International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 3 Issue: 2 10-Mar-2014,ISSN_NO: 2321-3337

Classified Average Precision for Retrieving Information using Feedback sessions

Bhuvaneswari.S[1], Shobana.P[2] ,Vaishnavi.V[3],Ramya.P[4]
Final Year, M.E, Department of CSE, Arunai engineering college, Tiruvannamalai,

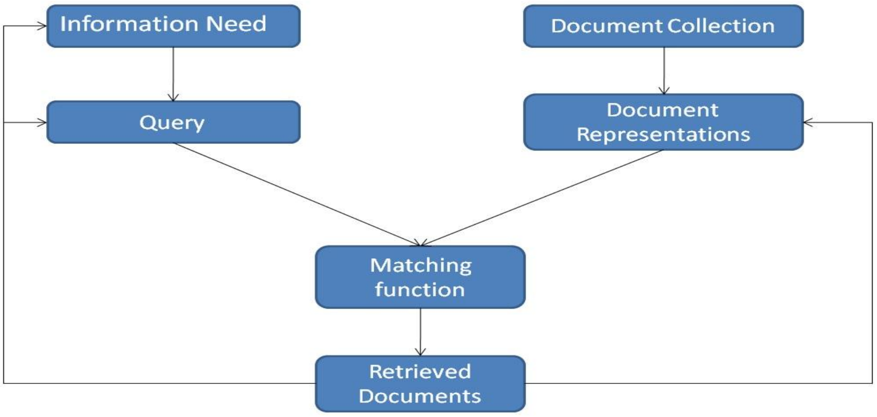
ABSTRACT*—Information Surfing is one of the vital phenomenon in today's world. Users prefer to surf internet by their queries to clarify their known uncertain information. Search engines do not often bring the user required information and does not fulfill the request completely. Hence it is necessary to infer and mine user specific interest about a topic. Using Internet the user collects the required information through the search engine. To provide the best result by the internet, the user search goal has to be analyzed. The feedback sessions are clustered to find out special user search goals for a query and the Pseudo-documents for it. The user search goals are understand using Classified Average precision (CAP) algorithm.*

Keywords— user search goals, feedback sessions, pseudo-documents, classified average precision

1,INTRODUCTION

While searching in web, the queries are submitted to search engines to symbolize the information of the users. Sometimes the queries are not precisely represent the users' specific information needs because many confusing queries are covered for a broad topic and different users want to get the information on the different aspects while submitting the 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. Therefore, it is necessary and possible to capture different user search goals in information retrieval. We 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.

Information retrieval is the activity of obtaining resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text (or other content-based) indexing. The meaning of the term information retrieval (IR) can be quite broad. Every time you look up information to get a task done could be considered as IR. E.g. getting a credit card out of a wallet to type in the card number From an academic point of view the following definition is more useful: Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text)that satisfies an information need from within large collections (usually stored on computers). Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.



INFORMATION RETRIEVAL

The final goal of an IR system can be described as the representation, storage, organization of, and access to information items. This section provides a brief description of the different resources, components and tasks involved in an information retrieval system. A global, abstract view of these elements is displayed. This overview of the IR process aims to introduce the main components that are developed in our semantic retrieval model.

2,CLASSIFICATION, PREDICTION AND CLUSTERING

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other cluster. Dissimilarities are assessed based on the attribute values describing the objects, often, distance measures are used. In this paper we use k-means clustering technique for constructing pseudo documents. K-means clustering is a centroid based technique. Classification and prediction are two forms of data analysis that be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large whereas classification predicts categorical labels, prediction models continuous-valued functions. It uses the preprocessing technique such as data cleaning, relevance analysis, data transformation and reduction. It provides the accuracy, scalability, robustness, speed and interpretability. Some of the related approaches:

- Agglomerative clustering of a search engine query log
- Bringing order to the web: automatically categorizing Search results
- Context-aware query suggestion by mining click-through And session data
- Query recommendation using query logs in search engines Problem definition
- Relevant term suggestion in interactive web search based on Contextual information in query session logs

We cluster pseudo-documents by K-means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set K to be five different values (i.e., $1; 2; \ldots ; 5$) and perform clustering based on these five values, respectively. The optimal value will be determined through the evaluation criterion.

