International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

A Novel Technique for Inferring User Search using Feedback Sessions

Suraj Gothankar[1], Pooja Kharat[2], Karishma Panen[3], Prof. Harish Barapatre[4]

[1,2,3]Final Year student, Department of Computer Engineering, Y.T.I.E.T., Karjat, Maharashtra, India
[4]Assistant Professor, Department of Computer Engineering, Y.T.I.E.T., Karjat, Maharashtra, India

---

Abstract:- Web search engines have become a resource pool to gain information. We submit queries to search engines to retrieve search results. These queries basically depict information needs of users. But there are situations in which these queries do not represent user's specific information needs as varied users have different needs with respect to same query. There are many search engines like Google, Yahoo, Amazon that retrieve search results for users based on their ranking algorithm. It should also be noted that while inferring user search goals, fulfillment of user's desire is very important. So a Novel Technique to Infer User Search Goal using Feedback Session is proposed. The basic step of Inferring user search goals is generation of feedback session through user click-through logs. Feedback session distinguishes between what user requires and what the user does not require. Keywords in the query will determine whether a document satisfies the user's needs. The initial search results will later be restructured based on users' needs and manually Average Precision would be used to examine the performance of users. The main objective of this technique is to improve information retrieval by generating feedback sessions and rank the results list based on user's domain of interest. These re ranked results will be categorization based which will be dependent on the keywords generated.

Keywords: User search goals, feedback session, Reranked search results, Average Precision, Categorization.

I. INTRODUCTION

Web Mining is a thriving & expanding research field. It combines the domain of Internet and Data Mining. The Web Mining research encapsulates areas of Computer Science such as Database Technology, Artificial Intelligence and Information Retrieval. Web Mining is classified as:

a. Web Structure Mining
b. Web Content Mining
c. Web Usage Mining.

Web structure mining gives much stress on discovering how to model the underlying link structures of the Web. Web content mining gives much emphasis on Information Retrieval and Discovery of relevant information from the Web. Web usage mining is a class that predicts user's behaviour on the web based on their usage pattern.

Web is a large Dynamic, ever-changing, innovative field that provides huge amount of information. Data present on the web is increasing at a large extent and most of data on web is fetched with the help of web crawler. To look up information in web, users submit their queries to search engines. But some queries do not represent user's specific information needs. The query maybe very broad and varied users have different needs with respect to same query.

There are many search engines like Google, Yahoo, Amazon that retrieve search results for users based on their ranking algorithm. The results are identified by web users by going through the list that are accompanied by title and snippets. It becomes very exhausting task since multiple subtopics get mixed together. Therefore, it becomes necessary and vital to capture different user search goals in information retrieval. These goals are defined as clusters of information needs for a query. It is essential that a search engine satisfies users need to obtain proper information.

User search goals can be divided into three main categories:

a. Query classification
b. Search result reorganization
c. Session boundary detection

In Query Classification, user goals are inferred by predefining some specific classes on which classification is performed. It becomes difficult and nearly impossible to find an appropriate predefined search goals class as users needs vary with respect to different queries. . In the second class, search results are reorganized by analysing the clicked URLs directly from user click-through logs. Now some URLs of query are small in size. This is also one of the main drawbacks of this class. [1] However, here user feedback is not considered. This results in noisy search results being analysed even if not clicked by any user. Therefore, this technique is imprecise in inferring user search goals. In the third class, session boundaries are detected. Session boundaries establish a common context based on user sessions and frequency of user activities. But this method only identifies whether a pair of queries belongs

to the same category and does not try to find the semantics of goal.

Taking into consideration all the drawbacks of existing system a novel technique is proposed that will infer user search goals with feedback sessions and generate re ranked search results that are categorized according to predefined keywords.

II. RELATED WORK

A lot of work has been investigated on user search goals. It is as below:

In Zheng Lu et.al proposed two methods for inferring user search goals. The first method deals with clustering of search results. In this method, a program is created that submits the queries to search engine. First 100 results are crawled up including the title and snippets. Then each search result is mapped to feature vector. The search results are clustered by means of K-means clustering and the most optimal K is selected by means of Classified Average Precision Criterion. But in this technique, search results are clustered. However, this technique resulted in noisy URLs which were randomly clicked by users also getting selected, thus degrading the performance.

