IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A NOVEL APPROACH TO INFER USER SEARCH GOALS FOR OPTIMIZE RESULT

Dhanashri Ingale*, Dr. M. M. Kshirsagar

Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

DOI: 10.5281/zenodo.49797

ABSTRACT

Search engine is one of the most significant applications for internet users. Different users may have different search targets when they submit broad-topic to a search engine. Most times, search engine does not deliver what user needs. So to produce best relevant results, there is necessity to analyse user goals behind searching. Inference and analysis of it can be very useful in improving quality of a search engine's results. In this paper, a framework is projected to examine user search areas for a query by clustering the feedback sessions. Feedback sessions are made from user click-through logs and can more accurately predict the information requirements of users. Post feedback session formations, pseudo-documents are generated which better illustrate feedback sessions for clustering. For clustering, bisecting K-means algorithm is used. After cluster formations, restructuring of web search results is done by calculating smallest distance. At the end, to evaluate the performance of inferring user search goals, a new criterion "Classified Average Precision (CAP)" is proposed.

KEYWORDS: User search goals, feedback sessions, pseudo-documents, clustering.

INTRODUCTION

Nowadays, internet is very popular among users for efficient information retrieval because it gives information very fast also easy to use and contains large amount of information. In the area of web mining which is also an application of data mining techniques, more importance is given to fast and most relevant extraction of information. Web mining is mainly divided into three categories, web usage mining, web structure mining and web content mining. In web usage mining it find the requirements of web user on the internet. In the web structure mining, to show the hyperlink structure of internet graph theory is used. Web content mining involves three steps i.e. mining, extraction and integration of beneficial data from web page content. With the wide range of information available on the internet, searching relevant information of a user's interest is very slow and tedious task. In web search applications user enters the desired query in the website to get the accurate information and that result get appears in the list format. These web users need to go through that long list by examining the titles, tags and snippets to recognise their requirements. This is a time consuming task. Many times is difficult to satisfy users information needs as different users have different information needs for a particular query. For example, when the query "cricket" is submitted to a search engine, some users want to get information about Cricket game, while some others want to learn about Cricket insect. The requirement of the information may vary for each user and to achieve the goal for different user is still becomes difficult. As similar results are offered by the search engines to different user, to avoid this difficulty many researchers developed some techniques. The analysis of user search goal improves the relevancy of information. To recognize goal the inferring technique is useful and to check its relevancy analysis of result is done. The inference and examination of user search goals will have tons of benefits which increases the search engine relevancy. Some advantages are summarized as follows. First is restructuring of web search results [1], [2] in this groups of search results is created which having the same search goal; therefore, users who have different search goals able find what they want. Second advantage involves the use of keywords which represents user search goals and able to use in query recommendation [3], [4]; thus query representation is help users to create their queries more exactly. Third advantage is the distribution of user search goals. This is useful in applications such as re-ranking web search results which contain totally different user search goals.

LITERATURE SURVEY

Research in web log mining has been subject of interest for researchers from many years. Lots of earlier works has been inspected on problem of analysing user query logs [5], [6]. Query logs information is useful and used in many different ways, such as to infer search query intents, to classification of queries, to provide context-based search, to help in personalization, to suggest query substitutes and to identify frequently asked questions (FAQs).

Document illustration model is presented by B. Poblete et al. [7] with the use of implicit user feedback which is collected from search engine queries. The main goal of this is to achieve higher results by non-supervised task i.e. with the help of clustering and labelling, through the incorporation of usage information obtained from search engine queries.

H. Cao, Hang Li et al. [8] introduces a context-aware query suggestion approach which they represented is in two steps. First step is online model learning, to deal with information thinness clustering concept is used, they cluster a click-through data by which queries are summarized into concepts. To capture user's search context mapping is done with the help of query sequence to a sequence of concepts. Second step, from session information, a concept sequence suffix tree is constructed because the query suggestion model.

R. Baeza-Yates et al. [9] proposed a technique which gives a list of associated queries, when any query give in to a search engine. Then with the help of clustering method it uses the content of historical preferences of users registered within the query log of the search engine.

