IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Novel Approach to Infer User Search Goal for Query by Clustering Its Feedback Session

Ku. Sushama K. Deotale[1], Prof. M. S. Khandare[2]

M.E. III[rd] Sem, Dept of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India[1]

Lecturer, Dept of Computer Engineering, Jagadambha College of Engineering and Technology, Yavatmal, India[2]

Abstract: For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

Keywords: User search goals, feedback sessions, pseudo-documents, restructuring search result, clustering, classified average precision (CAP).

I. INTRODUCTION

In web search applications queries are given to search engines for getting user search needs. Then web search engine provides the results for the query entered by the user. But sometime queries are unable to express the exact needs of the user because different queries may represent the different aspects. Users are usually giving some keywords representing their interests in their minds. Such keywords do not match with the results produced by the search engines.

We cannot guess the user behaviour exactly. For example, when the query "The apple" is submitted to a search engine, some users want to learn fruit apple, while some others want to learn the apple iphone, iPod etc.

1.1 Information retrieval (IR):-

It is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections. Web search engines are the most visible IR applications as follows:

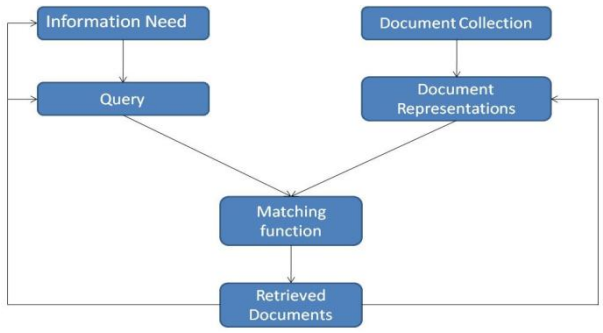

Fig: 1 Information Retrieval

However, sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. Therefore, it is necessary and potential to capture different user search goals in information retrieval.

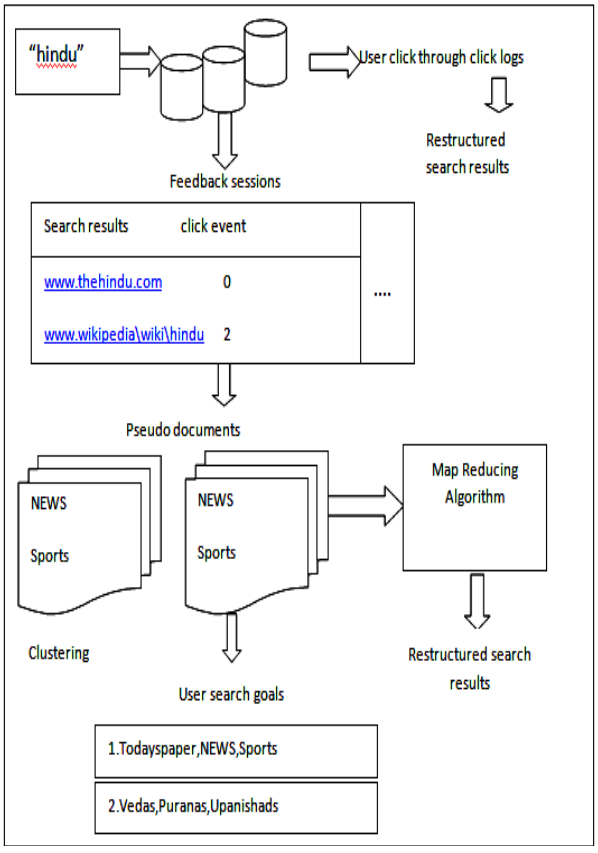

Fig: 2 framework of our approach

DOI 10.17148/IJARCCE.2016.53258

This system define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.Consider the following example of the query "hindu"

When the user type the word" the hindu", then it may be related to newspaper "The hindu" or it may related to the" religion hindu". Our system provides him the efficient result providing separate links as above.

