ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Data Mining Framework for Web Search Personalization

Geetha S.
4th Sem M. Tech CST
UMIT/ SNDT Women's university,
Mumbai, India

Prof.RachanaDhannawat
Asst Professor, CST Department,
UMIT/ SNDT Women's university
Mumbai, India

Abstract: Web users may have diverse search objectives when they search by giving queries and submitting it to a search engine. The inference and analysis of user search goals can be very useful for providing an expected result for a user search query. The below mentioned framework will give the data identified with the client objectives by analysing search engine query logs. User would practically benefit if search results could be displayed in different categories with manual rating so that user defined highest rated URLs are displayed at the top. It will help user to save effort and time in retrieving the required information. For achieving this feedback sessions are constructed from user click-through logs and pseudo-documents are generated. Pseudo-documents represent the feedback sessions for clustering. Using "Classified Average Precision (CAP)" criterion the performance of inferring user search goals can be evaluated.

Keywords: feedback session, Pseudo- documents, CAP, VAP, Risk

I. INTRODUCTION

Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining technique used to extract knowledge from Web data. Web data consists of Web content data with text, image, record, Web structure data with hyperlinks, logs and Web usage data with http logs, app server logs. Analyzing and exploring regularities in click through records which comprise of URL's, snippets, time interval, click sequence for electronic commerce, enhance the quality and delivery of internet information services to the end user, and improve Web server system performance. User search goals are words that define the group of results in a better way. For example, if a query "mouse" is entered for search we may get results for biological info of mouse as well as mouse as a part of computer. Both the results are displayed as intermixed with each other. Here, first user search goal is mouse as a rodent and another goal is hardware mouse .User search areas are considered as bunch of information generated from pseudo documents. Pseudo documents comprises the words those are more significant to the entered query and from these words some most noticeable words are chosen by K-means clustering algorithm to form user search goals .

II. RELATED WORK

The representation of user search objective can be done with some keywords. Using feedback sessions the user search goal inference can be performed. Each of the clicked URLs and the unclicked ones before the last click are considered as user implied feedbacks and taken under concern to build feedback sessions and map this to pseudo documents. The pseudo-documents will supplement the URLs with further textual contents as well as the titles and snippets. Based on these pseudo-documents, user search goals will be learned and denoted with some keywords [1]. The domain knowledge can be represented with different models. Ontology based model and automatically generated semantic network model can symbolize the domain knowledge through domain terms, Web-pages, and the relations between them. Integration of domain knowledge and Web usage knowledge can be used to create a conceptual prediction model. This model can automatically generate a semantic network of the semantic Web usage knowledge. Web –page candidates can be generated by knowledge bases queries [2].A bottom-up approach can be used to study the web dynamics of end users. Web dynamics include the web-related data browsed, collected, tagged, and semi-organized by users. In hybrid bottom-up search engine the produced search results based exclusively on user provided web-related data and their distribution among users. While comparing bottom-up search engine with the traditional one it has been observed that a bottom-up search engine starts from a fundamental part consisting of the most interesting part of the Web and incrementally progresses its ranking, coverage, and accuracy. The bottom-up approach can be integrated with PageRank, resulting in a new page ranking algorithm that can uniquely chain link analysis with users' preferences [3].

Information overload issue can be reduced by organizing the Web content into thematic hierarchies such as Web Directories and listings of topics which are organized and overseen by humans. A Web directory, permits users to find Web sites related to the subject they are interested in, by beginning with wide categories and steadily constricting down, choosing the category most related to their interests. The size and the complexity of the Web directory can be overwhelming cancelling out the gains that were expected with respect to the information overload problem. It is difficult to navigate to the information of interest to a particular user if it resides deep in the directory [4]. The uncertainty analysis for a Web event can help Websites to endorse suitable Webpages of Web events to their visitors. The identification of active and attractive part of a Web event can be done through the uncertainty analysis of the keyword system of a Web event. The semantic uncertainties of keywords in a Web event can help to recommend appropriate Webpages to users. The semantic representation a Web event and utilization of different levels of uncertainties of a Web event are difficult tasks. The non-content based methods and content-based methods can be used for webpage recommendations [5].

