Inferring user Search Goals with Feedback Sessions using STC

Asha P [1], Ambily Balaram [2]

[1] Computer Science and Engineering Department, Calicut University, India

Abstract: ***The information explosion on the internet makes high demand on search engine. But the existing search engines are not satisfying the user's needs. Most of the search engines returns irrelevant document in response to the user search query. Sometimes the search query may be imperfect description. For such imperfect query the search engine cannot retrieve relevant document to the user's needs inferring user search goal can improve search engine relevance and user search experience. In this paper propose a new method to infer user search goal by analysing search engine query logs. For better performance construct a feedback session from click through logs of search engine and map these sessions in to corresponding pseudo documents for clustering. By using clusters organize the search results according to the user relevance. By using CAP algorithm optimize the number of clusters and evaluate the performance of inferring user search goal.***

Keywords: User search goals, feedback session, pseudo documents, suffix tree, classified average precision

1. Introduction

As the web contents grow, it makes difficult to manage and classify its information. The huge collection of information on the internet has placed high demands on search engines. Most of the existing search engine does not satisfy the user's needs. Most often it returns thousands of documents in response to the users search query. Many of the returned documents are irrelevant.

Queries are represents information needs of the users in web search applications. Sometimes these queries are not exactly represents their information needs. Since the queries have more than one meaning so this is the main reason for the retrieval of irrelevant documents in search engine application. This is a complicated problem which effect the search engine relevance and users search experience. The inference and analysis of information needs of the user have lot of advantages in improving search engine relevance and users search experience. Due to the usefulness, so many works have been done to infer the user search goals. They can be classified in to three 1.query classification 2.search result reorganization 3.session boundary detection.

In the first class, infer user goals by performing Query classification by using some predefined specific classes. In second class people reorganize the search results by using click through data. In third class people try to detecting session boundaries. In this method only identifies whether a pair of queries belongs to the same goal or mission. It does not care about what the goal is. In the previous methods generally cluster top hundred search results or directly cluster different clicked URLs. These methods are not efficient.

In this paper we propose an efficient method .Based on the web log records of search engine construct a feedback session consist of clicked URLs and click sequence.Based on these Sessions generate corresponding pseudo documents with some keywords used in the URLs. Then cluster these pseudo documents.These clusters organize the result documents. By using performance evaluation method, optimize the number of clusters.

2. Related works

By inferring search goal of a user can improve the quality of a search engine. The previous studies have mainly focused on using manual query log investigation for inferring user search goal so Uichin Lee,Zhenyu Liu [2]describe whether and how can automate this goal identification process. For this purpose, here propose two types of features. Past user-click behavior and anchor-link distribution.

A)Past User-Click Behavior

In this feature user's goal for a given query may be learned from how users in the past have interacted with the returned results for the same query. For the navigational query the past users should have mostly clicked on a single website.For the informational query the users should have clicked on many results related to the query. The click distribution defines how frequently users click on various answers Given a query, its click distribution is constructed as follows: First sort the answers to the query in the descending order of the number of clicks they receive from all users. Then create a histogram where the ith bin corresponds to the number of clicks accumulated on the ith answer. In third step normalize the frequency values so that these values add up to 1. The goal for that query by investigating how that clicks distribution is skewed toward rank one. Main practical issue in using the user-click behavior is that a search engine needs to accumulate enough user clicks for a given query when it clicks multiple times.

B) Anchor-Link Distribution

The anchor is a piece of text surrounded by a pair of <A HREF..> tags in an HTML page. In this method use a notion anchor link distribution. Anchor link distribution for a query is computed as follows: Locate all the anchors appearing on the Web that have the same text as the query, extract their destination URL's. Then Count how many times each destination URL appears in this list and sort the destinations in the descending order. Then create a histogram where the frequency count in the ith bin is the number of times that the ith destination appears. Finally normalize the frequency in each bin so that all the frequency values add up to 1.Link spam and mirror sites are the main problem in this method. It distort anchor link distribution and introduce undesirable noises.

