ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Letter Pair Similarity Classification and URL Ranking Based on Feedback Approach

P.T.Shijili[1]

P.G Student, Department of CSE, Dr.Nallini Institute of Engineering & Technology, Dharapuram, Tamilnadu, India[1]

ABSTRACT: Search engine is one of the most important applications in today's internet. For an ambiguous query/topic, different users may have different search goals, so the search engine doesn't satisfy user information needs properly on the diverse aspects upon submission of same query/topic. The examination of user search goals can be very valuable in improving search engine importance and user knowledge. A major deficiency of generic search engines is that they follow the ''one size fits all'' model and are not adaptable to individual users. Here propose a framework that enables large-scale evaluation of personalized search. User interest is employed in the clustering process to achieve personalization effect. The goal of personalized IR (information retrieval) is to return search results that better match the user intent. Then these user search goals are used to restructure and reordered the web search results by using URL ranking process and search history process.

KEYWORDS: User search goals, feedback sessions, URL ranking, search history

I. INTRODUCTION

In web based search applications, user submits the query to search engine to search efficient information. The information needs of different user may differ in various aspects of query information. This becomes difficult to achieve user information needs. Sometimes ambiguous queries may not exactly represented by users so it results in less understandable to search engine. To achieve the user specific information needs many ambiguous uncertain queries may cover a broad topic and dissimilar users may want to get information on different aspects when they submit the same query. User information need is to desire and obtain the information to satisfy the needs of each user. To satisfy the user information needs by considering the search goals with user given query For example, when user submits a query "apple" to search engine, some users are interested to know information about fruit and some users want to know information about mobile phone. Therefore, it is necessary to discover different user information search goals. User information need is to desire and obtain the information to satisfy the needs of each user. To satisfy the user information needs by considering the search goals with user given query, cluster the user information needs with different search goals. Because the interference and evaluation of user search goals with query might have a numeral of advantages in improving the search engine significance and user knowledge. So it is necessary to collect the different user goal and retrieve the efficient information on different aspects of a query. Capture different user search goals in information retrieval outcome becomes changes than the normal query based information retrieval

Evaluation and analysis of user search goals has many advantages .First Reorganize web search results according to user search goals by grouping search results with same information need. This can be useful to other users with different search goals to find easily what they want. Second, query recommendation by using user search goals depicted with some keywords. This can be helpful to other users to form their query more effective. Third, Reranking web search results according to different user search goals.

User search goal analysis is important to optimize search engine and effective query results organization. When query is submitted to search engine, the returned web pages of search results are analyzed. X. Wang and C-X. Zhai [7] learns

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

interesting aspects of similar query/topic from web search logs which consists clicked web pages URLs and organize search results accordingly. Their approach may results in limitation, as the different clicked URLs for a query/topic may be small in number. Hua-Jun Zeng et.al [8] suggested a query based search results for user goal and the rank list of documents return by a certain web search engine. But this method only produces the result with higher level of the documents only and it doesn't make the results for all search based user goals

Clustering search results is an efficient method to systematize search results, which allows a user to find the way into applicable documents quickly. In this paper, to discover different user search goals for a query and depict each search goal with some keywords automatically. To discover the user information automatically at different point of view with user given query and collects the similar search goal result with URL first we collect similar feedbacks sessions from user click through logs of different search engines. Then, map feedback sessions to pseudo-documents which reflects user information needs. At last, At last, k means clustering algorithm can be used to cluster these pseudo-documents for inferring user search goals and depicting them with some meaningful keywords. Then these search goals can be used to restructure the web search results. After that previously restructured web search result can be reorder by using the URL ranking process and u the search history of individual user.

The rest of the paper is organized as follows: Section II contains literature survey about related work. Section III contains description of the proposed system. Finally paper is concluded in the Section IV.

