ISSN 2277-3061

Improve Enterprise Search with Pattern Matching Approach and Web Usage Mining techniques for E-Commerce Websites

ArchitGoel, NehaVerma
Student, Dept. of IT, VIPS, affiliated with GGSIP University, New Delhi, India
archit.goel4@yahoo.com
Asst. Professor, Dept. of IT, VIPS, affiliated with GGSIP University, New Delhi, India
nehaverma583@yahoo.co.in

ABSTRACT

With the extensive expansion in the number of E-commerce websites, applying Web Usage Mining techniques to improve business is imperative. Also, employee as well as visitor satisfaction is important for an enterprise. This satisfaction is usually depended upon both, the effectiveness and efficiency of the search technology and how that information is published.

Keywords

Enterprise Search; Web Usage Mining and E-Commerce.

Academic Discipline

Computer Science

Sub-Disciplines

Data-mining

Council for Innovative Research
Peer Review Research Publishing System

Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 12, No.2
editor@cirworld.com
www.cirworld.com, member.cirworld.com

ENTERPRISE SEARCH

Enterprise search may be interpreted as search of textual materials owned by an organization, including search of their external Website, company's intranet, and any other textual document that they hold such as email, database records and shared documents by the use of information retrieval tools [20]. This information may be structured or unstructured. Documents are produced by a variety of sources, perhaps in many different languages, and generally without formatting standards. Because of these complexity of the typical enterprise information spaces, only a few of the search engines available on the market are able to work with the wide range of databases, email formats, document formats, typical of medium scale enterprises.

Most enterprise search also includes other searches done by the employees of that enterprise. Employees usually have access to a "desktop search" facility. Dumais [18] report an extension of this type to include search of all the documents previously viewed by the user. In general, we can include all the searches done by employees under this label i.e. "workplace search". This label covers not only search of enterprise information, information held on the desktop, and information previously viewed, but also search of information held external to the organization, such as the WWW. [23] Illustrates the diversity of sources accessed by employees.

Major Problems of enterprise search

- No specified search set i.e.: data can be from enterprise's intranet, database, user's hard disk or WWW.
- Different formats and no effective and relevant ranking of the heterogeneous collection of data.
- Searching the conversations that took place within emails.
- Different privileges for employees.Refer [20] for a detailed explanation.

Architecture of Enterprise Search

Gathering-> Extracting -> Indexing

Gathering: Company's intranet, databases, emails or IMs used by the organization etc. Publication of multiple copies of the same content at differentlocations, difficulty of identifying content that has changed recently and duplicate detection are some of the problems faced.

Extraction: The use of XML formats by different office suites is potentially a major step for ease and accuracy of filtering, as it helps in provided a homogenous set of data.

Indexing: Indexing for enterprise search is usually no different than indexing done for web data. Use of hyperlinking with anchor text is sometimes done.

Softwares

Companies like Fast Search and Transfer, Autonomy are know for their softwares for enterprise search. Also, IBM, Oracle and Google have also developed products similar and intended for this market.

NEED

The basic and most important need is that- the Best Possible Answer at Rank One i.e. if we query a current Web search engine for the name of a institute, such as "VIPS" or a mathematical term, then the best answer page is the Institute's homepage or the definition of the concept should appear in the at least the top five search sets.

ENTERPRISE SEARCH FOR E-COMMERCE WEBSITES

Some businesses such as catering suppliers, retailers and travel agents etc. rely on e- commerce websites for part or most of their revenue. A typical e-commerce site provides product search, coupled with query suggestions i.e. in the search field and automatic generation of recommendations. E-commerce sites are sometimes custom-built database applications (like Google and Microsoft) but they may also be built on enterprise search tools with the relevant capabilities. Autonomy and FAST are well known in this.

