IJASRM

International Journal of Advanced Scientific Research and Management, Volume 3 Issue 10, Oct 2018

www.ijasrm.com

ISSN 2455-6378

Enhanced Feedback Sessions Based Data Search Goals using Soft Clustering Technique

T. Thamodaren

Senior Librarian, Dept of Library and Information Science, Thanthai Periyar EVR Govt. Polytechnic College,
Vellore – 2, Tamil Nadu, India

Abstract

A data search structure is any data that allows the efficient retrieval of specific items from a set of items, such as a specific record from a database. For broad categories, informational queries, navigational queries and transactional queries based on different users may have different search goals when they submit it to a search engine. The assumed and analysis of user's search goal can be used to develop search engine quality and user experience. In this paper, the proposed method of assumption of user search and analysing search engine queries. First, form a framework to detect different user's search goal for a queries by clustering the proposed enhanced feedback sessions. Feedback sessions are constructed from user click through log register and can efficiently reflect the information needs of users. Second, implementing the generate pseudo documents to better represent the feedback sessions for soft clustering. Finally, applied a new criterion "Classified Average Precision (CAP)" to enhance the performance of assumption of user's search goal. Examined results are presented using user's click through log register from a commercial search engine to validate the effectiveness of the proposed methods.

Keywords: *User's data search goal, enhanced feedback sessions, pseudo documents, restructuring search results, and classified average precision*

1. Introduction

Accurate measuring the semantic similarity between words is an important issue in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation; require the ability to accurate measure the semantic similarity between concepts or entities. In information retrieval, one of the main issues is to retrieve a set of documents that is semantically related to a given user's queries [Huang, Chien & Oyang, 2003]. Efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation (WSD), textual entailment, and automatic text summarization.

Semantically related words of a particular word are listed in manually created general purpose lexical ontology's such as WordNet (lexical database). In WordNet (lexical database), a synset contains a set of synonymous words for a particular sense of a word. However, semantic similarity between entities made changes overtime and across the domains. For example, apple a word is frequently used with computers on the web. In this sense of apple is not listed in most general purpose word finder. A user, who searches for apple on the web, might be interested in this sense of apple computer based and not as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining ontologies to capture these new words and senses is costly if possible.

The proposed an automatic method to estimate the semantic similarity between words or entities used in web search engines. Because of the vastly numerous documents and the highly growth rate of the web, it is the time consuming to analyze each document individually. Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful content of source provided by most web search engines. Page count of queries is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the queried word might appear many times on one page.

In this paper, the aim to detecting the number of diverse user's, search goal for a query and

depicting each goal with some keywords automatically. First, proposed the novel method of user's search goal for queries by soft clustering feedback session. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click through log register. Then, the proposed a novel optimization method to map feedback sessions to pseudo documents which can efficiently reflect user information needs. At last, the soft cluster these pseudo documents to assumption user's search goal and find them with some keywords [Cao, Jiang, Pei, He, Liao, Chen & Li, 2008].

Since the evaluation of clustering is also an important issue, Also proposed, a novel evaluation criterion Classified Average Precision (CAP) to evaluate the performance of the restructured web search results. We also demonstrate that the proposed evaluation criterion can help us to optimize the parameter in the clustering method when assuming user's search goal.

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

1. Form a framework for assumption different user's search goal for queries by soft clustering feedback sessions. The demonstrated that the soft clustering feedback session is more efficient than soft clustering without feedback session search results or clicked URLs directly. Moreover, the distributions of different user's search goal can be obtained conveniently after feedback sessions are clustered.

2. Implement the 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's. Thus, it can tell what the user's search goals are in detail.

3. A new criterion Classified Average Precision (CAP) to evaluate the performance of user search goal assumption based on restructuring web search results. Thus, it will be determine the number of user's search goal for a query.