II. LITERATURE SURVEY

The information in query logs has been used in many different ways, such as to infer search query intents or user goals, to classify queries, to provide context during search, to facilitate personalization, to suggest query substitutes and to identify frequently asked questions (FAQs).
Effective organization of search results is critical for improving utility and relevance of any search engine. Clustering search results is an effective way to organize search results which allows a user to navigate into relevant documents quickly. Generally all existing work [3], [7] perform clustering on a set of top ranked results to partition results into general clusters, which may contain different subtopics of the general query term. However, this clustering strategy has two deficiencies which make it not always work well. First, discovered clusters do not necessarily correspond to the interesting aspect of a topic from user-oriented perspective. Second, cluster labels are more general and not informative to identify appropriate clusters.
Wang and Zhai [2] clustered queries and learned aspects of these similar queries [18], which solves the problem in part. However, their method does not work if we try to discover user search goals of one single query in the query cluster rather than cluster of similar queries.
H-J Zeng et.al [3] proposed a query based method to cluster search results. For a given query, the rank list of documents return by a certain Web search engine, it first extracts and ranks most salient phrases as candidate cluster names, base on a regression model learned from pervious training data. Candidate clusters are formed by assigning documents to relevant salient phrases and the final cluster are generated by merging these candidate clusters. But this method only produces the result with higher level of the documents only and it doesn't make the results for all search based user goals.
T. Joachims [5] proposed an approach to automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking. Taking support vector machine (SVM) approach, for learning ranking functions in information retrieval. Preceding studies encompass mainly focused on manual query-log investigation to recognize Web query goals. U. Lee et al. [11] studied the "goal" at the back based on a user's Web query, so that this goal can be used to get better the excellence of a search engine's results. Their proposed method identifies the user goal automatically with no any explicit feedback from the user. User may issue number of queries to search engine in order to achieve information need/tasks at a variety of granularities.
R. Jones and K.L. Klinkner [7] proposed a method to detect search goal and mission boundaries for automatic segmenting query logs into hierarchical structure. Their method identifies whether a pair of queries belongs to the same goal or mission and does not consider search goal in detail. Zamir et al. [8] used Suffix Tree Clustering (STC) to identify set of documents having common phrases and then create cluster based on these phrases or contents. They used documents snippets instead whole document for clustering web documents. However, generating meaningful labels for clusters is most challenging in document clustering. So, to overcome this difficulty, in [3], a supervised learning method is used to extract possible phrases from search result snippets or contents and these phrases are then used to cluster web search results ,improve the retrieval quality [8], [9].

2.1 DISADVANTAGES OF EXISTING SYSTEM:

1. What users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.
2. Analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitations 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 analyzed as well. Therefore, this kind of methods cannot infer user search goals precisely.
3. Only identifies whether a pair of queries belongs to the same goal or mission and does not care what the goal is in detail.

In our work, we consider feedback sessions as user implicit feedback and propose a novel optimization method to combine both clicked and unclicked URLs in feedback sessions to find out that users really require and what they do not care.

III. ANALYSIS OF PROBLEM

In this paper, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords.

3.1 ADVANTAGES OF PROPOSED SYSTEM:

To sum up, our work has three major contributions as follows:

DOI 10.17148/IJARCCE.2016.53258

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

1 We propose a framework to infer different user search goals for a query by clustering feedback sessions. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.
2 We propose a novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.

We propose a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.

IV. PROPOSED WORK

1. Proposed work deals to infer different user search goals for a query by clustering feedback sessions:- Queries are submitted to search engines to represent the knowledge required by the users. However, generally queries might not specifically represent user's specific information requirements since several ambiguous queries could cover a broad topic and completely different users might want to induce info on different aspects once they submit a similar question. For instance, once the question "the sun" is submitted to a hunt engine, some users wish to find the homepage of a U.K newspaper, whereas some others wish to find out the natural knowledge of the sun.

2. The distributions of different user search goals can be obtained conveniently after feedback session are clustered: - Restructure internet search results per user search goals by grouping the search results with a similar search goal users with totally different search goals will simply notice what they require. User search goals depicted by some keywords will be used in question recommendation. The distributions of user search goals may be helpful in applications like re-ranking internet search results that contain totally different user search goals. As a result of its quality, several works concerning user search goals analysis are investigated. They will be summarized into 3 classes: question classification, search result reorganization, and session boundary detection.

