# Advanced Local Search Engine to Infer User Search Goals with Feedback Session

Engineering

**KEYWORDS :** Feedback session, pseudo document, click through logs.

| | |
|---|---|
| **R.Bhavadharani** | Assistant Professor,Department of CSE,Panimalar Engineering College,Chennai,India |
| **S.Yamuna Devi** | Assistant Professor,Department of CSE,Panimalar Engineering College,Chennai,India |
| **P.Abitha** | Assistant Professor,Department of CSE,Panimalar Engineering College,Chennai,India |
| **P.Bharathi** | Scholar,Department of CSE,R.M.K Engineering College, Chennai,India |

## ABSTRACT

*When user attempts to search for a query which covers a broad topic and ambiguous, each user will have different search goal. By analyzing search engine query log, different user search goals can be inferred. In order to identify different user search goals search engine query logs are analyzed which will improve user experience and search engine query relevance. Feedback session is constructed from user click-through logs which can efficiently reflect the information needs of users. Different user search goal for a query is identified by clustering the proposed feedback session. For the better representation of feedback sessions for clustering, pseudo-document is constructed from the proposed feedback session. Finally, a framework is proposed to evaluate the performance of inferred user search goals.*

## I. INTRODUCTION

The one important in information retrieval and web mining is to accurately measure the semantic similarity between words of the query [2]. Web mining applications require the ability to perfectly measure the semantic similarity between concepts or entities. One main problem in information retrieval [6] is to retrieve a set of documents that are closely relevant to a given query. Efficiently estimating the semantic similarity between words is crucial for several natural and language preference tasks Word Sensing Disambiguation (WSD), textual entailment and automatic text summarization.

Semantically similar words of a particular word are listed in manually created general-purpose lexical ontology's such as WordNet. Such asynset contains a pair of words for a particular sense of word. But, entities changes over time and semantic similarity across domains. For example, Ring is often associated with Bus topology. However, this sense of ring is not listed in most general-purpose thesauri or dictionaries. If a user searches for ring on the web, he might be in this sense of ring and not as a jewel. At regular intervals new words are being created but also new senses are assigned to existing words. A method to maintain ontology's for capturing these words and senses are costly.

A method is proposed to be optimization for an estimating the similarity between words or entities in the web search engine. It is time consuming to analyze each document separately due to vastly numerous documents and high growth rate of web. For this reason, most web search engines use page counts and snippets. Estimates of number of pages that contain the query words are called page count of a query. In general, page count may not necessarily be equal to the word frequency to the might appear many times on one page.

The main objective is to discover the number of different user search goals for a query and automatically representing each goal with some keywords. First and foremost, user search goals for a query are inferred by clustering our proposed feedback sessions. The feedback session will contain a series of both clicked and unclicked URLs and ends with last URL that was clicked in a single session from user click-through logs. To efficiently reflect information need of user, proposed feedback session is mapped to pseudo documents. Finally, a cluster pseudo documents to infer user search goals of user and depicting each of them with some keywords. An evaluation criterion is proposed to evaluate the performance of the restructured web search results and can help us to optimize the parameter in clustering method when inferring user search goals.

Our work has three major parts:

First, a framework is proposed to find diverse search goals by clustering feedback sessions. It is more efficient to cluster feedback session than to cluster search results or clicked URLs directly. After the feedback sessions are clustered, different user search goals are distributed.

To efficiently represent the information need of a user, an optimization method is developed to combine the enriched URLs in a feedback session into pseudo-document. By this way, user search goals are found in a detailed manner.

An evaluation criterion is proposed to evaluate the performance of user search goal based on restructuring web search results. By this way, the number of user search goals for a query is determined.

## II. ARCHITECTURAL DESIGN

The framework of our approach consists of two parts. In the upper part, first all the feedback session of a query are extracted from user click-through logs and then mapped to pseudo-document and depicting each of them with some closely related keywords. As the exact number of search goal are not known in advance, several different are tried and the optimal value will be determined by the feedback from the bottom part. In the bottom part, based on the user search goals inferred from the upper part, the original search results are restructured. An evaluation criterion is proposed to evaluate the performance of restructuring search results and also to select the optimal number of user search goals in the upper part. This evaluation will be used as feedback.

