A CUSTOMIZED CONTENT OPTIMIZATION USING USER SEARCH BEHAVIOR ANALYSIS

Priyadharsini.M [1] *Kathiresan.V*[2]

ABSTRACT

Search and content optimizations are very popular nowadays due to intensive growth of networks and data. There is a need to satisfy the user in the searching era. There are several techniques have been proposed to identify users search interest based on their searching history. The current proposal deals with the analysis of user search goals with the effective feedback. Exiting analytical process for individual interest mining from personalized weblog is a tedious process, because the existing techniques considered only the "click" based priority. The proposed system considers total number of clicks; unclicked URLs and time spend by the user in a particular page and links. Based on these parameters the personalization has been proposed. The Implementation of existing algorithm for web usage mining and analyzing the user feedback has a main drawback which is feedback collection issue. In this paper, three factors are analyzed which are personal interest, clicked and unclicked link similarity, and personal search sequence. The above factors are blended into a cohesive personalized search model and content optimization based on data mining techniques. This paper proposes an implicit and explicit model for analyzing the user search goals effectively.

IndexTerms—Feed Back sessions, Session clustering, Pseudo documents, LSC

I. INTRODUCTION

In real world all the internet users fully depending on the search engine, search engine is not a small, search engines having the lot of information's, all data's stored with semantic model, most of the existing works concentrate to make the relevancy between the data's, through this works generate the well-defined semantic data world in large scale data sources. Data search and information gathering on the Internet has placed high demand on search engines. User interest based search engine creations is very difficult, existing works fully concentrate the click based sessions values only, behavior based user data search multi point analysis give the minimum accuracy, The requirement of the information may differ from each and every user and also achieve the user goals are still becomes difficult. To achieve the user specific information s requirements many uncertain queries may full fill a broad topic and dissimilar users may want to get information on different angles when they submit the same query. End user information need is to desire and obtain the information to satisfy the needs of each user. We group the user information needs with different search goal .Because the interference and examination of user search goals with query might have a numeral of advantages by improving the search engine significance and user knowledge. So it is necessary to collect the different user

[1]Research Scholar, Department of MCA RVS Arts and Science College, Sulur, priyadharsini195@gmail.com

[2]Assistant Professor, Department of MCA RVS Arts and Science College, Sulur, kathirsujith@gmail.com

goal and retrieve the efficient information on different aspects of a query. Capture different user search goals in information retrieval outcome becomes changes than the normal query based information retrieval.

II. Literature Review

Application of data mining techniques to the World Wide Web, referred to as Web mining [1], has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining in this paper is the process of information discovery from sources across the World Wide Web. The second, called Web usage mining, is the process of mining for user browsing and access patterns. This defines Web mining and presents an overview of the various research issues, techniques, and development efforts. This briefly describe WEBMINER, a system for Web usage mining, and include the paper by listing research issues. Relational clustering based on a new robust estimator with application to Web mining

Mining typical user profiles [2] and URL associations from the huge amount of access logs is an important component of Web personalization. In this paper this defines the notion of a user session as being a temporally compact sequence of Web accesses by a user. This also defines a dissimilarity measure between two Web sessions that captures the organization of a Web site. To cluster the user sessions based on the pair wise dissimilarities, this introduce the relational fuzzy c-maximal density estimator (RFC-MDE) algorithm. RFC-MDE is robust and can deal with outliers that are typical in this application. This show real example of the use of RFC-MDE for extraction of user profiles from log data, and compare its performance to the standard non-Euclidean fuzzy c-means.

Large volumes of data are gathered automatically by Web servers and collected in access log files [3]. Analysis of server access data can provide significant and useful information. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data and is targeted towards applications. It mines the secondary data derived from the interactions of the users during certain period of Web sessions. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities. In this paper, this applied Kohonen's SOM (Self Organizing Map) to pre-processed Web logs of our university Web server logs (http://www.um.ac.ir/) and extract frequent patterns. Result of this paper would be useful for our university Web site owner.