2. Architecture Framework

The framework of the approach consists of two parts divided by the dashed line. In the main part, all the feedback sessions of a query are first extracted from user's click through log register and mapped to pseudo documents and depicted with some keywords. Since it's not know the exact number of user's search goals in advance, many different values are tried and the optimal value will be determined by the feedback from the last part.

In the last part, the original search results are restructured based on the user's search goals assumption from the main part. Then, it evaluates the performance of restructuring search results by the proposed evaluation criterion Classified Average Precision (CAP). And the evaluation result will be used as the feedback to select the optimal number of user's search goals in the main part. Fig. 1 shows General Architecture

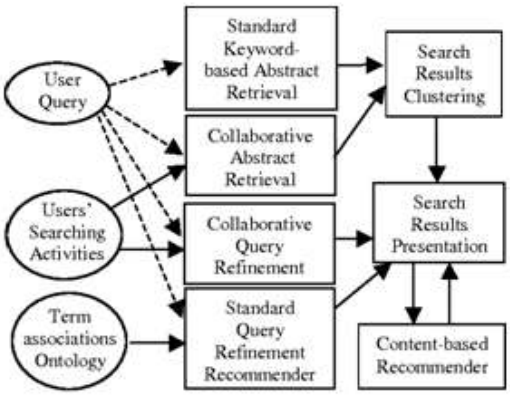

Fig.1 General Architecture

3. Demonstration of Feedback Sessions

a) Ambiguous Query

Queries are submitted to search engines to represent the information needs of user's. However, sometimes queries may not exactly represent user's 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. For example, when the query "the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. An Ambiguous query user's click through log register shown in Fig. 2

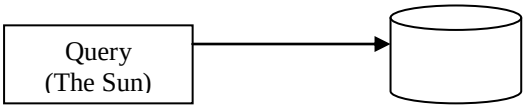

Fig. 2 An Ambiguous Query User click through log register

b) Restructure web search results

In need to restructure web search results according to user's search goals by combine the search results with the same search goal user's with different

search goals can easily obtain what they want. User's search goals represented by some keywords can be utilized in queries recommended. The distributions of user's search goals can be useful in applications such as re-ranking web search results that contain many user search goals. Due to be enhancing, many works about user search goals analysis have been investigated. They can be summarized into three classes: query classification, search result reorganization, and session boundary detection.

c) **Feedback Sessions**

The feedback session consists of both clicked and unclicked URLs and ends with the last in a single session. It is informed 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's feedback. Feedback session can tell what a user required and what he/she does not care about. Hence there are many variety feedbacks sessions in user click through log register [Yates, Hurtado & Mendoza, 2004]. Therefore, for assumption user's search goal, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly. Fig. 3 shows A User's Feedback Management system

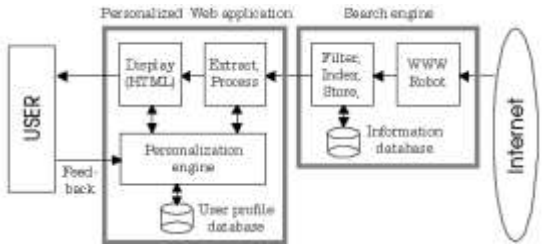

Fig. 3 A User Feedback Management system

d) Pseudo document

In this paper, in need to map feedback session to pseudo documents user search goals. The building block of a pseudo document has two steps.