3. Proposed work deals the method to combine the enriched URLs in feedback session to form a pseudo-document, which can effectively reflect the information need of user:-The feedback session consists of each clicked and unclicked URLs and ends with the last URL that was clicked in a very single session. It's impelled that before the last click; all the URLs are scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click ought to be a region of the user feedbacks. Feedback session will tell what a user needs and what he/she does not care for. Moreover, there are measures of lots of numerous feedback sessions in user click-through logs. Therefore, for inferring user search goals, it's a lot of economical to investigate the feedback sessions than to investigate the search results or clicked URLs directly.

4. The pseudo-documents are then clustered to infer user search goals: - In this, map feedback session to pseudo documents User Search goals. The building of a pseudo-document includes 2 steps. One is representing the URLs within the feedback session. Uniform resource locator in a very feedback session is depicted by a little text paragraph that consists of its title and piece. Then, some matter processes are enforced to those text paragraphs, like remodeling all the letters to lowercases, stemming and removing stop words. Another one is Forming pseudo-document supported uniform resource locator representations. So as to get the feature illustration of a feedback session, we tend to propose an improvement methodology to mix each clicked and unclicked URLs within the feedback session.

5. Analysis of Result. :-The pseudo documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo documents, user search goals can then be discovered and depicted with some keywords. The pseudo documents are clustered using the fuzzy clustering algorithm. The cluster labels are discovers precisely. Finally, classified average precision is formulated to evaluate the performance of user search goal inference. The restructured web search result is produced for every user search query. The result produced is efficient for user.

4.1 System architecture:

While considering the system architecture we consider three things as follows:-First, we can restructure the web search result as per the user search goals by grouping the search result with the same search goal; thus, users with different search goals can easily find what they want.

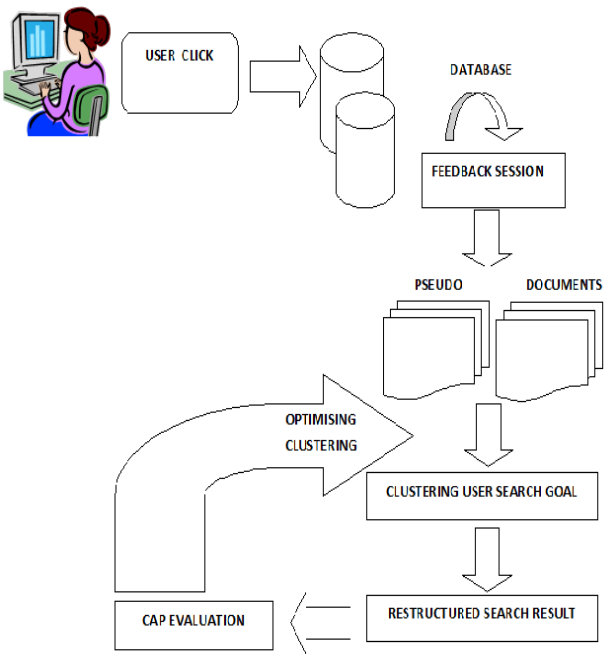

Figure 3: system architecture

DOI 10.17148/IJARCCE.2016.53258

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 3, March 2016

Second, user search goals represented by some keywords can be utilized in the query recommendation; thus, the suggested queries can help users to form their queries more precisely. Third, the distributions of user search goals can also be useful in applications such as reranking web search results that contain different user search goals.

In the following system architecture, basic operations involved in proposed approach to discover user search goals/intents by clustering pseudo-documents are described. Figure 3 shows the architecture of the system. 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 clustering these pseudo-documents and depicted with some keywords. Since we do not know the exact number of user search goals in advance, several different values are tried and the optimal value will be determined.

The original search results are restructured based on the user search goals inferred from the above procedure. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals.

4.2 Dataflow diagram:

Dataflow diagram shows following steps:-
Step 1: Insert any query for searching.
Step 2: Collect the log details of user search history.
Step 3: Depend upon the users click and unclick URL's create a feedback session document.
Step 4: From the number of feedback sessions collected from multiple user create pseudo documents
Step 5: Divide this pseudo Documents temporary groups.
Step 6: Then by using K-means Algorithm create final clusters of collected URL's according to subject
Step 7: Forward this clusters for Classified Average Precision (CAP.)
Step 8: End.