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, it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. We will introduce how to restructure web search results by inferred user search goals at first. Then, the evaluation based on restructuring web search results

3,INFERRING USER SEARCH GOALS BY CLUSTERING PSEUDO-DOCUMENTS

In recent years, many works have been done to infer the so- called user goals or intents of a query. But in fact, their works belong to query classification. Some works analyze the search results returned by the search engine directly to exploit different query aspects. However, query aspects without user feedback have limitations to improve search engine relevance. Some works take user feedback into account and analyze the different clicked URLs of a query in user click-through logs directly, nevertheless the number of different clicked URLs of a query may be not big enough to get ideal results. With the proposed pseudo-documents, we can infer user search goals. In this section, we will describe how to infer user search goals and depict them with some meaningful keywords. Each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is Ffs.

3.1 User search goals

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. Some advantages are we can restructure web search results according to user search goals by grouping the search results with the same search goal.

Thus, users with different search goals can easily find what they want. Second, user search goals represented by some keywords can be utilized in 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.

3.2 Pseudo-documents:

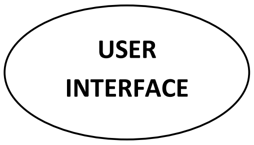
In this paper, we propose a novel way to map feedback sessions to pseudo-documents. For a query, users will usually have some vague keywords representing their interests in their minds. They use these keywords to determine whether a document can satisfy their needs. However, although goal texts can reflect user information needs, they are latent and not expressed explicitly. Therefore, we introduce pseudo-documents as surrogates to approximate goal texts. Thus, pseudo-documents can be used to infer user search goals.

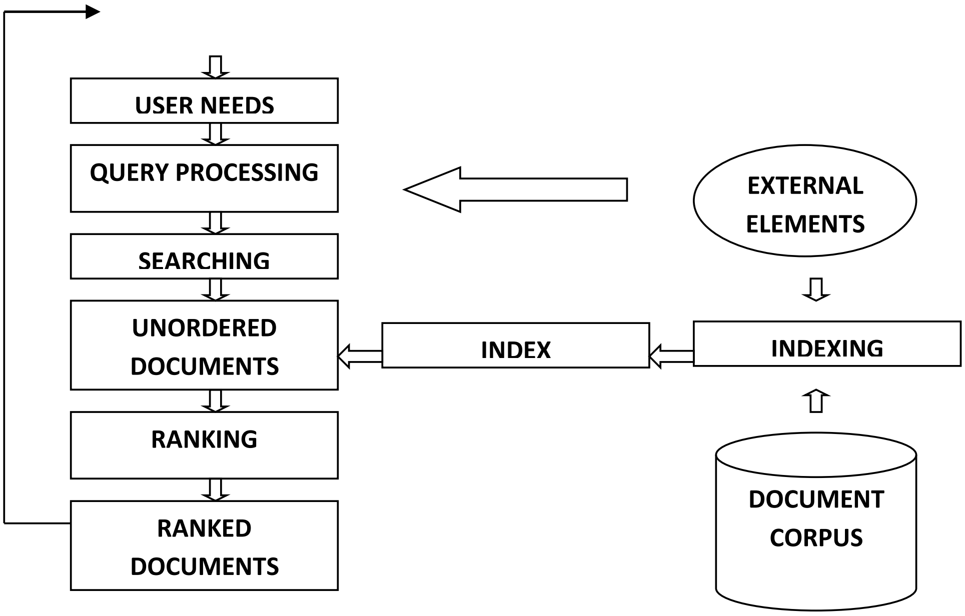
3.3 Restructuring search results:

Risk and VAP are used to evaluate the performance of restructuring search results together. Each point represents the average Risk and VAP of a query. If the search results of a query are restructured properly, Risk should be small and VAP should be high and the point should tend to be at the top left corner. We can see that the points of our method are closer to the top left corner comparatively. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP.

3.4 Classified average precision:

We also propose 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 inferring user search goals. we propose this novel criterion “Classified Average Precision” to evaluate the restructure results. Based on the proposed criterion, we also describe the method to select the best cluster number.





