Web Mining-Types,Applications,Challenges and Tools

Ms. Gaikwad Surekha Naganath , Ms. Mali Supriya Pralhad

***Abstract*— The World Wide Web is the area where extremely huge amount of information is available and the useful information is mined. Web mining is process of discovering and extracting useful information from extremely large web data. The web is rapidly updating and expanding. In such case web mining is becoming a challenging task. It has to handle different communities, different external interfaces etc. In this paper we are focusing on web mining process and its types. We also discuss these types in detail, technologies used for it and applications of it. This paper also covers the challenges and limitations of web mining and tools used for it.**

***Index Terms*— Data Mining, Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, User Profiles.**

Data Mining:
In atomization and computerization huge data repositories are used. Finding knowledge form raw data available in huge data repositories using some patterns is a challenging task. Data mining refers to the discovering or extracting knowledge from extremely large data repositories by using patterns. Different techniques are available for data mining such as Classification, Sequential Patterns, Clustering, Regression, Deviation Detection, Association Rules etc. Data mining technique has its application in web mining [3].

Web Mining:
Web mining is the application of data mining technique. The data mining and web mining differs in following points:

- In data mining, processing records from databases is large work as compared to processing web pages in web mining.
- Data mining is related to structural data processing whereas web mining has to process semi-structured or even unstructured data.
- In data mining data to be mined is generally private whereas data is public in web mining [7].

Web mining is the process of extracting and discovering knowledge from web data. Web data consists of:

- Web content –text, image, records, etc.
- Web structure –hyperlinks, tags, etc.
- Web usage –http logs, app server logs, etc

Web mining has 4 phases such as information collection, preprocessing, pattern discovery and pattern analysis as shown in figure 1.1.

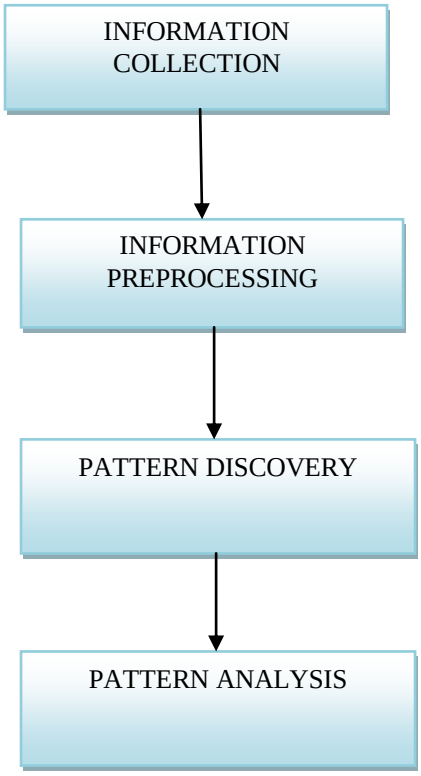

Fig.1.1 Web Mining Process

Web Mining Categories:
The web mining has following four categories as shown in figure 1.2:

- Web Content Mining
- Web Structure Mining
- Web Usage Mining
- User Profiles

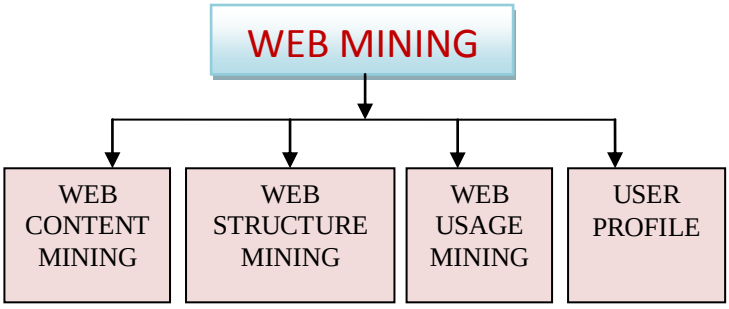

Fig.1.2 Categories of Web Mining

ISSN: 2278 – 1323

Web Content Mining:
Web content mining refers to the process of discovering and extracting useful information from contents of web documents. The contents of web page may consists of text, audio, video, images, links and structured records such as tables ,lists, graphs etc. The content mining has several steps such as collect the information, parse it, analyze it and then produce the useful result. This type of web mining is also called as Web Text Mining because text content is most popularly extracted. The two technologies are generally used for this type of mining: IR (Information Retrieval) and NLP (Natural Language Processing). The information retrieval should be performed from two views because the web page may contain semi-structured data or unstructured data. The most popular and basic techniques used in this web mining are: Classification, Association Rules, Clustering, Relevance of content, Concept hierarchy creation, Topic identification, tracking and drift analysis. There exists significant work while extracting information from images present in web pages. It requires image processing related techniques. [1]
Applications:
Web content mining is widely used to identify the topics, for processing user related queries, categorize documents of web, search web pages from several different servers, to list relevant documents in a collection, to show the documents or to hide the documents based on some rank. [5]