II. LITERATURE SURVEY

Web mining is the application of data mining technique it is used extract a knowledge from Web data. Web data is Web structure data and Web usage data. Since many years, research in web log mining has been subject of interest. Many previous works has been investigated on problem of analyzing user query logs. The information in query logs has been used in many different ways, such as to infer search query intents or user goals, to classify queries, to provide context during search, to facilitate personalization, to suggest query substitutes

Clustering search results is an effective way to organize search results which allows a user to navigate into relevant documents quickly. Generally all existing work perform clustering on a set of top ranked results to partition results into general clusters, which may contain different subtopics of the general query term. However, this clustering strategy has two deficiencies which make it not always work well. First, discovered clusters do not necessarily correspond to the interesting aspect of a topic from user-oriented perspective. Second, cluster labels are more general and not informative to identify appropriate clusters.

H. Chen and S. Dumais [2] developed a user interface that organizes web search results into hierarchical categories. Automatic text classification technique (SVM classifier) was used to classify arbitrary search results into existing category structure on-the-fly. This approach has advantage of known category labels information, for classifying new items into the category structure and to help user to quickly focus on task relevant information. A user study compared new category interface with the traditional ranked list interface of search results, which showed that category interface is superior in both subjective and objective manner

T.Joachims [6] proposed an approach to automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking. Taking support vector machine (SVM) approach, for learning ranking functions in information retrieval.

T. Joachims et al. [3] did a lot of work on examining the reliability of implicit feedback generated from clickthrough data in www search. The author proposes strategy to automatically generate training examples for learning retrieval functions from observed user behavior. The user study is intended to examine how users interrelate with the list of ranked results from the Google search engine and how their behavior can be interpreted as significance judgments. Implicit feedback can be used for evaluating quality of retrieval functions [5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

User may issue number of queries to search engine in order to achieve information need/tasks at a variety of granularities. R. Jones and K.L. Klinkner [4] proposed a method to detect search goal and mission boundaries for automatic segmenting query logs into hierarchical structure. Their method identifies whether a pair of queries belongs to the same goal or mission and does not consider search goal in detail.

III. PROPOSED SYSTEM

In this paper, user search goals are conditional by clustering these pseudo-documents and depicted with some keywords. Then evaluate the performance of restructuring search results by evaluation criterion CAP, VAP, AP and Risk

i.click through data

User clicks represent implicit relevance feedback. In this framework, user clicks are recorded in user click through data. User uses clickthrough data stored in user logs to simulate user experience in web search. In general, when query is issued, the user usually scans links to documents in a result list from first to last.

ii. Feedback session

Feedback sessions are considered as users' implicit feedback. In general, a session for web search is a sequence of consecutive queries to satisfy single information and some clicked results. The proposed feedback session consists of both clicked and unclicked URLs for a particular query in a single session and ends with last clicked URL. This shows that before last clicked URL, all the URLs are scanned and evaluated by user. In each feedback session clicked URL (visited link) tells users information need and unclicked URL (unvisited link) tells what users do not want. There are large numbers of diverse feedback sessions in user clickthrough log. So it is efficient to examine feedback sessions for inferring user search goals than to examine clicked URLs or search results directly.

iii.Generating Pseudo Documents

As URLs alone are not informative enough to tell intended meaning of a submitted query. To obtain rich information, we enrich each URL with additional text content by extracting the titles and snippets of URLs appearing in feedback session. Thus, each URL in feedback session is represented by small textual content which contains its title and snippet. Then some text preprocessing is done on those textual contents, such as transforming all letters to lowercase, eliminating stop words (frequent words) and word stemming by using porter algorithm [16]. Lastly, TF-IDF [1] vector of URL's titles and snippets are formed respectively as,

$$T_{ui} = [\, t_{w1,}\, t_{w2,} \ldots\ldots\ldots,\, t_{wn} \,]^T$$
$$S_{ui} = [\, s_{w1,}\, s_{w2,} \ldots\ldots\ldots,\, s_{wn} \,]^T \quad (1)$$

where T_{ui} and S_{ui} are TF-IDF vectors of URL's title and snippet, respectively. u_i is i^{th} URL in feedback session. W_j is the j^{th} term in the enriched URL. The t_{wj} and s_{wj} denotes j^{th} term in the URL's title and snippet respectively. Feature representation F_{ui}, of i^{th} enriched URL is weighted sum of T_{ui} and S_{ui}.

