International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Review Article

A Summarized Review on Web Usage Mining

Tayyaba Ashraf [Å*] and Imran Ashraf[Å]

[Å]IT Department, University of Gujrat, Gujrat Pakistan

Accepted 10 August 2014, Available online 25 Aug 2014, **Vol.4, No.4 (Aug 2014)**

Abstract

Incremental growth in the use of web based services and systems have led to generation of such tremendous amounts of data which is beyond imagining. This data plays a vital role in determining the factors like user's interests, priorities and product or services usage trends. This knowledge enables organizations to evaluate the effectiveness of their strategies and quality of services/products provided and leads them for further refinement. Solicitation of data mining approaches to process huge volume of data available on net is termed as web mining. Web usage is a further phase in web mining which discovers data about the use of internet. This paper aims at providing a review of phases and techniques involved in web usage mining.

Keywords: *Web-mining, Preprocessing, Sequential Patterns, Proxy Level Patterns.*

1. Introduction

In this rapidly growing age of information technology data has gained crucial importance for every organization. It is the most valuable asset for organizations in this era. Due to rapid emergence of electronic data management methods this age is called Information age (Goebel e Gruenwald, 1999).Each organization has a huge volume of data and it is very difficult and often impossible to handle that data without any computer based application. In addition to data management, analysis of such big collection of data is also a huge problem. Today's databases contain a huge volume of data that manual analysis and valuable decision making is not possible. In many cases a lot of independent fields need to be analyzed at a same time to get accurate results (Goebel e Gruenwald, 1999).Therefore humans require support to improve their analysis ability. The need for automated extraction of relevant data from a huge volume of data is widely recognized now. It leads to discover more efficient techniques for this purpose.

This review paper aims to collect and analyze the major approaches which have appeared in web about extraction of web data and provides an overview on mining phases which are most prominent regarding this and most recent trends in it. This paper is divided into five sections. After introduction section 2 gives an overview of related work and discusses in brief data mining in general. Section 3 is about web data, its usage and its structure. Section 4 highlights data foundations for web. Section 5 discussed the approaches used in web usage mining. In the end conclusion is given.

*Corresponding author: **Tayyaba Ashraf**; **Imran Ashraf** is working as Lecturer in CS & IT Department

2. Terminologies and Background

This era is of computer (Khushboo, Vekariya e Mishra, 2012) and electronic information (Han e Kamber, 2006); every sphere of life is based on accurate and timely available data. As a result a huge collection of data is produces in the field of science, medical, marketing and finance etc. (Anwer, Rashid e Hassan, 2010). Automated systems are required for systematize summarization, exploration, and classification of available data. It is helpful for management to take timely and related decisions. A lot of research areas like mathematics, artificial intelligence and meditation are involved to develop such automated systems (Gibson, Kleinberg e Raghavan, 1998; Pei *et al.*, 2000; Kohavi e Provost, 2001; Anwer, Rashid e Hassan, 2010; D.S.Deshpande, 2012; Seerat e Azam, 2012; Shelke, Deshpande e Thakre, 2012).

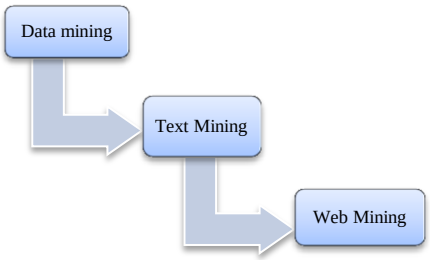

Fig.1 A hierarchical view of web usage mining

A large number of applications are presented to store and extract data from huge collections. Such computer based tools and methods are topic of discussion about Knowledge extraction in Database and Text mining is an interdisciplinary field which is used in different areas like

machine learning, information retrieval, statistics and data mining (D.S.Deshpande, 2012) to make mining algorithms (Gibson, Kleinberg e Raghavan, 1998). Some researchers say that text mining is a technique to find the new knowledge from huge volume of initial text using many algorithms. Web mining is another field in text mining which is used to mine the web data in form of web usage mining, web content mining and wed structure mining.

2.1 Data Mining

Data Mining means the analysis of large volumes of data stored in databases and data warehouses (Han e Kamber, 2006). It is multidisciplinary field which is also known as knowledge discovery in database and is used to extract knowledge from different areas such as knowledge-based systems, artificial intelligence, data visualization and high performance systems. According to another definition Data mining is referred as the phase to discover new trends in data. The process must be involuntary or (more commonly) semiautomatic. The patterns discovered must be meaningful in that they provide some benefits, mostly an economic advantage. The data is unvaryingly present in extensive quantities (Witten e Frank, 2005). This age is an age of Internet and world has become a global village .The speed and ease with which business applications are running on the internet is a very prominent force in the growth of electronic commerce.

