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

A Novel Approach for Restructuring Web Search Results by Feedback Sessions Using Fuzzy clustering

R.Dhivya[1], R.Rajavignesh[2]

(M.E CSE), Department of CSE, Arasu Engineering College, kumbakonam[1]

Asst. Professor, Department of CSE, Arasu Engineering College, kumbakonam[2]

ABSTRACT: Ambiguous queries are submitted in search engine by different users with different search goals. The analysis of user click through logs can be useful in finding the precise search results. The user click through logs contains the information about the user search information. By analyzing the user click through logs the feedback sessions are constructed. The pseudo documents are generated by representing the feedback sessions for clustering. The fuzzy c-means clustering algorithm is used for clustering those pseudo- documents. A novel approach for user and query dependent feedback sessions for user search results. The CAP is formulated to evaluate the performance of user search goal inference. This can be very useful in improving search engine efficiency.

KEYWORDS: User Search goals, feedback sessions, pseudo-documents, classified average precision

I. INTRODUCTION

Web mining is one of the applications of data mining techniques to discover knowledge from the web. In web search, users are submitted queries to the search engines to get relevant information. But many search engines results are not informative and failed to produce results according to the user search goals. Users are usually giving some vague keywords representing their interests in their minds. Such keywords do not match with the results produced by the search engines. Many works about user search goals analysis should be carried out. Some users give ambiguous queries to the search engines (e.g. Apple, jaguar, the sun etc.) they get mostly the irrelevant results. User search goals are classified as Navigational and Informational, the queries that seek a single website or webpage and queries that reflect the intent of the user to perform a particular transaction respectively.

Many related works have been carried out according to the web search applications and the user search goals. In previous works, clustering is done on a set of top ranked results. The user search logs information is not analyzed and the feedback sessions are not considered. Analyzing the clicked URLS only from the web search logs. They only identify whether a pair of queries belong to the same goal or mission and does not care about what the goal is in detail. Semantic based web search for a particular query and the similarity between the words are carried out. Various algorithms such as star clustering algorithm, k-means clustering algorithm are used for clustering the pseudo documents but it also does not cluster the relevant information according to the user search goals. In clustering the cluster labels discovered are also not informative. User search goal is the information on different aspects of a query that users wants to obtain. Information need is a user's wish/desire to obtain the relevant information to satisfy his need. To cluster web search results, the URLs are analyzed by extracting the titles and snippets. But all those works produced noisy results and does not obtain the user search goals precisely. When more irrelevant and relevant results are produced by the search engines it is time consuming to browse.

In this paper, the user submits the query into the browser. The search engine searches the relevant information according to the user query. The user actions are stored in the user click through logs. From the user click through logs each

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

and every session is analyzed and generates the feedback session. The feedback session contains both the clicked and unclicked URLs and the last clicked URL in a single session. The feedback session contains the URLs and the click sequence. By analyzing the feedback sessions, the pseudo documents are generated. The pseudo documents contains the keywords that are most clicked in a session. Likewise the pseudo documents are clustered using the clustering algorithm. The user search goals are obtained according to the feedback sessions. The restructure result is produced for the user query based on the user search goal. The CAP evaluation can be done for each user search goal and the clustering can be done to find the optimal number of users.

Fig.1 examples of the different user search goals and their distributions for the query "the sun" by our experiment.

II. FRAMEWORKS

Fig 2 shows the framework of our approach. Queries are submitted to search engines to represent the information needs of users. Ambiguous queries contain one or several polysemous terms. Query ambiguity is one of the main reasons for poor retrieval results (difficult queries are often ambiguous). User Click-through data log contains data about the interactions between users and Web search engines. It is one of the most extensive (yet indirect) surveys of user experience. The user search information's are stored in the user click trough logs. . From the user click through logs the feedback sessions are constructed. The proposed feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. The feedback sessions is based on a single session, although it can be extended to the whole session.

The feedback session contains the URLs with the click sequence. A novel optimization method to combine the enriched URLs in a feedback session to form a pseudo-document. This can effectively reflect the information need of a user. The URLs are enriched from the feedback sessions based on the click sequence. The enriched URLs with more value in click sequence are mapped to pseudo-documents. The pseudo documents are depicted with some keywords based on the URL. At last, cluster these pseudo documents to infer user search goals and depict them with some keywords. For clustering the pseudo-documents the fuzzy c-means clustering algorithm is used. The clustered pseudo documents are stored in the user search goals. From the user search goals the restructured results are produced. A new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus the restructured web search result is produced. This proposed novel criterion "Classified Average Precision" to evaluate the restructure results.

III. ANALYZING USER CLICK THROUGH LOGS

The user click through logs is analyzed for each session to propose a feedback sessions. The feedback session is the better representation for the user click through logs. It is more efficient than analyzing the user click through logs directly. For a single query each and every session is analyzed and represents the feedback session. The feedback session is

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

based on a single session although it can be extended to the whole session. An ambiguous query is that it gives more than one meaning. So the precise results according to the user search goal are difficult to obtain.