The Information Retrieval process

4, SYSTEM ANALYSIS

4.1 Existing system:

In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent user 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 user wants to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun.

4.2 Proposed system:

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. The approach to generate pseudo-documents to better represent the feedback sessions for clustering.
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.

5, SEMANTIC SEARCH

Any IR system is based on a logic representation of user information needs, and the information supplied by the information objects in the search space, in such a way that the comparison between queries and potential answers takes place in the ideal model. The various logic representations proposed in the area respond, on the one hand, to the requirement of being efficiently process able by an IR system, and necessarily entail some information loss. This is clear, for instance, in the representation of information needs by a simple list of keywords, as is the case in currently dominant technology in both research and industry. Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable dataspace, whether on the Web or within a closed system, to generate more relevant results. Author Seth Grimes lists "11 approaches that join semantics to search", and Hildebrand et al. provide an overview that lists semantic search systems and identifies other uses of semantics in the search process. Semantic search systems consider various points including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results. Major web search engines like Google and Bing incorporate some elements of semantic search.

6, MATHEMATICAL BASIS

Set-theoretic models represent documents as sets of words or phrases. Similarities are usually derived from set-theoretic operations on those sets. Common models are:

- Standard Boolean model
- Extended Boolean model
- Fuzzy retrieval

Algebraic models represent documents and queries usually as vectors, matrices, or tuples. The similarity of the query vector and document vector is represented as a scalar value.

- Vector space model
- Generalized vector space model
- (Enhanced) Topic-based Vector Space Model
- Extended Boolean model
- Latent semantic indexing aka latent semantic analysis

Probabilistic models treat the process of document retrieval as a probabilistic inference. Similarities are computed as probabilities that a document is relevant for a given query. Probabilistic theorems like the Bayes' theorem are often used in these models.

7, ANALYZE THE ADVANTAGES OF CLUSTERING FEEDBACK SESSIONS

Clustering feedback sessions namely pseudo- documents is better than the other two methods when inferring user search goals. With the introduction of feedback sessions, we will have a lot of advantages. Some advantages are summarized as follows:

1) Feedback sessions can be considered as a process of resampling.

If we view the original URLs in the search results as original samples, then feedback sessions can be viewed as the "processed" or "resampled" samples which differ from the original samples and reflect user informa- tion needs. Without resampling, there could be many noisy URLs in the search results, which are seldom clicked by users. If we cluster the search results with these noisy ones, the performance of clustering will degrade greatly. However, feedback sessions actually "resample" the URLs and exclude those noisy ones. Therefore, our method is much better than resampling.

2) *Feedback session is also a meaningful combination of several URLs*

A framework to infer different user search goals for a query by clustering feedback sessions. Clustering of 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. Feedback sessions are introduced that is to be analyzed to infer user search goals rather than search results or clicked URLs. User search goals are inferred for a query by clustering its feedback sessions represented by pseudo-documents. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. A new criterion "Classified Average Precision (CAP)" is used 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.

Therefore, it can reflect user information need more precisely and there are plenty of feedback sessions to be analyzed. the solid points represent the clicked URLs mapped into a 2D space and we suppose that users have two search goals: the star points belong to one goal and the circle points belong to the other goal represents a feedback session which is the combination of several clicked URLs. Since the number of the different clicked URLs may be small, if we perform clustering.

VIII CONCLUSION

We suggested 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 unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Second, we map feedback sessions to pseudo documents to approximate goal texts in user minds. our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online.

REFERENCES

[1] 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.

[2] T. Joachims, “Evaluating Retrieval Performance Using Click- through Data,” *Text Mining,* J. Franke, G. Nakhaeizadeh, and Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[3] 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.

[4] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, “Accurately Interpreting Clickthrough Data as Implicit Feed- back,” *Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’05),* pp. 154-161, 2005.

[5] R. Jones and K.L. Klinkner, “Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs,” *Proc. 17th ACM Conf. Information and Knowledge Manage- ment (CIKM ’08),* pp. 699-708, 2008.

[6] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating Query Substitutions,” *Proc. 15th Int’l Conf. World Wide Web (WWW ’06),* pp. 387-396, 2006.

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

[8] 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.

[9] 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.

[10] 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.