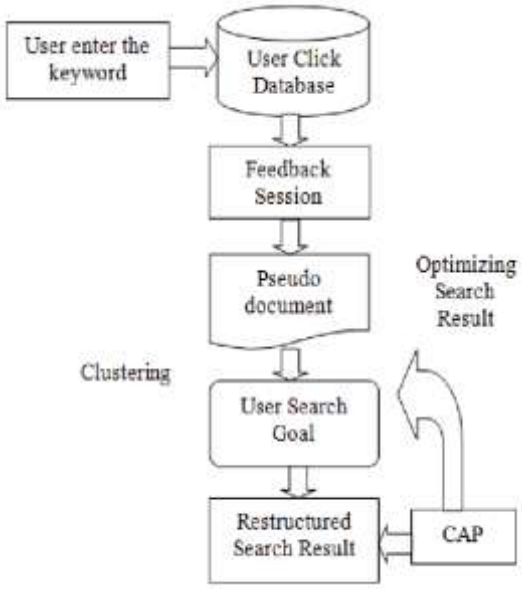

Fig. 4 Methodology of Pseudo document

First provide the URLs in the feedback session. URL in a feedback session is represented by a small text paragraph that consists of its title and snippet. Then, some textual process is implementing to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Second to form the pseudo document based URL representations. In order to produce the feature represent in the feedback session. The proposed an optimization method to combine both clicked and un-clicked URLs in the feedback session. Fig. 4 shows the methodology of pseudo document

e) User Search Goals

The cluster pseudo documents by Algorithm Fuzzy c-means (FCM) clustering which is simple and effective [Sangeetha & Nalini, 2015]. Since its not know the exact number of user search goals for each queries, set number of clusters to be five different values and perform clustering based on these five values, respectively. After clustering all the pseudo documents, each cluster can be considered as one user search goal. The center point of a cluster is computed as the average of the vectors of all the pseudo documents in the available cluster.

4. Soft clustering

1) A Soft self constructing algorithm (Data Mining Process)

Feature clustering is a powerful method to reduce the dimensionality of feature vectors for text

classification. In this paper, the proposed an Soft Similarity based self constructing algorithm for feature clustering. The words in the feature vector of a document set are grouped into clusters, based on similarity test. Words are similar to each other which are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. When all the words have been fed in, a desired number of clusters are formed automatically. Then it has one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words available in the cluster [Thamer Jawad & Ahmed Iraq, 2016].

By this algorithm, the derived membership functions match near with and describe properly the real distribution of the training data. Besides, the user need not specify the number of extracted features in advance, and trial and error methodology for determining the appropriate number of extracted features can then be avoided [Gupta, Goyal & Oberoi, 2012].

An experimental result shows that the proposed method can run faster and obtain better extracted features than existing methods. Soft clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" simple but "Soft" in the same sense as Soft logic registers.

2) Explanation of clustering

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as possible as similar, and items in different classes are as dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes, where the similarity measure controls how the clusters are formed. For examples the measures that the clustering includes distance, connectivity, and intensity.

In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In soft clustering, data elements can belong to more than one cluster, and combined with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Soft clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters [Beeferman & Berger, 2000].

One of the most widely used Soft clustering algorithms is the Fuzzy C-Means (FCM) Algorithm. The FCM algorithm attempts to partition a finite collection of n elements into a collection of c Soft clusters with respect to some given criterion. Given a finite set of data, the algorithm returns a list of c cluster centres and a partition matrix, where each element uij tells the degree to which element xi belongs to cluster cj. Like the k means algorithm, the FCM aims to minimize an objective function. The standard function is:

$$u_k = \frac{1}{\sum_j \left(\frac{d(center_k,x)}{d(center_j,x)}\right)^{2/(m-1)}}$$

which differs from the k means objective function by the addition of the membership values uij and the fuzzifier m. The fuzzifier m determines the level of cluster fuzziness. A large value (m) result shows in smaller membership (uij) and hence, fuzzier clusters. In the limit m = 1, the memberships uij converge to 0 or 1, which implies a crisp partitioning. In the absence of experimentation or domain knowledge, m is commonly set to 2. The basic FCM Algorithm, given n data points (x1, X2 ... Xn) to be clustered, a number of c clusters with (c1, c2...cc) the center of the clusters, and m the level of cluster fuzziness.

3) Soft c-means clustering

In soft clustering, every point has a degree of belonging to clusters, as in Soft logic, rather than belonging to just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different Soft clustering algorithms is available.

Any point x has a set of coefficients giving the degree of being in the kth cluster uk(x). With Soft c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster. The degree of belonging, uk(x), is related inversely to the distance from x to the cluster center as determined on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The soft c-means algorithm is simply similar to the k means algorithm

- Choose a number of clusters.
- Defined randomly to each point coefficients for being in the soft clusters.