This paper [4] describes an approach for automatically classifying visitors of a web site according to their access patterns. User access logs are examined to discover clusters of users that exhibit similar information needs; e.g., users that access similar pages. This may result in a better understanding of how users visit the site, and lead to an improved organization of the hypertext documents for navigational convenience. More interestingly, based on what categories an individual user falls into, this can

dynamically suggest links for him to navigate. In this paper, this describe the overall design of a system that implements these ideas, and elaborate on the preprocessing, clustering, and dynamic link suggestion tasks. This present some experimental results generated by analyzing the access log of a web site.

Designing a web site is a complex problem [5]. Logs of user accesses to a site provide an opportunity to observe users interacting with that site and make improvements to the site's structure and presentation. This proposes adaptive sites: web sites that improve themselves by learning from user access patterns. Adaptive webs can make popular pages more accessible, highlight interesting links, connect related pages, and cluster similar documents together. An adaptive web can perform these self-improvements autonomously or advise a site's webmaster, summarizing access information and making suggestion. In this paper this define adaptive web sites, explain and formalize several kinds of improvements that an adaptive site can make, and give examples of applying these improvements to existing sites.

One of the main activities [6] of web users, known as "surfing", is to follow links. Lengthy navigation often leads to disorientation when users lose track of the context in which they are navigating and are unsure how to proceed in terms of the goal of their original query. Studying navigation patterns of web users is thus important, since it can lead us to a better understanding of the problems users face when they are surfing. This derive Zipf's rank frequency law (i.e. an inverse power law) from an absorbing Markov chain model of surfers' behavior assuming that less probable navigation trails are, on average, longer than more probable ones. In our model the probability of a trail is interpreted as the relevance (or "value") of the trail. This apply our model to two scenarios: in the first the probability of a user terminating the navigation session is independent of the number of links he has followed so far, and in the second the probability of a user terminating the navigation session increases by a constant each time the user follows a link. This analyze these scenarios using two sets of experimental data sets showing that, although the first scenario is only a rough approximation of surfers' behavior, the data is consistent with the second scenario and can thus provide an explanation of surfers' behavior.

The query associated with a repeat search may differ from the initial query but can nonetheless lead to clicks on the same results. This paper explores repeat search behavior through the analysis of a one-year b query log of 114 anonymous users and a separate controlled survey of an additional 119 volunteers. The study demonstrates that as many as 40% of all queries are re-finding queries. Re-finding appears to be an important behavior for search engines to explicitly support, and it explores how this can be done. It demonstrates that changes to search engine results can hinder re-finding, and provide a way to automatically detect repeat searches and predict repeat clicks. The advantage of process is Same Query retrieved by using the information re retrieval and disadvantage is Repeated Queries are not identified by the methodology

In another work [7] analysis of contextual information in search engine query logs enhances the understanding of b users' search patterns. Obtaining contextual information on b search engine logs is a difficult task, since users submit few numbers of queries, and search multiple topics. Identification of topic changes within a search session is

an important branch of search engine user behavior analysis. The purpose of this study is to investigate the properties of a specific topic identification methodology in detail, and to test its validity. The topic identification algorithm's performance becomes doubtful in various cases. These cases are explored and the reasons underlying the inconsistent performance of automatic topic identification are investigated with statistical analysis and experimental design techniques.The main idea of using query patterns and time intervals in identifying topic shifts is valuable, but there are indications of some problems in the application of this idea Generalization, specialization and reformulation classes are used to replace the modified class.

Query logs record [8] the queries and the actions of the users of search engines, and as such they contain valuable information about the interests, the preferences, and the behavior of the users, as well as their implicit feedback to search engine results. Mining the wealth of information available in the query logs has many important applications including query-log analysis, user profiling and personalization, advertising, query recommendation, and more. In this paper the *query-flow graph*, a graph representation of the interesting knowledge about latent querying behavior. The query-flow graph is an outcome of query-log mining and, at the same time, a useful tool for it. A methodology that builds such a graph by mining time and textual information as well as aggregating queries from different users. Using this approach build a real-world query-flow graph from a large-scale query log and demonstrate its utility in concrete applications, namely, *findinglogical sessions*, and *query recommendation.* Hover that the usefulness of the query-flow graph goes beyond these two applications. The main advantage of this process is two key applications in usage mining that are supported by the query-flow graph. And drawback is the problem of query recommendation.

