International Journal of Instrumentation Control and Automation

Volume 1 | Issue 1 Article 8

April 2011

A Review Of Trends In Research On Web Mining

Manoj Pandia
Department Of Computer Science, SIT; Bhubaneswar, Orissa, India, manoj_pandia@rediffmail.com

S. K. Pani
P.G. Department Of Computer Science, RCMA; Bhubaneswar, Orissa, India,
Subhendu_pani@rediffmail.com

S.K. Padhi
Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India, Sanjaya2004@yahoo.com

L. Panigrahy
Department of Computer Science and Engineering; Konark Institute of Science and Technology; Bhubaneswar, Orissa, India, mynamelingaraj@gmail.com

R. Ramakrishna
Department of Computer Science and Engineering; Konark Institute of Science and Technology Bhubaneswar, Orissa, India, boostram@gmail.com

Recommended Citation

Pandia, Manoj; Pani, S. K.; Padhi, S.K.; Panigrahy, L.; and Ramakrishna, R. (2011) "A Review Of Trends In Research On Web Mining," *International Journal of Instrumentation Control and Automation*: Vol. 1 : Iss. 1 , Article 8.
DOI: 10.47893/IJICA.2011.1007
Available at: https://www.interscience.in/ijica/vol1/iss1/8

A Review Of Trends In Research On Web Mining

[1]Manoj Pandia, [2]Subhendu Kumar Pani, [3]Sanjay Kumar Padhi,
[3]Lingaraj Panigrahy, [3]R.Ramakrishna

[1] Department Of Computer Science, SIT; Bhubaneswar, Orissa, India
[2] P.G. Department Of Computer Science, RCMA; Bhubaneswar, Orissa, India
[3] Department of Computer Science and Engineering; Konark Institute of Science and Technology Bhubaneswar, Orissa, India
E-mail: manoj_pandia@rediffmail.com; Subhendu_pani@rediffmail.com; sanjaya2004@yahoo.com; mynamelingaraj@gmail.com;boostram@gmail.com

Abstract
In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task.So, the World Wide Web is a fertile area for data mining research.Web mining is a research topic which combines two of the activated research areas: Data Mining and World Wide Web. Web mining research relates to several research communities such as Database, information Retrieval and Artificial intelligence, visualization.This paper reviews the research and application issues in web mining besides proving an overall view of Web mining.

Keywords: *Web, Data mining, web usage mining, web content mining, web structure mining*

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. Furthermore, we survey some of the

emerging tools and techniques, and identify several future research directions.

This paper has been organized as follows. The next section presents an overview of classification of web mining. Techniques on the web mining are discussed in section 3. Section 4 discusses application area of web mining. Section 5 focuses on research direction. Section 6 concludes the paper.

2. WEB DATA MINING

2.1 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].

2.2 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.

2.3 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.

2.4 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.

3. WEB MINING TECHNIQUES

Traditional data mining techniques can also be used for web mining, such as classification, clustering, association rule mining, and visualization. In web mining, classification algorithms can be used to classify users into different classes according to their browsing behavior, for example according to their browsing time. After classification, a useful classification rule like "30% of users browse product/food during the hours 8:00-10:00 PM" can be discovered. The difference between classification and clustering is that the classes in classification are predefined (supervised), but in clustering are not predefined (unsupervised). The criterion by which items are assigned to different clusters is the degree of similarity among them. The main purpose of Clustering is to maximize both the similarity of the items in a cluster and the difference between clusters [12].The association rule technique can be used to indicate pages that are most often referenced together and to discover the direct or indirect relationships between web pages in users' browsing behavior [11]. For example, an association rule in the web usage mining area could take the form "the people who view web page index.htm and also view product.htm the support=50% and the confidence=60%". Visualization is a special analytical technique in web mining that allows data and information to be understood or recognized by human eyes by using graphical and visualized means to represent data, information and analysis results [13].In web structure mining, it usually plays an important role in illustrating the structure of hypertexts and links in a website or the linking relationship between websites. For the other two types of web mining technique, visualization is also an ideal tool to model the data or information. For example, a graph (or map) can be used for web usage mining to present the traversal paths of users or a graph may show information about web usage. This approach enables the analyst to understand and efficiently interpret the results of web usage mining.

Association Rules: After transactions are detected in the preprocessing phase, frequent item-sets are discovered using the A-priori algorithm [e.g. 13]. The support of item-set I is defined as the fraction of transactions that contain I and is denoted by $\sigma(I)$.

A hypergraph is an extension of a graph where each hyperedge can connect more than two vertices. A hyperedge connects URLs within a frequent item-set. Each hyperedge is weighted by the averaged confidence of all the possible association rules formed on the basis of the frequent item-set

that the hyperedge represents. The hyperedge weight can be perceived as a degree of similarity between URLs (vertices).

Sequential Pattern: patterns are used to discover frequent subsequences among large amount of sequential data. In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users sessions frequently. The typical sequential pattern has the form[15]:the 70% of users who first visited A.html and then visited B.html afterwards ,in the same session,have also accessed page C.html.Sequential patterns might appear syntactically similar to association pattern mining.

