web. The model is based on the topology of hyperlinks with or with out the description of the link. 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. 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. This model can be used to categories web pages and is useful to generate information such as similarity and relationship between different web sites. Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

3) Web Usage Mining

Web usage mining tries to make sense of the data generated by the web surfer's sessions or behaviors. While the web content and structure mining utilize the real or primary data on the web .web usage mining mines the secondary data derived from the interaction of the users while interacting with web. The web usage data include the data from the web server logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries and any other data as the result of interaction. Typically, the usage mining is defined as a three-phase process: data preprocessing, pattern discovery, and pattern analysis. **Fig 3**demonstrates such architecture. In this section, we present an overview of the detailed process.

- **Data Preprocessing**: retrieves raw data from the Web resources, and automatically selects and preprocesses the retrieved data. It includes any kind of transformation of the original raw data .
- **Pattern Discovery:** discovers knowledge from the pre-processed data. Machine learning and data mining procedures are carried out at this stage.
- **Pattern Analysis and Applications:** validates and post-processes the discovered patterns.

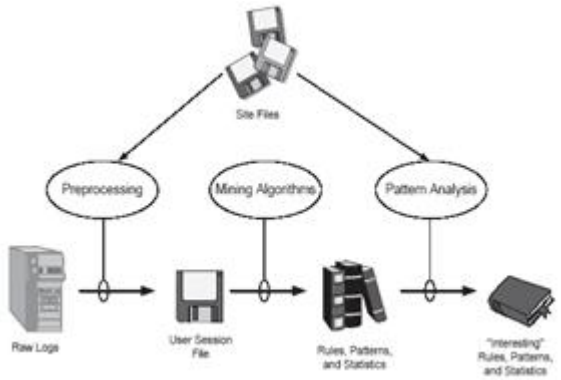

Figure 3: Web Usage Mining

4. Challenges in Web Mining

The web poses great challenges for resource and knowledge discovery based on the following observations −

- **The web is too huge** − The size of the web is very huge and rapidly increasing. This seems that the web is too huge for data warehousing and data mining.
- **Complexity of Web pages** − The web pages do not have unifying structure. They are very complex as compared to traditional text document. There are huge amount of documents in digital library of web. These libraries are not arranged according to any particular sorted order.
- **Web is dynamic information source** − The information on the web is rapidly updated. The data such as news, stock markets, weather, sports, shopping, etc., are regularly updated.
- **Diversity of user communities** − The user community on the web is rapidly expanding. These users have different backgrounds, interests, and usage purposes. There are more than 100 million workstations that are connected to the Internet and still rapidly increasing.
- **Relevancy of Information** − It is considered that a particular person is generally interested in only small portion of the web, while the rest of the portion of the web contains the information that is not relevant to the user and may swamp desired results.

5. Conclusion

In this paper, we have studied the web mining technique, which is the actually the application of data mining techniques to discover patterns from the World Wide Web including Web documents, hyperlinks between documents, usage logs of web sites, etc. Now in today advanced world, web becomes an important part of many of all organizations, businesspersons and daily individuals. As a web data is of very much different formats, we have studied the characteristics of web data. As it is very much important to mine particular data from web,

References

[1] K.D. Satokar, S.Z.Gawali, "Web Personalization Using Web Mining", International Journal of Engineering Science and Technology, Vol. 2, Issue 3, 2010, pages 307-311.
[2] K.Poongothai, M.Parimala, Dr. S.Sathiyabama, "Efficient Web Usage Mining with Clustering", International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, pages 203-209, 2011.
[3] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the Word Wide Web", Proceedings ICTAI, 1997.
[4] JiaweiHan and MichelineKamber, "Data mining, concept and techniques", http://www.cs.sfu.ca
[5] J. Srivastava, P. Desikan and V. Kumar, "Web Mining: Accomplishments & Future Directions", National Science Foundation Workshop on Next Generation Data Mining, 2002.
[6] Manoj Pandia, Subhendu Kumar Pani, Sanjay Kumar Padhi, Lingaraj Panigrahy, R.Ramakrishna, "A Review Of Trends In Research On Web Mining", International Journal of Instrumentation, Control & Automation, Volume 1, Issue 1, 2011, pages 37-41.
[7] Li Mei, Feng Cheng, "Overview of Web Mining Technology and Its Application in E-commerce", 2nd

International Conference on Computer Engineering and Technology, 2010, Volume 7, pages 277-280
[8] http://web.syr.edu/~dxing/Files/DataMining/WebMining.pdf
[9] https://en.wikipedia.org/wiki/Web_mining.