3.1USER CLICK THROUGH LOGS

User Click-through data log contains data about the interactions between users and Web search engines. It is one of the most extensive (yet indirect) surveys of user experience. For researchers it helps to understand human interaction with IR results. The user click through logs contains all the user actions. It contains the session id, query term, position of the URL, click sequence and the URL.

3.2 FEEDBACK SESSIONS

The feedback sessions is discovered from each and every session from the user click through logs. The feedback sessions consists of the URLs that users visited and unvisited. Using the click sequence, the order in which the URLs are visited by the users the feedback sessions are generated. The feedback sessions consists of URLs that contains the URLs from first URL and up to the last visited URL. The feedback session is based on the users browsing actions that are stored in user click through logs according to the particular query.

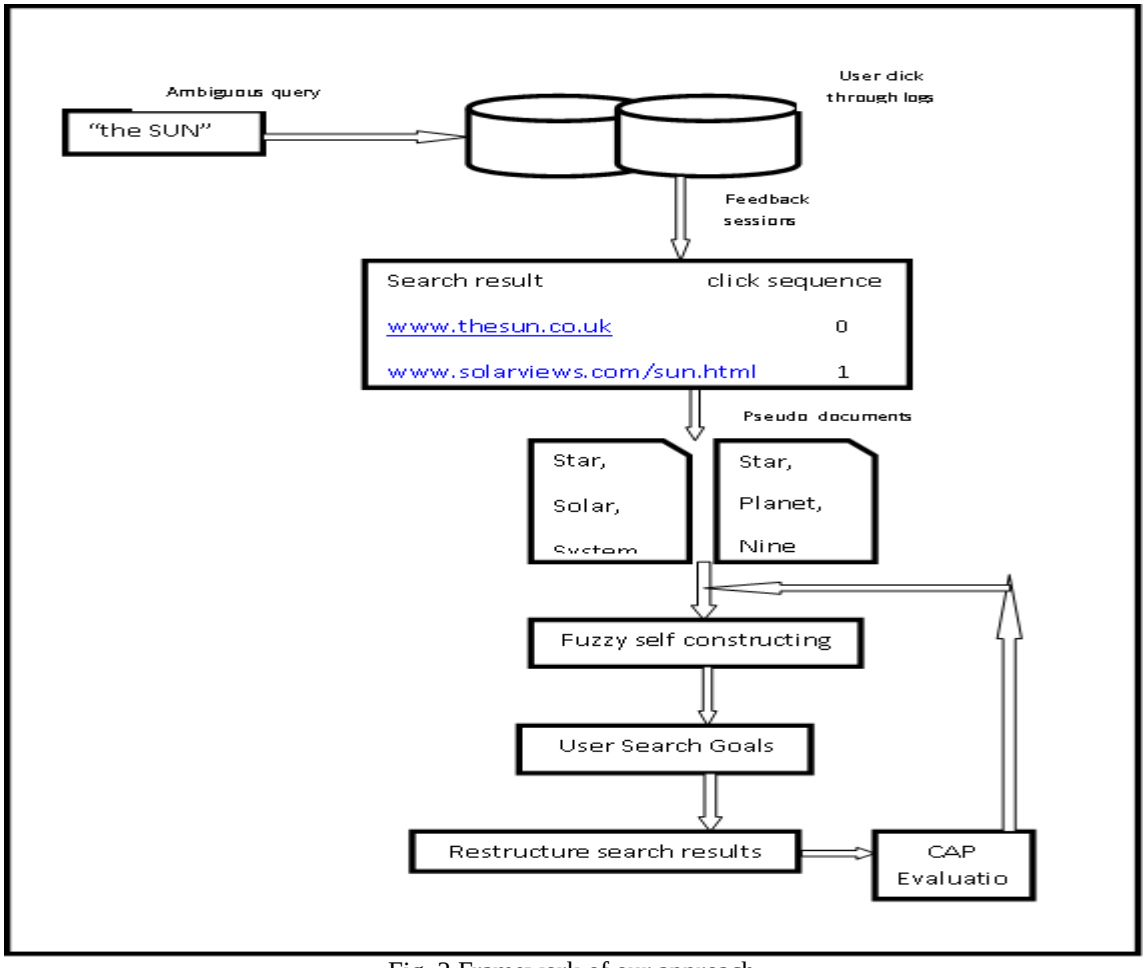

Fig. 2 Framework of our approach

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

3.3 PSEUDO DOCUMENTS

The pseudo documents are not the legitimate documents. The URLs in the feedback sessions are enriched by some format. The URLs are formatted by removing the stop words and the stemming words. It is the icon of showing the information about the whole document by some keywords. The documents are created by the number of occurrences of the keywords. The keywords which are having the more frequency are grouped together. The pseudo documents contain the keywords that are retrieved from the URLs in the feedback sessions. Using the Meta tag information the URLs are enriched. The Meta tag contains the most important keywords about the entire document information.