Specially understanding user behavior and click movements on web has brought vendors and end customers closer to each other. Due to this it has become possible for vendor to customize his product according to the requirements of end users. This phenomena is referred as Mass Customization (Srivastava *et al.,* 2000). The above stated scenario leads to web mining.

2.2 Web Mining

Web mining is a process in which techniques of data mining are applied to discover patterns from the web. Web mining can be categorized as web usage mining, web structure mining and web content mining (Wikiepedia, 2014). Web structure mining (Gibson, Kleinberg e Raghavan, 1998) and web content mining (Spertus, 1997) are not discussed here. Web Mining taxonomy is provided in (Wu *et al.,* 2004) which also describes the architecture of Web Miner systems (Mobasher *et al.,* 1996; D.S.Deshpande, 2012).

3. Web Data

A major step in knowledge discovery in database (Fayyad, Piatetsky-Shapiro e Smyth, 1996) is to create a data set for data mining tasks. Web mining contains different kinds of data which is collected from client-side, server side or proxy servers. Each type is different from other on the base of data source, implementation methods and data location. Following types of data is used in web mining (Laender *et al.,* 2002):

Content: It is the actual data contained on web pages. Mostly it includes text and graphics.

Structure: This type of data explains the arrangement of the contents. It contains the explanation of HTML or XML tags within a given page. Inter page structure information is a type of hyperlink.

Usage: It describes the usage samples of web pages such as page references, data, IP addresses etc.

User profile: It provides the demographic bio data of users such as customer profile and registration information.

4. Data Foundations

The data will be collected about the usage of web site and web traffic generated by a single user to multi-access patterns.

4.1 Server Level Collection

A web server keeps record of visitor's behavior. The movements of web users are recorded in log files which have different formats. Although, it is best for the storage of user's behavior but some time it does not provide accurate information due to facts such as caching in web environment or the usage of POST methods. Packet sniffing is also used for collecting user's data. Packet sniffing uses TCP/IP packets to examine data. Web server also uses other efficacies such as CGI scripts to handle data.

4.2Client Level Collection

Client-side data collection is used to collect data on client side. It requires user cooperation for enabling the functionality of Java applications or accepts the functionality of a modified browser. Its advantage over server side is that it amends the session identification and caching problem. However actual views time of page is not managed well in client side. The most problematic step in this client level collection is to convince users to use modified browser for their tasks. This problem can be solved by giving incentives to users similar to programs offered by Netzero (Netzero, 2014) and Alladvantage (Alladvantage, 2014) that give reward to visitors for clicking on advertisements.

4.3Proxy Level Collection

A web proxy is an intermediate type of caching to reduce the load time of web pages and to minimize network traffic (Cohen, Krishnamurthy e Rexford, 1998). Their performance depends on ability to predict future requests correctly.

5. Approaches/methods in Web usage mining

Web usage mining consists of three phases for analysis of web usage. Below is a brief description of each of these phases.

5.1 Preprocessing

It is the initial phase which converts the usage, structure and contents from data source into data abstractions which are necessary for pattern discovery.

5.1.1 Usage Preprocessing

It is the most difficult task because the data which is required for this task is difficult to find. Client side tracking mechanism is helpful for this purpose. It involves many problems such as (Lu, Yuan e Lu, 1996):

An IP address with multiple Server Sessions: ISPs work with a pool of proxy servers which maintains information about the user access. One proxy server may have many users who visit the site at the same time.

*Multiple IP address/Single server Session***:** If a single user accesses the internet by using multiple IP addresses it is difficult to track it.

*Multiple agents with Single user***:** A user who uses different browsers from the same machine is treated as a different user.

After each user is identified as through different ways, their click stream is divided into multiple sessions and a thirty nine minute time is used as a method for breaking a user click into sessions. Thirty minute time out is based on the result of (Catledge e Pitkow, 1995). The content server arranges the definition of each session when it is embedded in URL.

Along with access to each user's session information it is also necessary to have access to content server information.

