Data Mining Techniques: Elucidation to Improve the Data Stream Analysis in Web Usage Mining

[1] A.Swarna, [2]MD.Zaheer Ahmed
[1]Associate Professor, [2]Assistant Professor
Computer Science and Engineering
Vidya Jyothi Institute of Technology, Hyderabad, India

Abstract: Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories. Data Mining functionalities should allow mining of multiple kinds of patterns that accommodates different user expectations and applications, discover patterns at different levels of granularity hints , specifications ,queries to focus the search for interesting patterns. Data objects are linked together to facilitate interactive access. Users traverse from one object to another via links .Such Data allows more challenging opportunity for Data Mining. Understanding user access patterns allows better web design and better marketing strategy. This is called as Web Usage Mining or Web Log Mining. Web pages are highly unstructured. Web Page Analysis – Ranks web pages – helps in easy retrieval of relevant content pages when key word search is made. Data stream is in which data flows in & out of the platform (or window) dynamically. Traffic analysis is based on click stream data and it is easily being performed to improve the efficiency of e-bank services. Click stream data is one of the most important sources of information in websites usage and customers behavior in banks e-services. Many number of web usage mining scenarios depends upon the available information in the data warehouse. To improve the efficiency of e-banks services, banks need data mining techniques to improve their activities. Web usage mining and data mining techniques are integrated different stages for processing. Mostly banks use analytical mining techniques which include pattern discovery phases. A solution is defined for better performance, acquiring new customers, fraud detections in real time and customer purchase pattern over time.

Keywords: **Data mining techniques, Web mining, Data stream, Web Usage structure, Web content.**

I. INTRODUCTION

With the continued growth and proliferation of e-commerce, Web services, and Web-based information systems, personalization has emerged as a critical application which is essential to the success of a Web site. It is now common for Web users to encounter sites that provide dynamic recommendations for products and services, targeted banner advertising, and individualized link selections. Indeed, nowhere is this phenomenon more apparent as in the business-to-consumer e-commerce arena. The reason is that, in today's highly competitive e-commerce environment, the success of a site often depends on the site's ability to retain visitors and turn casual browsers into potential customers. Automatic personalization and recommender system technologies have become critical tools, precisely because they help engage visitors at a deeper and more intimate level by tailoring the site's interaction with a visitor to her needs and interests.

Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users. The experience can be something as casual as browsing a Web site or as (economically) significant as trading stocks or purchasing a car. Principal elements of Web personalization include modeling of Web objects (pages, etc.) and subjects (users), categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. The actions can range from simply making the presentation more pleasing to anticipating the needs of a user and providing customized information.

In the context of Web personalization and recommender systems, the use of semantic knowledge can lead to deeper interaction of the visitors or customers with the site. Integration of domain knowledge allows such systems to infer additional useful recommendations for users based on more fine grained characteristics of the objects being recommended, and provides the capability to explain and reason about user actions. Recent work in Web usage mining has focused on the extraction of usage patterns from Web logs for the purpose of deriving marketing intelligence. Despite the advantages, usage-based personalization can be problematic when little usage data is available pertaining to some objects or when the site content may change regularly.

Web usage mining will reduce the need for obtaining subjective user ratings or registration-based personal preferences. Web usage mining can also be used to enhance the effectiveness of collaborative ltering approaches [6, 16]. Collaborative filtering is often based on matching, in real-time, the current user's role against similar records (nearest neighbors) obtained by the system over time from other users. However, as noted in recent studies [10], it becomes hard to scale collaborative altering techniques to a large number of items, while maintaining reasonable prediction performance and accuracy. One potential solution to this problem is to first cluster user records with similar characteristics, and focus the search for nearest neighbors

only in the matching clusters. In the context of Web personalization this task involves clustering user transactions identified in the preprocessing stage.

II.LITERATURE SURVEY

Web Content mining can be done by retrieving information from unstructured document such as free text and semi structured document such as hypertext documents. In unstructured documents mining can be done by using word positions in the documents, text classification, event detection and tracking, finding extraction patterns in the text documents. The method used for semi-structured documents are hypertext classification and clustering, learning relations between web documents, learning extraction pattern or rules, and finding patterns in semi-structured data [1]. Web content mining is being used in various different areas like In [7] web content mining is used for mining online news sites. Beyond analyzing the news, they focused on current society interest and measured the social importance of ongoing events. Dynamic Crawler was used for resource finding. For trend analysis they used domain independent statistical analysis. Four stages of dynamic news analysis are Resource Identification, preprocessing, Generalization and Analysis. In Resource Identification phase, dynamic web crawler downloads page from current URL and then filters the downloaded pages and analyze the identified news report. Steps are repeated till the queue of URL's are empty. In preprocessing stage the news are converted into structured format. Interesting trends among new topics are found in Generalization stage. In analysis phase user analyzes the pattern and the process is repeated until interesting news is found. According to this system crawler downloaded 350 web pages each day and only 130 were selected for further analysis during the period of two weeks. It has been shown by experimentation that enough differences on news topics after a period of two weeks have been found.

Another area where web content mining has been proved very useful is a web content suggestion system for distance learning and is described in [2]. Two ways of Suggestions are collaborative filtering and content based filtering. Collaborative suggestion clusters students into groups with similar behavior .Content based filtering provide web pages to the students who have navigation records. Web page navigation behavior is stored in personal records. Students who are new to attend the course will be having less navigation record so they are asked to poll the interest. Content Suggestion system works with the help of six components such as Student Assistant Agents, Student Identification Component, suggestion Generation Component, Suggestion Delivery Component, Data Warehouse.

REFERENCES

[1]Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations. Vol. 2, 1-15.
[2]Yang, C. Y., Hsu, H. H. and Hung, J. C. 2006. A Web Content Suggestion System for Distance Learning. Tamkang Journal of Science and Engineering. Vol. 9, No. 3, 243-254.
[3]Bassiou, N. and Kotropoulos, C. 2006. Color Histogram Equalization using Probability Smoothening. Proceedings of XIV European Signal Processing Conference
[4]Bharanipriya, V. and Prasad, K. 2011. Web content Mining Tools: A Comparative study. International Journal of Information Technology and Knowledge Management. Vol. 4. No 1,211- 215.
[5]Cooper, M., Foote, J., Adcock, J. and Casi, S. 2003. Shot Boundary Detection via Similarity Analysis. In Proceedings of TRECVID 2003 workshop.
[6] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.
[7] Torreblanca, A. M., Gomez, M. M. and Lopez, A. L. 2002. A Trend Discovery System for Dynamic Web Content Mining. Proceedings of the 11th International Conference on Computing.
[8]Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. Tapping into the Power of Text Mining. Communications of the ACM – Privacy and Security in highly dynamic systems. Vol. 49, Issue-9.
[9]Fayyad, U. M. 1995. SKICAT: Sky Image Cataloging and Analysis Tool. ACM Proceedings of the 14th International joint Conference on Artificial Intelligence. Vol. 2.
[10]Gedov, V., Stolz, C., Neuneir, R., Skubacz, M. and Siepel, D. 2004. Matching Web Site Structure andContent. ACM. Proceedings of the 13th International World Wide Web Conference on Alternate track papers and posters.