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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Algorithms for Inferring User Search Goals Using Feedback Session

Rutuja R. Dhamdhere[1], Gayatri A. Dukale[2], Snehal J. Divate[3], Rachana M. Kendre[4], Sadiya N. Khan[5], Prof. Jangam D. Y[6]

Student, Janwantrao Sawant Polytechnic Hadapsar, Pune India[1-5]

Professor, Janwantrao Sawant Polytechnic Hadapsar, Pune India[6]

Abstract: The aim of topic is to discover the number of different user search goals for a query and representing each goal with some keywords. We first infer user search goals for a query by clustering feedback sessions. For that, we use a concept of pseudo document, which is the revised version of feedback session. At the end, we cluster these pseudo-documents to infer user search goals and represent them with some keywords. Since the evaluation of clustering is also an important problem, we used evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. The clustering is done by bisecting k means where in the existing system it is done by k means clustering. The new algorithm increases the efficiency of result. After the segmented result formation, the result in the every segment is reorganized as per number of clicks of URLs. The link which is clicked more number of times will appear at first location in the segment. This reduces the time requirement for searching.

KEYWORDS: Classified Average Precision (CAP); Clustering; Feedback session; Pseudo-document; Segmented Result; User Goals

I. INTRODUCTION

Web mining is also one of the applications of data mining techniques to extract data from web. Web mining is basically divided into three types, web usage mining, web content mining and web structure mining. Web usage mining is used to find the requirements of user on the internet. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data. In the web structure mining graph theory is used to represent the hyperlink structure of internet. Web content mining is the mining, extraction and integration of useful data from web page content.

In the existing system, the user enters the desired query and result get appears in the list format. In which there is no bifurcation as per different goals of the query. For every user there may be several goals for several users. So that time required to find the exact result increases.

Inferring and analysis are two important aspects to improve the user search results. Every time when user enters a query he has different goals in mind. To identify that goal the inferring technique is used and to check its relevance it performs the analysis of result. When the user enters the query “paper” the search engine will give different results. The results may be based on the links which gives the details of papers or links related to newspapers. In this, the search engine doesn’t know about the user goal therefore it gives the different links of different domains. So this method does not satisfy the user requirements. Therefore there is necessity to find out the user interest and distribute the results as per goals. To categorize the goals the inferring technique is used. In addition to this, the organization of segmented result is also necessary. This organization will keep the previously clicked queries at first location.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

The need of the proposed method is to find the exact goal of the query. This will improve the result and help the user to find the exact document they want. For this proposed method concepts of feedback session and pseudo document is used. Bisecting k mean clustering is used to divide the result into the different categories as per there domains. To organise the categorized result, number of clicks of links are stored in the feedback session. As per these number of clicks the results are reorganised in the segment. So that the link which is clicked more number of times will appear at first location in the categorized list.

II. RELATED WORK

In resent year, many works has been done for user search goals analysis. This work can be summarized into three classes: Classification of query, search result reorganization and Session boundary Detection. Some works belong to query classification, this work have been done to infer the so-called user goals or intents of a query. Some works analyze the search results returned by the search engine directly to use different query aspects. But query aspects without user feedback have limitations to improve search engine relevance. Some works consider user feedback and analyze the different clicked URLs of a ambiguous query present in user click-through logs directly, but the number of different clicked URLs of that ambiguous query may be not be sufficient to get ideal results. Wang and Zhai proposed framework that restructure web search results according to user search goals by grouping the search results with the same search goal. for example the query “car” is clustered with some other queries, such as “car rental,” “used car,” “car crash,” and “car audio.” Some other works introduce search goals and missions to detect session boundary hierarchically . But, their framework only identifies whether a pair of queries belong to the same goal. Limitation of this work is that he does not care what the goal is in detail. Other works perform analysis of the search results returned by the search engine when a query is submitted Here user feedback is not considered, lot off noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot be use for infer user search goals precisely. Other works make center of attraction is tagging queries with some predefined concepts to improve feature representation of queries Zamir et al. used Suffix Tree Clustering (STC) to identify set of documents having common phrases and then form cluster according to these phrase.
For clustering web documents They used documents snippets instead whole document. Generating meaningful labels for clusters is most challenging in document clustering, and this problem is solve in [8], in this work a supervised learning method is used to extract possible phrases from search result snippets and these phrases are then used to cluster web search results.

