IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661, ISBN: 2278-8727Volume 7, Issue 3 (Nov. - Dec. 2012), PP 24-29
www.iosrjournals.org

A Study: Web Data Mining Challenges and Application for Information Extraction

T. Sunil kumar[1], Dr. K. Suvarchala[2]
[1]*(Department of CSE, Krishnamurthy Inst. Of Tech. and Engineering, India)*
[2]*(Department of CSE & Informatics, Bankathlal Bhadruka College of IT, India)*

Abstract : *Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and various explorations. The aim of this paper is to explore the role of data mining for information extraction in web content, structure and usages mining in current web models, and the outlines the process of extracting patterns from data. This paper also present data mining primitives, from which data mining query languages can be designed. Issues regarding how to integrate a data mining system with a database or data warehouse are also discussed. In addition to studying a classification of data mining systems, and it's challenging research issues for building data mining tools of the future.*

Keywords *- Data Mining, Information Extraction, Knowledge discovery, Web mining*

I. Introduction

Web is a vast repository of information which grows at a fast pace. The intense growth of information evolves many new challenges for Web researchers, which include among other things, high data dimensionality and highly volatile and constantly evolving content. Due to this, it has become increasingly necessary to create new and improved approaches to traditional data mining techniques can be applied for the Web mining. Automatically extracting useful information is a key challenging issues in web data mining. The billions of Web pages created are generated dynamically by underlying Web database service engines using HTML or XML[1][18]. However, searching, comprehending, and using the semi structured information stored on the Web poses a significant challenge because this data is more sophisticated and dynamic than the information that commercial database systems store.

The mining data varies from structured to unstructured. Data mining mainly deals with structured data organized in a database while text mining mainly handles unstructured data [12]. Web mining lies in between and copes with semi structured data and/or unstructured data. Web mining calls for creative use of data mining and/or text mining techniques and its distinctive approaches. Mining the web data is one of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web and we can easily get overwhelmed with data [2]. As the Web reaches its full potential, however, we must improve its services, make it more comprehensible, and increase its usability. As researchers continue to develop data mining techniques, we believe this technology will play an increasingly important role in meeting the challenges of developing the intelligent Web.

II. Web Data Mining

Data mining can be viewed as a result of the natural evolution of information technology. The database system industry has witnessed an evolutionary path in the development of data collection, creation, data management (including data storage and retrieval, and database transaction processing), and advanced data analysis (involving data warehousing and data mining) [8]. The research and development in database systems since the 1970s has progressed from early hierarchical and network database systems to the development of relational database systems (where data are stored in relational table structures), data modeling tools, and indexing and accessing methods. In addition, users gained convenient and flexible data access through query languages, user interfaces, optimized query processing, and transaction management. Efficient methods for on-line transaction processing (OLTP)[3], where a query is viewed as a read-only transaction, have contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.

Application-oriented database systems, including spatial, temporal, multimedia, active, stream, and sensor, and scientific and engineering databases, knowledge bases, and office information bases, have flourished. Issues related to the distribution, diversification, and sharing of data have been studied extensively.

Heterogeneous database systems and Internet-based global information systems such as the World Wide Web (WWW) have also emerged and play a vital role in the information industry [8][16].

The Web and other repositories have an immense and dynamic collection of pages and information that includes countless hyperlinks and huge volumes of access and usage information provides a rich and unprecedented data mining source. However, the Web also poses several challenges to effective process resource and knowledge discovery as shown in Figure -1.

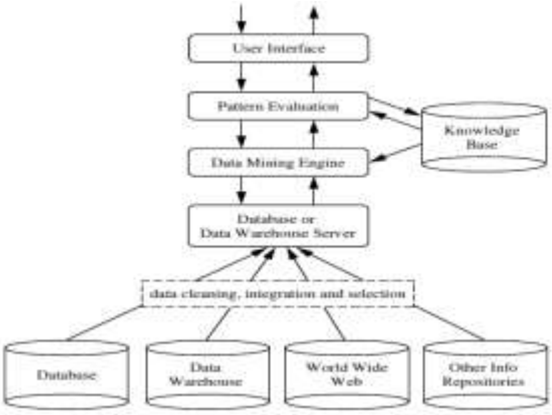

Figure -1 A Framework of Data Mining Process

To accessing information from web currently users choose various approaches. Most of the approaches are based on the following:

