A Study of Web Mining

Prashant Sharma[1],

[1]Department of Information Technology, Medicaps Institute of Technology, Indore-452001

[1]Prashant.sharma.cse17@gmail.com

Abstract

The hunt for knowledge has led to new discoveries and inventions. With the materialization of World Wide Web, it became a core for all these discoveries and inventions. Web browsers became a tool to make the information available at our finger tips. WWW stands for World Wide Web, and it is an advanced information retrieval system. As years passed World Wide Web became weighed down with information and it became hard to retrieve data according to the need. Web mining came as salvage for the above problem. Web content mining is a subdivision under web mining. This paper deals with a study of different techniques and pattern of content mining and the areas which has been influenced by content mining. The web contains structured, unstructured, semi structured and multimedia data. This survey focuses on how to apply content mining on the above data. It also points out how web content mining can be utilized in web usage mining.

Keywords

Web Content mining, Web Usage Mining, Structured Data, Unstructured Data, Semi-structured Data, and Multimedia Data.

1. INTRODUCTION

The advancement in the technology covered faster communications. The previous decade experienced a dramatic development in computer technology, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet. And this technique was named as web mining. Web mining is further classified into three: They are Web content mining, Web Structure mining, Web Usage mining [1]. Using the objects like text, pictures, multimedia etc. content mining is done in the web. In Web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users. And the study of this navigational pattern will trace out the interest of the users [1].

The World Wide Web (Web) is popular and interactive medium to disseminate information today. [2]The Web is huge, diverse, dynamic, widely distributed global information service center. Users could encounter following problems when interacting with the Web:

a) Finding relevant information

Most people use some search service when they want to find specific information on the Web. A user usually inputs a simple keyword query and a result is a list of ranked pages.[2] This ranking is based on their similarity to the query. Today's search tools have some problems: Low precision and low recall, mainly because of wrong or incomplete keyword query. This leads to irrelevance of many search results.

b) Creating new knowledge

This problem is data-triggered process that presumes that we have a collection of Web data and we want to extract potentially useful knowledge from these data [2].

c) Personalisation of information

People differ in the contents and presentations they prefer while interacting with the Web.

d) Learning about consumers or individual users

This is a group of sub-problems such as mass customizing information to intended consumers, problems related to effective Web site design and management, problems related to marketing and others.

The structure of the paper is as follows: Section 2 presents the overview of web mining, web content mining, Section 3 deals with Conclusion, section 4 present the future scope & section 5 represents the references.

2. OVERVIEW OF WEB MINING

Web mining means to discover the information from World Wide Web and it also find out its usage patterns [3].Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining should be decomposed into these subtasks:

- **Resource finding:** The task of retrieving intended Web documents.
- **Information selection and pre processing:** Automatically selecting and pre processing specific information from retrieved Web resources.
- **Generalization:** Automatically discovers general patterns at individual Web sites as well as across multiple sites.
- **Analysis:** Validation and/or interpretation of the mined patterns.

Resource finding is the process of retrieving data from text sources available on the Web such as electronic magazines and newsletters or text contents of HTML documents. Information selection and pre processing step is transformation process retrieved in information retrieval (IR) process from original data. These transformations cover removing stop words, finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc. Data mining techniques and machine learning are often used for generalization. In information and knowledge discovery process, people play very important role. This is important for validation and/or interpretation in last step.

2.1 CATEGORIES OF WEB MINING

Web mining is categorized into three areas of interest based on part of Web to mine:

1. Web content mining

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

2. Web structure mining

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

3. Web usage mining

- Mines secondary data derived from user interactions

2.1.1 WEB CONTENT MINING

Web content mining is the process of extracting useful information from the content of Web documents. Logical structure, semantic content and layout are contained in semi-structured Web page text. Topic discovery, extracting association patterns, clustering of Web documents and classification of Web pages are some of research issues in text mining. These activities use techniques from other disciplines - IR, IE (information extraction), NLP (natural language processing) and others [3] [7]. Automatic extraction of semantic relations and structures from Web is a growing application of Web content mining. In this area, several algorithms are used: Hierarchical clustering algorithms on terms in order to create concept hierarchies, formal concept analysis and association rule mining to learn generalized conceptual relations and automatic extraction of structured data records from semi-structured HTML pages. Primary goal of each algorithm is to create a set of formally defined domain ontologies that represent Web site content. Common representation approaches are vector-space models, descriptive logics, first order logic, relational models and probabilistic relational models. Structured data extraction is one of most widely studied research topics of Web content mining. Structured data on the Web are often very important as they represent their host pages' essential information. Extracting such data allows one to provide value added services, e.g. shopping and meta-search. In contrast to unstructured texts, structured data is also easier to extract. This problem has been studied by researchers in AI and database and data mining.