III. EXISTING SYSTEM

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. In the first class, people try to infer user goals and intents by predefining some specific classes and performing query classification accordingly. However, since user needs changes for different queries,so that finding suitable search goal is very difficult. In the second class, people try to reorganize search results. But this may involve many noisy search results that are not clicked by any users. In the third class, aim of people is to detect the session boundry. However, this only identifies whether pair of queries belongs to the same goal or not. In the existing system k means algorithm is used for clustering, in which the result depends on the k value. If the value of k is large then it will take exponential time to find the final cluster

DISADVANTAGES

1. OUser’s goal is not identified.
2. Many Noise search result will be shown in which user is not interested

IV. PROPOSED SYSTEM

Our system contains four different phases. First is feedback session which is combination of clicked and un-clicked URL’s. Second is a pseudo document that represents the feedback sessions in more meaningful manner. Third is clustering, that clusters these pseudo documents in appropriate user search goal’s. For the clustering bisecting k means is used. This algorithm gives the better results than k means algorithm. Fourth phase is organization of clustered data. And finally CAP method to evaluate the performance of our clustering.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Advantages

1. User's goal get identified

2. Data will be shown according to user's interest.

Process Summary: 1] User Enter Query 2] Search Cluster Data is present or not for query if present show cluster data with Google data otherwise show only Google data. 3] User click on interested URL after that Generate Feedback Session based upon clicked and un-clicked URL. 4] Get Title, Snippets, URLs, Click URL Count, Unclick URL Count in Feedback Session 5] Separate Title and Snippet From Feedback Session. Remove Duplicate title and snippet 6] Generate Pseudo Doc. 7] Apply K-Means clustering algorithm to these pseudo documents. 8] Organize the segmented result by considering the number of clicks. 8] Implement AP, VAP, Risk, CAP 9] Show different search goals to user in a segmented format.

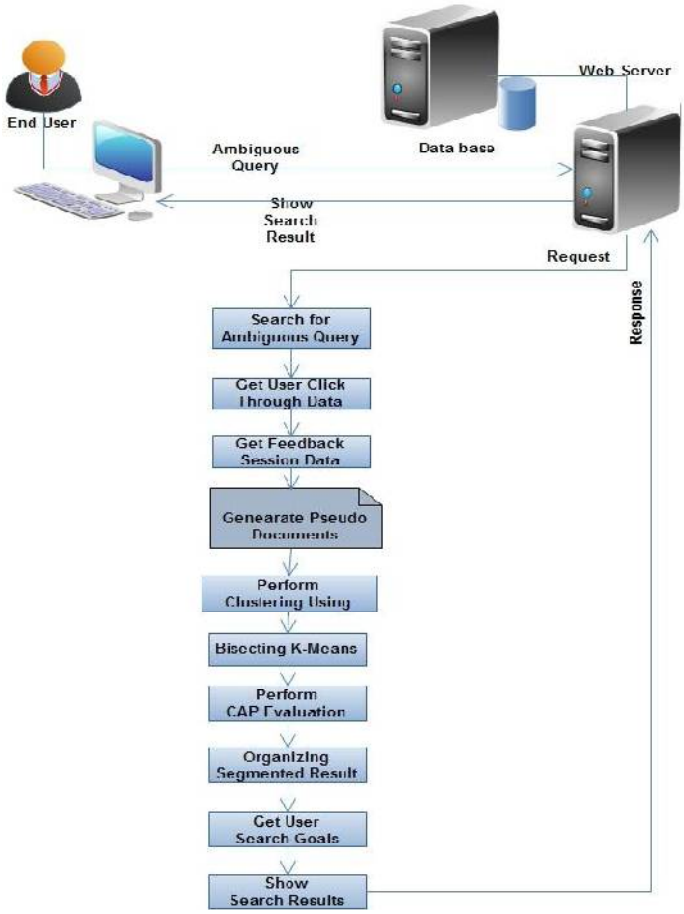

Fig: System Architecture

The entire system is divided into 4 parts.

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Phase 1:
The first part is feedback session. Feedback session collects the data from googles database. The feedback session consists of title and snippet. Every URLs title and snippet are represented by Term Frequency-Inverse Document Frequency(TF-IDF). It saves the both clicked and un-clicked URLs up to last clicked URL.

Phase 2:
In the step 2 the pseudo document is formed by using the feedback session. In the pseudo document both the clicked and un-clicked URLs are considered. Some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Sum of term frequency and inverse document frequency is stored as a feature representation of document in F.

Phase 3:
The next step is to find out the user goal by applying clustering algorithm on pseudo document. The similarity between documents is checked by using cosine function. Distance is calculated from that cosine function for clustering algorithm. After clustering, every cluster is considered as a different user goal.

Phase 4:
In this step, the evaluation of clustered is done. For the evaluation method of Classified Average Precision (CAP) is used. To calculate this CAP, the values of Average Precision and Risk is required

V. CONCLUSION

In this topic, a new approach is proposed in this of inferring user search goals by using the feedback session and pseudo document. In the feedback session both the clicked and the un clicked URLs ones before last click are stored. Pseudo document is made from mapping of feedback session. By performing clustering operation on this pseudo document will result into finding the user search goals which are depicted by keywords. To find out the user search goals the bisecting k means algorithm is used over the k means clustering. In the proposed work it will rearrange every segment as per the number of clicks of URLs in previous usage. So that the link which has the highest number of clicks will get appear at first position in the segment. Finally, criterion of 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] 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.
[2] 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.
[3] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, Hang Li,” Context-aware query suggestion by mining click-through and session data” ISBN: 978-1-60558-193-4,Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. (SIGKDD ’08), pp. 875-883, 2008.
[4] 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.
[5] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng,”A New Algorithm for Inferring
User Search Goals with Feedback Sessions”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013
[6] U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05),pp. 391-400, 2005.