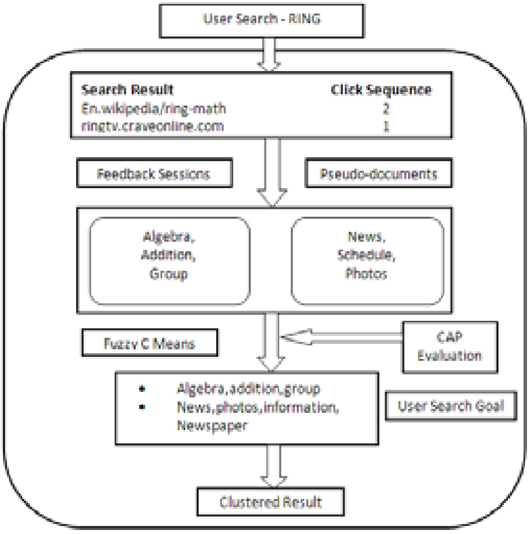


Fig. 2.1 Framework of our approach

## III. ILLUSTRATION OF FEEDBACK SESSIONS

### 1. Ambiguous query

Usually user will submit query to the search engine to satisfy their information need. Many ambiguous queries may cover a broad topic and different users may want to get information on different aspect at different time instant when they actually submit the same query [2]. Sometimes queries having an exactly represent user problem specific information need. For example(Fig. 3.1), when the query "Ring" is submitted to a search engine, some users might want to locate variety design of ring, while others want to locate the home page of boxing, some other want to see Japanese horror film Ring.

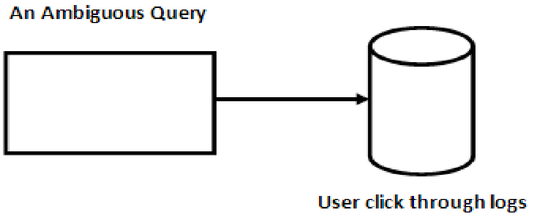


Fig. 3.1 Click through logs

### 2. Feedback session

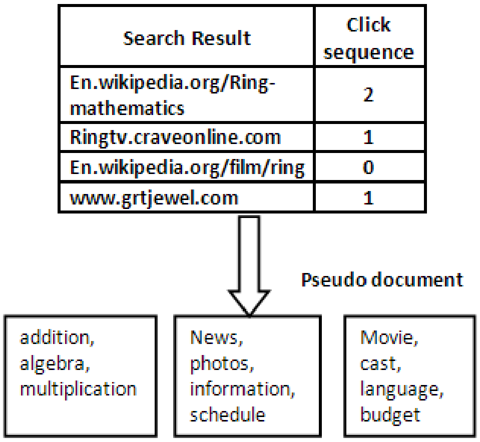
The feedback session is constructed by combining both clicked and unclicked URLs and also ends with last URL that was clicked in a single session. It is assumed that all the URLs have been scanned and evaluated by user before the last click. The clicked URLs together with the unclicked one before the last click should also be considered as user feedback. Feedback session can explain what a user needs and what he/she does not care about. In user click-through logs there are plenty of different feedback sessions. Therefore, it is more efficient to analyze the feedback session rather than analyzing the search results or clicked URL directly for determining user search goals.

| Search Result | Click sequence |
|---|---|
| En.wikipedia.org/Ring-mathematics | 2 |
| Ringtv.craveonline.com | 1 |
| En.wikipedia.org/film/ring | 0 |
| www.grtjewel.com | 1 |

Fig.3.2 Feedback session

### 3. Pseudo document

The feedback session need to be mapped to produce pseudo document. Pseudo document is built in two steps. At first, URL in a feedback session is represented by small text paragraph stating its title and snippet. After that, some textual processes such as transforming all the letters to lowercase, stemming and removing stop words are implemented to those textual paragraphs. Second one is forming pseudo document based on URL representations. So an optimization method is proposed to combine both clicked and unclicked URLs in the feedback session which efficiently reflect the feature representation of a feedback session.



| Search Result | Click sequence |
|---|---|
| En.wikipedia.org/Ring-mathematics | 2 |
| Ringtv.craveonline.com | 1 |
| En.wikipedia.org/film/ring | 0 |
| www.grtjewel.com | 1 |

Pseudo document
addition, algebra, multiplication
News, photos, information, schedule
Movie, cast, language, budget

Fig.3.3 mapping feedback session to pseudo document

### 4. Clustering pseudo document

The pseudo-documents are clustered by FCM clustering which is simple and effective. As the exact number of user search goals for each query is not known in advance, a set of number in clusters to be five different values and perform clustering based on these five values, respectively. After clustering all the pseudo-documents, each one of the cluster is considered as one user search goal. The centre point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster.