In the sponsored search model [9], search engines are paid by businesses that are interested in displaying ads for their site alongside the search results. Businesses bid for keywords, and their ad is displayed when the keyword is queried to the search engine. An important problem in this process is keyword generation: given a business that is interested in launching a campaign, suggest keywords that are related to that campaign. This problem by making use of the query logs of the search engine. Identify queries related to a campaign by exploiting the associations' between queries and URLs as they are captured by the user's clicks. These queries form good keyword suggestions since they capture the "wisdom of the crowd" as to what is related to a site. Formulate the problem as a semi-supervised learning problem, and propose algorithms within the Markov Random Field model. Perform experiments with real query logs, and demonstrate that the algorithms scale to large query logs and produce meaningful results.The advantage of this process is Identifying an appropriate set of keywords for a specific advertiser. And disadvantage of process is requires minimal effort from the part of the advertisers.

Contextual information provides an important basis for identifying and understanding users' information needs. The previous work in traditional information retrieval systems has shown how using contextual information could improve retrieval performance. With the vast quantity and variety of information available on the b,

and the short query lengths within b searches, it becomes even more crucial that appropriate contextual information is extracted to facilitate personalized services. Hover, finding users' contextual information. is not straightforward, especially in the b search environment where less is known about the individual users. It presents an approach that has significant potential for studying b users' search contexts. The approach [10] automatically groups a user's consecutive search activities on the same search topic into one session. It uses Dempster–Shafer theory to combine evidence extracted from two sources, each of which is based on the statistical data from b search logs. Drawback of this scheme is Problem of paucity of information about users' search contexts within the b and advantage is Applications of automatic session identification, such as adaptive information retrieval systems.

III. Study of Problem

This section represents the problem of organizing and maintaining the users click through logs over a web server. The user goal identification using the click through logs not only affects the storage and also affects the performance of the search engine if there is no proper system introduced.This deals the query clustering, user history organizing, result re arranging and page ranking oriented problems. The system has concentrated on the problem of web content extraction according to the query and desired results of the user. The system should automatically maintain and organizes the user click through logs and clicked citations together for creating pseudo documents. The followings are the main problem deals with the existing system.

- Deals the problem of organizing a user's histories queries into groups in a dynamic and automatic system.
- It deals the problem of automatically identifying query clusters which will helpful for a number of different search engine components and applications.
- Web extraction and citation finding according to the user query over a semantic search engine may produce less effective results and performance.
- Web structure identification is a main drawback in web mining, where each web page has its own template and may different in the outline.

Session identification, query clustering and reconstruction of clusters are very tedious. If the link is a non-uniform format the existing system suffered to store the histories effectively.

This thesis deals the above problem and finds how to make use of the relations between the concepts to retrieve a more precise and smaller result set.

IV. Proposed Methodology

In this proposed system an innovative approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents and implicit and explicit user feedbacks. Initially the system introduces feedback sessions which are categorized into two categories one is implicit and explicit feedbacks to be analyzed to infer user search goals rather than search results or clicked URLs.

Both the clicked URLs and the un-clicked ones before the last click are considered along with the time spent by the user in every click as user implicit feedbacks and taken into account to construct feedback sessions. Consequently the feedback sessions can suggest user information desires more efficiently. The next process of the proposal is mapping the feedback sessions to pseudo documents to estimate the goal of user.

The proposed system provides an effective way to store, organize and manipulate the user search histories with effective structure mining techniques.

The semantic based search engine produces a query matched results than the priority or important content based.

