Web Mining – Evolution and Applications

Mohd. Furkan

Research Scholar (Computer Science), CMJ University, Shillong, Meghalaya

Abstract— Web mining has unparallel opportunities for growth in the world of today and in this paper we provide an insight into the evolutions as well as applications of it. In the first part of the analysis a brief introduction of the topic along with the definition as well as the evolution mechanisms is illustrated in details. Then the advantages along with the applications of web mining is also explained in a detailed way. In the final analysis the future of web mining along with a summary is provided towards the end of the paper.

Index Terms— Data Mining, Web mining.

--♦-------------------------------------

1. INTRODUCTION

With unprecedented growth of information resources on the World Wide Web it has become quiet necessary for the users to adopt automated tools to filter the information as per their requirements. Two different schools of thought were adopted in defining it. First was the process centric approach which defines web designing as a sequence of tasks and the second was data centric approach where web designing is defined with regards to the type of web data that was used with regards to the mining process. In fact web mining has been the latest topic of research under dispute and has drawn interest from various communities. The difficult part in this regard is that no common ground has been created with regards to web mining. The researches have to engage more in a process of discussion to define it further (Wyld et al, 2011).

Web mining is an umbrella which is used to describe the three different types of data mining which are content mining, structure mining as well as usage mining. With the explosion of information mainly due to the online boom, the World Wide Web has become a powerful tool to store as well as disseminate information. Due to the complex nature of data, web data research has encountered a lot of challenges in the form of heterogeneous structure as well as scalability issues. As a result of this the users of web face the tricky situation of the overflow of information and generally encounter the following problems mainly (Xu, Zhang and Li, 2011).

- Finding relevant information- when a user looks for information on the web he resorts to query based search. The major problems in this regard is that a host of irrelevant pages turn out in this regard and then the issue of recall also arises which is caused by the capability of indexing all available pages on the internet. In fact how to find relevant pages related to the query search has become one of the major topics of discussion in the field of web management in the last decade or so.
- Finding specific information- Most of the information search on the web relates to a specific keyword in question. Sometimes the result which arises is not in alignment with the ones required by the user due to the existence of homograph. The semantics of web data is hardly taken into context in the domain of web search.

In the world of today web mining finds its prominence in various diverse fields (Ozkan, Chaki and Nagamalai, 2010).

Some of the common fields where it finds its application are ecommerce as well as web personalization. The prime objective of web designing is to enhance the web catching performance on all counts. In the midst of this it needs to be kept in mind that for generating data in systematic web sites it is considered essential to generate a data corresponding to the dynamic object.

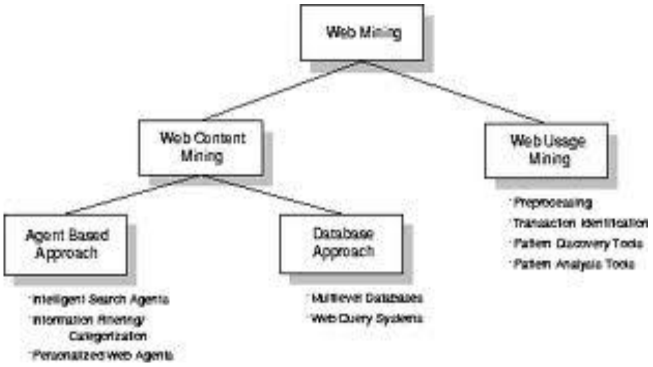

Figure 1: Web mining

2. WHAT IS WEB MINING AND EVOLUTION OF IT?

In the days gone by clustering was the most common form descriptive method of data analysis as well as data mining. It finds its prominence when there is a large number of data and it involves finding heterogeneous subsets also (Tuffery, 2008).

