ISSN 2047-3338

Research Challenges in Web Data Mining

Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik

***Abstract*— With the exponential growth of information in online, the World Wide Web (WWW) is a fruitful area of data mining. Mining data over the web using different tools and technologies of data mining is termed as web mining. Web mining could be used to solve the information overload problems directly or indirectly. This paper provides a brief overview both in terms of technologies and applications and outlines key future research directions.**

***Index Terms*— Data Mining, Information Retrieval, Knowledge Discovery, Data Mining Trends, Web Mining, Pattern Analysis and Pattern Discovery**

I. INTRODUCTION

THE popularity of World Wide Web (WWW) has made it a fertile ground for dissipating information. Due to the properties of huge, diverse and dynamic and unstructured nature of web data, web data research has encountered a lot of challenges for data mining principles, or web mining. The web mining filed encompasses a wide array of issues, aimed at extracting actionable knowledge form the web, and includes researchers from information retrieval, database technologies, and artificial intelligence [1]. In the present time, it is not easy task to retrieve the desired information because of more than 1 000 000 000 pages are indexed by search engines. So, this redundancy of resources has enhanced the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Data mining" [2].

The Web is changing fast over time and so is the user's interaction in the Web suggesting the need to study and develop models for the evolving Web Content, Web Structure and Web Usage. Therefore, the most important criterion to differentiate between a successful and an unsuccessful business based on mining data.

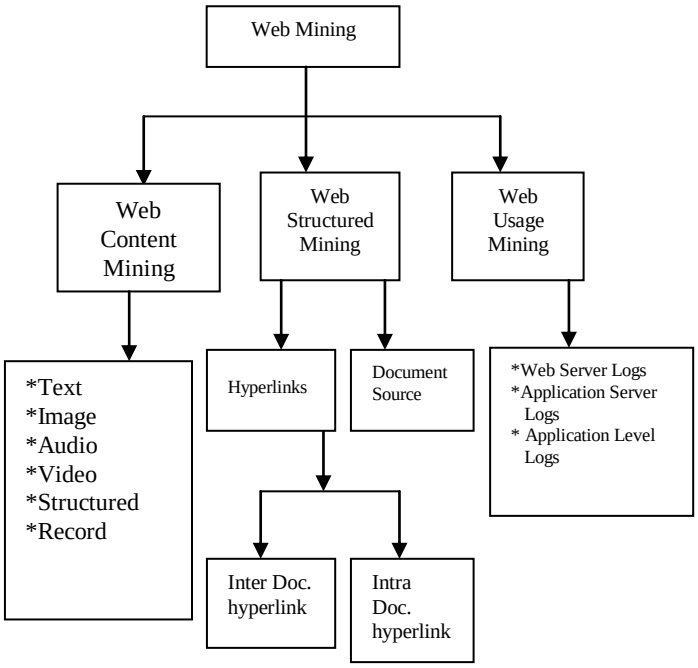

Fig. 1. Web Mining Taxonomy

Researchers have identified the broad categories of web mining [3]:

1) Web Content Mining is the application of data mining describes the discovery of useful information from the web contents, data and documents to content publish on Internet, usually as HTML (semi structured), plain-text (unstructured), or XML (Structured). Web content mining is especially representative to the attributes of text when it occurs in Web resources. Therefore, the aim of it on the discovery of patterns in large document collection, and in frequently the collections of document changing [4]. Further the methods of content mining will be used for ontology learning, mapping and merging ontologies, and instance learning [5].
2) Web Structure Mining operates on the web's hyperlink structure. The graph structure can provide information about a page's ranking [6] or authoritativeness [7] and

Md. Zahid Hasan is with the Green University of Bangladesh (GUB), Dhaka-1207, Bangladesh, phone: +8801672580748; e-mail: hasan.ice@ gmail.com).

Khawja Jakaria Ahmad Chisty, is with International Islamic University Chittagong (IIUC), Dhaka Campus, Bangladesh. He is now with the Department of Electrical and Electronics Engineering (e-mail: kja_chisty@yahoo.com).