- *Content or Keyword based* : Most of the search engine perform information search based on the keyword or content-directory browsing such as MSN, Google or Yahoo, which use keyword indices or manually built directories to find documents with specified keywords or contents.
- *Multilevel Deep Web Querying:* Information cannot be accessed through static URL links, as most of the information hides behind searchable database query forms that unlike the surface[15][16]. For example if a user searching for a movie, book or song, which information not remain on the index pages it need to go for multilevel web search to find the relevant information.
- *Dynamic Web Link Clicking:* Dynamically surfing the Web linkage links to a web resource presented by search engines.

The success of these approaches and techniques, especially with the more recent page ranking by search engines highly depends on the efficient web data mining which shows the great promise to become the ultimate information systems.

III. Limitation And Challenges In Web Data Mining

Web information presentation is a major challenge in current trends of information extraction. The traditional schemes for accessing the immense amounts of data that reside on the Web fundamentally assume the text-oriented, keyword-based view of Web pages. To achieve the required information we need a high potential web mining techniques to overcome the fundamental problems [2]. Firstly, we believe a data-oriented abstraction will enable a new range of functionalities. Second, at the service level, we must replace the current primitive access schemes with more sophisticated versions that can exploit the Web fully.

Current web search mining supports keyword, link address and content based web search, where data mining will play an important role. But these web search engines still cannot provide high-quality, intelligent services [4] because of several limitations in web mining which contributes to the problem[7].

A. Quality of keyword-based searches:

The quality of keyword-based searches suffers from several inadequacies such as a search often returns many answers, especially if the keywords posed include words from popular categories such as sports, politics, or entertainment. It overloaded keyword semantics and it can return low-quality results. For example, depending on the context, an apple could be a fruit, juice, company or computer and a search can miss many highly related pages that do not explicitly contain the posed keywords and, a search for the term data mining can miss many highly regarded machine learning or statistical data analysis pages.

B. Effective of deep-Web Extraction:

A research analysts estimated that searchable databases on the Web numbered more than 100,000. These databases provide high-quality, well-maintained information, but are not effectively accessible. Because current Web crawlers cannot query these databases, the data they contain remains invisible to traditional search engines. Conceptually, the deep Web provides an extremely large collection of autonomous and heterogeneous databases, each supporting specific query interfaces with different schema and query constraints. To effectively extract the deep Web, we must integrate these databases and implement efficient web mining approaches.

C. Self organized and constructed directories:

A content or type-oriented Web information directory presents an organized picture of a Web sector and supports a semantics-based information search [9], which makes such a directory highly desirable. For example, following organization links like Country > Sports > Football > Players makes searches more efficient. Unfortunately, developers construct such directories manually which limit coverage of these costly directories provide and developers cannot easily scale or adapt them.

D. Semantics-based query:

Most keyword- based search engines provide a small set of options for possible keyword combinations, essentially "with all the words" and "with any of the words." Some Web search services[12][9], such as Google and Yahoo, provide more advanced search primitives, including "with exact phrases," "without certain words," and with restrictions on date and domain site type.

E. Human activities feedback:

Web page authors provide links to "authoritative" Web pages and also traverse those Web pages they find most interesting or of highest quality [10][2]. Unfortunately, while human activities and interests change over time, Web links may not be updated to reflect these trends. For example, significant events—such as the 2012 Olympic or the tsunami attack on Japan can change Web site access patterns dramatically, a change that Web linkages often fail to reflect. We have yet to use such human-traversal information for the dynamic, automatic adjustment of Web information services.

F. Multidimensional Data analysis and mining:

Because current Web searches rely on keyword based indices, not the actual data the Web pages contain, search engines provide only limited support for multidimensional Web information analysis and data mining[14][19][20].

These challenges and limitation have promoted research into efficiently and effectively discovering and using Internet resources, a quest in which web data mining play an important role.

IV. Application Of Web Data Mining

Web data mining can successfully fix information extraction and the following features incorporated with the web mining program must be fixed if we need to use data mining successfully in creating Web intelligence[2][5].