$$F_{ui} = w_t T_{ui} + w_s S_{ui} = [f_{w1,}\, f_{w2,} \ldots\ldots\ldots,\, f_{wn}\,]^T \quad (2)$$

where w_t and w_s are weights of title and snippet respectively.Each term of F_{ui}, denotes importance of term in ith URL.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

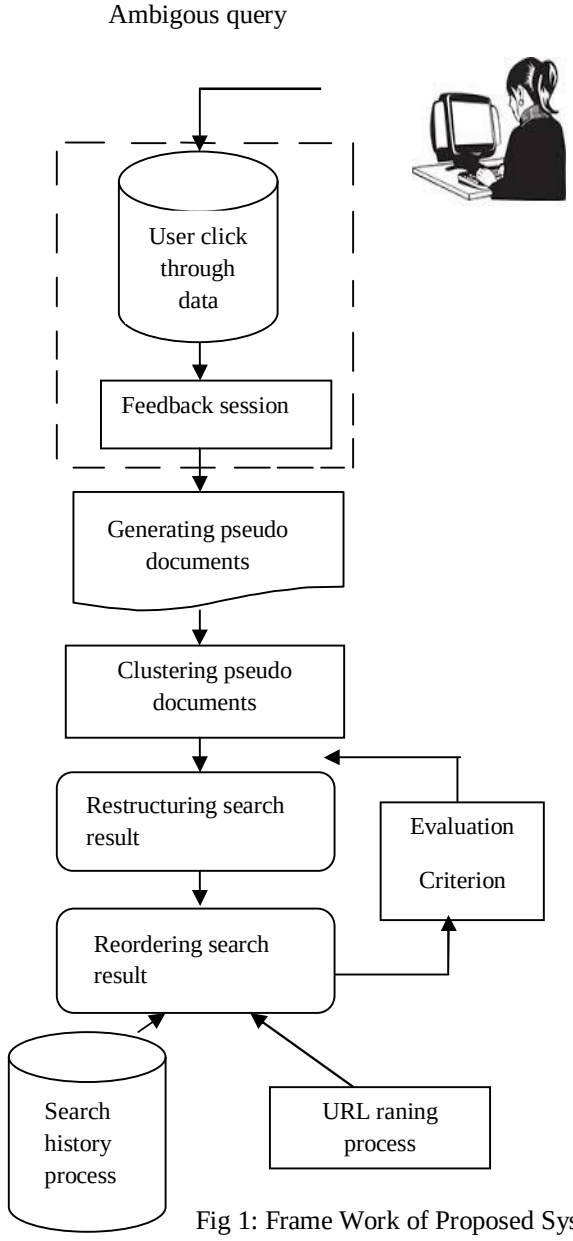

Fig 1: Frame Work of Proposed System

In order to obtain feature representation of a feedback session, optimization method is used to merge feature representations of each clicked and unclicked enriched URLs in the feedback session. Let F_{fs} be feature representation of a feedback session, F_{ucm} and F_{ucl} are feature representation of clicked and unclicked URLs respectively and $f_{fs}(w)$ is value for term w. F_{fs} should be such that sum of distance between F_{fs} and each F_{ucm} is minimized and sum of distance between F_{fs} and F_{ucl} is maximized.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

$$F_{fs} = [f_{fs}(w_1), f_{fs}(w_2), ..., f_{fs}(w_n)]^T \quad (3)$$

Each feedback F_{fs} session is represented by . This is nothing but pseudo-document which is used for discovering user intents or search goals. These pseudo-documents contain what user requires and what do not, which is used to learn interesting aspects of a query

iv Clustering Pseudo documents by K-means

The similarity between two pseudo-documents is computed as the cosine score of F_{fsi} and F_{fsj} , as follows:

$$Sim_{i,j} = \cos(F_{fsi}, F_{fsj}) = \frac{F_{fsi} \cdot F_{fsj}}{|F_{fsi}||F_{fsj}|} \quad (4)$$