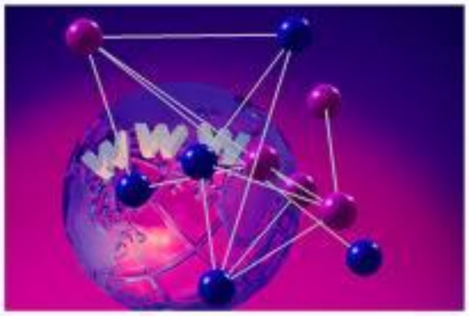

Figure 2: World wide web in web mining

The information needs of each and every user cannot be determined individually. As the information needs grow the business managers can aggregate and store the data effectively. On the other hand the managers also find it quiet difficult to interpret the mass data at the same time. This is the time where the concept of data mining has evolved big time. A lot of information is available on the internet for the users, but in most of the times it is observed that they face an information overload. Till now the search engines as well as the browsers cannot provide information specifically to meet the needs of the customers nor can effectively solve the problem of information overload on the internet. Web mining happens to be an emerging tool which analyzes the behaviour of the customers and in hindsight provides useful information to the customers to meet their needs also. It obtains data of importance from proxy servers, in the form of log files, transactions as well as user profiles. In fact according to the data as well as the mining process web usage mining provides an insight into the behavioural patterns of the users and a clear depth about their browsing behaviours also. On the other side of the coin the information needs of each and every individual tend to be ambiguous and it is very difficult to predict the exact consumer behaviour. In short the information needs of each and every individual cannot be defined correctly. For this precise reason the fuzzy theory is applied to understand the pedigree shift and patterns of each and every customer. Web mining theory could also be applied to understand the browsing patterns also (Ng et al, 2006).

Web mining is the synchronization of information which is collected by the traditional data mining techniques with information which is a part of the World Wide Web. Generally mining means extracting something from the core of the earth and the metal gold comes to our mind straight away. Web mining is a part of the customer relationship management or commonly referred to as CRM. It finds its utility in understanding the behaviour of the customers, study the effectiveness of a particular web site and gauge the success of a particular marketing campaign. It helps to look for the various patterns in data through structure, content as well as usage mining. Each of these has their own specific uses. Content mining evaluates the data collected by the various search engines whereas structure mining tries to understand the data involved in a particular structure. Web site mining evaluates the data submitted by the various users during the various transactions as well as the data collected by the forms submitted by the users. In fact whatever information which is collected through the web mining is evaluated with the help of software graphic applications or by using traditional data mining techniques such as clustering, classification etc.

Web mining is a younger discipline which concentrates on the analysis of the data pertinent to the web. The web mining methods are applied on usage data as well as web site content and they strive to make the web a better place to understand things in a better way. In a way they strive to enhance the usability factor and to promote the mutual satisfaction between the ecommerce platform as well as the potential customers. In the last few years the interest for the web as a medium of communication, business and interaction has led to the formation of new challenges and to dedicated as well as intensive research of sorts. In fact many of the infancy problems relating to the World Wide Web have been solved, but the widespread potential for the new as well as the improved uses, along with the misuses is presenting the web with more challenges (Nasraoui, Spiliopoulou and Srivastava, 2006)

The web mining involves a pre processing stage in which the original data is tailored to meet the specific requirements by mining techniques. In this regard the specific data may be converted or amalgamated with external data along with the data which is not of use may be dumped also. So a new version of mining technique involves the use of additional data components. Therefore the web mining techniques should avoid the obstruction the evolution of mining techniques (Masanes, 2006).

3. ADVANTAGES OF WEB MINING

Web mining seems attractive for the companies because of the host of benefits involved in it. It contributes to products by selling more products and at the same time minimizing costs. For this precise reason marketing intelligence is required and this intelligence can focus on marketing strategy as well competition analysis which focuses on building relationships with the customers. The different kinds of web data are categorized and clustered into the desired customer profiles. This not only helps the customers in providing personalized service to the existing customers but also helps in attracting potential customers.

The advantage of web mining is that it is more realistic than the probabilistic models. Among the various approaches of web mining content mining is the stand out among them. It involves the application of data mining techniques on the web. In the field of health care the use of web mining techniques has found widespread use. Some of the advantages can be depicted in the form of a table which are

- It acts as a virtual database in the form of link pages, where the content may be shared by multipurpose user applications. This is because of the fact that the web is a powerful tool for storing as well as sharing information that can be queried with a standard language

- The application of powerful web mining strategies based on previous patterns of web usage presents the integration of the web with artificial intelligence technologies. This results in the creation of new as well as interesting path for the users (Tan, 2010).