- Continued until the algorithm has converged (that is the coefficients' changes between two iterations are no more than, the given sensitivity threshold)
- Compute the centred for each cluster, using the formula mentioned.
- For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra cluster variance as well, but has the same issues as k means; the minimum is a local minimum, and the results depend on the origin choice of weights. Using a mixture of Gaussians along with the expectation maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Soft C-means is Soft K-means. Soft c-means has been a most important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise [Bhuvaneswari, Muneeswaran & Sakthi Priya, 2018].

5. Associated work

In present years, many works have been done to assumption then so called user goals or intents of queries. But actually, their works belong to query classification. Some works analyze the search results returned by the search engine directly to exploit various query aspects. However, query aspects without user's feedback have limitations to increase search engine quality. Some works take user's feedback into account and analyze the different clicked URLs of queries in user's click through log register directly; the number of various clicked URLs of a query may be not big enough to get ideal results. However, their method does not work if proposed method tries to detect user search goals of one single query in the query cluster preferably than a cluster of similar queries. However, their method only identifies whether a pair of queries belongs to the same goal or mission and does not care what the goal is in detail. A prior utilization of user click through log register is to obtain user's implicit feedback to enlarge training data when learning ranking functions in information retrieval. In this work, consider the feedback sessions as user's implicit feedback method and proposed a novel optimization method to combine both clicked and unclicked URLs in feedback sessions to find out what user's really required and what they do not care. One application of user search goals is restructuring web search results. There are also some related works focusing on organizing the search results. In this paper, the assumption user search goals from user click through log register and restructure the search results according to the assumed user search goals.

6. Conclusion

In this paper, a novel approach has been proposed to assumption user search goals for a query by clustering its feedback sessions represented by pseudo documents. First, introduce feedback sessions to be analyzed to assumption user's search goals preferably than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user's implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, map the feedback sessions to pseudo documents to approximate goal texts in user's minds. The pseudo documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo documents, user's search goals can then be detected and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal assumption. Experimental results on user click through log register from a commercial search engine demonstrate the effectiveness of the proposed methods.

Acknowledgement

I sincerely thank to all the faculty members in the Dept of Library and Information Science, Thanthai Periyar EVR Govt. Polytechnic College for who provided insight and expertise that greatly motivated to do the research, of this paper.

References

[1]. A. Sangeetha and C. Nalini, "Improved Algorithm for Inferring User Search Goals with Feedback Sessions", International Journal of Research in Computer Applications and Robotics, Vol 3 (Issue 2): Page No 111 – 118, (2015).
[2]. Aqeel Thamer Jawad, Shaymaa Taha Ahmed Iraq, "Using Feedback Sessions for Inferring User Search Goals", International Journal of Computer Science and Mobile Computing, Vol 5 (Issue 5): Page No 697 – 702, (2016).
[3]. Bhupesh Gupta, Sandip Kumar Goyal, Ashish Oberoi, "A Review on Query

Clustering Algorithms for Search Engine Optimization", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 2 (Issue 2): Page No 1 – 8, (2012).

[4]. C.K Huang, L.F Chien, and Y.J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Log register," J. Am. Soc. for Information Science and Technology, Vol 54 (Issue 7): Page No 638-649, (2003).

[5]. 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, Vol KDD '00: Page No 407-416, (2000).

[6]. 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 Vol KDD '08: Page No 875-883, (2008).

[7]. M. S. Bhuvaneswari, K. Muneeswaran, K. S. Sakthi Priya, "Fuzzy Clustering of Augmented Web User Sessions", International Journal of Pure and Applied Mathematics, Vol 118 (Issue 20): Page No 1153-1161, (2018).

[8]. R. Baeza Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Log register in Search Engines," Proc. Int'l Conf. Current Trends in Database Technology Vol EDBT '04: Page No 588-596, (2004).