IV. CLUSTERING OF PSEUDO DOCUMENTS

Inferred user search goals from the pseudo documents by using clustering algorithm. The fuzzy self- constructing is used for the clustering of similar pseudo documents. The similarities of the keywords are grouped together and form the user search goals. 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. 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. 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. Like the k- means algorithm, the FCM aims to minimize an objective function. For clustering of pseudo documents, the similarity of the documents is clustered using the fuzzy clustering. The same users in the same session have different goals at different times. It is inappropriate to capture such overlapping interests of the users in crisp clusters.

The fuzzy is used to discover different search goals. The similarity of the cluster is based on the centroid values. The search goals having least precision in one cluster may have to appear in another cluster with high precision. So discover different search goals for the users, the fuzzy clustering is used. The clusters are very informative and they are stored as the user search goals.

4.1 LABELING THE CLUSTERS

A label will be generated to describe what each cluster is about. A user can then view the labels to decide which clusters to look into. The best cluster will have the high precision. Generate more meaning full cluster labels using the past keywords that are given by the users during the search. The keywords are derived from the user search logs. Assuming that query words entered by users in the past that are associated with the current query can provide meaningful descriptions of the distinct aspects. Thus they can be better labels than those extracted from the ordinary contents of search results.

CLUSTERING RESULTS FOR ''IRAQ"
WAR, middle east, map, sadden Hussein, human rights, country, special report, guide, united nations, travel business
CLUSTERING RESULTS FOR ''RESUME"
Cover letter, job, resume writing, services, employment, professional resume, free resume, career, resume samples, experience

Fig. 3 clustering results for query "iraq" and "resume"

4.2 USER SEARCH GOALS

The clustering of pseudo-documents by fuzzy self-constructing clustering algorithm, which is simple and effective. Since we do not know the exact number of user search goals for each query. After clustering all the pseudo-documents,

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

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.

V. EVALUATIONS OF RESTRUCTURE SEARCH RESULTS

The evaluation of user search goals can be done using the CAP (CLASSIFIED AVERAGE PRECISION). The classified average precision is the calculation of precision of documents. Because from user click-through logs, we can get implicit relevance feedbacks, namely "clicked" means relevant and "unclicked" means irrelevant. A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence. VAP is the voted average precision which can be used for grouping the dissimilar documents for the particular user query search. Risk is the mapping of similar and dissimilar documents for the particular user query. If there is a similarity then the mapping value is 0 and if there is no similarity between VAP and risk then the mapping value is 1.

$$CAP=VAP*(1-Risk) \quad (1)$$

A single session	click sequence	rel(r)R_r/r
www.sun.co.uk	0	0
www.nineplanets.org	1	½
www.solarviews.com	2	2/3
en.wikipedia.org	0	0
www.thesunmagazine.org	0	0
www.space.com	0	0
en.wikipedia.org/the_sun(newpaper)	3	3/7
www.nasa.gov	0	0
www.nasa.go/worldbook	4	4/9

AP=1/4[1/2+2/3+3/7+4/9]=0.510

Class 1	click sequence	rel(r)R_r/r
www.nineplanets.org	1	1/1
www.solarviews.com	2	2/3
en.wikipedia.org	0	0
www.space.com	0	0
www.nasa.gov	0	0
www.nasa.go/worldbook	4	3/6

VAP=1/3[1/1+2/2+3/6]=0.833

Class 2	click sequence	rel(r)R_r/r
www.sun.co.uk	0	0
www.thesunmagazine.org	0	0
en.wikipedia.org/the_sun(newpaper)	3	1/3

Risk=1+1+1/c

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

VI. CONCLUSIONS

A novel approach for infer user search goals for an ambiguous query by clustering its feedback sessions represented by pseudo documents. The 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. Therefore, feedback sessions can reflect user information needs more efficiently. 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. The pseudo documents are clustered using the fuzzy clustering algorithm. The cluster labels are discovers precisely. Finally, CLASSIFIED AVERAGE PRECISION is formulated to evaluate the performance of user search goal inference. The restructured web search result is produced for every user search query. The result produced is efficient and time consuming for users.

REFERENCES

[1] Zheng Lu, HongyuanZha, Xiaokang Yang, Weiyao Lin, and ZhaohuiZheng, 2013 "A New Algorithm for Inferring User Search Goals with Feedback Sessions" Published by the IEEE Computer Society, pp. 502-522.
[2] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, 2008 "Context-Aware Query Suggestion by Mining Click-Through," Proc.14th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining (SIGKDD '08), pp.875-883.
[3] X. Li, Y.-Y Wang, and A. Acero, 2008 "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346.
[4] D. Shen, J. Sun, Q. Yang, and Z. Chen, 2006 "Building Bridges for Web Query Classification," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 131-138.
[5] X. Wang and C.-X Zhai, 2007"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.
[6] Giovanna Castellano, A. Maria Fanelli, Corrado Mencar and M. Alessandra Torsello 2007 "Similarity-based Fuzzy clustering for user profiling" Published by the IEEE Computer Society.
[7] Lazzerini,B. Marcelloni, F.; Cococcioni, M. "A system based on hierarchical fuzzy clustering for web users profiling" Published by the IEEE Computer Society, Print ISBN: 0-7803-7952-7