A. Web search-engine data Mining

For website optimization web crawls on indexes Websites, and builds and stores large keyword-based indices that help identify sets of Websites that contain specific keywords and phrases. By using a set of tightly restricted keywords and phrases, an experienced user can quickly identify appropriate documents[6]. However, current keyword-based search engines suffer from several deficiencies. First, a subject of any breadth can easily contain tens of thousands of records. This can lead to a look for website returning many document entries, many of which are only partially appropriate to the subject or contain only poor-quality materials. Second, many highly appropriate records may not contain keywords and phrases that explicitly define the subject, a trend known as the *polysemy* problem [11]. For example, the keyword and key phrase information exploration may turn up many Websites related to other exploration industries, yet fail to identify appropriate papers on knowledge discovery, mathematical analysis, or machine learning because they did not contain the information exploration keyword and key phrase.

Depending on these observations, we believe data mining should be integrated with the web search engine service to enhance the excellence of Web searches[13]. To do so, we can start by enlarging the set of look for search phrases to consist of a set of keyword and key phrase alternatives. For example, a look for the keyword and key phrase information exploration can consist of a few alternatives so that an index-based web search engine can perform a parallel search that will obtain a larger set of records than the search phrases alone would return. The search engine then can look for the set of appropriate Web records obtained so far to select a smaller set of highly appropriate and authoritative records to present to the user. Web-linkage and Web-dynamics analysis thus provide the basis for discovering high-quality records.

B. Web Link Structure Analyzing

Given a keyword and key phrase or subject, such as investment, we believe an individual would like to find web pages that are not only extremely appropriate, but trustworthy and of high quality[17]. Instantly determining trustworthy Websites for a certain subject will improve a Web search's excellence. The secret of power conceals in Website linkages. These hyperlinks contain quantity of hidden human annotation that can help instantly infer the idea of power. When a Web page's writer makes a web page link directing to another

Website, this action can be considered as an approval of that web page. The combined approval of a given web page by different writers on the Web can indicate the value of the site and lead normally to the development of trustworthy Web pages.

First, not every web page link symbolizes the approval for a search. Web-page writers make some links for other requirements, such as routing or to provide as compensated ads. Overall, though, if most hyperlinks operate as recommendations, the combined viewpoint will still control. Second, a power that belongs to a professional or aggressive interest will hardly ever have its Web page point to competing authorities' pages. For example, manufacture will likely avoid supporting a product by guaranteeing that no links to that product in their Websites appears.

These qualities of web link components have led researchers to consider another essential Website category: locations. A hub is just one Website or web page set that provides selections of links to authorities. Although it may not be popular, or may have only a few links directing to it, a hub provides hyperlinks to a selection of popular websites on a typical subject. These web pages can be list of suggested links on individual home pages, such as suggested referrals websites from a course home-page or a expertly constructed source list on a commercial site [21]. A hub unquestioningly confers authority status on websites that focus on a specific subject. Generally, a good hub points to many excellent authorities, and, on the other hand, a page that many good locations point to can be considered a good authority. Such a common encouragement relationship between locations and authorities helps users my own trustworthy. Websites performs development of high-quality Web components and sources.

Techniques for determining trustworthy Web pages and locations have led to the development of the PageRank1 and HITS3 methods [18]. Some over the counter available Web search engines, such as Google, are built around such methods. By assessing Web links and textual perspective information, these systems can generate better-quality look for results than term-index look for engines.

C. Automatically Classifying Web documents

Although Yahoo and similar Web listing service systems use human visitors to categorize Web records, inexpensive and improved speed make automated category highly suitable. Common category methods use good and bad illustrations as training sets, then determine each papers a category brand from a set of defined subject groups depending on pre classified papers illustrations. For example, designers can use Yahoo's taxonomy and its associated records as exercising and test places to obtain a Web papers category program. This program groups new Web records by giving groups from the same taxonomy. Developers can obtain great results using typical keyword-based papers category methods, such as Bayesian category, support vector machine, decision-tree introduction, and keyword and key phrase centered organization analysis to categorize Web records. Since hyperlinks contain high quality semantic signs to a page's subject, such semantic information can help achieve even better precision than that possible with genuine keyword-based category[12][18].