In response to a users search request, most of the search engines return a ranked list of web pages or web documents. The returned list consists of Web pages on different topics or different aspects of the same topic are mixed together. Hao Chen and Susan Dumais[3]propose a system with two components. A text classifier that categorizes web pages on-the-fly. A user interface that presents the web pages within the category structure It will help the user to manipulate the structured view .In this method a statistical text classification model is trained offline on a representative sample of Web pages with known category labels.At query time the new search results are quickly classified on-the-y into the learned category structure. The user interface compactly displays web pages in a hierarchical category structure.Some heuristics methods are used to order these categories. A Support Vector Machine (SVM) algorithm was used as the text classifier. It maximizes the margin between the two classes so this method is fast and efficient.

Organizing Web search results into clusters helps users to fast browsing of search results. But the traditional clustering techniques are inadequate. So Hua Jun and Qi Cai propose a method [4]renormalize the search result clustering problem as a salient phrase ranking problem. Here first extracts and ranks salient phrases as candidate cluster names, based on a regression model learned from human labeled training data. In second step the documents are assigned to relevant salient phrases to form candidate clusters. Then final clusters are generated by merging these candidate clusters. The algorithm is composed of the four steps:

- Search result fetching.
- Document parsing and phrase property calculation.
- Salient phrase ranking.
- Post-processing.

1) An HTML parser is analyzed the web pages and extract all possible phrases from the contents.These phrases are become the candidate clusters name.
2) The properties for each phrase such as phrase frequencies, document frequencies, phrase length, etc. are calculates during parsing .The higher probability that salient phrases occur in titles. Apply stemming to each word using Porter's algorithm
3) By using a regression model which is learned from previous training data, to combine these properties into a single salience score. The salience phrases are then ranked by the score in descending order. After salient phrases are ranked, the corresponding document lists constitute the candidate clusters. The salient phrases being cluster names.
4) In the post-processing, filter out pure stop words. Merge the clusters and phrases, to reduce duplicated clusters. This method is efficient. It generates shorter and readable cluster names, which enable users to quickly identify the topics of a specified cluster. But the main problems are the clusters do not correspond to the interesting aspects of a topic from the user's perspective. And the cluster labels generated are not informative to a user to identify the right cluster.

The organization of search results is a very important factor that can affect the utility of a search engine significantly. Clustering search results is an effective way to organize search results, which allows a user to navigate into relevant documents quickly. Generally two deficiencies of this approach make it not always work well: First the clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective. Second the cluster labels generated are not informative enough to allow a user to identify the right cluster. So Wang and Chengxiang [5] proposes a strategy for organizing the search result. First learn interesting aspects" of similar topics from search logs and organize search results based on these interesting aspects". Then generate more meaningful cluster labels using past query words entered by users. It helps to learn what specific aspects are interesting to users given the current query topic. Search engine logs are separated by sessions. Aggregate all the sessions which contain exactly the same queries together. Then apply the star clustering algorithm for clustering. It can suggest a good label for each cluster naturally. This method will improve the ranking baseline, especially when the queries are difficult or the search results are diverse. And also it can generate more meaningful aspect labels than the cluster labels generated. Query substitution means generating a new query to replace a user's original search query. The new query is strongly related to the original query.R.Jones and B.Ray propose a method [6] in which first identify a new source of data for identifying similar queries and phrases. Then define a scheme for scoring query suggestions, which can be used for other evaluations. In third step combine query increase coverage and effectiveness. But the main drawback is that it needs machine translation techniques.

The ranking technique has become a central research problem for informational retrieval and Web search. It directly influences the relevance of the search results, the quality of a search system and users search experience. For a given a query, first deploy a ranking function then it measures the relevance of each document to the query. Sorts all documents based on their relevance scores and present a list of top ranked ones to the user. But the essential problem of search technology is to design a ranking function that can best represent relevance. But a single ranking model could not used for diverse types of queries. A straightforward method is to add query difference as additional features into learning the single ranking function. This method requires high quality of both the new features and training data. So it usually does not effective in practice. Jiang and Xin li proposes a divide-and-conquer method [7] is improving the ranking relevance for all queries. First identify a set of ranking-sensitive query topics and divide the learning problem of one single ranking model for all queries into learning a set of sub-models for corresponding different query topics. Then conquer these learning problems by introducing a global loss function and exploring a unified approach to co-optimize all sub models. At testing time, select a set of ranking-sensitive query topics the new query most likely belongs to, and apply respective ranking models to ranking the documents. Then assemble these ranking results together to obtain the final rank for the new query. This whole framework is called ranking specialization.

