www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 8 August, 2014,Page No. 7879-7881

Study of Web Content Mining and Its Tools

Dr.P.Sujatha[1], G.Thailambal[2], R.Sheela Angalin Ruby[3]

[1]Associate Professor, School of Computing Sciences,
Vels University, Chennai, India,
sujinagi@gmail.com

[2]Assistant Professor, School of Computing Sciences,
Vels University, Chennai, India,
aishusri2009@gmail.com

[3] Research Scholar, School of Computing Sciences,
Vels University, Chennai, India,
sheelarosi@gmail.com

Abstract: ***This paper presents the details of searching or extracting information from the web. It also discusses main tasks involved in web mining. It mainly focuses on the types of Web Content mining such as Unstructured, Structured and Semi-structured types. Finally, some tools that are used for mining is also focused in this paper.***

Keywords: Web Mining, Web Content Mining, Structured Data extraction, Unstructured Data Extraction.

1. Introduction

Data is the prime source of processing in Computer. It will be in many forms and in many Places around the world. Raw data are wasted if it is not processed and there are many ways to access the data for processing. Each way has its own hurdles in reaching users' knowledge. People get the data in the earlier days through the Newspaper and from Different Medias. Now-a-days the modern gadget Computer is used to quench the thirst of people Knowledge. WWW is the only popular medium to disseminate information to people. Many barriers in the work of retrieving pave the way to do research in searching the Web with the term 'WEB MINING'. Web pages are viewed using a browser. Deep web is accessing databases through queries. Web mining is divided into four tasks:

i) **Resource Finding:** Retrieving the document related to our search.

ii) **Information selection and Pre-processing:** Selecting an appropriate document from the displayed documents.

iii) **Generalization:** Discovering patterns from individual websites

iv) **Analysis:** Checking the validity of the mined patterns.

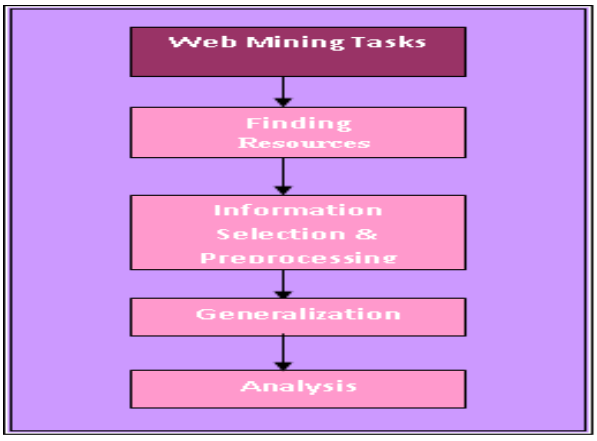

Figure 1: Tasks of Web Mining

Data mining is broadly classified as Web content mining (WCM), Web Structure Mining (WSM) and Web usage Mining (WUM). WCM is identifying data for the user from the text, image, audio or video data. WSM is identifying node and connection to find the structure of a web site. WUM is identifying user access patterns from user logs.

2. Related Works

Some of the related works of web content mining are discussed below.

Aidan Finn [4] discussed about Content classification for digital libraries and methods for content extraction from single source in which the content will be in a single body. Page is tokenized into either words or tags and then sectioned into 3 contiguous regions. To partition the document most tags are

placed into outside regions and word tokens into the center region. However this approach is well suited for single body documents and will not produce good results for multi body documents.

Mckeon detects largest body of text on a webpage by counting number of words. It is suitable for simple pages. But this algorithm does not produce accurate results for multibody documents such as random advertisement and image placement.

Rahman [7], [15] proposed a technique which uses structural analysis, contextual analysis, and summarization. In this technique, the structure of an HTML document is analyzed and decomposed into smaller subsections. Then the content of the individual sections is extracted and summarized. However, this proposal is yet to be implemented. This paper does not propose methods for content extraction and given only the prerequisites for doing so. For formatting web pages to fit on the small screens of cellular phones and PDAs, various approaches have been suggested. However, they conclude only reorganizing the content of the webpage to fit on a constrained device and require a user to scroll and hunt for content.

Buyukkokten [13-14] proposed a strategy in which a page can be shrunk or expanded like the instrument. Furthermore they also discussed a method to transform a web page into a hierarchy of individual content units called Semantic Textual Units [STU]. First, STUs are built by analyzing syntactic features of an HTML document, such as text contained within paragraph (<P>), table cell (<TD>), and frame component (<FRAME>) tags. These features are then arranged into a hierarchy based on the HTML formatting of each STU.

3. Classification of WCM:

Data in web may be in any form such as table as structured data, free-text as unstructured data and HTML documents as semi structured data. Web content mining is classified as two types IR (Agent-based) view and DB (Database) view. With the DB view the data is combined and organized in such a way that sophisticated queries can be used for searching data in a database. In IR view helps to retrieve information easily from a drawing source of web data. The main object used in Web Content Mining is "Text documents". The two main approaches in WCM are (a) Structured data extraction and Semi-structured (b) Unstructured data extraction.