However, since the back-linked web pages around a documents may be loud and thus contain unrelated subjects, innocent use of terms in a document's web page link community can lower precision [14]. For example, many personal home pages may have climate.com connected simply as a save, even though these web pages have no importance to the subject of climate. Tests have shown that combining solid mathematical designs such as Markov unique areas with pleasure brands can considerably improve Web papers category precision. As opposed to many other category techniques, automated category usually does not clearly specify adverse examples: We often only know which category a pre classified papers connected to, but not which records a certain category definitely limits. Thus, preferably, a Web documents category program should not require clearly marked adverse illustrations. Using positive illustrations alone can be especially useful in Web papers category, forcing some scientists to recommend a category method based on a enhanced support-vector-machine program.

D. Web Page Content and Semantic Structure Mining

Completely automated removal of Website components and semantic material can be difficult given the present restrictions on computerized natural-language parsing [9]. However, semiautomatic techniques can identify a large part of such components. Professionals may still need to specify what types of components and semantic material a particular web page type can have. Then a page-structure-extraction system can evaluate the Website to see whether and how a segment's content suits into one of the components. Designers also can evaluate individual reviews to enhance the training and evaluate procedures and enhance the quality of produced Website components and contents. Specific research of Website exploration systems shows that different types of web pages have different semantic components. For example, a department's home-page, a professor's home-page, and a job marketing web page can all have different components [15].

First, to identify the relevant and interesting framework to draw out, either an expert personally identifies this framework for a given Website category, or we develop techniques to instantly generate such a framework from a set of relabeled Website examples. Second, designers can use Website framework and content removal methods for automatic removal based on Website classes, possible semantic components, and other semantic information. Web page category identification allows to draft out semantic components and material, while getting such components allows validating which category the produced pages are part of. Such a connection mutually increases both procedures. Third, semantic page structure and content recognition will greatly enhance the in depth analysis of Web page contents and the building of a multilayered Web information base.

E. Dynamic Web Mining

Web mining can also recognize as dynamics web. How the Web changes in the perspective of its material, components, and accessibility styles. Saving certain pieces of traditional details related to these Web exploration factors helps in discovering changes in material and linkages. In this case, we can evaluate pictures from different time postage stamps to recognize the up-dates. However, as opposed to relational data source systems, the Internet's wide depth and large shop of details create it nearly difficult to consistently shop past pictures or upgrade records. These restrictions create discovering such changes generally infeasible. Mining Web accessibility activities, on the other hand, is both possible and, in many programs, quite useful.

With this strategy, customers can mine Web log information to discover Web page access styles. Assessing and discovering regularities in Web log information can enhance the quality and distribution of Internet information services to the end individual, enhance Web hosting server system performance, and recognize customers for electronic industry. A Web hosting server usually signs up a Web log entry for every Web page accessed. This accessibility includes the asked for URL, the IP address from which the request is started, and a time seal. Web-based e-commerce hosts gather many Web accessibility log details[10]. Popular Web sites can register Web log details that variety hundreds of mega bytes each day. Web log directories provide rich details about Web characteristics. Opening these details requires innovative Web log exploration techniques.

The success of such programs relies on what and how much legitimate and efficient knowledge we can find out from the raw details. Often, researchers must clean, reduce, and convert these details to recover and evaluate significant and useful details. Second, scientists can use the available URL, time, IP deal with, and Web page content details to create a multidimensional view on the Web log data source and execute a multidimensional

OLAP research to find the top customers, top utilized websites, most frequently utilized times, and so on. These results will help find out customers, marketplaces, and other organizations.

Third, exploring Web log records can expose organization styles, successive styles, and Web accessibility styles. Web accessibility routine exploration often requires taking further measures to obtain more individual traversal information. This information, which can include individual browsing series from the Web server's barrier pages along with related information, allows detailed Web log research. Researchers have used these Web log information to assess program performance, enhance program style through Web caching and page pre-fetching and changing, determine the characteristics of Web traffic, and to assess individual respond to site style. For example, some studies have suggested flexible Web sites that enhance themselves by learning from individual access styles. Web log research can also help develop personalized Web services for individual customers. Since Web log information provides details about particular pages' reputation and the techniques used to access them, these details can be incorporated with Web content and linkage framework exploration to help position Websites, categorize Web records, and develop a multilayered Web details platform.