This is a new document representation model based on implicit user feedback obtained from search engine queries. B.Poblete and B.Y Ricardo propose a model[8] to represent web documents.The main objective of this model is to achieve better results in non-supervised tasks, such as clustering and labeling. In this approach selects features using what seems more appropriate to refer to as a bag of query-sets idea. This representation is very simple and it reduces the number of features for representing the document set. It allows using all of the document features for clustering. There are two document models based on queries and clicked URLs.1.Query document model and 2.Query set document model. These models can be applied to organize documents within a website. The query document model consists of representing documents using query terms as features only. It reduces the feature space dimensions considerably. The query-set document model is an enhanced version of the query model. It uses frequent query-sets as features. It preserving the information provided by the co occurrence terms inside queries. This is achieved by mining frequent item sets or frequent query patterns. And every keyword in a query is considered as an item. Here patterns are discovered through analyzing all of the queries from which a document was clicked. The query-set model reduces the number of features needed to represent a set of documents .It improves more than 90 percentage discount the quality. And it reduces the computational cost. But the main drawback is it needs broader comparison with online directory.

The click graphs can improve query intent classification. All the previous works on query classification have primarily focused on improving feature representation of queries. X.Li and Y.Y.Wang propose a method[10] to investigate a completely orthogonal approach instead of enriching feature representation. A common challenge is that the sparseness of query features coupled with the sparseness of training data. To improve classification performance, mainly focus on an orthogonal direction to query feature enrichment. So drastically expand the training data in an automated fashion. This is achieved by leveraging click graphs, called bipartite-graph representation of click through data. The edges in a bipartite-graph are connecting between queries and URLs and are weighted by the associated click counts. This click graph contains a vast amount of user click information. It brings opportunities for semi supervised learning, which leverages both labelled and unlabeled examples in classification. Here use a principled approach that automatically labels a large amount of queries in a click graph. These queries are used in training content-based classifiers. The key idea is that manually label a small set of seed queries in a click graph. Then iteratively propagate the label information to other queries until a global equilibrium state is achieved. Finally regularize the learning with content-based classification. When a click graph is noisy such regularization would prevent a click graph from propagating erroneous labels. The main advantages are it can improve classification performance. This approach can jointly perform both graphs based and content based learning in a unified frame work. The impact of seed query is the main drawback. Query suggestion plays an important role in improving the usability of search engines. Some recently proposed methods can make meaningful query suggestions by mining query patterns from search logs, none of them are context aware. So Huanhuan and Daxin[11] proposes a novel context aware query suggestion approach by mining click-through data and session data. In the offline model learning step, to address data sparseness, queries are summarized into concepts by clustering a click-through bipartite. In the online query suggestion step, a user's search context is captured by mapping the query sequence submitted by the user to a sequence of concepts. This approach considers not only the current query but also the recent queries in the same session to provide more meaningful suggestions. And it outperforms two baselines in both coverage and quality. But this approach has larger coverage area. J.R Wen and J.Y. Nie propose a method [10] for clustering user queries of a search engine. In this method attempt to cluster similar queries according to their contents as well as user logs. The preliminary results show that the resulting clusters provide useful information for FAQ identification. In this method query clustering is based on user logs. The principles are as follows. 1) If users clicked on the same documents for different queries, then the queries are similar. 2) If a set of documents is often selected for a set of queries, then the terms in these documents are related to the terms of the queries to some extent. These principles are used in combination with the traditional approaches based on query contents (i.e. keywords).The density-based clustering method (DBSCAN) and its incremental version Incremental DBSCAN is used to satisfy meet the requirements.