S. M. Beitzel et al. [10] proposed an approach which examines classification effectiveness, they examines pre vs. post-retrieval classification effectiveness. These two were the previously unaddressed difficulties in query classification. And because of this, effect of training explicitly from classified queries vs. bridging a classifier trained victimization document taxonomy.

T. Joachims et al. [11] presented a complete study in which they address the trustworthiness of implicit feedback for web search engines. In this, detailed proof regarding the users‟ decision method is get combine by watching i.e. by eye tracking, with a comparison against specific relevance judgments.

Work is distributed into three classes on the basis of investigation associated to user search goals analysis, and these three classes are query classification, search result reorganization, and session boundary detection.

1) Query classification:

In the query classification, classification of query is done according to predefined classes and these classes are defined by people. Lee et al. in 2005 [12] proposed a concept which works in two steps, in first step they divide user goals in two parts i.e. Navigational and Informational and in second step they accordingly categorize queries into these two classes. Li et al. in 2008 [13] classify queries according to the product intent and Job intent. These two are the predefined intents. They [14] focus on tagging queries with some predefined concepts to improve feature representation of queries. However, since what users care about varies a lot for different queries, searching suitable predefined search goal classes is very hard and also not practical.

2) Search Result Reorganization

In the search result reorganization, search results are reorganized by people. To organize search results Wang and Zhai [15] first they directly analyse clicked URLs from user click-through logs and authors learn interesting aspects of queries or of comparable topics. Then in second step by using previous query words entered by users they produce cluster labels. As they only consider clicked URLs of a query, the number of different clicked URLs may be very small and this is the limitation of this method. Whenever any query is submitted to search engine it produce a long list of result and analysis is done on that search result without considering any feedback from users [16], [17] i.e. they did not consider any clicked URLs. Since in this method user feedback is not considered, many noisy, unwanted search results that are not clicked by any users may be analysed as well. For this reason, this kind of methods cannot infer user search goals precisely.

3) Session boundary detection:

In the Session boundary detection, main goal is to detect session boundaries. Jones and Klinkner [18] in 2008 introduced the notion of query substitution by crating query to exchange users original search query. They produced hierarchically segment query logs by predicting goal and mission boundaries. By this their method only identifies whether a pair of queries belong to the same goal. But unable to care what the goal is in details.

EXISTING SYSTEM

1) No feedback:

In very previous systems there is not any feedback session. Whenever user submitted any particular query to the search engine, it gives ranked list of all the related documents.

2) Relevance feedback:

In this type, feedback sessions are introduced. Once feedback documents are collected extracting informative terms from them. Advantage of this type of method is, it suggests more precise terms for the user"s query. These feedback sessions give more relevant information from previous system [19][20].

3) Collective feedback

To improve usability, query suggestion technique is provided by search engines. Google, Yahoo etc. provides this technique. In this they produce result as collective feedback. In collective feedback, whenever user submitted any query to the search engine the logged database is searching for feedback. Instead of considering feedbacks from the same user on other queries in his/her history, collective feedback consider from the other users on the same query.

Problems in existing system:

1) What users care about varies from query to query; finding suitable predefined search goal classes is very difficult and impractical.

2) Whenever a user"s feedback is not considered, lot of irrelevant, noisy search results which are not clicked by the user may also be unnecessarily analysed as well. This is the main reason system"s inability to infer user search goals accurately.

3) In some systems, feedback is considered but only the clicked URLs directly from user click-through logs. This is useful in organizing the search results. Disadvantage of this method is that here only clicked URLs are considered and number of these clicked URLs may be very small; and thus, not a lot feedback data to analyse.

4) Existing systems does not take into account the details to identify goal; only recognizes whether a pair of queries goes to the same goal or mission.

PROPOSED SYSTEM

Complete system design is as shown in figure 1.

