IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 12, December 2015.
www.ijiset.com
ISSN 2348 – 7968

Web Search Result Optimization using Association Rule Mining Algorithm

G.Pratibha[1], Dr. Nagaratna Hegde[2]

[1] Assistant Professor, Department of Computer Science and Engineering, Matrusri Engineering College Hyderabad, Telangana State, India

[2] Professor, Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, Telangana State, India

Abstract

The web is a hub with enormous information where large number of data resources such as documents, images, videos or other multimedia can be retrieved. In this area of context, several information retrieval technologies have been developed to assist users to fulfill their searching needs on web. The most commonly used search engines by users are Google, Yahoo, Amazon, Bing, Ask, Search and so on. The search engines allow users to find relevant resources needed by setting up their query patterns and reviewing a list of URLs. In this paper a Search Result Optimization Method (SROM) for Search Engine Optimization (SEO) by updating page rank, query recommendation and query reformulation are studied and implemented. It allows the user explore queries registered in the search engine's query logs in order to know how users search and also design algorithms to improve the correctness of the results suggested to users. The studied method starts by exploring the query logs for query clusters and identifies query sessions and then examines query logs to discover relationship among keywords, pages, and queries within clusters using association rule mining algorithms such as an Apriori algorithm and Automated Apriori algorithm.

Keywords: *Apriori Algorithm, Automated Apriori Algorithm, Clustering, Page, Keyword, Rank Improvement Algorithm, Search Result Optimization Method (SROM) and Uniform Resource Locator (URL).*

1. Introduction

World Wide Web (WWW) is a great deal with a huge information repository of resources in documents, images, video etc format growing day by day very quickly. These distinctive characteristics of web raised several new challenges for Internet and Web researchers. In this regard, identifying and clustering most relevant data needed by the user automatically. This has become a most critical issue to process such available huge quantities of data. They are:

·How do we know which information is most relevant to the user?

·How can we understand the user requirement based on entered query?

Due to these issues, it's become necessary to investigate retrieval techniques, ideas which are associated with web and data mining known as query log, association rules, clustering and web applications and tools. In this paper, we tried to summarize these ideas used by most of web applications and tools.

Web Mining (WM) is based on the data mining application techniques such as clustering, classification, and associating to search query patterns from the web. The web mining can be classified into

1. Web Structure Mining,
2. Web Content Mining, and
3. Web Usage Mining

A web search engine is an application designed to extract useful information on web. The search engine allows the user to evoke specific content meeting criteria and retrieving a listing of resources in terms of URLs that match those criteria. Search query is that the regular expression of the user who gives it as an input to the search engine in his natural language. Query log is a text file consisting of a consequence requests. Clustering is a collection of objects which are grouped based on the similarities.

Association rules mining are used to find all rules that meet user defined restriction on minimum support (MS) and confidence with respect to a given dataset items. The most commonly used association rule finding algorithm that search the frequent items set strategy is Apriori algorithm. This algorithm is best suitable to work with a large scale of data sets. Page Rank algorithm is a link analysis algorithm that assigns a numerical weighting to every part of document collection by measuring its importance inside the data set. In this paper, a novel method to optimize search result has been studied and

ISSN 2348 – 7968

implemented which uses association rule mining technique, and clustering.

2. Literature Review

In the section, we reviewed some similar articles and studied the other authors views and approaches on working with various algorithms for relevant information retrieval based on user query patterns The technologies for the massive information systems are available today includes the internet, web search, portals, agents, collaborative filtering [1], which uses data mining Association Rules to create massive item-set as a keywords for sites. The recent research have proposed ranking methodologies [2] to use linkage web structure, and query log instead of content, to improve the search result.. The previous research work concentrated on search engines [3] like Google, Yahoo. But there are still many issues, where user is provided with irrelevant and non-relevant pages in the top most based on their rankings. Nowadays, the results of the search engines are based on the user query and least bothers the way how the query is expressed. The search engines filter the long result list to find his desired content which leads to information overkill (overabundance) problem [4]. The search query logs maintains records about users searching behavior [5, 6]; such analysis has found helpful in many different circumstances like query recommendation [7, 8] and document ranking [9].Most of the work focus on query recommendation based on query similarity count [8, 10] that can be used for clustering of queries [7, 11]. Baeza-Yates et al. [7] and Wen et al. [11] have been presented a solution to build factitious queries and worked on a clustering methodology for query recommendation that occurs in four notions of query distance: query keywords; Keywords string matching; common URLs clicked; and measure the distance of the clicked documents for searching. Jones et al. have been focused on query substitution [12] notation. In this paper, we identified the purpose of web log mining is to improve search engines performance.