V. Conclusion

Data mining for Web information extraction will be an important research in Web technology. To makes it possible to fully use the immense information available on the Web one must overcome many mining challenges before we can make the Web a richer, friendlier, and more intelligent resource that we can all share and explore. Many promising data mining methods can help achieve effective Web mining. But using data mining to find a user's profile patterns can further enhance these services. Although a personalized Web service based on a user's history could help recommend appropriate services, a system usually cannot collect enough information about a particular individual to warrant a quality recommendation. Either the traversal history has too little historical information about that person, or the possible spectrum of recommendations is too broad to set up a history for any one individual. For example, many people make only a single book purchase, thus providing insufficient data to generate a reliable pattern. So, customizing service to a particular individual requires tracing that person's Web history to build a profile, then providing intelligent, personalized Web services based on that information. Collaborative filtering can be effective because it does not rely on a particular individual's past experience but on the collective recommendations of the people who share patterns similar to the individual being examined. This approach generates quality recommendations by evaluating collective effort rather than basing recommendations on only one person's past experience. Indeed, collective filtering has been used as a data mining method for Web Data Mining and effective result presentation in future.

References

[1]. Ramakrishna, Gowdar et al ”Web Mining: Key Accomplishments, Applications and Future Directions”, in the International Conference on Data Storage and Data Engineering 2010.

[2]. Jiawei Han,Kevin,Chen-Chuan Chang "Data Mining for Web Intelligence" IEEE International Conference on Data Mining, 2002.

[3]. S. Chaudhuri and U. Dayal, “An Overview of Data Warehousing and OLAP Technology,” SIGMOD Record, vol. 26, no. 1, 1997, pp. 65-74.

[4]. S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Proc. 7th International World Wide Web Conf. (WWW98), ACM Press, New York, 1998, pp. 107-117.

[5]. J. Srivastava et al., “Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data,” SIGKDD Explorations, vol. 1, no. 2, 2000, pp. 12- 23.

[6]. Qingyu Zhang and Richard s. Segall,” Web mining: a survey of current research, Techniques, and software”, in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720

[7]. Q. Yang and X. Wu, 10 challenging problems in data mining research, International Journal Information Technology Decision Making 5(4) (2006) 597–604

[8]. Kosala and Blockeel, “Web mining research: A survey,” SIGKDD:SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000

[9]. Semantic Web Mining: State of the art and future directions” Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4, Issue 2, June 2006, Pages 124-143

[10]. Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su,” Web usage mining with intentional browsing data” in international journal of Expert Systems with Applications 34 (2007) 1893–1904

[11]. Andrei Broder. A taxonomy of web search. SIGIR Forum, 36(2):3–10, 2002.

[12]. Ricardo Baeza-Yates and Alessandro Tiberi. “Extracting semantic relations from query logs” proceeding for ACM SIGKDD international conference on Knowledge discovery and data mining, 2007.

[13]. Ricardo Baeza-Yates. Web usage mining in search engines. In Web Mining: Applications and Techniques, Anthony Scime, editor. Idea Group, 2004.

[14]. N. Barsagade, Web usage mining and pattern discovery: A survey paper, Computer Science and Engineering Dept., CSE Tech Report 8331 (Southern Methodist University,Dallas, Texas, USA, 2003).

[15]. R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004), pp. 155–163

[16]. P. Kolari and A. Joshi, Web mining: Research and practice, Comput. Sci. Eng.July/August (2004) 42–53

[17]. B. Liu and K. Chang, Editorial: Special issue on web content mining, SIGKDD Explorations 6(2) (2004) 1–4

[18]. Semantic Web Mining:State of the art and future directions” Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4, Issue 2, June 2006, Pages 124-143

[19]. Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng.”Fundamentals of content based image retrieval”www.cse.iitd.ernet.in/~pkalra/siv864/Projects/c h01_Long_v40-proof.pdf

[20]. Zhang, Z. Chen, M. Li and Z. Su, Relevance feedback and learning in content-based image search, World Wide Web 6(2) (2003) 131–155.

[21]. L. Chen, W. Lian and W. Chue, Using web structure and summarization techniques for web content mining, Inform. Process. Management: Int. J. 41(5) (2005) 1225– 1242