In the second method different clicked URLs are clustered to infer user search goals. User click-through logs are created by combining different single sessions. These URLs, which were very few were selected along with their titles and snippets. A feature vector is assigned to every clicked URL. Finally, results are clustered. But in this method, instead of feedback sessions the clicked URLs are clustered. So the number of URLs clicked differently is reduced. Thus, precise segmentation becomes difficult.

In user search logs were generated from feedback session and accordingly a framework was proposed. This resulted in Pseudo-document being created which represented users search interest. An innovative approach known as Pattern Matrix was used. It consisted of documents and patterns were calculated as inputs. And later the documents were clustered. In this technique a new Clustering method called as semantic clustering was used to cluster documents. And by using TF*IDF matrix user search goal text were analyzed. Though TF*IDF is a good weighting algorithm, it clustered documents that had similar keywords so it identified near identical documents but it failed miserably for all those topics that shared one keyword even if they shared the same topic. Also for effective semantic clustering, the clusters should be correctly specified by the programmer because in this approach the clusters were created on demand and were not specified by the user.

K.N. Vimal Shankar et.al [3] proposed a framework in which user profile was created for each individual and it got authenticated by client. The user interest was registered in the database and according to the change in interest of user the database got updated. Feedback sessions were generated and they were clustered dependent on the keywords. The original search results were then restructured and they were based on user search goals. In in this approach, keywords that are also known as Goal texts determined whether a document could satisfy the user's needs. But goal texts were latent and were not expressed explicitly. That is why Pseudo-documents were created. Feedback session tells what the user requires and what the user does not require. It depends on the number of clicked and un-clicked URL. Certain situations may arise where the user might not click some URLs if they appeared similar to previously clicked ones. The "unclicked "URLs reduced the weight of some terms in pseudo-documents which resulted in incorrect results.

Based on the study of related work, the following gaps were identified:

a. Some URLs might get skipped by users as they appeared similar to previous ones. Also the unclicked URLs could wrongly reduce the weight of some terms in –pseudo-documents.

b. TF*IDF, a good weighting algorithm clustered documents that had similar keywords but it failed miserably for topics that shared one keyword even if they shared same topic.

c. There were many noisy URLs in the search results that were seldom clicked by users. When search results were clustered with noisy ones its performance degraded greatly.

d. In situation where the number of clustered URLs that were clicked by user appeared small, to segment them precisely was a difficult task.

III. PROPOSED FRAMEWORK

Based on the loopholes identified in the existing systems, a novel technique is proposed in this paper that will infer the user search goals by clustering the feedback sessions and then perform categorization based re ranking with the help of Keywords predefined in database. The proposed system represented its block diagram as shown below in Fig.1:

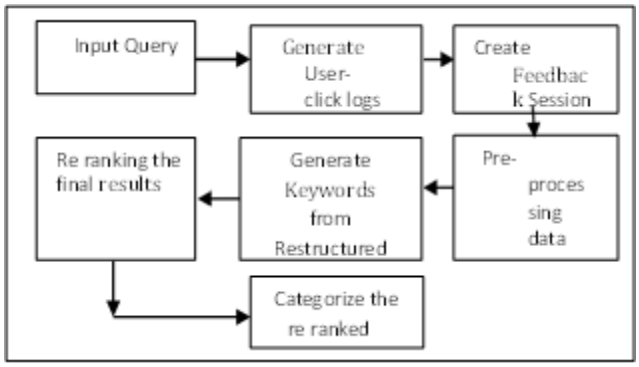

Fig.1. Block Diagram of proposed system

The components of proposed solution are as follows:

1. INPUT QUERY

For feedback session to be generated it is essential to have user-click patterns. In this the user enters any vague search term in search engine.

2. GENERATE USER-CLICK LOG

The user searches for a term in search engine like Google and Yahoo. The links that are generated while searching for the term will be stored in database. The logs would be created through the users click pattern. The user click log for a query is represented below in Fig.2:

Fig.2. User Click Log

3. CREATE FEEDBACK SESSION

In this module, a feedback session is created which consists of both clicked and unclicked URLs which are in a way part of single session. All the URLs before the last click will be scanned and evaluated. The clicked URLs are those which are needed by users. The unclicked URLS are on the other hand those which are not needed but evaluated once by users. So, the unclicked ones will also become a part of user feedback.