5.1.2 Content preprocessing

In Content preprocessing text, scripts, images and other type of files are converted to a form which is helpful in web usage mining process. Mostly it includes clustering or classification (Pentland, Picard e Sclaroff, 1996). The content of a site can be used for many other purposes such as filtering input to or output from algorithms etc. The results of an algorithm can be used to limit the patterns about a subject or class of product. Page views can also be classified to in term of their uses (Pirolli, Pitkow e Rao, 1996; Cooley, Mobasher e Srivastava, 1999). The usages of page views are to spread information, collect information from user, and filter the session and to allow navigation. The contents of a static view page can be easily preprocessed by reformatting information, parsing HTML or running other algorithms. The preprocessing of dynamic pages is a difficult process. Content may be revised on regular basis before preprocessing the contents of each page should be assembled through a template or by an http request. If only a portion of page view is preprocessed, then the output clustering may be lopsided.

5.1.3 Structure preprocessing

A site structure is created by creating hyperlinks among different pages views. It is maintained in the same way as the contents of a site. The structure of dynamic pages is difficult to maintain as compared to static pages.

5.2 Pattern Discovery

In pattern discovery, methods developed from other fields are used to make patterns. Hypertext links between pages can be used to create structure of a web site (Dhar e Tuzhulin, 1993). The contents of a site act as a guideline to create structure. Dynamic contents create more problem than static pages. The order of item selection is not taken into consideration but a server session is an ordered sequence of pages as the user has requested.

5.2.1 Statistical Analysis

It is the most common method to extract knowledge about users of a website. By analyzing traffic load to a website many facts can be determined such as numbers of times a user visited a website, viewing time, length of a path through a site etc. structure analysis enhance the performance of a system by error tracking and error finding.

5.2.2 Association Rules

Association rules can be used to relate multiple pages which are requested together in a one session. These pages may not be directly hyperlinked. Association rules may be helpful in pre-fetching documentation for users.

5.2.3 Clustering

In clustering a group of items having same characteristics are grouped together. In web usage two types of clustering is common: usage clusters and page clusters. In usage clustering a group of users having same interests are grouped together while in page clustering similar pages are grouped together.

5.2.4 Classification

In Classification data items are mapped to predefined classes (Fayyad, Piatetsky-Shapiro e Smyth, 1996). Classification is performed by using different algorithms. For example classification on logs of server lead to discover rules such as about the interests of a particular age groups and their demographic information etc.

5.2.5 Sequential Patterns

Sequential pattern is used for future predictions by finding inter session patterns in the way that presence of a set of items is followed by another pattern. Analysis that can be performed on the basis of sequential Pattern is trend analysis, similarity analysis etc.

5.2.6 Dependency Modeling

This model is helpful to show the dependency among multiple variables in the web domain. Many probabilistic learning techniques can be implemented to see the behavior of different users.

5.3 Pattern Analysis

Pattern Analysis is the last phase of web usage mining. In this phase irrelevant or uninterested rules are filtered out, and the exact methodology is applied. Its most common form consists of a query mechanism such as SQL (Zaiane,

Xin e Han, 1998; *(Auth.).* 2007). Visualization techniques such as signings color to values or grouping patterns are helpful in highlighting trends or patterns in data. In this aspect content and structure information is helpful to filter patterns on the basis of pages type, content type and usage.

6. Results and Discussions

This paper intends to concisely describe the process of data extraction from web data and phases which are involved in each step. It is evident from literature that each phase which is involved in web usage mining has its prose and crones. These phases range from collecting behavior of a single user to a group of users and from client to proxy and server level information is gathered throughout. The study also sought to learn that there are many user behavioral issues which are difficult to handle such as convincing user to adopt new approaches for web surfing or to convince them to use a changed version of a browser to measure variation in user's behavior. Along with this if same user logins using multiple browsers on the same machine then there are issues for his correct identification. But literature review indicated that there are solutions of such problems also. Incentives can be given to users for adopting new policies and trends in order to make them convince and to measure the efficacy of a new product.

Conclusion

This paper provides comprehensive information about phases of web mining. With the growth in web usage and web based applications the significance of web based extraction has gained numerous importance. So this paper intends to provide the description of each phase in order to understand its approaches and methods. The data extracted from each of these phases can be used for multiple purposes specially to improve the standard of a site according to requirements of visitors. When new trends and changes are applied, their impact can be found by measuring visitor's behavior and improvements can be performed on the basis of measurements.

References

*(AUTH.)., Z.-H. Z.(*2007*)*, Advanced data mining and applications. In: *REDA ALHAJJ*;H*ONG GAO, et al*, third international conference, ADMA, China.

Alladvantage. (2014) http://www.alladvantage.com. last accessed on 20th, May, 2014.

Anwer, N.; Rashid, A.; Hassan, S. (2010) Feature based opinion mining of online free format customer reviews using frequency distribution and Bayesian statistics. Networked Computing and Advanced Information Management (NCM), 2010 Sixth International Conference on, IEEE. p.57-62.