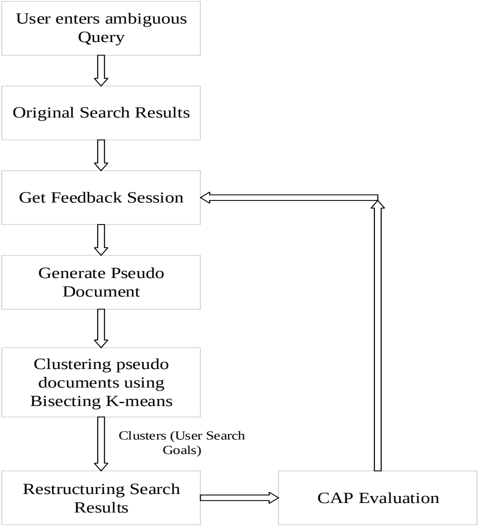
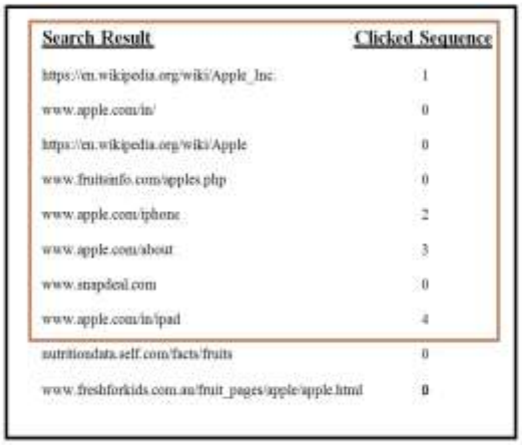

User enters ambiguous Query
Original Search Results
Get Feedback Session
Generate Pseudo Document
Clustering pseudo documents using Bisecting K-means
Clusters (User Search Goals)
Restructuring Search Results
CAP Evaluation

Figure 1: System Architecture

First user enters the query in the system. Original search results will be displayed. From this search results user clicked some URLs, by using this data system makes a feedback session. After that pseudo document is created by mapping feedback session to it and clustering is performed on these pseudo documents. At last for the evaluation of system CAP evaluation criteria is used after restructuring of result. Detailed explanation of all the phases are as follows:

1) Feedback Sessions:

Feedback sessions can be defined as a users‟ implicit feedback. In general, to satisfy user information need there is a sequence of consecutive queries in a session. Here to infer user search goals for a query, single session is considered. Feedback session consist of a combination of both clicked and unclicked URLs up to the last clicked URL in a single session. All the unclicked URLs before the last clicked URL is considered in a feedback session in a single session because those URLs also has been browsed and analysed by the user. As this feedback session contains both clicked and unclicked URLs, clicked URLs shows what exact information user wants and the unclicked URLs shows what information the user do not need. Feedback session does not contain URLs that are present after the last clicked URL feedback because it is not sure whether the user has looked over those URLs or not.



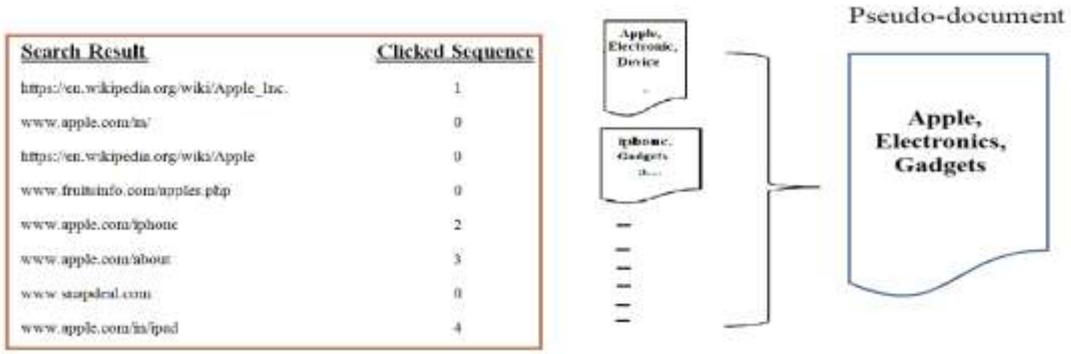
Search Result	Clicked Sequence
https://en.wikipedia.org/wiki/Apple_Inc	1
www.apple.com/in/	0
https://en.wikipedia.org/wiki/Apple	0
www.fruitsinfo.com/apples.php	0
www.apple.com/iphone	2
www.apple.com/about	3
www.snapdeal.com	0
www.apple.com/in/ipad	4
nutritiondata.self.com/facts/fruits	0
www.freshforkids.com.au/fruit_pages/apple/apple.html	0

Figure 2: Feedback Session

In Fig.2, the left part lists shows 10 search results of the query “apple” and the right part is a user’s click sequence. Here “0” represents unclicked URLs and other number represents clicked URLs. The single session includes all the 10 URLs in Figure.2, while the feedback session only includes the eight URLs in the upper rectangular box. Out of eight URLs four are clicked URLs and four are unclicked URLs in this example.