Web Structure Mining:
The web structure mining is the process of discovering and extracting structural information from the web. This mining focuses on the structure within the web documents and structure of hyperlinks within the web. Mining the structure of web document mainly focused on extracting Document Object Model (DOM) of web document. The structure of web document is generally tree structured format that describes HTML or XML tags. The web document object may contain HTML or XML tags, word appearances, anchor text, links etc. Secondly, mining the structure of hyperlinks is related to discovering the hierarchy of hyperlinks present in the website of a particular domain. Hyperlink is nothing but the link which connects to the location within the same web page or to the location of another web page. The hyperlink that connects to the different parts within same page is called "intra-document hyperlink" and the hyperlink which connects to the two different web pages is called as "inter-document hyperlink". Discovering hyperlinks structure within the website is also called as "Hyperlink Analysis". The two algorithms are mostly used for this type of mining that are: Page Rank and HITS.[1]
Applications:
Web structure mining is mainly used for extracting the information such as which pages are duplicate pages, to find the relate pages, for deciding which page should be added to the collection, to categorize the web pages, for ranking the web pages etc.[5]

Web Usage Mining:
Web usage mining also called as web log mining is the process of discovering meaningful information from web data usage. It focuses on predicting the user behavior while surfing the web. The user interest, comparison between user expectations and availability and user's behavior is extracted via this type of web mining. This data is extracted from server logs. Web usage mining mainly focuses on serially analyzing the pages visited by user during a particular session, analyzing the web clicks etc. The automatically generated server logs, client-side cookies, referrer logs, agent logs, user profiles, meta data are typical resources for obtaining information related to web usage. [2]
Applications:
Web usage mining has its application in improving websites design, intrusion detection, for predicting user's interest, for analyzing website performance, for excellence of e-commerce, for identifying suspicious activities, to find out primary places for advertising, for social network analysis, for the analysis of traffic etc. [5]

User Profile:
This mining provides information about users of the web site. It consists of user registration data and profile information of every web user. It is used to understand the user's behavior, decision making etc. by recording click streams.

Limitations and challenges in web mining:
- As web is extremely huge and rapidly increasing it becomes a challenging task to mine the web.
- It becomes hard to handle unstructured, non-standard, heterogeneous and irregular data patterns.
- Organizing hardware and software for such a complex and extremely large processing is also challenging.
- The information source available is rapidly changing source which gives challenge for web mining.
- The web users belong to the different backgrounds, with different purpose and with different interest. Web mining needs to handle this diversity.
- Web mining has to handle variety of external interfaces. [5]

Web Mining Tools:
Different web mining tools are available for example:
- Web Content Extractor (WCE)
- Screen-scraper
- Web Info Extractor (WIE)
- Mozenda
- Automation Anywhere 6.1 (AA) [6]

Conclusion:
This paper describes the basic concept of data mining. Then we discussed process of web mining, its types in detail, the different techniques used for it and applications of each type of mining. Further we also discussed how web mining became challenging process, limitations of web mining and tools available for it.

References:

[1] Kosala, Raymond; Hendrik Blockeel (July 2000). "Web Mining Research: A Survey" (PDF).

[2] Wang, Yan. "Web Mining and Knowledge Discovery of Usage Patterns".

[3] Data Mining: What is Data Mining?,www.anderson.ucla.edu/faculty/jason.frand/teacher / technologies/palace/datamining.html

[4] Dunham, M.H.: Data Mining: Introductory and Advanced Topics. Prentice Hall, Pearson Education Inc. (2003)

[5] J. Srivastava , P. Desikan , V. Kumar, "Web Mining – Concepts, Applications and Research Directions", Studies in Fuzziness and Soft Computing, Volume 180, pp. 275–307, (2005).

[6] Bharanipriya, V., Prasad, V.K., Web Content Mining Tools: A comparative Study, International Journal of Information Technology and Knowledge Management. Vol. 4, No 1, pp. 211-215 (2011).

[7] M. Spiliopoulou. Data mining for the Web. In Proceedings of Principles of Data Mining and Knowl edge Discovery, Third European conference, PKDD'99, P588-589

First Author:
Ms.Gaikwad Surekha Naganath
BIGCE, Solapur University,
B. E. (I.T.),M.E.(C.S.E.-appeared)

Second Author:
Ms. Mali Supriya Pralhad
B.E.(C.S.E.)

ISSN: 2278 – 1323