2.1.2 WEB STRUCTURE MINING

Web structure mining uses the hyperlink structure of the Web to yield useful information, including definitive pages

specification, hyperlinked communities identification, Web pages categorization and Web site completeness evaluation. Web structure mining can be divided into two categories based on the kind of structured data used:

1. **Web graph mining:** The Web provides additional information about how different documents are connected to each other via hyperlinks. The Web can be viewed as a (directed) graph whose nodes are Web pages and whose edges are hyperlinks between them.

2. **Deep Web mining:** Web also contains a vast amount of noncrawable content. This hidden part of the Web is referred to as the deep Web or the hidden Web [3] [6]. Compared to the static surface Web, the deep Web contains a much larger amount of high-quality structured information. Most of mining algorithms, that are improving the performance of Web search, are based on two assumptions.

(a) Hyperlinks convey human endorsement. If there exists a link from page A to page B, and these two pages are authored by different people, then the first author found the second page valuable. Thus the importance of a page can be propagated to those pages it links to.

(b) Pages that are co-cited by a certain page are likely related to the same topic, opularity or importance of a page is correlated to the number of incoming links to some extendt, and related pages tend to be clustered together through dense linkages among them.

Web information extraction has the goal of pulling out information from a collection of Web pages and converting it to a homogeneous form that is more readily digested and analyzed for both humans and machines [4]. The result of IE could be used to improve the indexing process, because IE removes irrelevant information in Web pages and facilitates other advanced search functions due to the structured nature of data.

It is usually difficult or even impossible to directly obtain the structures of the Web sites' backend databases without cooperation from the sites. Instead, the sites present two other distinguishing structures: Interface schema and result schema. The interface schema is the schema of the query interface, which exposes attributes that can be queried in the backend database. The result schema is the schema of the query results, which exposes attributes that are shown to users.

2.1.3 Web Usage Mining

and related pages tend to be clustered together through dense linkages among them. Web information extraction has the goal of pulling out information from a collection of Web pages and converting it to a homogeneous form that is more readily digested and analyzed for both humans and machines. The result of IE could be used to improve the indexing process, because IE removes irrelevant information in Web pages and facilitates other advanced search functions due to the structured nature of data [4]. It is usually difficult or even impossible to directly obtain the structures of the Web sites' backend databases without cooperation from the sites. Instead, the sites present two other distinguishing structures: Interface schema and result schema. The interface schema is the schema of the query interface, which exposes attributes that can be queried in the backend database. The result schema is the schema of the query results, which exposes attributes that are shown to users.