Now a day's data storage and references are very huge in size. Providing filtered data is more important. The proposed system provides an effective summarization and organization of user histories, which is implemented in the web search. This proposed web search engine provides the most semantic relativity between the given words and terms, and it will generate the semantic measures automatically and it also performs the user history organizations effectively. This kind of summarization, organization and extraction techniques improves the efficiency of the user search over the internet.

The proposed system is an automatic method to evaluate, estimate the semantic similarity between words or entities using web search engines, text snippets and a lexical pattern extraction schemes. That considers word subsequences in text snippets. The system proposed a new algorithm which is named as "LSC" (Link Semantic Clusters) algorithm which combines more features of web mining concepts.

Terms and web structures and snippets are the useful information sources provided by most web search engines while searching and then the system trains a two-class method to classify synonymous and non-synonymous word pairs.

Both novel pattern extraction algorithm and pattern clustering algorithm outperforms well in the case of page counts for given words with the text snippets.

Advantages:

- Link seeking and semantic cluster helps to reduce the risk in user search goal detection.
- Fast and improves the accuracy.
- Reduces the clustering difficulty.

The following algorithm represents the overall steps involved in the proposed system.

Algorithm: LS

Input: The user click through log

Output: Reconstructed re ranked results

Steps:

1. Read the user query Q.
2. Split the query Q into pattern P
3. Calculate the weighted token by identifying the frequency F.
4. Pass the F as pattern to the server
5. Apply pattern matching algorithm
6. Get the results
7. Update the result cluster RC from step s

a. Sequence identification from click through logs

b. Extract user goal from the above step 7.a and search semantic contents over all query groups LS(Rc)

c. Find best match and update the cluster (Rc)

8. Apply data re construction and summarization process

9. Update the results

10. Return the re ranked results

The above algorithms used a semantic self-organizing system, which is based on the sequence mechanism.

Link seekingscheme eliminates a number of the most common web link extraction which is typically found in web results. It places tags in classes depending on their semantic annotations.

In the upper part, all the feedback sessions of a query are first extracted from user click-through logs and mapped to pseudo-documents. Then user search goals are inferred by session clustering. These pseudo-documents arerepresented with some keywords in the cluster. Several different values are tried and the optimal value will be determined by the feedbackfrom the bottom part. In the bottom part, the original search results are restructured based on the user search goals inferred from the upper part. Then, this evaluates the performance of restructuring search results by the proposed LS algorithm. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the upper part.

V. Experiments

Inthis section, wewillshow experimentsofour proposedalgorithm.The system used a dynamic dataset, which can be any number of user weblog created by website.Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting). The first step in the Web usage mining process consists of collecting the pertinent Web data, which will be analyzed to offer useful information about the users' behavior.

Generally, a session for web search is a series of successive queries to satisfy a single information need and some clicked search results. In this chapter, this focuses on inferring user search goals for a particular query based on their previous query and its content. Therefore, the single session contain only one query is introduced, which distinguishes from the predictable session. Meanwhile, the feedback session in this chapter is based on a single session, although it can be extended to the whole session.

Table1: User click through logs

5/20/2014 1:56:15 PM	1	http://www.sunnetworks.com	Anu
5/20/2014 2:56:15 PM	2	http://www.wikipedia.org/sun	Anu
5/20/2014 3:16:15 PM	3	http://www.wikipedia.org/moon	Anu
5/20/2014 3:46:15 PM	3	http://www.wikipedia.org/star	Anu
5/20/2014 3:58:15 PM	4	http://www.wikipedia.org/solarsystems	Anu

VI. Object Evaluation and Comparison

To evaluate the performance of the proposed schemes, execution time and storage are the main measurement of performance evaluation. Without loss of generality, this defines processing delay and clustering delay for deployed clustering. Processing delay indicates the execution time for clustering to produce frequent items and corresponding interest before page load. Goal detection delay is also evaluated by measuring time spent on processing time on clustering frequent items and interest in the proposed schemes. Another criterion is cost evaluation. Cost evaluation involves storage and computation aspects.The performance of this proposed work LS using session clustering and pseudo document scheme is compared with two existing approaches method1 and method2. The figure below shows the results and comparison of the proposed system.