It needs to be understood that web design and all the technologies related to the web are dynamic as well as ever changing in nature. So the most important factor is to keep the structure of the programme as simple and flexible as possible. In this regard the log structure has to be considered as a dynamic source of information with regards to the structure as well as the content of the data in the first place (Perner, 2006)

More and more transactions are becoming digital in the world of today. This is not only happening from the supplier's side but also from the customer's side. Companies today are in a position where they can collect a large amount data quiet easily. By using the mechanism of web mining the companies can easily understand as well as predict the behaviour of the customers. In this regard the tool of web analysis is also very important as all the visitors to the website leave some information in the form of storage files. So the future definitely holds good in the coming days.

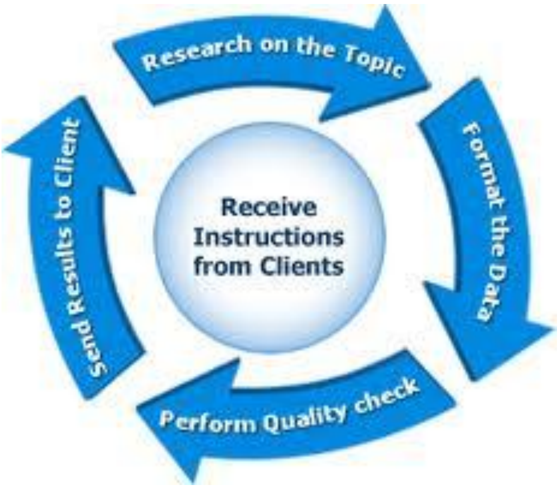

Figure 3: Web mining in providing customer centric solutions.

4. APPLICATIONS OF WEB MINING

As already illustrated in the earlier part of the discussion web mining can be classified into three major domains. The web is one of the biggest plethora's of information and the data mining in web based behaviour improves the system performance as well as enhances the quality of information to the end user. It can also help the organization to determine the lifetime value of the customers and cross marketing strategies across products. In the midst of this even though the web is unstructured and dynamic, it provides the users the options of semi structured documents. Examples are linguistic conventions as well as web directories (Sumathi and Sivanandam, 2011).

The web mining is an application of data mining. The widespread use of resources on the web have promoted the need for developing automated mining techniques giving rise to web mining. Web mining is the application of data mining techniques on the web in order to discover useful patterns and can be divided into three major categories (Tatnall, 2007). A glance at the techniques is as follows

- Web usage mining- This includes techniques in assisting the users in locating web documents. This relates to pages that meet certain specific criteria's

- The second relates to discover information based on the web site structure data

- The third category focus on analyzing web access logs along with other uses of information concerned user interactions on the web.

In fact the close relation between web mining and web personalization has motivated too much research work in the area. It is a complete process and consists of primary data relating to data collection, data processing

The last decade or so has seen significant growth in DM technology and applications in particular. In fact the web has created a host of information available for DM and this has created new opportunities. One of the areas which have emerged is web mining which is global information available from the retrievable websites. The web is the pandora of information and offers unprecedented opportunities for knowledge discoveries. It needs to be kept in mind that the discovery process is more difficult because of lack of the uniformity between websites along with the prevalence of structured as well as unstructured information. The web mining applications are becoming more and more feasible. The new challenge which lies is to take specific content from the browser and incorporated into the data sheet or an email alert.

It is very difficult to replicate data mining as an automated process. The prime reason for this is that web is not constructed as a database and does not involve standard protocols (Meghabghab and Kandel, 2008)

Web mining is a fascinating approach to the web and needs to make some sense, and how people use and build it. The results of web mining in turn shape the web and its usage. Therefore the basic understanding of this field is the key component of digital technology and a deeper understanding is a commendable extension to the knowledge (Melucci and Baeza-Yates, 2011)

5. FUTURE OF WEB MINING

The future of web mining depends to a great extend on the development of the semantic web as the role of technology is bound to increase in the field of education, entertainment as well as the government sectors. In fact the future of web mining hinges on the challenges which are bound to evolve in the days to come. Lot of research as well as recent developments have contributed to this big time. Web mining has contributed immensely to the way in which information is organized, structured, disseminated along with being retrieved. This in a way allows the flourishing of the web mining applications that extends the areas beyond web mining. This has profound impact on the business as well as the social environment business. Researchers have a huge role in managing the challenges emerging out of web mining (Wang, 2006).

