Integration of Web mining and web crawler: Relevance and State of Art

Subhendu kumar pani
PGDept. Of Comp. Application,
RCMA
BBSR,Orissa,India

Deepak Mohapatra
PG Dept of .Management
RCEM
BBSR,Orissa,India

Bikram Keshari Ratha
PG Dept. Of Comp. Application,
UTKALUNIVERSITY
BBSR,Orissa,India

Abstract:- **This study presents the role of web crawler in web mining environment. As the growth of the World Wide Web exceeded all expectations,the research on Web mining is growing more and more.web mining research topic which combines two of the activated research areas: Data Mining and World Wide Web .So, the World Wide Web is a very advanced area for data mining research. Search engines that are based on web crawling framework also used in web mining to find the interacted web pages. This paper discusses a study on crawlers and related research issues on web mining. A theoretical framework is also suggested.**

KEY WORDS*:* **Data mining, Web mining, web data, crawler**

1. Introduction

Internet is the shared global computing network. It enables global communications between all connected computing devices. It provides the platform for web services and the World Wide Web. Web is the totality of web pages stored on web servers. There is a spectacular growth in web-based information sources and services. It is estimated that, there is approximately doubling of web pages each year. As the Web grows grander and more diverse, search engines also have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated. In Internet data are highly unstructured which makes it extremely difficult to search and retrieve valuable information. Search engines define content by keywords.

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities. Many organizations and corporations provide information and services on the web such as automated customer support, on-line shopping, and a myriad of resources and applications.web based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts etc., are becoming common practice and widespread. The WWW is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child sharing music files with friends to a senior receiving photographs and messages from grandchildren across the world. It is typical to see web pages for courses in all fields taught at universities and colleges providing course and related resources even if these courses are delivered in traditional classrooms. It is not surprising that the web is the means of choice to architect modern advanced distance education systems.

There are several important issues, unique to the Web paradigm that comes into play if sophisticated types of analyses are to be done on server side data collections. These include the necessity of integrating various data sources such as server access logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior. We devote the main part of this paper to the discussion of issues and problems

that characterize Web usage mining. Web is the totality of web pages stored on web servers. There is a spectacular growth in web-based information sources and services. It is estimated that, there is approximately doubling of web pages each year. As the Web grows grander and more diverse, search engines have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated.

This paper has been organized as follows. The next section presents an overview of classification of web mining.Crawler Based Search Engine issues are discussed in section 3. Section 4 discusses datasources. Section 5 concludes the paper.

II. WEB DATA MINING

A. OVERVIEW

The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services [5]. This area of research is so huge today partly due to the interest in e-commerce. This phenomenon partly creates confusion what constitutes Web mining and when comparing research in this area. Similar to [5], we suggest decomposing Web mining into these subtasks, namely

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

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

3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.

4. Analysis: Validations and/or interpretation of the mined patterns.

We should also note that humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and/or interpretation in step 4.So, interactive query-triggered knowledge discovery is as important as the more automatic data triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans. Thus, Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. It implicitly covers the standard process of knowledge discovery in databases (KDD) [2].We could simply view web mining as an extension of KDD that is applied on the Web data. From the KDD point of view, the information and knowledge terms are interchangeable[3].There is a close relationship between data mining, machine learning and advanced data analysis[4].Web mining is often associated with IR or IE.However,web mining or information discovery on the web not the same as IR or IE[1].

B. Web Content Mining

Web content mining describes the automatic search of information resources available online [6], and involves mining web data contents. In the web mining domain, web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in web documents. The web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of web data forces the web content mining towards a more complicated approach.

The web content mining is differentiated from two different points of view [7]: Information Retrieval View and Database View.R.kosla et al [8] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structures between the documents for document representation. As for the database view, in order to

have the better information management and querying on the web, the mining always tries to infer the structure of the web site of to transform a web site to become a database. Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources.

C. Web Structure Mining

Most of the web information retrieval tools only use the textual information, while ignore the link information that could be very valuable. The goal of web Structure mining is to generate structural summary about the web site and web page. Technically, web content mining mainly focuses on the structure of inner-document, while web Structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different web sites. Web structure mining can also have another direction-discovering the structure of web document itself. This type of structure mining can be used to reveal the structure (schema) of web pages; this would be good for navigation purpose and make it possible to compare/integrate web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in web pages by providing a reference schema. The detailed works on it can be referred to[9]

The structural information generated from Web structure mining includes the follows: the information measuring the frequency of the local links in the Web tuples in a web table containing links that are interior and the links that are within the same document: the information measuring the frequency of web tuples in a web table that contains links that are global and the links that span different web sites.web structure mining has a nature relation with the web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the web. Its quiet often to combine these two mining tasks in an application.