WUM with E- Commerce

Web usage mining is focused on learning about Web users and their interactions with Web sites by studying the web server logs (i.e. Access logs, Referrer logs, agent logs), client-side cookies, user profiles [22] and/or user ratings as well as other sources. The motive of this type of mining is to find users' access models automatically and efficiently such as frequent access paths, frequent access page groups etc. Web usage mining provides the support for the web site design and helps in improving customer relationships and other business decision [18]. For history of this refer [26]. The use of artificial intelligence techniques has provided useful methods for this.

Data Collection

Data for Web Usage mining is mostly collected from: Web server logs, client data and middle data (agent server data and packet detecting). Previous searches done by the employee, user profiles etc. also need to be collected along with information about the employees who previously did the same query or worked on the project that is being searched.

Data Cleaning

Eliminating irrelevant items and filtering of raw data is the primary purpose of this. Irrelevant records in web access log should be eliminated during data cleaning. Since the main motive for WUM is to gather user's travelpatterns, therefore data cleaning can be performed such as: The records of graphics, videos and the format information have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record. For websites that are built dynamically, the URLs should be considered as they provide information about the type of file, document it corresponds to.

User Identification

User's identification is to identify who and which pages are accessed in the website. A session is the series of web pages accessed by the user in a single access. Every website is accessed by the visitor through an IP address, which has a corresponding domain name, and these are linked through the Domain Name System (DNS). Using DNS a corresponding IP address can be found. Some information can be discovered by this method, but this solely cannot be used for mining. A visitor's IP address can be converted into a domain name by using the DNS system in reverse, which is called reverse DNS lookup.

The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log, or the same user may have different IP addresses. To identify a user, different approaches are taken.

To distinguish user sessions, following measures could be undertaken:

a. Use of cookies. Cookies usually are randomly assigned IDs that a Web server gives to a Web browser the first time that the browser connects to a Web site. Cookies are independent of IP addresses, and work well on sites with a substantial number of visitors from ISPs.

b. On subsequent visits, the Web browser sends this same ID back to the Web server, effectively telling the Web site that a specific user has returned.

c. The different IP address distinguishes different users; many different cases for this are provided here. [24]

Pattern Analysis

Web site administrators and company heads are extremely interested in questions like "How are people using the site?" "Which pages are being accessed most frequently?" "Which pages are accessed by returning customers" etc. These questions require the analysis of the contents of the pages and the their hyperlinking. This analysis might include:

1. The frequency of visits per document i.e. product on web page

2. Most recent visit per page

3. Who is visiting which pages?

4. Frequency of use of each hyperlink

5. Which documents or webpages are accessed in a session?

6. Which pages are accessed by the user after purchasing a product?

The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, and Usability Analysis. [25]

Path Completion

There are some reasons that result in path's incompletion that can result in some accesses not recorded in the access log file, and the number of URLs recorded in the logs may be less than the real ones because of local and agent cache, "post" technique in the 'form' part of the website and browser's "back" button. Using proxy servers also produces the difficulties for path completion because users can access the pages in the proxy servers caching without leaving any record in server's access log.

Techniques

The various techniques used include: Association Rules, Sequential Patterns and Clustering Techniques. [23]

For E-Commerce Website Administrators

- Track the product i.e. the number of times the product is displayed jpeg, links or text
- Times user clicks on a product in a specific session
- Watch the shopping carts: add/remove or changes in features of that product

- Separate the web pages (can be done when designing the website) into specific categories and then search on the basis of the URL- forms, info, products of same genre, by using classification.
- Post advertisements of items on the website of the same category that user purchased recently. If user dismisses the advertisement a specific number of times, ask him/her before removing it permanently.
- Create widgets of accessories or products, dynamically on a product's page, that other customers also bought after buying that specific product.

PROPOSED SYSTEM FOR ENTERPRISE SEARCH

Using the system, improved version of page rank algorithm i.e. using the web dictionary method and some additions to it like the search domain as WWW, enterprise's intranet, emails, databases, previous searched items etc., and applying it for the enterprise search [19]:

1. Module 1- Create a Search Domain

Using the architecture of enterprise search i.e.: gathering, extracting and indexing, we get a search set. This also includes the previous search results from the previous queries made by the employee along with the employees' location, context and profile.