3. Working of Apriori and Automated Apriori Algorithm

The key principle of Apriori is the subset of any frequent item sets must also be frequent. It means if an item sets does not meet minimum support (MS) expectations, and then item is not frequent. If an item is added to any item sets, then the resulting item sets cannot occur more frequently. Therefore, union of items and their sets is not frequent either. The Apriori algorithm finds frequent item sets frequently with range from 1 to n-item sets and further used to generate association rules.

Apriori algorithm works as follows:

Step 1: In first pass, algorithm determines frequent item sets by simply counting item occurrences.
Step 2: The singleton items are combined to form two member candidate item-sets.
Step 3: Supporting values of these candidates are then determined by scanning the data sets again. This step also considers the candidates with threshold higher than support values.
Step 4: In next pass, algorithm creates item sets of three members. Repeat this process until all frequent item sets are accounted.
Step 5: These item sets are then used to generate association rules which have threshold values less than or equal to confidence values.
Step 6: It first creates the rules for frequent item sets and then for subsets is created recursively.

Associative classification method uses association rules for classification of data. Association rule mining discovers the hidden and interesting relationship between the database items based on the support and confidence thresholds. A new approach has been proposed in [18, 19] to identify suitable supported thresholds for frequent item set generation without user consultation. The Associations will be found more strong rules by Automated Apriori algorithm when compared to Apriori algorithm. The Automated Apriori algorithm steps are as follows:

Step 1: Calculate each item support.
Step 2: Arrange items in ascending order based on their support.
Step 3: Calculate ms for each item.
Step 4: Generate all frequent item sets
Step 5: Calculate cumulative support (cs) of each item sets
Step 6: Calculate mini support (ms) of each item sets
Step 7: Selection most frequent item sets
Step 8: Generate strong association rules from frequent item sets

4. Optimization System Implementation

The most challenging goal of W3C is to design search engines that allow users to find resources which are semantically connected to their queries. The huge volume of data in the web and the complexity among most commonly used terms still poses a big problem to achieve its goal Figure 1 depicts the proposed system. In this paper, we studied about search result optimization method to explore the users query registered in search engine logs which in turn used to learn how users perform search

ISSN 2348 – 7968

activities. After that, we need to design algorithms which can contribute to the improvement in getting high precision values for the user queries. The studied system follows the following steps to accomplish the above tasks:

- Module M0 shows the users search behavior query log as a text file.
- Module M1 creates a clusters based query log's text file obtained from Module M0.
- Module M2 is used to apply the association rule algorithms on the clusters obtained from M1. This module also creates an association of page, query, and the keywords.
- Module M3 is used to improve the page rank of web pages obtained from Module M2.

Associate the outputs of previous modules to optimize relevant page searching.

4.1 Implementation of Proposed Optimization System on Java platform

The Apache Tomcat 5.5, MySQL Server 3.5, and JDK1.7.0 are used to implement the purposed system's modules for new search engine development. The steps followed are:

Step 1: Firstly, create a folder with a name Search-Engine in C:\Program Files\Apache Software Foundation\Tomcat 5.5\webapps\Search-Engine.

Step 2: Run the search engine page "http:// localhost: 8080/Search-Engine/Link_Registration.jsp" on any browser for registration of new websites. This registration page includes link or URL name, link descriptions, and link keywords are shown in Figure 2(a),

Step 3: Run search query page "http://localhost : 8080/ Search-Engine/." and enter search queries, which you want to perform search.

Step 4: Check the search results pages for a user query for relevancy.

Step5: Check the page rank of each resultant URLs by using http://localhost:8080/Search-Engine/ Check_Page_Rank.jsp.