3. Framework of the Work

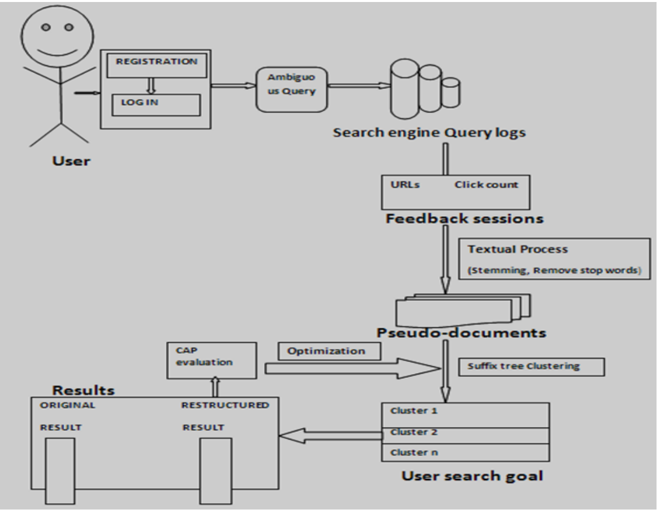

Figure 1: Framework of the approach

A Web server usually registers a log entry, or Weblog entry, for every access of a Web page. It includes the URL requested the IP address from which the request originated and a time stamp. Based on the Weblog records construct the feedback session. Because Weblog data provide information about what kind of users will access what kind of Web pages. This session consists of URLs and click sequence and it focus on user search goals. By using only a feedback session could not understand the user search goals exactly. Based on these feedback session constructs the pseudo document for analyzing the accurate result. This pseudo document consists of key words of URLs in the feedback session. This is called as enriched URLs. The enriched URLs are clustered. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have a high similarity in comparison to one another but are very dissimilar to object in other clusters. After constructing the clusters the Web search results are restructured based on the documents collection detail. In general here propose a novel approach to infer user search goals by analyzing search engine query logs. First, 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, propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Suffix Tree Clustering is used here for better clustering .Finally, by using a new criterion Classified Average Precision (CAP) evaluate the performance of inferring user search goals. Figure 1 shows the frame work of this project is shown above.

The proposed project mainly consist of 5 modules. They are:

Module 1: Crawler(Search)
In this module user enter their query in search field. Here search engine BING's API is integrated. For a new query, when user click on search button 50 original search results will obtained. Generate feedback session for the old query from user click through log.

Module 2: Stemming
In the stemming module remove endings, stopwords and connectors from retrieved snippets and construct optimal document or pseudo documents.

Module 3:Similarity
In this module calculate the frequency of each terms and similarity between the documents. SCAM algorithm is used for similarity calculation. These frequency scores are used for ordering the documents

Module 4: Clustering
The fourth module deals with clustering. Clustering is used for common snippet foundation. Here STC is used for clustering.

Module 5: Result reorganization
In this module reorganize the search results based on the evaluation criterias.

3.1 Suffix tree clustering algorithm details

Clustering as an unsupervised machine learning method, is an effective data mining technique that has been comprehensively studied and extensively applied to a variety of application areas. The Suffix Tree Clustering (STC) algorithm groups the input texts according to the identical phrases they share . The rationale behind such approach is that phrases, compared to single keywords, have greater descriptive power. This results from their ability to retain the relationships of proximity and order between words. A great advantage of STC is that phrases are used both to discover and to describe the resulting groups. The Suffix Tree Clustering algorithm works in two main phases: base cluster discovery phase and base cluster merging phase. In the first phase a generalized suffix tree of all texts' sentences is built up using words as basic elements. After all the sentences are processed, the tree nodes contain information about the documents in which particular phrases appear. Using that information documents that share the same phrase are grouped into base clusters of which only those whose score exceeds a predefined Minimal Base Cluster Score are retained. In the second phase of the algorithm, a graph representing relationships between the discovered base clusters is built based on their similarity and on the value of the Merge Threshold. Base clusters belonging to coherent subgraphs of that graph. They are merged into final clusters. A clear advantage of Suffix Tree Clustering is that it uses phrases to provide concise and meaningful descriptions of groups.

