Content Acquisition Using Hidden Web Mining

Sonali K. Shelke

Department of Computer Science and Engineering, Deogiri Institute of Engineering & Management Studies, Aurangabad (Maharashtra), India

***Abstract* -Web mining is a heavily researched area in the field of data mining and web mining with wide range of applications. This paper contains the three different categories of web mining – web content mining, web structure mining and web usage mining. Some specific tools are available for extracting useful knowledge from web. Further a brief description on web mining with region extraction algorithms and hidden web retrieval seen as future research area for hidden web mining.**

***Keywords:* Web Mining; Web Content Mining; Web Structure Mining; Web Usage Mining; Hidden Web Mining etc.**

I. INTRODUCTION

Web mining is the application of data mining techniques to extract useful knowledge from Web data net, including Web pages, hyperlinks between documents, web sites logs and web pages, etc. In addition, a community of researchers interested in the area has been formed, largely through the successful series of WebKDD workshops, since 1999[1].Web mining forums, and the Web Analytics workshops, which have been held in conjunction with the SIAM data mining conference[1] . In this paper, the data-centric view of Web mining which is represented in following figure.

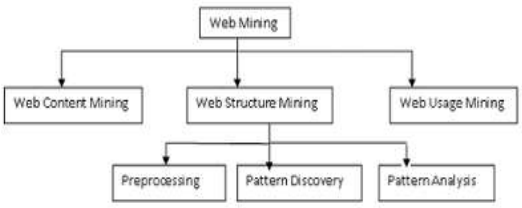

Figure 1: Overview of web mining Componets.

The main objective of this review of web minning will focus on the content extraction basedon data records and data region through the tools and region extractor. As various survey results in statement that , "there is explosive growth of the web databases",the question remains in mind that how to extract the specific contents or the region through the web documents and how to retrive the extracted content from the hidden web document. The primary solution comprises following approaches: use of web extraction tools and the algorithms such as Mining Data Record(MDR) and VIPS for vision based page segmentation.The secondary solution comprises use of Innovative vision based page segmetation (IVPS) methods for segmenting pagefor hidden web retrieval.

1.1. Drawbacks in the existing approaches

- The explosive growth of the web has introduced a heavy demand on networking.
- Knowledge discovery consumes a lot of System Resources & web servers [2].
- Users experience a heavy time loss due to widespread information .
- Hence, an obvious solution in order to improve the quality of Web services would be the increase of bandwidth, but such a choice involves increasing economic cost.
- Web caching scheme also exhibits three significant drawbacks: If the proxy is not properly updated, a user might receive stale data, and, as the number of users grows, origin servers typically become bottleneck.
- Main drawback of systems which have enhanced pre-fetching policies is that some pre-fetched objects may not be eventually requested by the users. So the network traffic gets over congested.

II. WEB MINING OVERVIEW

Web mining uses the technique of data mining into the documents on the World Wide Web. The process of web mining includes extraction of information from the World Wide Web through the conventional practices of the data mining and putting the same into the website features. The mining process, includes three types of mining - web content mining, Web structure mining, Web usage mining[3].

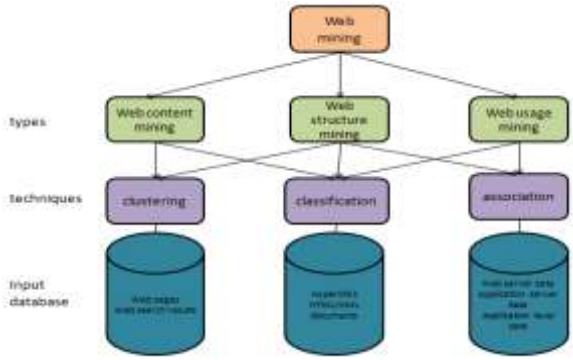

Figure 2: Categories of Web Mining

A] Web Content Mining-Web content mining, basically extracts the useful information from web documents with the help of content mining tools [7]. Many pages are having open to access on the web. These pages are content of web. Searching the information and open access search pages is also retrieval content of web. End accurate result is defined the result pages content mining.

B] Web Structure Mining- One can define web structure mining in terms of graphs. The web pages are representing as nodes (root) and Hyperlinks represent as edges(sublinks). Basically it is shows the relationship between user and respective web. The main objective of web structure mining is generating structured summaries about information on web pages/webs. It is shown the link one web page to another web page [4].

C] Web Usage Mining- It is discovery of meaningful patterns from data generated by client server transaction on one or more web nets. A web is a collection of inter related files based on necessary data on one or more web servers. It is automatically generates the data which is stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content and sites.

III. METHODS & TECHNIQUES

Web mining uses region and content extractors to extract the useful content from the web document. In every research area, there is the significance of algorithmic approach towards mining of the web document.

There are interesting algorithms, available for the extraction purpose. Some tools for the web data scrapping are also useful for the required retrieval of content. Some of the tools are briefly described with their use.