Catledge, L. D.; Pitkow, J. E. (1995), Characterizing browsing strategies in the World-Wide Web. Computer Networks and ISDN systems, v. 27, n. 6, p. 1065-1073, 1995. ISSN 0169-7552.

Cohen, E.; Krishnamurthy, B.; Rexford, J. (1998), Improving end-to-end performance of the Web using server volumes and proxy filters. ACM SIGCOMM Computer Communication Review, ACM. p.241-253.

Cooley, R.; Mobasher, B.; Srivastava, J. (1999), Data preparation for mining world wide web browsing patterns. Knowledge and information systems, v. 1, n. 1, p. 5-32. ISSN 0219-1377.

D.S.Deshpande. (2012), A survey on web data mining applications. emergind trends in computer science and information technology, international journal of computer applications.

Dhar, V.; Tuzhulin, A. (1993), Abstract-driven pattern discovery in databases. Knowledge and Data Engineering, IEEE Transactions on, v. 5, n. 6, p. 926-938. ISSN 1041-4347.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996), From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37. ISSN 0738-4602.

Gibson, D.; Kleinberg, J.; Raghavan, P. (1998), Inferring web communities from link topology. Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space---structure in hypermedia systems: links, objects, time and space---structure in hypermedia systems, ACM. p.225-234.

Goebel, M.; Gruenwald, L. (1999), A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explorations Newsletter, v. 1, n. 1, p. 20-33. ISSN 1931-0145.

Han, J.; Kamber, M. (2006), Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann. ISBN 0080475582.

Khushboo, N.; Vekariya, S. K.; Mishra, S. (2012), Mining of Sentence Level Opinion Using Supervised Term Weighted Approach of Naïve Bayesian Algorithm. International Journal of Computer Technology & Applications, v. 3, n. 3. ISSN 2229-6093.

Kohavi, R.; Provost, F. (2001), Applications of data mining to electronic commerce. Springer. ISBN 1461356482.

Laender, A. H. *et al*. (2002), A brief survey of web data extraction tools. ACM Sigmod Record, v. 31, n. 2, p. 84-93. ISSN 0163-5808.

Lu, H.; Yuan, S.; Lu, S. Y. (1996), On preprocessing data for effective classification. ACM SIGMOD'96 Workshop on Research Issues on Data Mining and Knowledge Discovery, Citeseer.

Mobasher, B. *et al*. (1996), Web mining: Pattern discovery from world wide web transactions. Technical Report TR96-050, Department of Computer Science, University of Minnesota.

Netzero. (2014), http://www.netzero.com. last accessed on 12th July,2014.

Pei, J. *et al.* (2000), Mining access patterns efficiently from web logs. In: (Ed.). Knowledge Discovery and Data Mining. Current Issues and New Applications: Springer. p.396-407. ISBN 3540673822.

Pentland, A.; Picard, R. W.; Sclaroff, S. (1996), Photobook: Content-based manipulation of image databases. International Journal of Computer Vision, v. 18, n. 3, p. 233-254. ISSN 0920-5691.

Pirolli, P.; Pitkow, J.; Rao, R. (1996), Silk from a sow's ear: Extracting usable structures from the web. Proceedings of the SIGCHI conference on Human factors in computing systems, ACM. p.118-125.

Seerat, B.; Azam, F. (2012), Opinion Mining: Issues and Challenges (A survey). International Journal of Computer Applications (0975–8887) Volume.

Shelke, N. M.; Deshpande, S.; Thakre, V. (2012), Survey of techniques for opinion mining. International Journal of Computer Applications (0975–8887) Volume.

Spertus, E. (1997), ParaSite: Mining structural information on the Web. Computer Networks and ISDN Systems, v. 29, n. 8, p. 1205-1215. ISSN 0169-7552.

Srivastava, J. *et al*. (2000), Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter, v. 1, n. 2, p. 12-23. ISSN 1931-0145.

Wikiepedia. (2014), http://en. wikipedia.org/wiki/Web_ Mining. last accessed on 27th, Jun,2014.

Witten, I. H.; Frank, E. (2005), Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann. ISBN 008047702X.

Wu, S.-T. *et al*. (2004), Automatic pattern-taxonomy extraction for web mining. Web Intelligence, 2004. WI. Proceedings. IEEE/WIC/ACM International Conference on, 2004, IEEE. p.242-248.

Zaiane, O. R.; Xin, M.; Han, J. (1998), Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on, IEEE. p.19-29.