Nur-E-Zaman Ayshik is with the Electrical and Electronics Engineering Department, University of Bangladesh University of Engineering and Technology (BUET), Bangladesh (e-mail: nez.ayshik@gmail.com).

enhance search results through filtering. Web structure mining and Web content mining are often worked together to exploit the content and the structure of hypertext [5]. Similarly, the search engine Google owes its success to the PageRank algorithm, which focuses that the relevance of a page increases with the number of hyperlinks to it from other pages, and in particular from other relevant pages [8]. Indeed, some researchers included both under the notion of Web content mining [9].

3) Web Usage Mining is the automatic discovery of user interactions with a web server, including web log, click streams and database transactions at a website or a group of related sites [10] .Web usage mining focuses on privacy concerns and is currently the topic of extensive debate. The knowledge gathered from Web usage mining can be very useful in many Web applications such as Web caching, Web perfecting, intelligent online advertisements, and in addition to construct Web personalization. Most of the research efforts for modeling personalization systems are clustering pages or user session, association rule generation, sequential pattern generation and Markov models [11].
4) Text mining, also known as text data mining or knowledge discovery from textual databases, refers to process of extracting interesting and non-trivial patterns or knowledge from the large volume of text documents. In the text mining systems, are based on natural language processing where the integration function integrated with the products for knowledge distillation [12].

The objective of this paper is to provide an outline of web data mining and to give a perspective of some important research contributions in web mining, with a goal of providing a broad overview rather than in-depth analysis.

II. Research Challenges

A. Web Metrics and Measurement

From the experimental viewpoint of human behaviorist's, the web is the perfect experiment appliance. There are number of web measurement or web analytics techniques have the ability of measuring for human experimental behaviorists. The measurement ways are hits, page views, visits or user sessions and find the unique visitor regularly used to measure the user impact of various proposed changes. An example of operational metrics such as site visits and visit/buy ratios, as well as on financial metrics such as revenue and profit. Research should to be done in developing the right set of web metrics, and their measurement procedures, so that various web phenomena can be studied.

B. Process Mining

When Process mining aims at extracting information form event log to capture the business process as it is being executed. For example in market based data collected at the point of sale in any store provides only the process end result. The overall goal of any online store to maximize probability of reaching final state (Complete purchase) or maximize expected sales from each visit is shown in Fig. 2. Whenever the click stream data provides the opportunity for making the decision process itself and extracted the knowledge from where it can be used for optimizing and influencing the process. Research needs to be carried out in:

i) Process mining starts by accumulating information about the process as they take place. So, extracting process models from usage data.

ii) It's important to understand how different parts of the process model impact various web metrics of interest and

iii) Process models can be changed in response to various changes that are changed, so how to change the models which can also change stimuli to the user [13].

C. Web Evaluation

Web mining has been explored to a vast degree and different methodologies have been proposed for a variety of applications including web search, classification, personalization, etc. The temporal behavior of web mining has been classified into three kinds of web data: web content mining, web structure mining, and web usage mining. The researchers have to be finding out in extracting temporal models of how web content, web structures and web communities, authorities, etc are evolving. Large institutions and organizations archive usage data from the web sites.

From a statistical approach of web log data to determine the browser type of our visitors. With the availability of these data, there is a large scope to develop techniques for analyzing of how the web evolves over time. The popular two browsers of Chrome and Mozilla Firefox as can be observed from the right-hand side and pie chart on the left depicts percentages of the total hits per browser, in Fig. 3 representing the volumes of accesses for a web browser [14].

D. Optimization of Web Services

Data mining is basically classified into web mining; web content, web structure and web usage mining; the last two are used to solve the website structure the optimization problem.