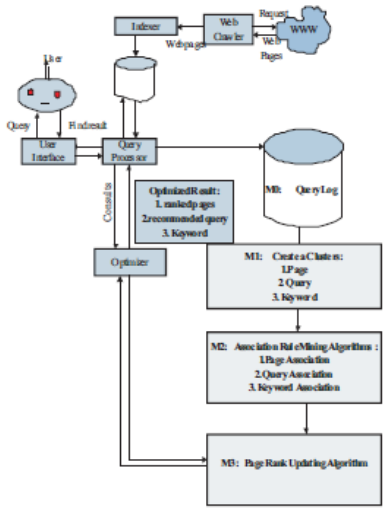

Fig 1: Optimization System Architecture

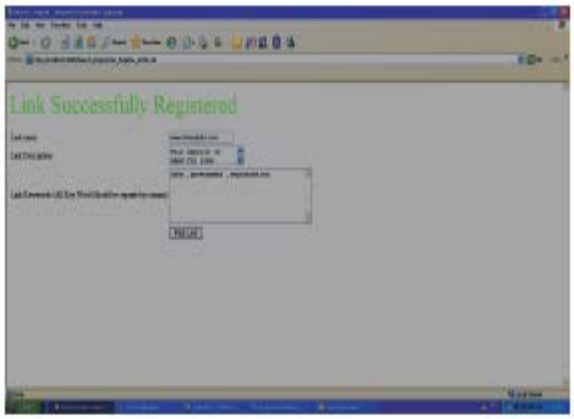

Fig 2: Link Registration of websites

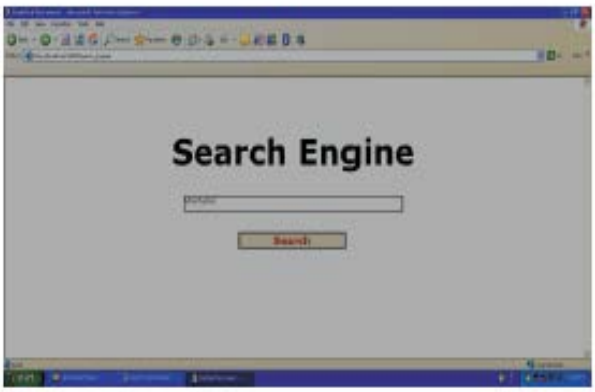

Fig 3: Search Query for jobs for B.Tech graduates

ISSN 2348 – 7968

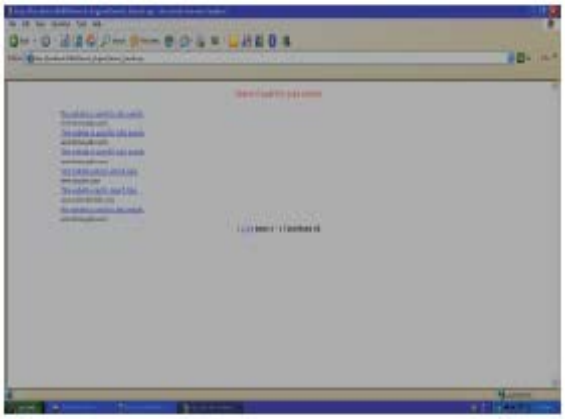

Fig 4: Search Result URL links

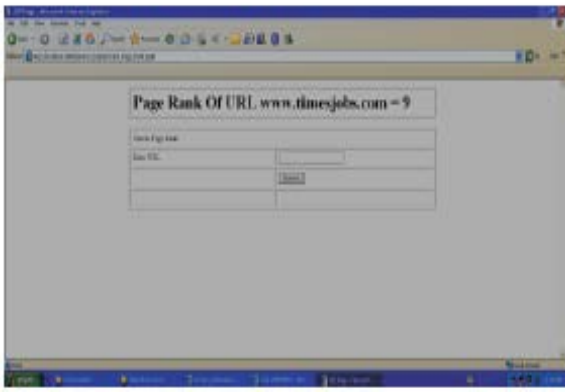

Fig 5: Page rank of www.timesjobs.com

5. Conclusions

The proposed system studied in this paper uses clustering and association rule discovery data mining concept to achieve search engine's result optimization by the means of effective page reranking and query recommendations. By the new approach of search engine optimization, many advantages can be achieved like returning relevant pages with high rank, recommendations for semantically related queries.