Consider the following example wherein a feedback session inside a single session is displayed below in Fig. 3a:

Fig.3. A Single session consisting of Feedback session

4. PRE-PROCESSING DATA

For each distinct query, in each feedback session the following pre-processing steps are carried out-

a. Stop words Removal: Removes frequently occurring keywords like a, an, the, on, over etc. from each pseudo-document.

b. Non-words Removal: Removes non-words i.e. symbols like @, #, %, etc. from each pseudo- document.

c. The extraction and removal of stem words like “ing”, “ly” at the end of each word was done using Porter’s stemming algorithm.

The Pseudo-document generated for one of the clicked URL is shown below in Fig.4:

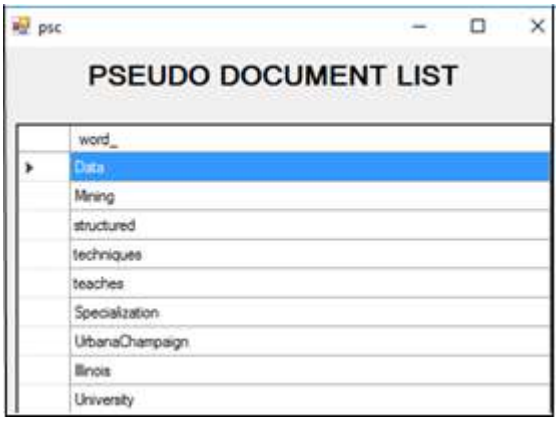

Fig.4. List of Pseudo-document

5. GENERATE KEYWORDS FROM RESTRUCTURED WEB SEARCH RESULTS

In the above table, a “0” in the click sequence indicates “unclicked URL” All the 10 URLs are used to construct one session. The URLs in the box construct a feedback session. As the URLs are scanned from top to bottom, not only the four clicked URLs but also the six unclicked ones are browsed and evaluated. The clicked URLs give

information about user's needs and un -clicked ones tell us which links to ignore. Thus it effectively analyzes feedback session and infers search goals.

After the data is preprocessed the keywords would be generated from restructured web search results. Top 10 queries would be selected from the application and the links would be shown to the user. Finally based on the user-click pattern the results will be stored in database. And Feedback session would be represented in more meaningful manner.

6. RERANKING SEARCH RESULTS

Every time the user performs click-through it will keep on adding in database. If the user again performs similar click- through there will be reranking performed. There are many algorithms used in past for reranking like Graph algorithm, Pattern Similarity Matching algorithm, Relevancy algorithm.

The search results will be recorded. And the reranking will also be performed based on users past history. The restructured results is shown below in Fig. 5a :

Tid	Title	Query	link	Contlink	Userid
1	Data mining - Wikipedia	data Mining	https://en.wikipedia.org/wiki/Data_	6	1
2	Data Mining | Coursera	data Mining	https://www.coursera.org/specializa	6	1
3	What is data mining? - Definition fro	data Mining	http://searchsqlserver.techtarget.co	0	1
4	Data Mining	data Mining	http://www.investopedia.com/term	6	1
5	Data Mining and Modeling - Researc	data Mining	https://research.google.com/pubs/D	6	1
6	What is data mining? | SAS	data Mining	https://www.sas.com/en_us/insight	6	1
7	Data Mining | Sloan School of Mana	data Mining	https://ocw.mit.edu/courses/sloan-s	6	1
8	Data Mining and Knowledge Discove	data Mining	http://link.springer.com/journal/106	6	1
9	Data Mining: Practical Machine Lear	data Mining	https://www.amazon.com/Data-Min	0	1
10	Analytics, Data Mining, and Data Sci	data Mining	http://www.kdnuggets.com/	0	1

Fig.5. Restructured results

7. CATEGORIZE RERANKED SEARCH RESULTS

For categorization, certain categories will be pre-defined in database. Depending upon the Keywords generated the reranked search results will be categorized. Again there are many categorization algorithms used in past like Deep Classifier, Search Result Record (SRR) grouping, Support Vector Machine [8][9]. One out of these algorithms can be used for categorization whichever is found feasible and easy to understand by the specified end –user.