Construction of user profile enhances the user profile by means of background knowledge. The enhanced user profile can be used to retrieve focused information and suggesting

good Web pages to the user based on his background knowledge and search query. Classification needs to be done for web pages into particular category along with its probability of belonging to that particular category [6]. An innovative approach has been presented for efficient personalized web search using a method which tries to learn the behavior of user search to make search result relevant to the user. In the calculation of performance measure of personalized web search it has been experimentally verified that the personalized search showing good performance over generic search engine. Creation of User profile has been suggested with two modes. Gathering of all the user previous search history need to be done and categorize the search domain into defined clusters in particular domain of DMOZ categories[7].User profile can represent user's interest. It can also be used to infer their intention when presenting new queries. User profile creation can be done in two modes like manual creation by user and programmed profile generation using user search pattern. Then user query is matched with related classification, where it belongs to stored local database. Different clustering algorithms can be used to classify the local database in different module so that user requests can be further matched with its profile and group to show effective search result. Collaborative filtering and k-nearest neighbour partition clustering can be applied for efficient user personalization [8].User profiling can be initiated by the identification of relevant search term for particular user from previous search pattern by examining web log file maintained in the server. The terms can be attached to user's ambiguous query later. This approach precedes the user's search result and re-ranks the retrieved result by identifying interest value of user on retrieved links. We can also find the user curiosity on retrieved links by examining the user interest value generated from Vector Space Model and real rank of that link. Current interest of user can also be identified by suggesting some relevant and irrelevant keywords to user. Since users have diverse background on same query, it is very difficult for some informative query to identify user's current objective. User issued queries and user-selected snippets/documents are categorized into concept hierarchies that are collected to generate a user summary. When the user issues a query, each of the returned snippets/documents is also categorized. The documents are re-ranked based upon how well the file groups match user interest profiles [9].Based on user navigation pattern user profile creation can be done. The user navigation behaviour is identified and data priority is given while accessing the content in the web. Grouping and ordering techniques are used for discovering user navigation pattern. Click through data can be used that is documented in search engine logs to simulate user practices in Web search. In the existing approaches only certain behavior of user is identified. The prediction and identification of user expectation is not accurately defined. In the adaptive approach for creating behaviour profiles the prediction is not done based on the user transaction. The ant based clustering algorithm can be used for clustering and number of visits made to a single webpage can be calculated. From this data most frequently viewed web pages are identified. Personalized search can be used for faster retrieval of search results. The results shown based on users interest from previous clicks by ranking. Thus the time required for finding results using personalized search is lesser than normal search [10]. In order to achieve personalized web search user profiling methods can be employed. SVM method, k-nearest neighbor method and Rocchio method are some of the common methods used for user profiling. A domain data set can be constructed for calculating the effectiveness of these methods. The effect of increase in number of training documents on personalized search performance can also be measured and analyzed [11].Web users generally issue queries to search engines to draw appropriate information on particular topics. Some search engines customize their results according to user preferences. In existing scenario the results are provided based on ranking algorithm [12].With the growing number of Web pages and users on the Web, the number of requests submitted to the search engines are also growing quickly. Therefore, the search engines needs to be more effective in its process. Web mining techniques are used by the search engines to retrieve appropriate documents from the web database and provide the necessary information to the users. The search engines become very successful and popular if they use efficient Ranking mechanism. Google search engine is very successful because of its PageRank algorithm. Page ranking algorithms are used by the search engines to provide the search results by considering the applicability, importance and content score and web mining techniques to order them according to the user interest. Ranking algorithms may depend on the link structure of the documents or the actual content in the documents [13].

III. FRAMEWORK OF APPROACH

There are four modules in this system
1) Capturing feedback sessions
2) Automated categorization
3) Building pseudo-documents
4) Restructuring based on web search results

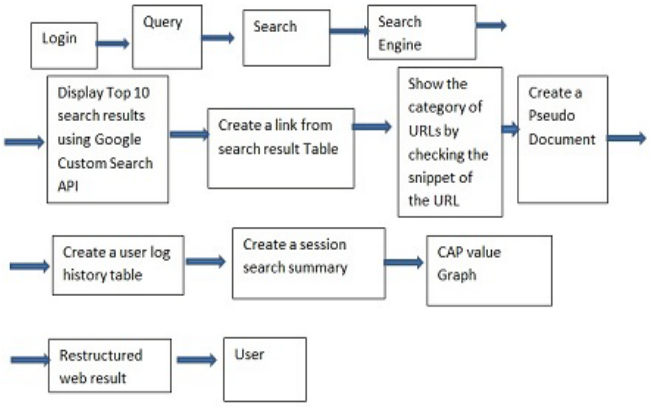

Fig 1: Flow of the System

I) Click through Log

The major outcome of the system depends on user feedback for clustering the obtained results. Once the user fires the query, the unstructured results are obtained which needs to be clustered as per user feedback. So URLs are to be clicked on to create the binary vector and record the click sequence for evaluation.