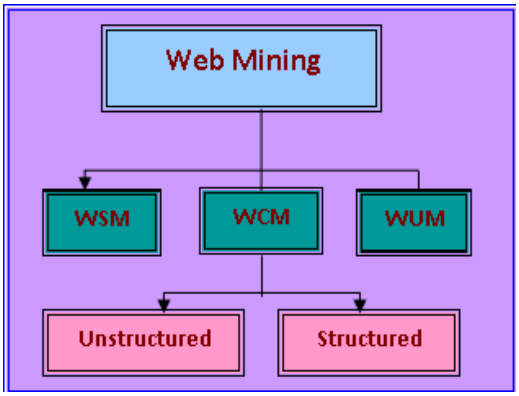

Fig.2 Classification of Web Mining

3.1 Structured data extraction:

A large amount of information on the Web is contained in regularly structured data objects. Such Web data records are important because they often display the essential information on their host pages, e.g., lists of products and services. Two approaches used are: Wrapper Induction and Automatic data extraction. Many techniques have been used to retrieve appropriate information from the page. First, Classification of websites on top of web page demands new algorithms developed. Secondly, crawling of web page gives a large number of relevant data, Clustering is a method used to group the set of related information for better understanding. Crawlers used to traverse through structured data divides into internal which traverses through internal pages and external web crawler which traverses through unknown website.

3.2 Semi-Structured data:

The techniques used are OEM (Object Extraction Model) in which a group is formed which contains relevant information and stored in OEM, top down extraction in which complex objects are converted to atomic objects and web extraction language in which web data converted to structured data in the form of tables.

3.3 Unstructured Data extraction:

Web content data is unstructured data and the research around it is knowledge discovery in texts (KDT). Some of the techniques are,

- ***Information Extraction:*** Keyword and phrases are identified and searched inside the text
- ***Information Visualization:*** It utilizes feature extraction and key term indexing which builds an image in large images.
- ***Topic Tracking:*** In this method, user interest is checked from other profiles of user and topics selected according to that,
- ***Summarization***: Length of the document is reduced which helps the user to decide whether they need to read the document,
- ***Categorization***: Main themes are identified and number of words in that document is counted used for ranking that page.
- ***Clustering***: Similar documents are grouped which helps the user to select the topic of interest.

4. Web Content Mining Tools:

WCM tools help the user in downloading the information for users in an easier way. Some of the tools are,

- ***Web content Extractor:***
  This tool is used by book readers to extract details about books, businessman extract and collect price and real estate details, Journalists information, online information, Job seeking details.
- ***Web Info extractor:***
  It extracts unstructured data as well as tabular data to a file, Monitors web pages and retrieve new content from any kinds of file types.
- ***Automation:***
  Automates complex tasks, web recorder, automate scripts, powerful task scheduling and auto-run scheduled tasks.

- ***Screen Scrapper:***
  This tool allows us to Mine data on products and download them to a spreadsheet.
- ***Mozenda:*** Agents are set up to extract data in a regular fashion and circulate it to several destinations.
- ***SAS Enterprise Miner:*** User friendly GUI to the SEMMA (Sample, Explore, Modify, Model, Assess) process
- ***QL2 Software:*** Data extraction using SQL query like language (WebQl).
- ***Wizsoft Software:*** Software developed based on mathematical algorithms used in business sectors.
- ***Website Parser:*** Helps to gather information from any website quickly, which helps for the business owner or retail sites.

5. Conclusion:

The web is a largest data repository in the world. As the information is dynamic everyday new contents needs to be added on the web page. The way of searching information also differs in respect to people knowledge, the content of the page and necessity of them. New algorithms and techniques should be developed to cope with the growth of data repository. Data extraction should not completely depend on algorithms or techniques. Many tools have been developed for extracting information from the data warehouse which also contains some pros and cons that yet to be considered for further research.

References:

[1] Arvind Kumar Sharma1, P.C. Gupta, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 8, October 2012.

[2] V. Bharanipriya1 & V. Kamakshi Prasad2, "WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY", International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[3] Mrs. Bhanu Bhardwaj, "Extracting Data through Webmining," International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 3, May - 2012 ISSN: 2278-0181.

[4] Aidan Finn, Nicholas Kushmerick and Barry Smyth, "Fact or fiction: Content classification for digital libraries", In Proceedings of Joint DELOSNSF Workshop on Personalization and Recommender Systems in Digital Libraries (Dublin),No. 01/W03 18 – 20, June 2001.

[5] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, "Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition", 2002.

[6] Cooley, R.; Mobasher, B.; Srivastava, J.; "Web mining: information and pattern discovery on the World Wide Web", In Proceedings of Ninth IEEE International Conference. pp. 558 -567, 3-8 Nov. 1997.

[7] A. F. R. Rahman, H. Alam and R. Hartono, "Content Extraction from HTML Documents", in First International Workshop on Web Document Analysis (WDA2001), 2001.

[8] G. Shrivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011. 3rd International Conference on Electronics Computer Technology (ICECT), pp. 399-403, April 2011.

[9] B. Singh, H.K. Singh, "Web data Mining Research", In Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010.

[10] O. Etzioni, "The World Wide Web: Quagmire or gold mine." Communications of the ACM, Vol. 39 No. 11, pp. 65-68, Nov.1996.

[11] Qingyu Zhang & Richard S. Segall, "Web Mining: A Survey of Current Research", Information Technology and Decision Making, 7 (4), 683-720, 2008.

[12] R Kosala, H Blockeel-ACM SIGKDD Explorations Newsletter, 2000.

[13] O. Buyukkokten, H. Garcia - Molina and A. Paepcke, "Accordion Summarization for End - Game Browsing on PDAs and Cellular Phones, In Proceedings of Conference on Human Factors in Computing Systems, pp.213 – 220, 2001.

[14] O. Buyukkokten, H, Garcia - Molina and A. Paepcke, "Seeing the Whole in Parts: Text summarization for Web Browsing on Handheld Devices", In Proceedings of the 10th international conference on World Wide Web, pp. 652-662, 2001.

[15] A. F. R. Rahman, H. Alam and R. Hartono, "Understanding the Flow of Content in Summarizing HTML Documents", In International Workshop on Document Layout Interpretation and its Applications, DLIA01, 2001.