IV. PERFORMANCE ANALYSIS

As user search goals are not defined, inference of user goals becomes an important problem in Information Retrieval. Thus, we need to use a metric that evaluates the performance of user search goal inference objectively and effectively. The novel criterion "Average Precision" (AP) evaluates the restructured results. The formulae for AP is displayed as equation (1),

$$AP = \frac{1}{N^+} \sum_{r=1}^{N} rel(r) \frac{R_r}{r}$$

– AP is the Average of precisions computed in the point of each relevant document in the ranked sequence

– N + is the number of relevant (or clicked) documents in the retrieved ones, r is the rank

– N is the total number of retrieved documents

– rel () is a binary function on the relevance of a given rank

– Rr is the number of relevant retrieved documents of rank r or less.

This is the evaluation metric that will be compared individually against the existing techniques like Bayesian Classifier, Naïve Bayes Algorithm, and Bisecting K-means that clusters both the clicked URLs and search results for inferring user search goals.

V. CONCLUSION

In this work, we have presented a technique to infer user search goals with clustering of feedback session. Because previous works focused on user click feedback only for a single session, as user may search different concepts at different sessions [11]. Instead of relying on search results or clicked URLs, feedback sessions are used to infer user search goals. Feedback session are the URLs list obtained after query submission [10]. The search results are restructured and represented with some keywords. A new criterion AP i.e. Average Precision is used to evaluate the performance of user goals. The reranked search results are categorized dependent upon keywords. This further improves the performance of underlying search engine and proves itself to be very effective technique with better results to infer user search goal.

VI. REFERENCES

[1] Zheng Lu , Hongyuan Zha , Xiaokang Yang , Weiyao Lin ,Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 25, No. 3, pp.502-512 ,March 2013.

[2] Shine Sonali Bhaskar, Prof. Bharat Tides, "A New Approach and Compressive Survey on Restructuring User Search Results by using Feedback Session", IEEE International Conference on Computing Communication

Control and Automation (ICCUBEA), pp.479-484, February 2015.

[3] S .Preetha , K.N. Vimal Shankar, "Personalized Search Engines On Mining User Preferences Using Click through Data",IEEE International Conferences on Information Communication & Embedded systems (ICICES), pp. 01-06 , February 2014.

[4] Ms. D. Indumathi ,Dr. A.Chitra, "A Collaborative Search with Query Expansion and Result Re-ranking", IEEE World Congress on Information and Communication Technologies, pp.986-989, December 2011.

[5] Dasari Amarendra, Kaveti Kiran Kumar, "Inferring User Search Goals with Feedback Sessions using K-means Clustering Algorithm", International Journal of Computer Engineering In Research Trends (IJCERT) Volume 2, Issue 11, pp. 780-784, November 2015.

[6] Jiafeng Guo, Lin Ding, Gang Zhang, Yue Liu, Xueqi Cheng, "PSM: A new Re-ranking algorithm for Named-page", In Proc. of the Fifteenth Text Retrieval Conference, TREC 2006, Gaithersburg, Maryland, pp.1-5 November 14-17, 2006.

[7] Antonis Sidiropoulos, Yannis Manolopoulos, "Generalized comparison of graph-based ranking algorithms for publications and authors", The Journal of Systems and Software 79, Department of Informatics, Aristotle University, pp. 1679-1700, 16 January 2006.

[8] Sukanya S. Gawade , Gyankamal J.Chhajed, "Using Feedback Sessions for Inferring User Search Goals", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE),Vol. 3, Issue 6 , pp.7015-7019, June 2016.

[9] Dikan Xing, Gui-Rong Xue, Qiang Yang, Yong Yu1, "Deep Classifier: Automatically Categorizing Search Results", In Proc. First ACM International Conference on Web Search and Data Mining (ACM WSDM 08), Stanford, California, USA, pp. 139-148, February 11-12, 2008.

[10] Febna V, Anish Abraham, "User Search Goals Evaluation withFeedback Sessions", International Conference on Emerging Trends in Engineering , Science and Technology (ICETEST), Procedia Technology (24), pp.1256-1262, Available online at www.sciencedirect.com.