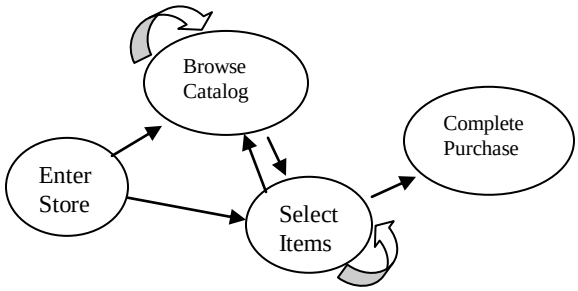

Fig. 2. Transition Diagram of Online Shopping Model

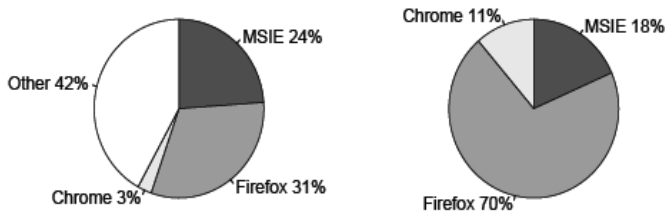

Fig. 3. Pie charts representing volumes of accesses for a web browser [14]

To make the robust, scalable and efficient services should be provided for the growing demand of web. So, web mining can be applied to better understand and behavior of these services. To improve various aspects of web services, developing of web mining methods logics are needed.

E. Fraud and Threat Analysis

Data mining technologies have advanced a great deal. The main problem issue is that, are they ready for detecting and/or preventing fraud activities and can we completely remove the false positive and false negatives? The challenge is to find how we can gather knowledge directed data mining to eliminate false positives and false negatives.

Another challenge of data mining is in real-time. The available tools of data mining have the ability to detect credit card violations and calling card violations. The research community should have a challenge to build a real time model. The challenge is necessary for many companies where they have interactions with up to millions of external parties. Figure 4 details the subgroups of internal, insurance, credit card, and telecommunications fraud detection which is very concerned for both the researchers and particular organization [15].

F. Counter Terrorism

Privacy is a major challenge with respect to data mining for counter-terrorism. In this scenario, the challenge is to extract the structure and usage patterns or mine useful information form data mining but at the same time maintain privacy. Different efforts are under way for privacy preserving data mining. There are various using techniques such as randomization, cover stories as well as multi party policy enforcement for privacy preserving data mining. That is while data mining could become a useful tool for counter-terrorism, there are many challenges need to be addressed [16].

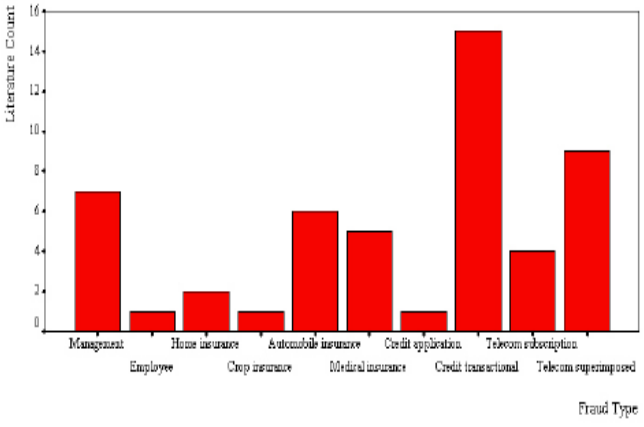

Fig. 4. Bar chart of fraud types from 51 unique and published fraud detection papers [15]

G. Semantic Web Mining

The research area of semantic web mining is targeted to combine two fast-developing research areas semantic web and web mining. The researchers are very much interested to improve from both areas of the results of web mining by exploring semantic structures in the web. The interesting thing of semantic web mining to create itself as the dependence between the semantic web and web mining increases. These research activities benefits many areas of industry such as 'e-activities', health care, privacy and security and knowledge management and information retrieval. So, the researchers need to be carried out in to explore the semantic structures in the web [17], [18], [19].