In short web mining is concerned with the emerging needs of the business and the future is intertwined with it in a big way. The development of the techniques should be fuelled by the gaps as well as improvement of the existing techniques. One has to understand the new horizons of web mining, the areas which are under research and how they can help in a business computing setting in the long run (Adomavicius and Gupta, 2009).

A website creates a lot of information which is more than itself. The real utility of a website is derived when the web traffic data is combined with corporate databases such as sales automation systems, inventory as well as accounting systems. When a correlation of all the resources is done the business can turn the information available at their disposal into important ones. So the fact which emerges is that web mining must coordinate and integrate all the data resources irrespective of the underlying software or the hardware in question. Information is vibrant and changes at a split second and on the other hand the corporate databases might be hundreds of giga bytes, so a mining system helps it to link with the other databases (Saxena, 2009).

In the world of today most of the data warehouses are used for summarization method, online analytical processing as well as multidimensional structures. But there has been a sea change in the domain of web mining as it is expected that organizations will be using data warehouses for sophisticated data analysis. This paves way for the development of Online data mining (ODM) in the coming days. In the coming days the technology will enable e commerce to do personalized marketing which will contribute to higher volumes of sales. In fact the agencies of the government are using this technology to classify threats as well as fight against tourism. The company can establish customer relationship by providing the customers what they exactly need. The needs of the customers can be understood in a better way and the company can take the requisite steps to provide ultimate in terms of products as well as services. All this augers well for the business in the days to come.

6. CONCLUSION

From all the analysis done as part of the subject matter under discussion it is quiet clear that web mining is an active research area. As the domain grows it will provide more opportunities for usage along with the chances of analyzing the data from the web and all the useful information being extracted from it is the urgent need of the hour. In the last five years of so the web mining has evolved leaps and bonds. The precise reason for this has been the efforts of the research community as well as the organizations that are practicing it also. Notable contributions in the domain of computer science have also been made in this regard also. The popularity of the web usage mining will continue to grow with the popularity of the www and this will have a significant impact on the study of the online behaviour pattern of the users.

Web mining has tremendous potential in the days to come, but one has to channelize its effectiveness in the best possible way and remove the loopholes associated with it. So the urgent need of the hour is to develop a user oriented as well as a concept oriented approach. So a multilingual approach is called for.

REFERENCES

1. Wyld, D., Wozniak, M., Chaki, N., Meghanathan, N., and Nagamala, D. (2011). Advances in Network Security and Applications: 4th International Conference, CNSA 2011. Chennai: Springer

2. Xu, G. (2011). Web Mining and Social Networking. London: Springer Özcan, A., Chaki, N., and Nagamalai, D. (2010). Recent Trends in Wireless and Mobile Networks: Second International Conference, WiMo. Berlin: Springer.

3. Tufféry, S. (2011). Data Mining and Statistics for Decision Making. Sussex: Wiley

4. Ng, H. (2006). AIRS 2006. Chennai: Springer

5. Nasraoui, O., Spiliopoulou, M., and Srivastava, J. (2006). Advances in Web Mining and Web Usage Analysis: 8th International Workshop on Knowledge Discovery on the Web, 2006. Chennai: Springer.

6. Masanès, J. (2006). Web Archiving. New York: Springer.

7. Tan, J., Payton, C. (2010). Adaptive Health Management Information Systems: Concepts, Cases, & Practical Applications. USA: Aspen.

8. Perner, P. (2006). Advances in Data Mining: Applications in Medicine, Web Mining, Marketing, Image and Signal Mining. Chennai: Springer.

9. Sumathi, S., and Sivanandam, S. (2006). Introduction to Data Mining and its Applications. Berlin: Springer.

10. Tatnall, A. (2007). Encyclopedia of Portal Technologies and Applications. UK: IGI Global.

11. Meghabghab, G., and Kandel, A. (2008). Search Engines, Link Analysis, and User's Web Behavior: A Unifying Web Mining Approach. Berlin: Springer.

12. Melucci, M., and Baeza-Yates, R. (2011). Advanced Topics in Information Retrieval. New York: Springer.

13. Wang, J. (2006). Encyclopedia of Data Warehousing and Mining. London: Idea.

14. Adomavicius, G., and Gupta, A. (2009).n Business Computing. UK: Emerald.