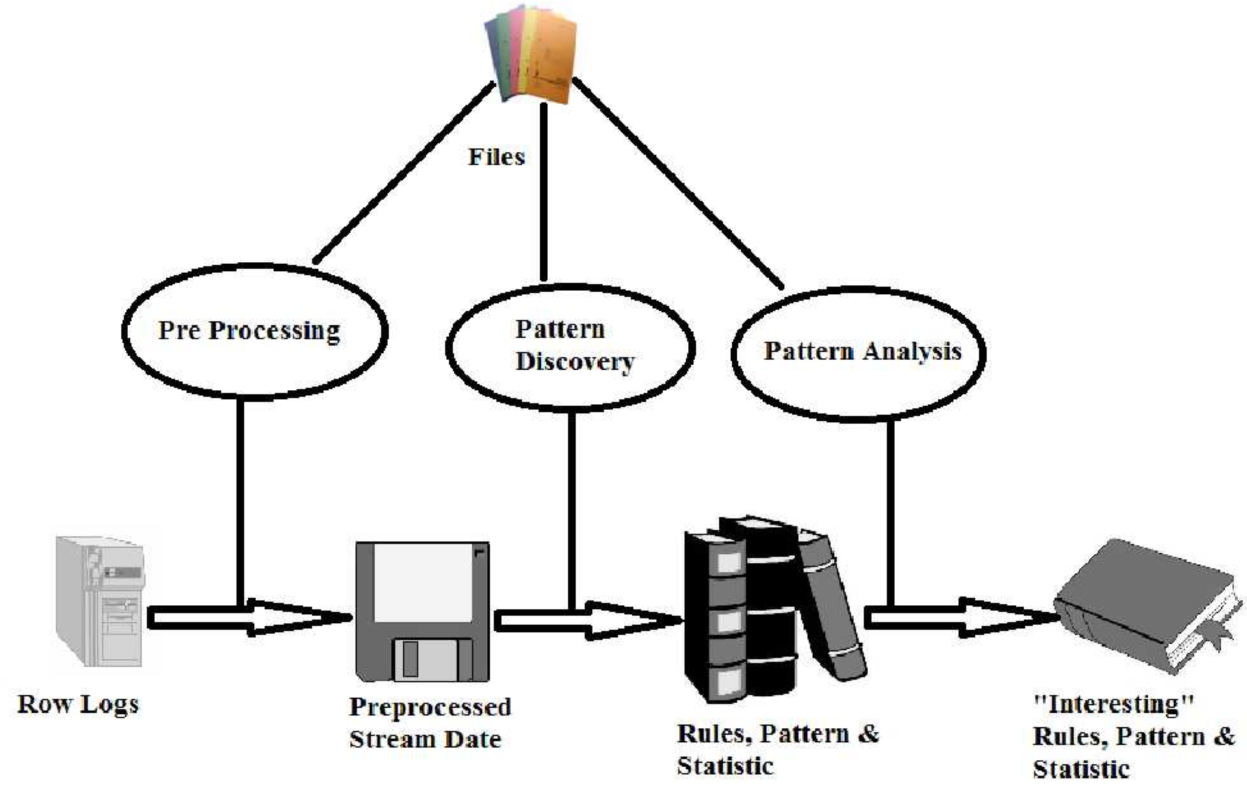

Figure 1: the process of Web Usage Mining

Pre processing is first phase of Web mining process. Usage, content and structure information contained in the various available data sources are converted for next step that is pattern discovery. Pattern discovery is based on methods and algorithms developed from several areas such as data mining, machine learning and pattern recognition [5]. This is used for understanding how users use some Web site. Pattern analysis is the final step in Web usage mining process. In this phase, uninteresting rules or patterns from the set found in pattern discovery are filtered. It turns discovered patterns, rules and statistics into knowledges. Knowledge query mechanism such as SQL is a form of pattern analysis. Loading usage data into a data cube in order to perform OLAP operations is another method. Highlighting overall patterns or trends in the data is usually done by some visualization technique, such as graphing patterns or assigning colors to different values.

3. CONCLUSION

This paper describes Web mining. It is an application of data mining techniques to extract knowledge from the content, structure, and usage of Web data sources. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web structure mining is the process of inferring knowledge from the 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. There are many applications of these techniques, for example search engines, Web analysis, Web agents, personalization services etc. New modifications and extensions of these techniques should be next topics in this area of research.

4. FUTURE SCOPE

This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce. This phenomenon partly creates confusion when we ask what constitutes Web mining and when comparing research in this area. This trend is likely to continue as Web services continue to flourish. As the Web and its usage grow, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value to be realized.

5. REFENCES:

[1] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.

[2] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648. Morgan Kaufmann Publishers, 2006.

[3] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, Second edition, p. 628-648.

Morgan Kaufmann Publishers, 2006.

[4] Srivastava, J., Cooley, R., Deshpande, M., Tan, P.: Web Usage Mining: Discovery and Applications of Usage [5] Patterns from Web Data. SIGKDD xplorations, 2000. Paper available on http://www.acm.org/sigs/sigkdd/explorations/issue1-2/srivastava.pdf (January 2007).

[5] Ahmed, S. S., Halim, Z., Blaig, R. and Bashir, S. 2008. Web Content Mining: A Solution to Consumers Product Hunt. International Journal of Social and Human Sciences 2, 6-11.

[6] Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining. World Academy of Science, Engineering and Technology 49.

[7] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. Tapping into the Power of Text Mining. Communications of the ACM - Privacy and Security in highly dynamic systems. Vol. 49, Issue-9.

[8] Bharanipriya, V. and Prasad, K. 2011. Web content Mining Tools: A Comparative study. International Journal of Information Technology and Knowledge Management. Vol. 4. No 1,211- 215.

[9] Cooper, M., Foote, J., Adcock, J. and Casi, S. 2003. Shot Boundary Detection via Similarity Analysis. In Proceedings of TRECVID 2003 workshop.

[10] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.