2. Module 2 – Implement Web Dictionary

This module splits user search string into various words. It then counts the length of each of the word to find the minimum (MIN) and maximum (MAX) number of characters among various keywords of search string. It will implement web dictionary by allowing only those words having length in between to that of MIN and MAX.

3. Module 3 – Relevancy of Web Page Using Content

This module will determine relevancy of document from its content. It will count frequency of various keywords of the search string to determine the value of FOUND, which represents the total frequency of all the keywords within the web dictionary and NFOUND, which determines the number of keywords not found in the web dictionary. The difference between these values will determine the relevancy of web page.

4. Module 4 – Relevancy of Search Results Using Time Spent Statistic

This module will determine relevancy of web page using previously spent time statistic by retrieving its value from database. It will calculate new value of time statistic by calculating average of previous value and new value.

5. Module 5 – Priority Calculation Module

This module will determine priority of web page by first calling MODULE 3at first and after then MODULE 4 so overall priority is determined by judging candidate web page twice, using two different modules.

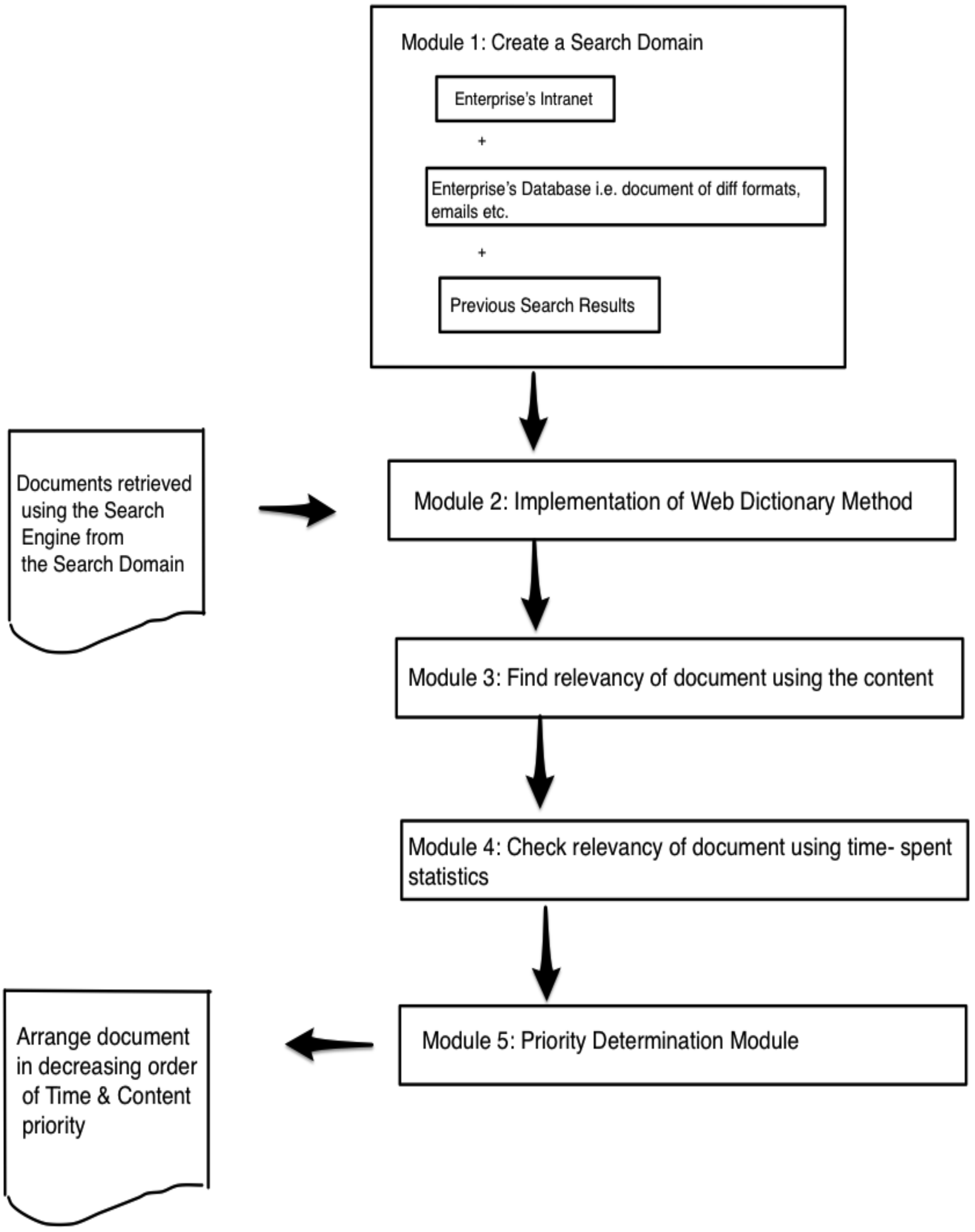
Module 1: Create a Search Domain
Enterprise's Intranet
+
Enterprise's Database i.e. document of diff formats, emails etc.
+
Previous Search Results
Documents retrieved using the Search Engine from the Search Domain
Module 2: Implementation of Web Dictionary Method
Module 3: Find relevancy of document using the content
Module 4: Check relevancy of document using time- spent statistics
Module 5: Priority Determination Module
Arrange document in decreasing order of Time & Content priority