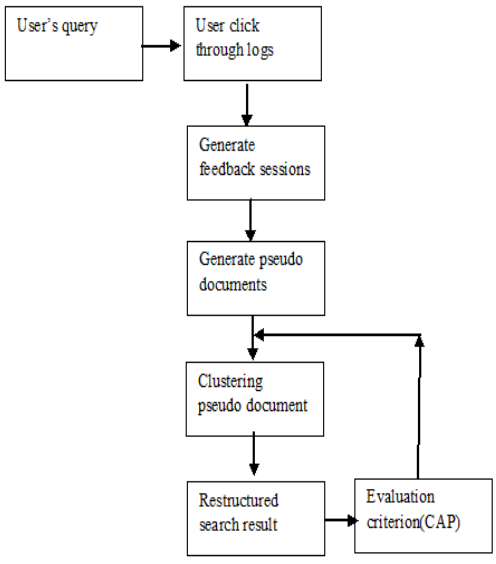

Fig 4 : Data flow diagram

4.3 Modules:

1. User search logs

The user enters the queries to the search engine. The queries are maintained as a log and the results will be produced based on the keywords. The search goals for a query and depicting each goal with some keywords automatically. The user's queries are saved. In web search environment, there are many abundant queries and user clicks. User clicks represent implicit relevance feedback. In this framework, user clicks are recorded in user clickthrough data. User uses clickthrough data stored in user logs to simulate user experience in web search. In general, when query is issued, the user usually scans links to documents in a result list from first to last. Clearly, the user clicks on the links to the documents that look relevant of informed choice and skips other documents. Therefore, the proposed approach utilize user click as relevance judgments to evaluate search precision since clickthrough data can be collected at low cost, it is possible to do large scale evaluation under this framework.

2. Feedback Sessions

The feedback session is formed by both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. Feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

3. Generating Pseudo Documents

The feedback sessions vary a lot for different clicks through and queries, it is not suitable to directly use the feedback sessions. Some procedure is needed to represent the feedbacks in a more efficient way. We can use binary vector method to represent a feedback session. The feedback sessions 1 as clicked and 0 as unclicked is denoted by using binary vector. The binary vector representation is not useful enough to tell the contents of user needs. Here we consider new approach to generate pseudo documents from feedback sessions.
Steps to construct pseudo documents:

i) Represent the URL in the feedback session:
It extracts the titles and snippets of the returned URL's from the feedback sessions. Each URL is represented as a small text paragraph containing title and snippet. Then some textual process is implemented as text paragraphs such as transforming all the letters to lower case stemming and removing stop words. And then TF-IDF vector of URL's titles and snippets are formed.
ii) Forming pseudo documents based on URL representations: In order to obtain the feature representation of a feedback session, we propose an optimization method to combine both clicked and unclicked URLs in the feedback session.

DOI 10.17148/IJARCCE.2016.53258

4. Clustering Pseudo Documents

Clustering is the process of grouping the data into classes or clusters. The Pseudo documents are clustered by using K means clustering algorithm .The K-means algorithm is simple and effective. The terms with the highest values in the centre points are used as the keywords to depict user search goals. The clustering is the process based on a term-weight vector representation of queries, obtained from the aggregation of the term-weight vectors of the clicked URLs for the query.

