Ultra Scientist Vol. 26(3)B, 243-248 (2014).

A Review on Web Mining

NEETU SAHU and PRAGYESH KUMAR AGRAWAL

Department of Physics Govt. Nutan Girls College, Bhopal (India)

(Acceptance Date 2nd November, 2014)

Abstract

Web Mining – i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage is the collection of technologies to fulfill this potential. Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). This paper provides a brief overview of the accomplishments of the field, both in terms of technologies and applications. In this paper, we first introduce the concepts related to web mining. The aim of this paper is to give overview of web data mining categories.

Index Terms—Web Mining, Web Content Mining, Web Structure Mining, and Web Usage Mining.

I. Introduction

Due to rapid growth of internet, websites appear and disappear, contents are modified and mastering of their organization becomes very difficult. Web Mining is a subset of Data Mining. Web mining is an application of data mining field, which extracts interesting and potentially useful patterns and hidden information from web documents and web activities[1,2].

II. Web Mining :

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following subtasks[3,4]:

1 Resource finding: the task of retrieving intended web documents.

2. Information selecting and pre-processing specific information from retrieved Web resources.

3. Generalization: automatically discovers general patterns from individual website as well as across multiple sites.

4. Analysis: validation and interpretation of the mined patterns.

III. Web Mining Classification :

Web mining can be categorized into three Types[5] as follows:

a) Web Content Mining
b) Web Structure Mining
c) Web Usage Mining

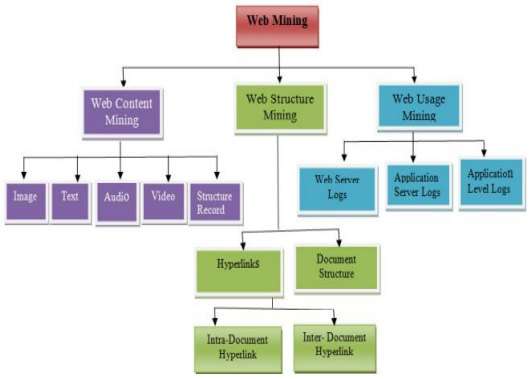

Fig. 1. Classification of Web Mining

a) Web Content Mining :

Web content mining is the automatic search of information resources available online[6]. It mines the E-contents of Web pages. It is a process of discovery of useful information from the web contents/ data/ documents. Web content consists of several types of data such as text, image, audio, video, metadata as well as hyperlinks. Currently research on mining different types of data is called multimedia data mining[4].Web content mining is differentiated from two different points of view: Information Retrieval View and Database View. Kosala *et al*. summarized the research works done for unstructured data and semi-structured data from information retrieval view[7]. Two main approaches are used in Web Content Mining:

(1) Unstructured text mining approach, and
(2) Semi-structured and Structured mining approach[8].

Unstructured Text Data Mining: Web content data is mainly unstructured text data. The research around applying data mining techniques to unstructured text is termed as Knowledge Discovery in Texts (KDT), or text data mining, or text mining[9].

Semi-structured and Structured Data Mining: Semi-structured data is a type of structured data that does not resemble to the formal structure of data models associated with relational databases or other forms of data tables. This type of data contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data.HTML is an uncommon instance of such intra-record structure[10]. The systems utilized for semi organized information mining are

- Object Exchange Model (OEM),
- Top Down Extraction, and
- Web Data Extraction dialect[8]

b) Web Structure Mining :

It mines the useful knowledge from hyperlinks that show the structure of the web. We can search useful web page which is the key technology used in search engines[9]. The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure

information from the web[11]. Based on kind of structure information used, structure mining can be divided into two types.

Hyperlinks :

A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different one. A hyperlink that connects to a different part of the same page is called an **Intra-Document Hyperlink**, whereas a hyperlink connecting two different pages is called an **Inter-Document Hyperlink**[12].

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. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents[12,13].

c) Web Usage Mining :

Web usage mining focuses on techniques that would predict user behavior where user interacts with Web[14]. As mentioned before, the deep-mined information in this class is the secondary information on Web because the results of interactions. These information might vary terribly wide however usually we tend to might classify them into the usage information that reside within the internet purchasers, proxy servers and servers[15]. The Web usage mining method may be classified into two ordinarily used approaches[16]. It may be considered as a three-phase method, consisting of the info preparation, pattern discovery and pattern analysis phases. Within the first phase, diary information is preprocessed so as to spot users, sessions, page views, and so on. Within the second phase, applied math strategies or data processing strategies (such as association rules, successive pattern discovery, clustering, and classification) are applied so as to notice fascinating patterns. These patterns are kept so that they can be additionally analyzed within third phase of online usage mining method[17]. Web usage mining itself can be classified further depending on the kind of usage data considered[11]:

Web Server Data :

User logs are collected by the web servers which typically include IP address, page reference and access time[11].

Application Server Data :

Commercial application servers such as Web logic and Story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and list them in application server logs[11].

Application Level Data :

New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories[11].

IV. Web Mining Processes :

There are four step involved in web mining process as depicted in Fig. 2.