D. Web Usage Mining

Web usage mining tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the web. It focuses on the techniques that could predict user's behavior while the user interacts with web. M. Spiliopoulou abstract the potential strategic aims in each domain in to mining goal as: predication of the user's behavior within the site , comparison between expected and actual web site usages, adjustment of the web site to the interests of its users. There are no definite distinctions between the web usage mining and other two categories. In the process if data presentation of web usage mining, the web site topology will as the information sources, which interacts web usage mining with the web content mining and web structure mining moreover the clustering in the process of pattern discovery is a bridge to web content and structure mining from usage mining.

There are lots of works have been done in the IR , Database, Intelligent Agents and topology, which provides a sound function for the web content, web structure mining . Web usages mining is a relative new research area, and gains more and more attentions in recent years. I will have a detailed introduction in the next section about mining, based on some up-to-date research works.

III. Crawler Based Search Engines

A crawler is a program that retrieves Web pages, commonly for use by search engine or a Web cache. These engines are software programs that crawl around the Web searching for deviations to Web pages, Web text and HTML tags-all this work is accomplish by the software Web information is indexed by the software.

Web crawlers [11] (also known as Web spiders, bots, robots, walkers and wanderers) are programs which downloads the documents from the internet.

The number of Web pages is increasing in a very fast rate [6]. This growth has urged the development of retrieval tools like search engines to get the information from WWW. Web crawling is one of main component in Web information retrieval. Web crawling is a program which traverses the World Wide Web (WWW) in a methodically, automated

manner to generate a copy of all the visited pages for latter processing by a search engine [12]. Due to limited bandwidth storage, and computational resources, and to the dynamic nature of the web, search engines cannot index every every Web page, and even the covered portion of the web cannot be monitored continuously for changes. In fact a recent estimate of the visible Web is at around 9.2 billion static pages as of March 2007[13]. This estimate is more than triple the 2 billion pages that the largest search engine, Google, reports at its Web site [14]. Therefore it is essential to develop effective agents to conduct real time searches for users.

Topic specific crawlers have become important tools to support applications such as specialized Web portals, online searching, and competitive intelligence. These crawlers are designed to retrieve pages that are relevant to the triggering topic. Generally employing single crawler to gather all pages is inevitably difficult. Therefore, many search engines often run multiple processes in parallel to perform the task. We refer to this type of spider as a parallel spider. This approach can considerably improve the collection efficiency.

In this paper, a novel crawling architecture is presented which can gather Web pages that are on topic of users interest. This domain collection can later be used to build specialized Web portals. The architecture comprises of multiple, parallel crawling. Web crawling is to improve the performance of the search engine and produce more comprehensive search in the WWW.

A .Architecture

A crawler must not only have a good crawling strategy, but it should also have a highly optimized architecture.

Shkapenyuk and Suel[15] noted that: "While it is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability."

Web crawlers are a central part of search engines, and details on their algorithms and architecture are kept as business secrets. When crawler designs are published, there is often an important lack of detail that prevents others from reproducing the work. There are also emerging concerns about search engine spamming ", which prevent major search engines from publishing their ranking algorithms.