Acknowledgments

I want to take this opportunity to thank all the people especially my guide Dr. Nagaratna Hegde who helped me during my conference sojourn. I understand that it is rather late to acknowledge their contributions, but as the saying goes, better late than never!

References

[1]. J. Han, M. Kamber (2000), Data Mining: *Concepts and Techniques Morgan Kaufmann.*

[2]. O. Etzioni (1996), "The World Wide Web: Quagmire or Gold Mine", *in Communication of the ACM*, 39 (11):65- *30 i-manager's Journal on Information Technology, Vol. 3* l *No. 1* l *December 2013 - February 2014* 68.

[3]. S.Brin et.al (1998), The anatomy of a large-scale hyper textual web search engine. *Computer Networks and ISDN systems*, pp: 107-117, 1998.

[4]. J. Srivastava, P. Desikan and V. Kumar (2002), "Web Mining: Accomplishments and Future Directions", *National Science Foundation Workshop on Next Generation Data Mining (NGDM'02).*

[5] G. Salton and M. McGill (1983), *Introduction to Modern information Retrieval.* McGraw Hill.

[6] S. Deerwester, S. Dumains, G. Furnas, T. Landauer and R. Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of American Society for Information Science.* 41(6): 391-407.

[7]. H. Ahonen, O. Heionen, M. Klemettinen and A. Verkamo (1998). Applying data mining techniques for descriptive phrase extraction in digital document collections. *In advances in Digital Libraries (ADL 98).* Santa Barbara California, USA.

[8]. W. W. Cohen (1995). Learning to classify English text with ilp methods. In *Advances in Inductie Logic Programming.* (Ed. L. De Raed)m IOS Press.

[9]. P. Desikan, J. Srivastava, V. Kumar, P.N. Tan (2002), "Hyperlink Analysis Techniques and Applications", *Army High Performane Computing Center Technical Report.*

[10]. K. Wang and H. Lui (1998), "Discovering Typical Structures of Documents: A Road Map Approach", in *Processding of the ACM SIGIR symposium on Information Retrieval.*

[11]. C. H. Moh, E.P. Lim, W. K. Ng (2000), "DTD Miner: A Tool for Mining DTD form XML Documents", *WECWIS*:144-151.

[12]. S. Chakrabrti (2000), Data Mining for hypertext: A tutorial Survey. *ACM SIGKDD Explorations*.1 (2):1-11.

[13]. J. M. Kleinberg (1998), Authoritative Sources in a hyperlinked environment. In *Proc. Of ACM SIAM symposium on Discrete Algorithms,* Pages 668-667.

[14]. S. Brin and L. Page (1998). The Anatomy of a Largescale hypertextual Web search engine. In *seventh international World Wide Web Conference*, Brishbane, Australia, 1998.

[15]. S. K. Madia, S. S. Bhowmick (1999), W. K. Ng, E. P. Lim: Research Issues in Web Data Mining. *DAWak*: 303-312

[16]. J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan(2000). " Web Usage Mining:Discover y and Applications of usage patterns form Web Data", *SIGKDD Explorations*, Vol 1, Issue 2.

[17]. R. Cooley. (2000), Web Usage Mining: *Discovery and Application of Interesting Patterns* from Web data. Phd thesis, Dept. of Computer Science, University of Minnesota, May 2000.

[18]. Kalaiselvi, C. Kanimozhiselvi, C.S. (2009), "Frequent Itemsets Generation using Collective Support Thresholds

ISSN 2348 – 7968

for Associative Classification, *Conference Proceedings RTCSP'09*, pp. 232-235, 2009.

[19]. Kanimozhiselvi, C.S. and Tamilarasi A. (2011), "Mining of High Confidence Rare Association Rules with Automated Support Thresholds", *European Journal of Scientific Research*, Vol. 52, No. 2, 2011.