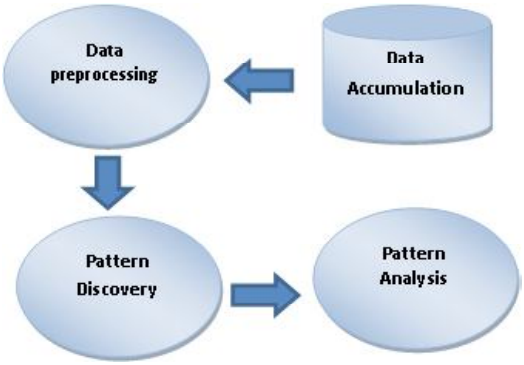

Fig. 2. Web Mining Process

They are:

a) Data Accumulation
b) Data Preprocessing
c) Pattern Discovery
d) Pattern Analysis

a) Data Accumulation :

In this step we collect data from different sources like web pages related to e-commerce. The main task is to get data from Web document[1]. Data may be obtained from e-mails, electronic documents, news group, weblog files and data transaction data in the databases.

b) Data preprocessing :

The real data that is to be composed generally have the features like employment and certainty. For mining the knowledge more successfully, preprocessing of the collected data is essential. Preprocessing provides precise, brief data for data mining. Preprocessing of Data, includes data cleaning, user recognition, user sessions recognition, access path supplement and transaction identification[18].

c) Pattern Discovery :

Pattern discovery gives us useful and ultimately explicable information and knowledge using mining algorithm. Some methods are classification analysis, association rule discovery, sequential pattern discovery, clustering analysis, and dependency modeling[1].

d) Pattern Analysis :

Pattern analysis is mostly concerned with selecting pattern we are interested in from the pattern set found by model pattern discovery algorithm. Main purpose is to search out a helpful model, rules and modes. We can get graphical user interface using visualization techniques for users[19].

V. Key Concepts :

In this section we briefly describe the key new concepts introduced by the Web mining research community.

Ranking metrics - for page quality and relevance:

Searching the Web involves two main steps: Extracting the relevant pages to a query and ranking them according to their quality. Ranking is important as it helps the user look for “quality” pages that are relevant to the query. Different metrics have been proposed

to rank Web pages according to their quality. We briefly discuss two of the prominent metrics[11].

- **Page Rank**: Page Rank is a metric for ranking hypertext documents based on their quality. The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined. The rank of a page *p* can thus be written as[12]:

$$PR(p) = d/n + (1-d) \sum_{(q,p)\varepsilon G} \frac{PR(q)}{outdegree(q)}$$

Here, n is the number of nodes in the grap and outdegree(q) is the number of hyperlinks on page q . Intuitively, the approach can be viewed as a stochastic analysis of a random walk on the Web graph. The first term in the right hand side of the equation corresponds to the probability that a random Web surfer arrives at a page p by typing the URL or from a bookmark, or may have a particular page as his/her homepage. Here, d is the probability that a random surfer chooses a URL directly, rather than traversing a link and 1-d is the probability that a person arrives at a page by traversing a link. The second term in the right hand side of the equation corresponds to the probability of arriving at a page by traversing a link.

VI. Conclusion

A survey of researches in the area of web mining has been presented in this paper. Three recognized types of web data mining are introduced generally. Possible applications and modifications have also been discussed. This review aims at providing a boost in this field by the introduction of newer concepts and approaches.

References

1. Jyoti Yadav, Bhawna Mallick, "Web Mining: Characteristics and Application in E-Commerce", Volume *1,* 2012, *IJECSE*
2. Li Mei , Feng Cheng , "Overview of WEB Mining Technology and Its Application in E-commerce*", 2010, IEEE*.
3. Srivastava, J., Cooley, R., Deshpande, M., and Tan, P-N. (2000) "Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, 1(2), 12-23.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch.4.
4. Raymond Kosala, Hendrik Blockeel "Web Mining Research: A Survey" Volume *2, Issue I,* July (2000).
5. Arun K Punjari "Data Mining Techniques", (2001).
6. S.K. Madria, S.S. Bhowmick, W. K., Ng, and E.P. Lim. Research issues in web data mining. In Proceedings of data ware housing and knowledge Discovery, first International conference, *DaWak'99*, pages 303-312 (1999).
7. Monika Yadav, Mr. Pradeep Mittal "Web Mining: An Introduction", Volume *3,* Issue *3,* March (2013).
8. Johnson, F., Gupta, S. K., "Web Content Minings Techniques: A Survey", International Journal of Computer Application. Volume *47* – No.*11*, p44, June (2012).
9. Abdelhakim Herrouz, Chabane Khentout Mahieddine Djoudi "Overview of Web

Content Mining Tools", Volume *2,* Issue *6,* (2013).

10. Mr. Akshay A. Adsod, Prof. Nitin R. Chopde," A Review on: W*eb Mining Techniques", Volume 10,* No. *3,* Apr (2014).
11. Mr. Dushyant Rathod," *A Review On Web Mining",* Volume *1,* Issue *2,* April (2012).
12. Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining - Concepts, Applications & Research Directions" chapter 3
13. JC.H. Moh, E.P. Lim, and W.K. Ng. DTD-Miner: A Tool for Mining DTD from XML Documents. WECWIS, (2000).
14. C.H. Moh, E.P. Lim, and W.K. Ng. DTD-Miner: A Tool for Mining DTD from XML Documents. WECWIS, (2000).
15. Arun Kumar Singh, Dheeraj Sharma, Avinav Pathak, "*Web Usage Mining*: A Concise Survey on Tools and Applications", Volume *74,* No.*1,* July (2013).
16. J. Srivastava, R. Cooley, M. Deshpande, P. Tan, Web usage mining: Discovery and applications of usage patterns from web data' SIGKDD Explorations newsletter, *1(2),* 12-23 (2000).
17. J. Borges, and M. Levene, Data Mining of User Navigation Patterns', Web Usage Analysis and User Profiling, San Diego, CA, *USA,* 31-39 (2000).
18. M. Eirinaki and M. Vazirgiannis, Web mining for web Personalization, *ACM Transactions on Internet Technology, 3(1),* 1-27 (2000).
19. S.Yadav, K. Ahanad, J. Shekar, *"Analysis of web mining applications Beneficial areas"* Volume *12* (2011).