Global Journal of Computer Science and Technology
Network, Web & Security
Volume 12 Issue 17 Version 1.0 Year 2012
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Web Usage Mining Architecture and Applications

By Mandeep Josan & Dr. G.N. Singh

Shri Guru Gobind Singh College, Chandigarh

Abstract - The WEBMINER is a system that implements parts of this general architecture. The first part is domain dependent application. The second part is the domain independent application. This includes pattern discovery and analysis as part of the system's data mining engine. The overall architecture for the Web mining process is depicted below:

GJCST-E Classification : H.2.8

Web Usage Mining Architecture and Applications

Mandeep Josan [α] & Dr. G.N. Singh [σ]

I. Introduction

The WEBMINER is a system that implements parts of this general architecture. The first part is domain dependent application. The second part is the domain independent application. This includes pattern discovery and analysis as part of the system's data mining engine. The overall architecture for the Web mining process is depicted below:

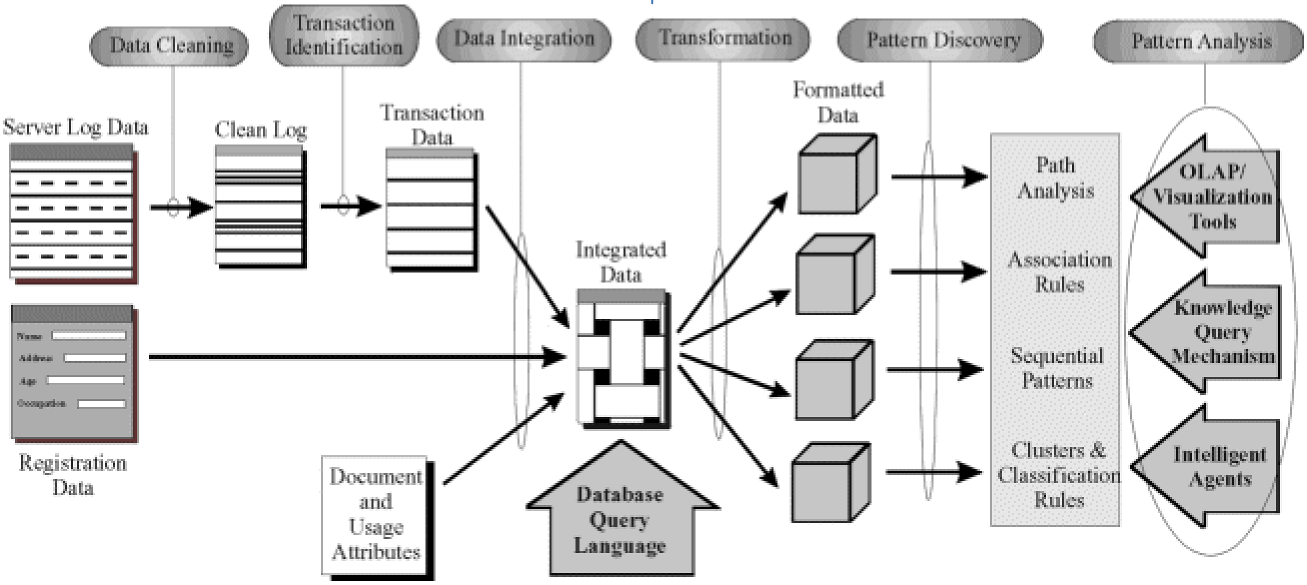

Figure : A General Architecture for Web Usage Mining

Just to briefly explain the above figure, data cleaning is the first step performed in the Web usage mining process. We have discussed some of the techniques to clean the log data. "Currently, the WEBMINER system uses the simplistic method of checking filename suffixes. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc".

After the data cleaning using one or a series of transaction identification modules into clusters. We have discussed a few techniques how to separate data into transactions. The "WEBMINER system currently has reference length, maximal forward reference, and time window divide modules, and a time window merge module".

"Access log data may not be the only source of data for the Web mining process. User registration data, for example, is playing an increasingly important role, particularly as more security and privacy conscious client-side applications restrict server access to a variety of information, such as the client user IDs". The data collected through user registration is then integrated with the access log data. There are also known or discovered attributes of references pages that could be integrated into a higher level database schema. The discovered attributes could include page types, usage frequency and link structures. "While WEBMINER currently does not incorporate user registration data, various data integration issues are being explored in the context of Web usage mining".

"In WEBMINER, a simple Query mechanism has been implemented by adding some primitives to an SQL-like language". This allows the user to specify his patterns of interest to the mining engine.

This information from the query is used to reduce the scope, and thus the cost of the mining process. The development of a more general query mechanism along with appropriate Web-based user

Author α : Shri Guru Gobind Singh College, Chandigarh.
E-mail : mandeep.jgill@gmail.com
Author σ : Department of Physics and Computer Science, Sudarshan Degree College, Lalgaon Distt. Rewa (M.P.) India.

interfaces and visualization techniques, are still in research.

a) BENEFITS

Let's have a look at some of the benefits you get from Web mining:

II. Match Your Available Resources to Visitor Interests

Resources can be products you "sell, information fragments you distribute online, banner ads from your client advertisers, e-mail fragments from a mailing list, or anything else" which is distributed online. "Metadata of these resources are then stored in a database. WebAnalyst helps learn visitor interests by collecting and analyzing information generated by interactions with your website, such as clickstream data, search requests, and cookies. WebAnalyst can use the gleaned knowledge to rank your resources by their relevance to the user's interests. Servicing a user request for information, with the best matching resources, results in a higher visitor-to-customer conversion rate for your e-business".