K-mean Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. After clustering all the pseudo-documents, each cluster can be considered as one user search goal. Clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

5. Evaluation Criterion and Restructuring Search Results

The results are restructured based on the evaluation of web search goals. This approach is called CAP (Classified Average Precision).Search engines will returns millions of search results so it is necessary to organize them to make it easier for users to find what they want. Restructuring Web Search Results it is an application of inferring user search goals. The inferred ones are represented by the feature representation of each URL in the search result. Then categorize them into a cluster centred by the inferred search goals. In this module implement the novel evaluation criterion classified average precision (CAP) to estimate the performance of the reorganized web search results. CAP is extended version of AP and VAP. In AP and VAP, we can't analyze the risks. If all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; however, VAP could be very low. Generally, categorizing search results into fewer clusters will make smaller Risk and bigger VAP, and more clusters will result in bigger Risk and smaller VAP. The intended CAP depends on both of Risk and VAP. Average Precision evaluates according to user implicit feedbacks [1]. It is the average of precisions computed at the point of each relevant document in the ranked sequence. The URLs in the single session are restructured into two classes. Voted AP is the AP of class including more clicks namely votes. Finally experimental result measures the clustering results with parameters like classified average precision (CAP), Voted AP (VAP), risk to avoid classifying search results and average precision (AP). The user search goals are represented as the vectors. So we perform categorization by choosing the smallest distance between the URL vector and user search goal vectors. By this way the results can be restructured according to the inferred user search goals.

CONCLUSION

In this Project, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. Feedback sessions are constructed by considering both the clicked URLs and the unclicked ones before the last click. We introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Therefore, feedback sessions can reflect user information needs more efficiently. Then we generate pseudo documents from feedback session related to user search goals. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. User search goals can be represented with some keywords by using pseudo documents. Lastly, a new criterion CAP is formulated to evaluate the performance of user search goal inference.

REFERENCES

1. Zheng Lu, Hongyuan Zha, Xiaokang, Yang, Weiyao Lin,and Zhaohui Zheng " A New Algorithm for Inferring User Search Goals with Feedback Sessions" IEEE transactions on knowledge and data engineering, vol. 25, no. 3, March 2013
2. P.Srinivasam, S.RajaNamakkal, "India Improved Search Goals with Feedback Session by Using Precision Values", International Journal of Computer Applications (0975 – 8887) Volume 118 – No.14, May 2015.
3. Bhushan Thakare ,Rohan Rawlani ,Sahil Pathak ,Dipali Salve, Ritesh Natekar, " A New Algorithm for Inferring User Search Goals with Feedback Sessions" ,International Journal of Computer Applications (0975 – 8887)Volume 118 – No.14, May 2015.
4. Charudatt Mane, Pallavi Kulkarni Charudatt Mane, "A Novel Approach to Discover User Search Goals Using Clickthrough Data", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 20-24 ISSN:0975-9646.
5. Pranali Dhondiram Desai, Prof.Wadne Vinod Subhashrao ," A New Approach To Discover User Search Goals Using Feedback Session ",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)Volume 3 Issue 11, November 2014.
6. M. Sulaiman, K.Ananthajothi, Syed Zubair Ahmed Hussainy, S.Jayakrishnan , "Classified Average Preciseness: CAP Algorithm for Finding Web Search Goals using Session Feedbacks ",(IJCTA), Int. J. Computer Technology & Applications ,Vol 5 (3),884-888 , ISSN:2229-6093.
7. Bhavesh Pandya, Charmi Chaniyara , Darshan Sanghavi , Krutarth Majithia ,"A New Algorithm for Inferring User Search Goals with Feedback Sessions", Int. Journal of Engineering Research and Applications, www.ijera.com ,ISSN: 2248-9622, Vol. 5, Issue 8, (Part - 2) August 2015, pp.30-33 .
8. Bhuvaneswari S., Shobana.P, Vaishnavi .V, Ramya. P. "Classified Average Precision for Retrieving Information using Feedback sessions", International Journal of Advanced Research in Computer Science Engineering and Information Technology Volume: 3 Issue: 2 10-Mar-2014, ISSN_NO: 2321-3337.
9. H.M .Sameera, N. Rajesh Babu, "Classified Average Precision (CAP) To Evaluate The Performance of Inferring User Search Goals", International Journal of Research in Computer and Communication Technology, Vol 3, Issue 11, November – 2014.
10. B. Saranya, G. Sangeetha, "An Effective Approach for Increasing the Efficiency of Web Searching With Feedback Sessions", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.
11. Mrs. V. Geetha and Ms. S. Santhosi ,"A novel approach for constructing user search goals with Feedback session", IJARRAS, Volume 1, Issue: 4.