And the distance between two feedback sessions is

$$Dis_{i,j} = 1 - Sim_{i,j} \quad (5)$$

Clustering 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; . . . ; 5) 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, as shown in

$$F_{center\ i} = \sum_{k=1}^{Ci} F fsk \quad [F_{fsk} \subset Cluster\ i] \quad (6)$$

where $F_{center\ i}$ is the ith cluster's center and Ci is the number of the pseudo-documents in the ith cluster. $F_{center\ i}$ is utilized to conclude the search goal of the ith cluster.The terms with the highest values in the center points are used as the keywords to depict user search goals.

v.Restructurng Web Search Result

Since search engines always return millions of search results, 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. As inferred user search goals are depicted with vectors in (6) and feature representation of each URL in search result is calculated by (1) and (2). Then categorize each URL into a cluster centered with user search goals/intents by selecting smallest distance between user search goal vectors and URL vectors.

vi.Reordering Web Search Result

Restructuring web search results is an application of inferring user search goals. Reordering of web search result is based in URL ranking process and using the search history process.Specific personalization method can be rerank relevent doccument for a user higher in result list,the user would be more satisfied.The bolded document that have been clicked by the user have been ranked.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

vii.Evaluation Criterion

The performance of restructured (clustered) web search results and original search results is evaluated by using parameters like Average Precision (AP) [1], Voted AP (VAP) which is AP of the class having more clicks, Risk to avoid wrong classification of search results and Classified AP (CAP). If user got correct classified results with higher CAP value, this value is used to optimize the no of clusters of user search goals.

1) Average precision (AP): It is calculated according to given user feedbacks. AP is the average of precisions computed at the point of each clicked document in the ranked sequence of user feedback.

$$AP= \frac{1}{N+}\sum_{r=1}^{N} rel(r)\frac{Rr}{r}$$

where N^+ is the number of clicked documents from total
retrieved documents in single user feedback session, *r* is the rank, *N* is the total number of retrieved documents, *rel()* is a binary function on the relevance of a given rank, and Rr is the number of relevant retrieved documents of rank *r* or less.
2) Voted AP (VAP): It is calculated for restructured search results classes i.e. different clustered results classes.It is same as AP and calculated for class which having more clicks i.e. the class user interested in.

$$VAP= \frac{1}{NC}\sum_{r=1}^{N} rel(r)\frac{Rr}{r}$$

where *NC* is the number of clicked documents from the
class having maximum number of clicks
3) Risk: Sometimes VAP will always be highest value because each URL from single session is classified into the single class no matter whether users have different search goals or not. So, there should be a risk to avoid wrong classification search results into too many classes. It evaluates the normalized number of clicked URL pairs that are not in the same class.

$$Risk=\sum_{i,j=1(i<j)}^{m} d_{i,j}$$

where m is number of clicked URLs and dij is 0 if pair of
clicked URLs belongs to same class otherwise dij is 1.
4) Classified AP (CAP): New criterion Classified AP (CAP) is extension of VAP by using above Risk. It combines AP of class having more clicks and risk of wrong classification. It is used to evaluate performance of restructured search results.

$$CAP= VAP \times (1\text{-}Risk)^{\gamma}$$

where γ is normalizing factor used to adjust influence of Risk on CAP.

IV. CONCLUSION

The proposed system can be used to improve discovery of user search goals for a query by clustering user feedback sessions represented by pseudo-documents. Using proposed system, the inferred user search goals/intents can be used to restructure as well as the web search results. So, users can find exact information needed as they want very efficiently and personalization method can be rerank relevant document for a user higher in result list, the user would be more satisfied. The discovered clusters can also be used to assist users in web search. Thus, users can find what they want conveniently.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

REFERNCES

1 R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.ACM Press, 1999.
2 H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.
3 T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay,"Accurately Interpreting Clickthrough Data as Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005
4 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 Management (CIKM '08), pp. 699-708, 2008.
5 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
6 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.
7 X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.
8 H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004
9 Porter, M. An algorithm for suffix stripping. Program, Vol. 14(3), pp. 130-137, 1980