II) TF - IDF Calculation

Once the clicked and unclicked URLs are recorded for current session, the terms from the URLs are needed to be counted for determining the relevance ratio of the terms to the clicked URLs. So term frequency and Inverse document frequency is

needed to be calculated for analyzing the term count and further pseudo document creation.

III) Pseudo Document Creation

Once the term frequency is computed major clustering criteria is to be decided which is done on basis of Higher TF values obtained for all terms in the documents. Higher 10 TF values are considered as pseudo document contents.

IV) Re-ranking of search results

After clustering the pseudo document, search results are re-ordered according to higher cluster which matches user need. Performance of the re-ordering assess by CAP evaluation.

V) CAP Evaluation

Basically CAP evaluation technique refers click through logs. The clicked URL reflects the relevant results used in calculation of CAP evaluation. CAP evaluation value indicates that cluster of restructured results is relevant to user search goals.

IV. SYSTEM IMPLEMENTATION

Existing System

We define user search objective as the information on different aspects of a query that user groups want to obtain. Information requisite is a user's particular need to obtain information to fulfill his/her needs. User search objective can be considered as the clusters of information needs for a query. The interpretation and scrutiny of user search objective can have a lot of advantages in improving search engine relevance and user experience.

Problems on existing system:

- What users pay attention about varies a lot for different queries, finding appropriate predefined search objective classes is very tough and unfeasible.
- Examining the clicked URLs right from user click-through logs to organize search results. However, this method has restrictions since the number of different clicked URLs of a query may be small.
- Since user feedback is not considered, many noisy search results that are not clicked by any users may be evaluated as well. Therefore, this kind of methods cannot deduce user search objectives specifically.
- Only recognizes whether a pair of queries belongs to the same objective and does not care what the objective is in detail.

New System

In this paper, the main aim is to discover the number of diverse user search goals for a query and representing each goal with some keywords automatically. A unique approach to understand user search objectives for a query by storing click through logs and forming feedback sessions has been created. Later, a different optimization method to map feedback sessions to pseudo-documents which can efficiently replicate user information needs has also been constructed. In the ending phase the clustering of the pseudo documents to depict user search objectives and represent them with some keywords has also been done.

ADVANTAGE:

- Clustering feedback sessions is more effective than clustering search results or clicked URLs directly.
- The spreading of different user search objectives can be obtained appropriately after feedback sessions are clustered.
- By merging the enriched URLs in a feedback session to form a pseudo-document, effective representation of the information need of a user is made.
- Detailed user search objective representation.
- Evaluation of the performance of user search objective inference based on restructuring web search results has been done with new criteria named CAP. Hence, the decision of number of user search goals for a query has been made.

To implement the desktop based application, Google API was downloaded and added to the reference library so that application could fetch URLs from search engines. As a first step, user has to register by clicking on 'Create New Account', fill-in all the details and thus create a user id and password.