Fig 1: Design of the proposed system

CONCLUSION

With the extensive growth of E-Commerce websites, application of web mining techniques and tools is imperative. It not only helps in improving the business but also helps in increasing customer satisfaction. Also, the major motive of enterprise is employee satisfaction, and this satisfaction is usually based on the quality and efficiency to search documents and information. In this paper, we provide a solution to the enterprise search problem by using a web dictionary method [19] and using a search domain in which the system is applied. Also, for E-commerce website developers and administrators some pointers are provided in order to help their business to excel.

REFERENCES

[1] A. Mendez-Torreblanca, M.Monte.” A Trend Discovery for Dynamic Web content Mining”, IEEE, Intelligence system, Vol. 14, pages 20-22, 2002.

[2] R.Khanchana and Dr. M. Punithavalli “ A Web Usage Mining Approach Based On New Technique in Web Path Recommendation Systems” , International Journal of Engineering Research and Technology, Vol. 2, January 2013

[3] Bing Liu, Kevin Chen- Chuan Chang, Editorial issue :” Special Issue on Web Content Mining”, SIGKDD Explorations, Volume 6, Issue 2.

[4] Bharanipriya& V. Kamakshi Prasad,” Web Content Mining Tools: a comparative study”, International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[5] G.K Gupta,” Introduction to Data mining with Case Studies”, PHI.

[6] G.Poonkuzhali, G.V.Uma, K.Sarukesi, “Detection and Removal of Redundant Web Content Through Rectangular and Signed Approach”, International Journal of Engineering Science and Technology,Vol. 2(9), 4026-4032, 2010.

[7] Hillolkargupta ,Aunpam Joshi, KrishnamoorthySivakumar and Yelena Yesha,” Data Mining – Next Generation Challenges and Future Directions ”,PHI, 2007

[8] J.W.Han, M.Kamber, “Data Mining: Concepts and Techniques “, New York Kaufmann publishers 2001.International Journal of Computer Applications (0975 – 8887) Volume 65– No.24, March 2013

[9] JaideepSrivastava, PrasannaDesikan, Vipin Kumar, “Web Mining-Accomplishments & Future directions”, Department of Computer science.