2) Mapping Feedback Sessions to Pseudo-Documents:

Since feedback session differs a lot for various click-through and queries, that’s why it is not suitable for direct use. Therefore some representation method is required to define it. Figure 3 shows a design of mapping feedback session to pseudo documents. There may any number of URLs present in single feedback session and each URL from that feedback session is represent by a little text i.e. its title and snippet. After that some pre-processing is done on those documents, such as removing stop words, stemming by using porter algorithm, transformation of letters from upper case to lower case etc. Term Frequency- Inverse Document Frequency (TF-IDF) vector is used for each URL‟s title and snippet representation. With some weights TFIDF vectors of the URL’s title and snippet are multiplied. In the pseudo document by adding the important terms, weight value is increased. These important terms are acquired on the basis of number of times the words occur in the content. So the TF-IDF vector values of the pseudo document also get affected. Sum of term frequency and inverse document frequency is stored as a feature representation of document in F. The resulting vector representation is used for clustering.



Search Result	Clicked Sequence
https://en.wikipedia.org/wiki/Apple_Inc.	1
www.apple.com/in/	0
https://en.wikipedia.org/wiki/Apple	0
www.fruitsinfo.com/apples.php	0
www.apple.com/iphone	2
www.apple.com/about	3
www.snapdeal.com	0
www.apple.com/in/ipad	4

Figure 3: Mapping feedback session to pseudo document

3) Clustering:

In this, clustering algorithm is applied on pseudo documents to find out the user search goal. The similarity between documents is checked by using Jaccard similarity coefficient function. Bisecting k-means clustering algorithm is applied on the basis of that calculated distance. After clustering algorithm is applied, every cluster is considered as a different user search goal. When Bisecting algorithm is applied, first cluster the documents using k-means algorithm and then on the result of k-means algorithm bisecting algorithm is applied.

Advantages of bisecting k-mean algorithm:

- For the large number of cluster, bisecting k-mean clustering algorithm is more efficient than the regular k-means clustering algorithm.
- Have advantage of not requiring a priori number of clusters, since the clusters are bisected at each step.
- Bisecting algorithm produce cluster of relatively uniform size whereas k-means algorithm produce cluster of non-uniform size.

4) Restructuring web search results:

Web search results are reorganized on the basis of discovered user search goals/intents. As inferred user search goals are depicted with vectors and feature representation of each URL in search result is calculated. Then categorize each URL into a cluster centred with user search goals/intents by selecting smallest distance between user search goal vectors and URL vectors.

EVALUATION

The performance of restructured (clustered) web search results is evaluated by using parameters like Classified Average Precision (CAP). For the calculation of CAP, the value of Average Precision and Risk is required.

1) Average precision (AP): AP is an average precisions computed at the point of each related document in the ranked sequence. AP=1/P+Σrel(r)Rd/r Where,

P: Total number of retrieved documents
r: Ranking of document
Rd: Number of relevant retrieved document of rank r

2) Voted average precision (VAP): It is calculated for restructured search results classes. VAP same as AP and calculated for the class which having more clicks. Calculation is done on the class for which user interested in.

3) Risk: Sometimes value of VAP always be highest because each URL from single session is categorized into the single class no matter whether users have different search goals or not. Because of this, there should be a risk to avoid wrong classification search results into too many classes. It estimates the normalized number of clicked URL pairs that are not belong in the same class.

$$Risk = \frac{\sum_{i,j=1(i<j)}^{m} d_{ij}}{C_m^2}$$

Where, m is number of clicked URLs and dij is 0 if pair of clicked URLs belongs to same class otherwise dij is 1.

4) Classified AP (CAP): VAP is extended to CAP by introducing combination of VAP and Risk. Classified AP can be calculated by using the formula, as follows:

$$CAP = VAP \times (1 - Risk)^{\gamma}$$

In the above equation, CAP selects the AP of the class that user is interested in (i.e., with the most clicks/votes) and takes the risk of wrong classification into account. And γ is used to adjust the influence of Risk on CAP. Finally, we utilize CAP to evaluate the performance of restructuring search results.