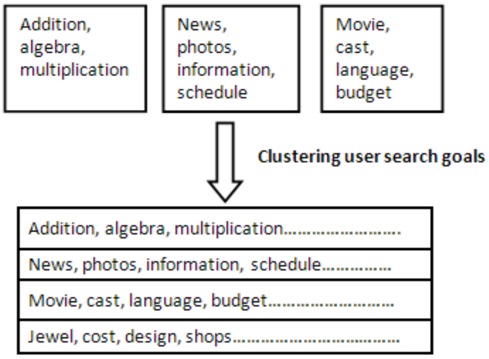

Addition, algebra, multiplication
News, photos, information, schedule
Movie, cast, language, budget
Clustering user search goals
Addition, algebra, multiplication.........................
News, photos, information, schedule...............
Movie, cast, language, budget..........................
Jewel, cost, design, shops...................................

Fig.3.4 clustering pseudo document

### 5. Restructure web search result

According to user search goals, the web results needed to be restructured by grouping the search results with the same search goal users with different search goals. User search goals represented by some keywords can also be used in query recommendation. The distribution of user search goals can also be useful in applications such as re-ranking web search results that contain different user search goals. They can be classified into search result reorganization, [2] query classification and session boundary detection.

## IV. FUZZY C-MEANS CLUSTERING

A fuzzy similarity based algorithm is proposed for feature clustering to reduce the dimensionality of feature vector for text classification. Based on similarity test, the words in the feature vector of a document set are grouped in a cluster. Similar words are grouped into the same cluster. Each cluster will have a membership function with statistical mean and deviation. Automatically the desired number of clusters is formed, when all the words have fed. The extracted feature of a cluster is a weighted collection of words contained in the cluster.

This algorithm closely matches with the derived membership function and also properly describes the real distribution of training data. The problem of determining the appropriate number of features can be avoided, even if the user did not specify the exact number of extracted feature in advance. Experimentally, our method can run faster and obtain better extracted features than other methods.

FCM algorithm partitions a finite collection of n elements into a collection of c fuzzy clusters according to the given criterion. The algorithm returns a list of c cluster centre's and a partition matrix when a finite set of data is given; where each element Wij tells the degree to which element Xi belongs to cluster Cj. FCM aims to minimize objective function like means algorithm. The objective function is:

$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(\mathrm{center}_k, x)}{d(\mathrm{center}_j, x)}\right)^{2/(m-1)}}.$$

where m determines the level of cluster fuzziness

## V. ASSOCIATED WORK

In recent years, so many works have been done for finding user search goals of the query. In fact, most of their work belongs to query classification. For finding different query aspect, some works directly analyze the search results returned by the search engine. However, it has limitation to improve search engine relevance without user feedback. Even though, if some works take user feedback and directly analyzing different clicked URLs in user click-through log of a query, the number of different clicked URLs of a query may not be big enough to get ideal results. Some works clustered queries and analyzed these similar queries, but which solved the problem in part. Their method was unable to work if it tried to discover user search goals of one single query in the query cluster rather than a cluster of similar queries. For example, the query "bike" is clustered with some other queries such as "used bike", "bike crash" and "bike audio". By this method, different aspects of the query "car" are able to be learned. However, the query "used bike" have different aspects in the cluster which is quite different to find by their method. Some works use search goals and mission for hierarchically finding session boundary. But, the method identifies only queries belong to the same search goal or mission and does not care about what the goal in detail. A feedback sessions are taken as user implicit feedback and also it propose an optimization method to find out what users really care and what they do not care by combining both clicked and unclicked URLs in a feedback sessions. According to the inferred user search goals, a user search goals is inferred from user click-through logs.

## VI. CONCLUSION

An approach is proposed to infer user search goals for a query by clustering its feedback sessions which in turn represented by pseudo documents. First feedback session is used for the analysis purpose to infer user search goals rather than search results or clicked URLs. The user implicit feedback is made by combining both clicked and unclicked URLs and also the one before the last click is taken into account to construct feedback sessions. The users information need can be reflected in an efficient way by feedback sessions. Second, pseudo document is produced by mapping the feedback sessions for approximately finding the goal text in user minds. The pseudo document can add additional textual contents such as titles and snippets to the URLs. According to the pseudo document, user search goals are found and also depicted with some keywords. Finally, an evaluation criterion is used for evaluating the performance of user search goal inference. Experimental results will show the effectiveness of our proposed approach by comparing the user click through logs from a commercial search.

## REFERENCES

[1] R. Baeza-Yates and B.Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data," Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[7] 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.