4. Conclusions

As the web contents grow it makes difficult to manage its information. So research for improving search engine relevance also important. In recent years so many works have been done to improve search engine relevance by through inferring user search goal. But these works could not provide a better result. So in this paper a new method has been proposed for inferring user search goal by through clustering the feedback session. For better performance here we use suffix tree clustering method. Then restructure the search result by using these clusters. By using CAP algorithm optimize the number of clusters

References

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin,Zhaohui Zheng,"A New Algorithmfor Inferring User Search Goals with Feedback Sessions,"IEEE Transaction On Knowledge and Data Engineering, Vol. 25, No 3,March2013

[2] U.Lee, Z. Liu, and J. Cho, "Automatic Identi_cation of User Goals in Web Search,"Proc14th Intl Conf. World Wide Web (WWW 05), pp. 391-400, 200

[3] H. Chen and S. Dumais, "Bringing Order to the Web:Automatically Categorizing SearchResults,"Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI 00), pp. 145-152, 2000.

[4] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma,"Learning to ClusterWeb Search Results,"Proc. 27th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval(SIGIR 04),pp. 210

[5] X.Wang and C.-X Zhai,"Learn fromWeb Search Logs to Organize Search Results, "Proc. 30th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR07),pp. 87-94,2007.

[6] R.Jones, B. Rey, O. Madani, W.Greiner,"Generating Query Substitutions,"Proc. 15thIntl Conf. World Wide Web (WWW 06), pp. 387-396, 2006.

[7] B. Poblete and B.-Y Ricardo,"Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents," Proc. 17th Intl Conf. WorldWide Web (WWW 08), pp. 41-50,2008.

[8] X. Li, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs,"Proc. 31st Ann. Intl ACM SIGIR Conf.ResearchDevelopment in Information Retrieval(SIGIR 08) pp. 339-346, 2008

[9] 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 Intl Conf. Knowledge Discoveryand Data Mining (SIGKDD 08), pp. 875-883, 2008.

[10] J.-R Wen, J.-Y Nie, and H.-J Zhang,"Clustering User Queries of a Search Engine "Proc.Tenth Intl Conf. World Wide Web (WWW 01),pp. 162-168, 2001.

[11] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Log," Proc. Sixth ACM SIGKDD Intl Conf .Knowledge Discovery and Data Mining (SIGKDD 00), pp. 407-416,2000

[12] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical WebQuery Classi_cation," Proc. 30th Ann. Intl ACM SIGIR Conf. Research and Development (SIGIR 07),pp. 783-784, 2007.

[13] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive WebSearch Based on Contextual Information in Query Session Logs," J. Am. Soc. for InformationScience and Technology, vol. 54, no. 7, pp. 638-649, 2003.

[14] T.Joachims, "Evaluating Retrieval Performance Using Clickthrough Data,"Text Mining, pp.79-96, Physica/Springer Verlag, 2003.

[15] T.Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining (SIGKDD 02), pp. 133-142,2002.

[16] T.Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Click through Data as Implicit Feedback,"Proc. 28th Ann. Intl ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR 05), pp. 154-161, 2005

[17] R.Jones and K.L. Klinkner, "Beyond the Session Timeout:Automatic Hierarchical Segmentation of Search Topics in QueryLogs,"Proc. 17th ACM Conf. Information a KnowledgeManagement(CIKM 08), pp. 699-708, 2008.

[18] D. Shen, J. Sun, Q. Yang, and Z. Chen, "Building Bridges forWeb Query Classi_cation,"Proc.29th Ann. Intl ACM SIGIR Conf Research and Development in Information Retrieval (SIGIR06),pp. 131-138, 2006.

[19] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine,"Proc. Tenth Intl Conf. World Wide Web (WWW 01),pp. 162-168, 2001

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

Author Profile

Asha P received B.Tech degree in computer science and engineering from KMCT College of engineering, Calicut (Calicut University).She is currently pursuing her M.Tech degree in the same college. Her research interest includes Data mining and Database management system.

Ambily Balaram received B.Tech degree in computer science and engineering from KMCT College of engineering, Calicut (Calicut University). . Her research interest includes Data mining,Web mining and Database management system.At present she is working as a Assistant professor in Department of Computer Science and Engineering, KMCT College of engineering, Calicut (Calicut University).