Table 1. Evaluation of queries

Query	Voted Average Precision (VAP)	Risk	Classified Average Precision (CAP)
Cricket	0.757	0.3	0.5897
Apple	0.755	0.2	0.6458
Bat	0.7961	0.33	0.6014
Jaguar	0.7596	0.25	0.6210
Sun	0.8734	0.28	0.6939

CONCLUSION

In this paper, user goals are inferred by clustering the feedbacks given by the user. Inferring search goal is very ensuring solution in order to get the related information to user query. First, feedback sessions are introduced which is considered as user implicit feedbacks and it contains combination of both clicked and unclicked URLs until the last clicked one instead of search results or clicked URLs. Use of feedback sessions satisfies user information desires more effectively. Second, pseudo document is made from mapping of feedback session to estimate goal texts in user minds. Next clustering is performed on this pseudo document and for clustering bisecting k-mean clustering algorithm is used. This algorithm helps to give more efficient result. Post clustering restructuring of web search result is done. Finally, classified average precision (CAP) is used to evaluate the performance. So, users can find most relevant information quickly and very efficiently. The system is less time consuming and efforts are minimised for searching targeted data.

REFERENCES

[1] X. Wang and C.-X Zhai, ―Learn from Web Search Logs to Organize Search Results,‖ Proc. 30th Ann. Int"l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR "07), pp. 87-94, 2007.

[2] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, ―Learning to Cluster Web Search Results,‖ Proc. 27th Ann. Int"l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR "04), pp. 210-217, 2004.

[3] R. Baeza-Yates, C. Hurtado, and M. Mendoza, ―Query Recommendation Using Query Logs in Search Engines,‖ Proc. Int‟l Conf. Current Trends in Database Technology (EDBT ‟04), pp. 588-596, 2004.

[4] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, ―Context-Aware Query Suggestion by Mining Click-Through,‖ Proc. 14th ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (SIGKDD ‟08), pp. 875-883, 2008.

[5] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (SIGKDD ‟02), pp. 133-142, 2002.

[6] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (SIGKDD ‟00), pp. 407-416, 2000.

[7] B. Poblete and B.-Y Ricardo, "Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Int‟l Conf. World Wide Web (WWW ‟08), pp. 41- 50, 2008.

[8] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (SIGKDD ‟08), pp. 875-883, 2008.

[9] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Int‟l Conf. Current Trends in Database Technology (EDBT ‟04), pp. 588-596, 2004.

[10] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int‟l ACM SIGIR Conf. Research and Development (SIGIR ‟07), pp. 783-784, 2007.

[11] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feedback," Proc. 28th Ann. Int‟l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ‟05), pp. 154- 161, 2005.

[12] U. Lee, Z. Liu, and J. Cho, ―Automatic Identification of User Goalsin Web Search,‖ Proc. 14th Int„l Conf. World Wide Web (WWW „05),pp. 391-400, 2005.

[13] X. Li, Y.-Y Wang, and A. Acero, ―Learning Query Intent from Regularized Click Graphs,‖ Proc. 31st Ann. Int„l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR „08), pp. 339-346, 2008. 14]

[14] D. Shen, J. Sun, Q. Yang, and Z. Chen, ―Building Bridges for Web Query Classification,‖ Proc. 29th Ann. Int„l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR „06), pp. 131-138, 2006.

[15] X. Wang and C.-X Zhai, ―Learn from Web Search Logs to Organize Search Results,‖ Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[16] H. Chen and S. Dumais, ―Bringing Order to the Web: Automatically Categorizing Search Results,‖ Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI „00), pp. 145-152, 2000.

[17] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, ―Learning to Cluster Web Search Results,‖ Proc. 27th Ann. Int„l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR „04), pp. 210-217, 2004.

[18] R. Jones and K.L. Klinkner, ―Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs,‖ Proc. 17th ACM Conf. Information and Knowledge Management (CIKM „08), pp. 699-708, 2008.

[19] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int‟l Conf. Knowledge Discovery and Data Mining (SIGKDD ‟08), pp. 875-883, 2008.

[20] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.