[10] ShohrehAjoudanian, and Mohammad DavarpanahJazi,” Deep Web Content Mining”, World Academy of Science, Engineering and Technology 49 2009.

[11] R.Cooley, B.Mobasher, J. Shrivastava,”Web mining: information and pattern discovery on the World Wide Web”, Department of computer science, University of Minnesota, USA.

[12] RajshreeShettar, Dr. Shobha G.,” Survey on Mining in Semi Structured Data”, IJCSNS International Journal of computer science and Network security”, Vol.7 No.8. Aug 2007.

[13] Soumenchakrabarti,” Mining the Web: discovering knowledge from hypertext data”, Elsevier.

[14] Chen, M.S., Park, J.S. and Yu, P.S., “Efficient Data Mining for Path Traversal Patterns”, IEEE Transactions on Knowledge and Data Engineering, March/April, 1998,pp 209-221.

[15] H.Jiang et al,” TIMERANK” A Method of Improving Ranking Scores by Visited Time”, In proceedings of Seventh International Conference on Machine Learning and Cybernetics, Kunming 12-15, July 2008.

[16] Milan Vojnovic et al, ”Ranking and Suggesting Popular Items”, In IEEE Transactions of KDE”, Vol 21,No. 8, Aug 2009.

[17] Chu-Hui Lee, Yu-Hsiang Fu “Web Usage Mining Based on Clustering of Browsing Features”, IEEE Eighth International Conference on Intelligent Systems Design and Applications, 2008, p. 281-28

[18] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. *“Stuff I’ve seen: a system for personal information retrieval and re-use”.* In Proceedings of ACM SIGIR ’03, pages 72–79, New York, NY, USA, 2003. ACM.

[19] NehaVerma, DheerajMalhotra, " *An Ingenious Pattern Matching Approach to Ameliorate Web Page Rank ",* International Journal of Computer Applications (0975-8887), Vol - 65, No. 24, March 2013, Published by Foundation of Computer Science, New York, U.S.A.

[20] D. Hawking.“*Challenges in Enterprise Search”*. In Proceedings of the Australasian Databases Conference ADC2004, pages 15–26, Dunedin, New Zealand, January 2004. Australian Computer Science Paper: http://es.csiro.au/pubs/hawking_adc04keynote.pdf.

[21] P. Thomas. “*Server characterization and selection for personal metasearch*”.PhD thesis, Australian National University, 2008.http://es.csiro.au/pubs/thomas_thesis.pdf

[22] Ken McGarry, Andrew Martin and Dale Addison. “*Data Mining and User Profiling for an E-Commerce System”*. School of Computing and Technology, University of Sunderland, St Peters Campus, St Peters Way, Sunderland, United Kingdom

[23] Federico Michele Facca and Pier Luca Lanzi. “*Recent Developments in Web Usage Mining Research”.*

[24] RajniPamnani, PramilaChawan. “*Web Usage Mining: A Research Area in Web Mining”* .Department of computer technology, VJTI University, Mumbai

[25] Jinguang Liu &RoopaDatla.“Web Usage Mining- Pattern Discovery and its applications”.

[26] David L. Banks and Yasmin H. Said. “Data Mining in Electronic Commerce.”Institute of Mathematical Statistics, 2006

[27] Martin Butler.“*The Business Value of Enterprise Search”. 2009*

AUTHOR’ BIOGRAPHY

She started her career with AVL India Software Pvt Ltd. as a software developer in embedded systems. She obtained her B.Sc in Computer Science from Delhi University. She obtained her MCA from GGSIPU and M.Tech from USIT, GGSIPU in Information Technology. She is J2EE certified professional from NIIT. She published and presented research papers in various National and International Conferences. Her area of specializations is Data Mining, OOP Programming with JAVA and C++, Data Structure.

He is a final year student of BCA in Computer Science from GGSIP University. His area of interests includes data mining, system programming, artificial intelligence and human computer interactions systems.