Singaporean Journal of Scientific Research

International Journal of Engineering (IJE)
Singaporean Journal of Scientific Research(SJSR)
Vol6.No.22014Pp.102-105
availableat:www.iaaet.org/sjsr
PaperReceived:10-03-2014
PaperAccepted:20-03-2014
PaperReviewedby:1Prof.Dr.Ponnambalam2. ChaiChengYue
Editor:Dr. BinodKumar

A KNOWLEDGE BASED APPROACH FOR KEYWORD SEARCH USING DATA MINING

Sasirekha.R and Lynsha Helena Pratheeba. H. P
Department of CSE
Magna College of Engineering
Chennai

ABSTRACT

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.

Index Terms—User search goals, feedback sessions, pseudo-documents, restructuring search results, classified average precision

1. INTRODUCTION

Accurately measuring the semantic similarity between words is an important problem 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 accurately measure the semantic similarity between concepts or entities. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. 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 ontologies such as WordNet. In WordNet, a synset contains a set of synonymous words for a particular sense of a word. However, semantic similarity between entities changes overtime and across domains. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries. A user who searches for apple on the web, might be interested in this sense of apple and not apple 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 not impossible.We propose an automatic method to estimate the semantic similarity between words or entities using web search engines. Because of the vastly numerous documents and the high growth rate of the web, it is time consuming to analyze each document separately.

Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines. Page count of a query 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.

2. PROPOSED CAP

A. Ambiguous Query

Queries are submitted to search engines to represent the information needs of users. 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. 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.

B. Restructure Web Search Results

We need to restructure web search results according to user search goals by grouping the search results with the same search goal users with different search goals can easily find what they want. User search goals represented by some keywords can be utilized in query recommendation. The distributions of user search goals can also be useful in applications such as reranking web search results that contain different user search goals. Due to its usefulness, 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 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.

D. Pseudo Document

In this paper, we need to map feedback session to pseudo documents User Search goals. The building of a pseudo-document includes two steps. One is representing 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 processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Another one is Forming pseudo-document 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.

E. USER SEARCH GOALS

We cluster pseudo-documents by FCM clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we 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 cluster.

3. ALGORITHM DETAILS

FUZZY CLUSTERING

- **A fuzzy 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, we propose a fuzzy 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 that are similar to each other 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. We then have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster.

By this algorithm, the derived membership functions match closely 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 for determining the appropriate number of extracted features can then be avoided. Experimental results show that our method can run faster and obtain better extracted features than other methods.

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic.

➢ Explanation ofclustering

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, 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. Some examples of measures that can be used as in clustering include distance, connectivity, and intensity.In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster. In fuzzy clustering (also referred to as soft clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels.

These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters.

One of the most widely used fuzzy 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 fuzzy 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 wij 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:

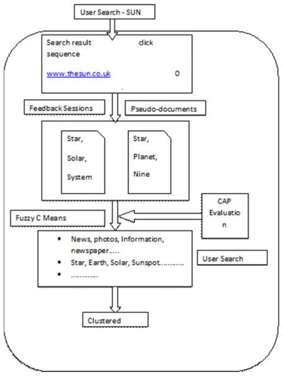
$$w_k(x) = \frac{1}{\sum_j \left(\frac{d(\mathrm{center}_k, x)}{d(\mathrm{center}_j, x)}\right)^{2/(m-1)}}.$$

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 m results in smaller memberships wij and hence, fuzzier clusters. In the limit m = 1, the memberships wij 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, . . ., xn) to be clustered, a number of c clusters with (c1, . ., cc) the center of the clusters, and m the level of cluster fuzziness with,

➢ Fuzzy c-means clustering

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely 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 fuzzy clustering algorithms is available.

4. OVERALL DESIGN



5. CONCLUSION

In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo documents. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the un clicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. 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. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click through logs from a commercial search engine demonstrate the effectiveness of our proposed methods.

REFERENCES

[1]. DanushkaBollegala, Yutaka Matsuo, and Mitsuru Ishizuka ”*A Web Search Engine-Based Approach to Measure Semantic Similarity between Words*” Knowledge and Data Engineering, IEEE Transactions on (Volume:23 , Issue: 7)July 2011.

[2]. YuzhongQu and Gong Cheng “*Falcons Concept Search: A Practical Search Engine for Web Ontologies*”Systems and Humans, IEEE

Transactions on (Volume:41 , Issue: 4)July 2011.

[3]. John B. Killoran "*How to Use Search Engine Optimization Techniques to Increase Website Visibility*" Professional Communication, IEEE Transactions. (Volume:56, Issue:1) March 2013.

[4]. AthanasiosPapagelis and Christos Zaroliagis "*A Collaborative Decentralized Approach to Web Search"*Systems and Humans, IEEE Transactions on (Volume: 42 , Issue: 5) Sep 2012.

[5]. Peng-Yeng Yin, Birbhanu, Fellow, Kuang-Cheng Chang, And Anlei Dong "*Long-Term Cross-Session Relevance Feedback Using Virtual Features"*Knowledge and Data Engineering, IEEE Transactions on (Volume: 20 , Issue: 3) March 2008.

[6]. Zheng Lu, Xiaokang Yang, Weiyao Lin, " *A New Algorithm for Inferring User Search Goals with Feedback Sessions"*Knowledge and Data Engineering, IEEE Transactions(Volume:25, Issue: 3) March 2013.