Clustering: techniques look for groups of similar items among large amount of data based on a general idea of distance function which computes the similarity between groups.Clustering has been widely used in Web Usage Mining to group together similar sessions [14]. Besides information from Web log files, customer profiles often need to be obtained from an on-line survey form when the transaction occurs. For example, you may be asked to answer the questions like age, gender, email account, mailing address, hobbies, etc. Those data will be stored in the company's customer profile database, and will be used for future data mining purpose

4. APPLICATION AREA OF WEB MINING

Web mining extends analysis much further by combining other corporate information with Web traffic data. This allows accounting, customer profile, inventory, and demographic information to be correlated with Web browsing, which answers complex questions such as:

- Of the people who hit our Web site, how many purchased something?
- Which advertising campaigns resulted in the most purchases, not just hits?
- Do my Web visitors fit a certain profile? Can I use this for segmenting my market?

Practical applications of Web mining technology are abundant, and are by no means the limit to this technology. Web mining tools can be extended and programmed to answer almost any question.

Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly. Also, the company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing campaign or marketing research, which will help the business to improve and align their marketing strategies timely.

For example, the company may have a list of goals as following:

- Increase average page views per session;
- Increase average profit per checkout;
- Decrease products returned;
- Increase number of referred customers;
- Increase brand awareness;
- Increase retention rate (such as number of visitors that have returned within 30 days);
- Reduce clicks-to-close(average page views to accomplish a purchase or obtain desired information);
- Increase conversion rate (checkouts per visit).

The company can identify the strength and weakness of its web marketing campaign through Web Mining, and then make strategic adjustments, obtain the feedback from Web Mining again to see the improvement. This procedure is an on-going continuous process.

5. RESEARCH DIRECTIONS

The techniques being applied to Web content mining draw heavily from the work on information retrieval, databases, intelligent agents, etc. Most of these techniques are well known and reported elsewhere, hence in this survey we have not focused on Web content mining. Hence, in this survey we have focused on Web Usage Mining, which is just starting as an area of research, and hence has a number of open issues. In the following we provide some directions for future research:

5.1 THE MINING PROCESS

The key component of Web mining is the mining process itself. As discussed in this paper, Web mining has adapted techniques from the field of data mining, databases, and information retrieval, as well as developing some techniques of its own, e.g. path analysis. A lot of work still remains to be done in adapting known mining techniques as well as developing new ones. Specifically, the following issues must be addressed:

1.New Types of Knowledge: Web usage mining studies reported to date have mined for association rules, temporal sequences, clusters, and path expressions. As the manner in which the Web is used continues to expand, there is a continual need to figure out new kinds of knowledge about user behavior that needs to be mined for.

2.Improved Mining Algorithms:the quality of a mining algorithm can be measured both in terms of how effective it is in mining for knowledge and how efficient it is in computational terms. There will always be a need to improve the performance of mining algorithms along with these dimensions.

3. Incremental Web mining: Usage data collection on the Web is incremental in nature. Hence, there is a need to develop mining algorithms that take as input the existing data mined from various logs can be integrated together into a more comprehensive model.

6. CONCLUSION

We survey the researches in the area of web mining. Three recognized types of web data mining are introduced generally. Web mining is a rapid growing research area. Web content mining is related but different from data mining and text mining. Web data are mainly semi-structured and/or unstructured. Web content mining requires creative applications of data mining and/or text mining techniques and also its own unique approaches

REFRENCES:

[1]G.Salton and M.McGill .introduction to modern information Retrival. McGraw Hill, 1983.

[2]U.Fayyad, G.Piatetsky-Shapiro, P.Smythfrom data mining to knowledge discovery: An overview. In advances in knowledge Discovery and data mining, pages 1-34.AAA Press, 1996.

[3] U.Fayyad, G.Piatetsky-Shapiro, and P.Smyth knowledge discovery and data mining: toward a unifying framework. In proceeding of the second int. conference on Knowledge Discovery and Data mining, pages 82-88, 1996.

[4] M.A.Hearst.Untangling text data mining. In proceedings of ACL'99: the 37th Annual meeting of the Association for computational Linguistics,

[5] O.Etzioni.The World Wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68,1996.

[6].S.K.Madria, S.SBhowmick, W.K, Ng, and E.P.Lim.Research issues in web data mining. In Proceedings of data ware housing and knowledge Discovery, first International conference, DaWak'99, pages 303-312, 1999.

[7]R.Cooley, B.Mobasher, and J.Srivastava.Web mining: Information and pattern discovery on the World Wide Web. In proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997.

[8]R.kosala, H.Blockeel.Web mining Research: A survey.

[9]S.K.Madria, S.S Bhowmick, W.K.Ng, and E.P.Lim.Research issues in web data mining. In proceedings of data warehousing and knowledge Discovery, first International Conference, DaWak'99,pages301-312,1999.

[10]Garofalaski,R.Rastogi,S.Shesadri,K.Shim,Datamining and the web:past,present and future,proceedings WIDM99,Kanas City,USA 1999

[11]J.Srivastava, R.Cooley, M.Deshpande, P.tan,web Usage Mining: Discovery and applications of usage Patterns from web data,SIGKDD Explorations,Vol.1,No.2,Jan.2000

[12]J.han and M.kamber, Data Mining Concepts and Techniques, Morgan Kaufmann Publisher, 2001.

[13] R. Agrawal, T. Imielinski, A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD Conference,1993.

[14] Jeffrey Heer and Ed H. Chi. Mining the structure of user activity using cluster stability.
In Proceedings of the Workshop on Web Analytics, Second SIAM Conference on Data Mining. ACM Press, 2002.

[15] Eleni Stroulia Nan Niu and Mohammad El-Ramly. Understanding web usage for dynamic web-site adaptation: A case study. In Proceedings of the Fourth International Workshop on Web Site Evolution (WSE'02), pages 53–64. IEEE, 2002.