The table2 shows the performance comparison of the proposed method with other existing approaches based on the four different metrics clustering delay, time, processing delay, number of iterations.

Table2. Comparison table

	Method I	Method II	LS
Clustering Delay	3702.09	2984.06	2108.08
Result Accuracy	89	90	94
Time	73.71	70.68	49.69
Number of iterations	64.52	57.81	48.21

VII. Conclusion

The proposed novelapproach LS scheme has been proposed to infer user search goals for a query by clustering its implicit feedback sessions represented by pseudo-documents.

Initiallythis introduces an implicit feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the un-clicked ones before the last click and time spent on each link are considered as user implicit feedbacks and taken into account to construct feedback sessions.

The feedback sessions extracted from user logs can reflect user information needs more efficiently. Then it locates feedback sessions to link semantic cluster which is related to pseudo documents to approximate goal texts in user minds. The pseudo LS-documents can enrich the URLs with additional textual contents including the titles, snippets and tags and labels. Based on these Meta cluster information user search goals can then be discovered and represented with some keywords. The main advantage of the proposed method is the implementation of link stay time of user in every link. This can exactly reflect the user search goal.

VIII. Future Enhancement

In future work, the user search techniques will be extended with some other semantic information. This can also be improvedwith session extraction and that will be introduced into the mining system so queries of similar meanings can be clustered and generalized. In addition, more log files of longer periods of time (such as months) are required to fabricate more reliable and more useful rules mining algorithm, which will improve further the performance of the web servers. Usage of most recent association mining will provide best results.

References

[1] Web mining: information and pattern discovery on the World Wide Web R. Cooley, B. Mobasher, and J. Srivastava, 8 Nov 1997.

[2] Nasraoui, O. Krishnapuram, R. Joshi, A. Missouri Univ., Columbia, MO 06 August 2002.

[3] Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, 29 September 2009.

[4] From user Access Patterns to Dynamic Hypertext Linking, TakWoon Yan, Matthew Jacobsen, Feb 1996.

[5] Adaptive Web Sites: Automatically Learning for User Access Pattern Mike Perkowitz and Oren Etzioni, Sep, 1997.

[6] Data Mining of User Navigation Patterns, Mark Levene and Jos´e Borges, August 29, 2000.

[7] Ozmutlu, H. Cenk, and FatihÇavdur. *"Application of automatic topic identification on excites web search engine data logs."* Information Processing &Management 41.5 (2005): 1243-1262.

[8] Boldi, Paolo, et al. *"Query suggestions using query-flow graphs."* Proceedings of the 2009 workshop on Web Search Click Data. ACM, 2009.

[9] Feng, Juan, Hemant K. Bhargava, and David M. Pennock. *"Implementing sponsored search in web search engines: Computational evaluation of alternative mechanisms."* INFORMS Journal on Computing 19.1 (2007): 137-148.

[10] U.Lee,Z.Liu,andJ.Cho, *"Automatic Identification of User Goals inWebSearch,"* Proc. 14thInt'lConf.World WideWeb(WWW'05), pp. 391-400,2005.

Authors Biography

Kathiresan.V is an Assistant Professor, Department of Computer Applications (MCA) in RVS College of Arts and Science, Coimbatore. He received his B.Sc., in 2003 and MCA in 2006 from Bharathiar University, Coimbatore. He obtained his M.Phil. in the area of Data mining from Periyar University, Salem in 2007. His research interest lies in the area of Data mining. He got Faculty Excellence Award from RVS College of Arts & Science for the Academic years 2007-08, 2008-09, 2009-10, 2010-11, 2011-12 & 2012-13 consecutively.

Priyadharsini M completed her undergraduate degree from Maharaja College of arts and science, and has also completed her post graduate Bharathiar University and is currently pursuing her M.Phil in Computer Science at RVS College of Arts &Science, Coimbatore, India. Her area of interest is Data Mining.