III. Increase the Value of Each Visitor

Upon carrying out collaborative filtering, we can predict what kind if information a visitor may be interested in, and the products she might consider purchasing. "These predictions used to present the visitor with related products and resources", and hence chances of them purchasing it. "This knowledge significantly increases the value of a customer for an e-business when used in individualized cross-selling and up-selling promotions, and thus increasing revenue."

IV. Improve the Visitor's Experience at the Website

"A sound combination of data and text mining techniques can help determine user interests - early in the process of the visitor's interaction with the website. This allows the website to act interactively and proactively and deliver the most relevant customized resources to the visitor". In the world of Internet, easy access to relevant information might make a difference between a profitable customer and lost opportunity. "By increasing the customer's satisfaction, you reduce attrition and build brand loyalty".

V. Perform Targeted Resource Management

Since, all visitors are different in buying behavior you may notice that some of them are your best potential customers, ready to click and buy, while others are prospecting for information, simultaneously familiarizing themselves with your brand. These prospecting customers may become "very important and profitable customers" in the future. Also not to forget there is another group of visitors who enjoy only free rides. "These folks will use promotional resources that you offer to the fullest extent, but will never purchase anything. All these visitors come through a single pipe to your website and are in a common queue for your website resources". It's best to your advantage if you can tell each type of visitor apart from the other. "Your website performance is limited and you might want to prioritize requests coming from your best prospects. If you are distributing promotional resources of high value, you might want to spend your promotional budget wisely by offering and delivering your promotional materials only to your best prospects - not to every Web surfer on the planet. WebAnalyst can work with load-balancing products to provide the best quality of service to your best customers".

VI. Collect Information in New Ways

"While for the majority of e-vendors the task of collecting data is just an intermediate step necessary for better targeting their marketing, for others this task might be the main motivation for creating a website itself". Traditional data collection methods like promotions, surveys, focus groups, etc. have many well known problems, including high cost, poor response rates and low accuracy. "Now imagine that you can offer your promotional items online through a content-rich website, where visitors can find useful information in addition to submitting their contact information and requesting the promotion. WebAnalyst can learn the visitor's preferences (at virtually no cost) based on the content that the user was browsing. Of course, WebAnalyst is designed to work hand-in-hand with your privacy management system, allowing you to collect valuable data while respecting the privacy of your visitors".

VII. Test the Relevance of Content and Web Site Architecture

Perhaps you would like to increase usability, or optimize your website for the eyes of your best prospects by taking close look at the website's content and architecture. "Log analyzers can help you visualize the most navigated paths through your website, averaged over all visitors. When optimizing your website structure, your main concern should be to improve experience of your most promising prospects, and not just everybody. Roughly 15% of your website visitors comprise really valuable prospects. The remaining 85% have little value to you other than sustaining the brand recognition traffic. Thus you have to segregate your least important prospects and subtract their contribution from the overall picture of the site navigation. What is left represents the real quality of your website. This is the picture that can help you really improve your bottom line".

VIII. Web Mining Applications

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.

References Références Referencias

1. Allan, J. and H. Raghavan. "Using part-of-speech patterns to reduce query ambi- guity." In *Proceedings of the 25th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, Tampere, Finland, 2002. ACM Press. http://doi.acm.org/10.1145/564376.564430.
2. Aslam, J. A. and M. Montague. "Models for metasearch." In *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284, New Orleans, LA, 2001.
3. Beitzel, S. M., E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman. "Temporal analysis of a very large topically categorized web query log." *Journal of the American Society for Information Science and Technology*, 2008
4. Beitzel, S. M., E. C. Jensen, A. Chowdhury, and D. Grossman. "Using titles and category names from editor-driven taxonomies for automatic evaluation." In *Proceedings of the 12th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 17–23, New Orleans, LA, 2003. ACM Press. http://doi.acm.org/10.1145/ 956863. 956868.
5. Chakrabarti, K., S. Chaudhuri, and S.-w. Hwang. "Automatic categorization of query results." In *Proceedings of the 2004 ACM SIGMOD International Confer- ence on Management of Data*, pages 755–766, Paris, France, 2004. ACM Press. http://doi.acm.org/10.1145/1007568. 1007653.
6. Chien, S. and N. Immorlica. "Semantic similarity between search engine queries using temporal correlation." In *Proceedings of the 14th International Conference on the World Wide Web (WWW)*, pages 2–11, Chiba, Japan, 2005. ACM Press. http://doi.acm.org/10.1145/1060745. 1060752.
7. Saraiva, P. C., E. S. de Moura, N. Ziviani, W. Meira, R. Fonseca, and B. Riberio- Neto. "Rank-preserving two-level caching for scalable search engines." In *Pro- ceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 51–58, New Orleans, LA, 2001. ACM Press. http://doi.acm.org/10.1145/383952. 383959.
8. Sebastiani, F. "Machine learning in automated text categorization." *ACM Computing Surveys*, 34(1):1–47, 2002 http://doi.acm.org. ezproxy.gl. iit.edu/ 10.1145/505282.505283.

This page is intentionally left blank