Fig 2: Query Submission

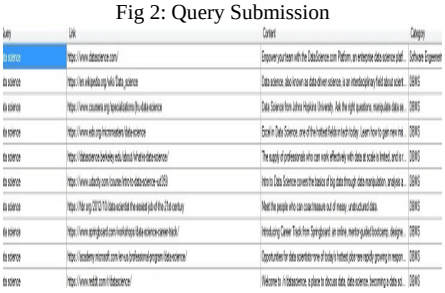

Fig 3: Display of Original Results

Fig 4 :User Search History

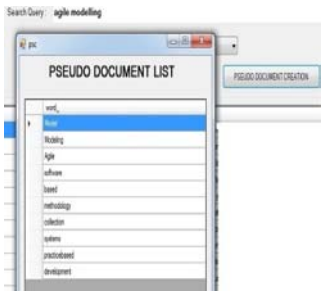

Fig.5: Pseudo document

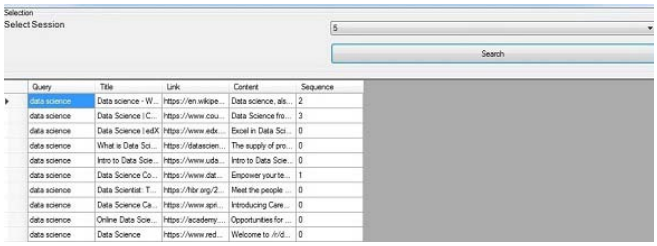

Fig 6: Session Summary

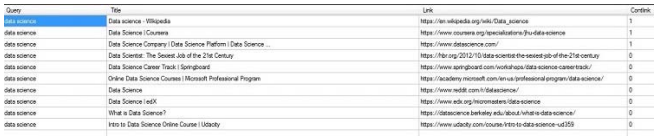

Fig 7 : Restructured Result

V. RESULT

User search goals evaluation is of great concern. Therefore, an evaluation method based on restructuring web search results is used to assess whether user search goals are inferred properly or not. This approach is called CAP (Classified Average Precision). CAP is evaluated based on VAP and Risk. VAP selects the Average Precision of the class that user is interested in with the most clicks/votes. Risk is the factor for wrong classification taking into account. If all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0. Considering one login/logout as a single session, usage logs will be captured and feedback sessions will be generated to restructure and optimize display of search results. The performance of the application will be evaluated using Classified Average Precision (CAP) factor which depends on both of Risk and VAP.

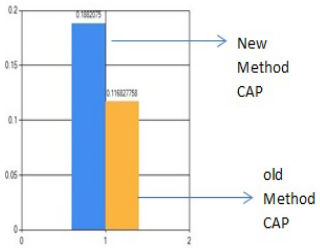

Fig. 8: Comparison of CAP values

VI. CONCLUSION

In this paper, a unique methodology has been introduced to gather user search goals for a query by analyzing its feedback sessions represented by pseudo-documents. First, we present feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo-documents to approximate goal texts in user minds. The pseudo-documents can enrich him URLs with additional textual contents including the titles and snippets. The manual rating will restructure the results as per User requirement irrespective of number of visits. Thus, this desktop application can infer user search goals and save user effort and time. Hence, users can find what they want conveniently.

VII. REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, And Zhaohui Zheng,"A New Algorithm For Inferring User Search Goals With Feedback Sessions", pp.502-513, 2013 IEEE

[2] Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", pp.2574-2587, 2014 IEEE.

[3] Athanasios Papagelis and Christos Zaroliagis,"A Collaborative Decentralized Approach to Web Search", pp 1271-1290, 2012 IEEE.

[4] Dimitrios ierrakos and George Palioura, "Personalizing Web Directories with the Aid of Web Usage Data", pp 1331-1343, 2010 IEEE.

[5] Junyu Xuan, Xiangfeng Luo, Guangquan Zhang, Jie Lu, and Zheng Xu,"Uncertainty Analysis for the Keyword System of Web Events", pp 829-842,2016 IEEE

[6] Rakish Kumar and Aditi Sharan, "Personalized web search using browsing history and domain knowledge", pp. 2161-2174, 2014 IEEE.

[7] Anoj Kumar and Mohd. Ashraf, "Efficient Technique for personalized web search using users browsing history" , 2015 IEEE.

[8] Anoj Kumar and Mohd. Ashraf, "Personalized Web Search Engine using Dynamic User Profile and Clustering Techniques " , 2015 IEEE

[9] Kamlesh Makvana, Pinal Shah "A Novel Approach to Personalize Web Search through User Profiling and Query Reformulation" , 2014 IEEE.

[10] Josna Jojo and Dr.N.Sugana,"User Profile Creation Based On Navigation Pattern for Modeling User Behavior with Personalized Search ",2013 IEEE

[11] C Liang, "User Profile for Personalized Web Search", pp. 1847-1850, 2011 IEEE.

[12] Geetha.S, Rachana Dhannawat, "Predicting User goal in personalized web search and improvement using categorization", pp.01-04,2017 IOSR-JCE.

[13] Ashutosh Kumar Singh, Ravi Kumar P,."A Comparative Study of Page Ranking Algorithms for Information Retrieval " 2009 IEEE