A] Tools for web mining

Some of useful tools used for Web Mining are:

- *QL2 Software*: Specializes in web data harvesting and extraction using SQL-like query language (WebQL).
- *Screen Scraper*: Products and services for web site data extraction. Flagship product, screen-scraper, provides a GUI to define links to follow and information to extract, and works with several programming languages and platforms [5] [6].
- *SAS Enterprise Miner*: an integrated suite which provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process.
- *SPAD*: provides powerful exploratory analyses and data mining tools, including PCA, clustering, interactive decision trees, discriminate analyses, neural networks, text mining and more, all via user friendly GUI.
- *Website Parser:* A web site parser tool is a program that will allow you to gather information from many web sites and web pages throughout the Internet. The tool goes through the targeted sites and is able to grab large amounts of data, through the parameters that you have created. This data can be used in XLS, CSV, XML, and TSV files for later use. Being able to gather huge amounts of information quickly and easily is an invaluable tool for any business owner or retail site [5] [6].
- *Web Extractor Software*: Web extractor software may be one of the smartest software tools to invest in. The cutting edge technology may be used in a variety of settings. It has been effectively utilized by law enforcement, researchers, and several businesses by extracting vital information from specific websites. Data extraction, screen scraping, and web crawling may only be a few of the features available [5].
- *Mozenda-Mozenda:* is a Software as a Service (SaaS) company that enables users of all types to easily and affordably extract and manage web data. With Mozenda, users can set up agents that routinely extract data, store the data, and publish data to multiple destinations. Once information is in the Mozenda systems users can format, repurpose, and mashup the data to be used in other online/offline applications or as intelligence [5].

B] Algorithmic Methods and framework.

1. MDR: Mining Data Records.

Mining Data Records known as MDR algorithm is designed and intended to extract data records from web usage logs. It is built on the following hypothesis – "a data region contains a repetitive structure in a document, such that each repetitive structure contained within the data region is a data record, and that data records are usually resides in tables and forms of the web document source" [8]. DOM (Document Object Model) tree is used as input for the MDR.

2. VIPS: Vision-Based Page Segmentation:

Vision-Based Page Segmentation for short VIPS is also intended to extract the region based on visual section of the webpage. VIPS is also builds the hypothesis that the web designer must provide the visual cues that will be useful for people to recognize the visual sections of the web content [8]. For example horizontal, vertical rule as well as special fonts, panels associated with colors, frames etc. VIPS takes DOM tree as the input for extraction the data region including multiple visual features related to the web document.

3. IVPS: Innovative Vision-based Page Segmentation:

An Innovative vision- based Page Segmentation framework intended to extract the information from hidden web pages. IVBPS framework helps to extract useful

information from hidden web document that will be required by user [9]. As web database getting huge day by day, it allowing the researcher to focus on extracting data from hidden web documents is known as "Hidden Web Mining".

IV. CONCLUSIONS

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past ten years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations pursuing research in web mining that are practicing it. In this paper a brief description on types and tools of web mining, and outlined some promising areas of future research with hidden web mining through web content extractor. This review provides an initial point for fruitful discussion towards web mining methodologies in association with hidden content extraction tools.

REFERENCES

[1] R. Kosala, H. Blockeel ,2000: 'Web mining research: A survey', ACM SIGKDD Explorations, **Vol. 2** ,1-15,
[2] A.Rastogi, S.Gupta, S. Agarwal, N. Agarwal,2012: 'Web Mining : A comparative study', *IJCER International Journal of Computational Engineering Research* ,**2,**325-331.
[3] Web mining blog: http://googleblog.blogspot.com/2008/07.
[4] ACM Portal: http://portal.acm.org/portal.cfm.
[5] A. Herrouz, 2013: 'Overview of Web Content Mining Tools', *IJES*, **Volume 2**, Issue 6,ISSN: 2319 – 1813 ISBN: 2319 – 1805.
[6] www.psl.cs.colubia.edu/classes/cs6125s10/presentations/presentation_hemanth_murthy.ppt
[7] Ananthi.J,2014 : 'A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites', *IJCSIT Journal*,5, 4091-4094
[8] Hassan A. S. and Rafael C, 2013: 'A Survey on Region Extractors from Web Documents', *IEEE transactions on knowledge and data engineering*, **25**, 1960-1981.
[9] L.Wei, X. Meng, and W. Meng, 2006: 'Vision-Based Web Data Records Extraction,' Proc. Int'l Workshop Web and Databases (WebDB).
[10] Kopal Maheshwari, 2013: 'Advance Frameworks for Hidden Web Retrieval Using Innovative Vision-Based Page Segmentation' , *IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661*, p- ISSN: 2278-8727,Volume 12, Issue 3 (Jul. - Aug. 2013), PP 52-58.