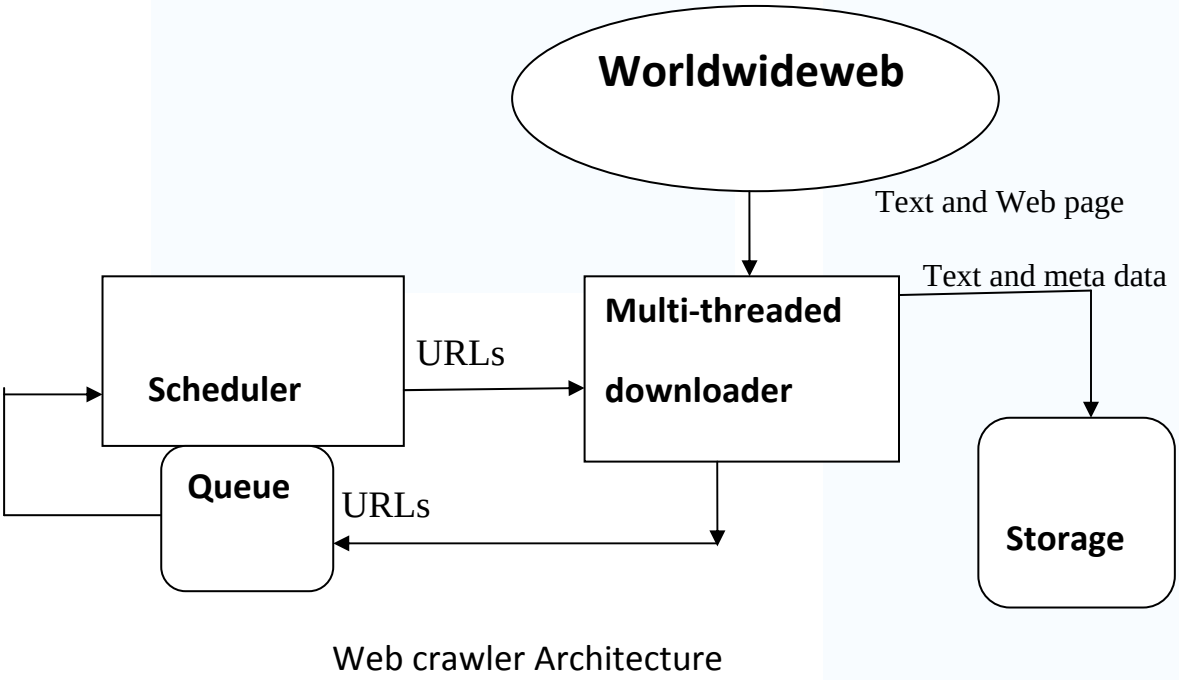
IV Data Sources

Web Usage Mining applications are based on data collected from three main sources [5]: (i) web servers, (ii) proxy servers, and (iii) web clients.

*The Server Side***:** Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format [2], Extended Log Format [3], and LogML [7]. When exploiting log information from web servers, the major issue is the identification of users' sessions' .Apart from web logs, users' behavior can also be tracked down on the server side by means of TCP/IP packet sniffers.

*The Proxy Side***:** Many internet service providers (ISPs) give to their customer Proxy Server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers.

*The Client Side***:** Usage data can be tracked also on the client side by using JavaScript, java applets [8], or even modified browsers [10]. These techniques avoid the problems of users' sessions' identification.



Web crawler Architecture

V. Conclusion

we survey the researches in the area of web mining. Three recognized types of web data mining are introduced generally, and a theoretical framework of web crawler in web mining is discussed.

REFERENCES

[1] Configuration File ofW3C http, 1995. http://www.w3.org/Daemon/User/Config/.

[2] O.Etzioni.The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65.

[3] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating e-commerce and data mining: Architecture and challenges. In WEBKDD 2000 - Web Mining for E-Commerce – Challenges and Opportunities, Second International Workshop, August 2000.

[4] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating e-commerce and data mining: Architecture and challenges. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001). IEEE Computer Society, 2001.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan.Web usage mining: Discovery and applications of usage patterns from web data.SIGKDD Explorations, 1(2):12–23, 2000.

[6] Kurt D. Fenstermacher and Mark Ginsburg. Mining client-side activity for personalization. In Fourth IEEE International Workshop on Advanced Issues of Ecommerce and Web-Based Information Systems (WECWIS'02), pages 205–212, 2002.

[7] John R. Punin, Mukkai S. Krishnamoorthy, and Mohammed J. Zaki. Logml: Log markup language for web usage mining. In R. Kohavi, B. Masand, M. Spiliopoulou,and J. Srivastava, editors, WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA August 26, 2001. Revised Papers, volume 2356 of Lecture Notes in Computer Science, pages 88–112. Springer.

[8] Cyrus Shahabi and Farnoush Banaei-Kashani. A framework for efficient and anonymous web usage mining based on client-side tracking. In R. Kohavi, B. Masand,M. Spiliopoulou, and J. Srivastava, editors, WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers, volume 2356 of Lecture Notes in Computer Science, pages 113–144. Springer, 2002.

[9] Boris Diebold and Michael Kaufmann. Usage-based visualization of web localities. In Australian symposium on Information visualization, pages 159–164, 2001.

[10] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN Systems, 27(6):1065–1073, 1995.

[11]Heydon and .Najork.Mercator:"A scalable, extensible Web crawler". World wide web2 (4):219-229,1999.

[12]F.C.Cheong.,InternetAgents:Spiders,wanderers,Brokers,andBots,NewRiders,publising,Indianpolis,Indiana,USA,1996.

[13] Google. http://www.google.com

[14] S.chakraborti, M.van den Crawling: A new approach to topic-specific web resource discovery". In the 8th International WorldW ide WebConference, 1999..

[15] Shkapenyuk, V. and Suel, T. (2002). In Proceedings of the 18th International Conference on Data Engineering (ICDE), pages 357-368, San Jose, California. IEEE CS Press.