III. Conclusion

As seen in web metrics and measurements that various web phenomena can be studied in developing the right set of web metrics and their measurement procedures. The goal of process mining of any online enterprise used to maximize expected sets from each visit which is depicted in Fig. 2. As because large institutions and organizations achieve usage data from the website so a model is to be researcher out in extracting temporal models of Web content, Web structure, Web communities, authorities, etc. are involving. Optimization of Web services are needed to make the robust scalable and efficient services should be provided for the growing demand of the Web. For fraud and threat analysis, knowledge directed data mining to eliminate false positive and false negative is to be developed efficiently. For Counter terrorism, many challenges are needed yet to be addressed to make data mining to become a useful tool. Research is to be carried out is to explore the semantic Web structure in the Web for getting benefits in many areas of the industries.

REFERENCES

[1] Sourav S. Bhowmick, Wee-Keong Ng, Ee-Peng Lim, "Information Coupling in Web Databases", In Proceedings of the 17th International Conference on Conceptual Modeling, 1998.

[2] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE transactions on neural network, Vol. 13, No. 5, September 2002, pp.1163-1177.

[3] Jebaraj Ratnakumar, "An implementation of web personalization using web miming techniques", Journal of Theoretical and Applied Information Technology, Vol. 18. No.1, 2010, pp.67-73.

[4] J. Allan, editor. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic Publishers, Norwell, MA, 2002.

[5] Gerd Stummea, Andreas Hothoa, Bettina Berendtb,"SemanticWeb Mining: State of the art and futuredirections", Journal of Web Semantics, Vol. 4, Issue 2, June 2006, pp. 124–143.

[6] Larry Page, Sergey Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bring Order to the WebStanford Digital Library Technologies Project , Jan. 1998.C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[7] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Proc. 9th Ann. ACM–SIAM Symp. Discrete Algorithms, ACM Press, 1998, pp. 668–677.M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

[8] L. Page, S. Brin, R. Motwani, and T.Winograd. The PageRank citation ranking: Bringing order to the web. In Proceedings of the 7th International World Wide Web Conference, pages 161–172, Brisbane, Australia, 1998.

[9] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. In Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97). IEEE Computer Society, Nov 1997.

[10] Mirela Pater ; Daniela E. Popescu ; Daniela Maştei, Pattern discovery techniques in Web mining, Journal of Computer Science and Control Systems, Vol.1, Issue-1, 2008, ISSN 18446043, pp. 77-81.

[11] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, (2000) "Web usage mining: Discovery and applications of usage patterns from Web data", SIGKDD Explorations, Vol. 1, No. 2, pp. 12-23, 2000.

[12] Ah-hwee Tan,"Text Mining: The state of the art and the challenges, In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, 1999.pp. 65-70.

[13] Wil van der Aalst, "Process Mining: Making Knowledge Discovery Process Centric", ACM SIGKDD Explorations Newsletter, Vol. 13, Issue 2, New York, NY, USA , December 2011, pp. 45-49.

[14] Olga Baysal, Reid Holmes, and Michael W. Godfrey, "Mining Usage Data and Development Artifacts", Proc. of the 2012 IEEE Working Conference on Mining Software Repositories (MSR-12), Zürich, Switzerland, May 2012.

[15] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions", IEEE transactions on neural network, Vol. 13, No. 5, September 2002, pp.1163-1177.

[16] Bhavani Thuraisingham, "Data Mining for Counter-Terrorism", Chapter-3, MITRE Corporation, Burlington Road, Bedford, MA.

[17] Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., Stumme, G., "A Roadmap for Web Mining: From Web to Semantic Web", Web Mining: From Web to Semantic Web, Vol. 3209/2004, 2004, pp. 1–22.

[18] Gerd Stumme ,Andreas Hotho , Bettina Berendt, "Semantic Web Mining", Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Vol.4 Issue 2, June, 2006, pp. 124-143.

[19] M. Manuja&D. Garg, *"Semantic Web Mining of Un-structured Data: Challenges and Opportunities", International Journal of Engineering (IJE), Vol. 5, Issue 3, 2011, pp. 268-276.*