ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Inferring User Search for Feedback Sessions

Sharayu Kakade[1], Prof. Ranjana Barde[2]

PG Student, Department of Computer Science, MIT Academy of Engineering, Pune, MH, India[1]

Assistant Professor, Department of Computer Science, MIT Academy of Engineering, Pune, MH, India[2]

ABSTRACT- Due to the incredible increase in use of internet surfing in today's world, user may not get the most exact search outcome which they have chosen for their queries to clarify their known tentative information. Search engine may not often bring the user ideal information and does not satisfy the user request completely. This paper proposes a new structure to verify different user search goals for a user given query by clustering feedback sessions. User click logs are used to construct feedback sessions which will efficiently return the information needs of users. System will create pseudo-documents to represent the feedback sessions better for clustering. Classified Average Precision (CAP) an innovative standard is also put forward in order to calculate performance of inferring user search goals. An Administration View for minimizing the risk is introduced. The proposed clustering algorithm is experimented with to show the effectiveness of clustering approach. The scope of the proposed system is to organize the search results in an appropriate way so that user searching process can be done effectively.

KEYWORDS: Click logs, feedback sessions, pseudo documents, CAP (Classified Average Precision), Search Result Restructuring

I. INTRODUCTION

In web applications, when user searches any keyword or phrase, queries are submitted to search engines to retrieve the information needs by the users. However, many times queries may not exactly represent user's specific information needs since many ambiguous queries may cover a broad topic and different users may expect information on different aspects when they submit the same kind of query. For example, when the keyword "The Apple" is submitted to the search engine, some users want the information on Apple products, while some other users want information on fruit Apple. User search goals can be considered as the clusters of information requirements for a query. The aim is to discover the number of diverse user search goals for a query.

If we imagine world from the perspective of a search engine, our view towards the user behavior will be the set of queries user's submit. Search engine designers often consider and adapt this perspective, studying these query streams of queries and trying to optimize the engines based on factors such as the overall length of a query. This perspective prevents us from looking beyond the query to get why the users are performing their searches in the first place; the 'why' of user search behavior is actually essential to satisfying the user's information needs. Searching is merely a means at the end, a way to satisfy the user with the underlying goal they are expecting to achieve. In fact, in some cases the same query might be used for different goals. For example, the query The Apple might have been used in any of the situation (assuming that it may be product/fruit/film etc.). Therefore, it is essential and potential to pick up the different user search goals in information retrieval. User search goals can be defined as the information on different aspects of a query that user groups want to obtain. Information need is a users desire to obtain particular information to satisfy his/her need. User search goals can be considered as the set 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. Due to its efficiency, many works about user search goals analysis have been invented.

II. RELATED WORK

This paper briefly survey previous work on inferring user search goals. Previous work focused on manual query log analysis to find out user search goals. Uchin Lee, Zhenyu Lyu, and Junghoo Cho suggested automatic identification of

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

user search goals [2]. They proposed two types of goal identification tasks: user click behavior and anchor link distribution.

Doug Beeferman and Adam Berger introduced a technique to discover clusters of similar queries and similar URLs [3].They viewed the dataset of user query and selected URLs from offered results as a bipartite graph in which one of the side represents queries and other side as URLs. The agglomerative clustering algorithm is applied to the vertices of graph to recognize related queries and URLs. Thus, it helps users in web search with the discovered clusters.

An approach was developed which is using the web log data to identify and improve user navigation pattern by prediction. Prediction is done by clustering and classification from the web log data[4]. A hit is defined as a request to view a HTML document or image or any other document. The web log data are automatically created and can be obtained from either client side server or proxy server or from an organization database. Each entry in the web log data include details like the IP address of the computer making the request, user ID, date and time of the request, a status Field indicating if the request was successful, size of the file transferred, referring URL, name and version of the browser being used. Weblog file is created and maintained by web server.

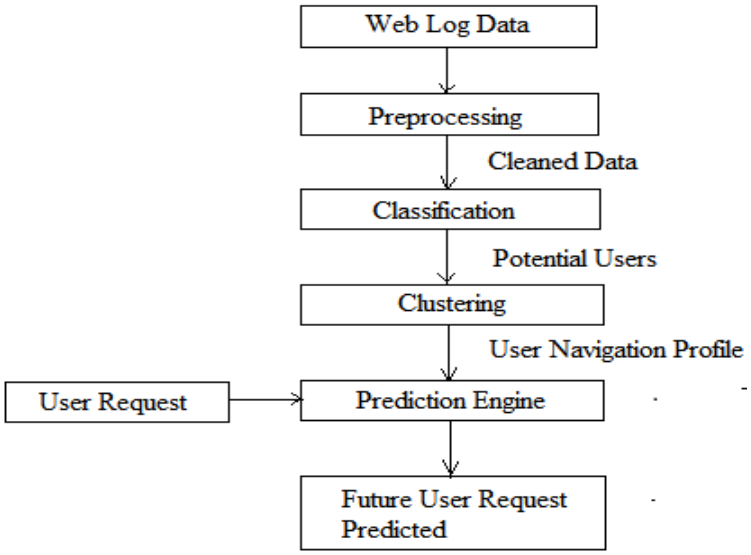

Figure 1. PUCC Model

This model processes the web log data so that unformatted log data is converted in to a form so that it can be used for mining. This pre-processing includes cleaning user identification and session identification. To identify or separate the potential users, decision tree classification algorithm is used. A Graph partitioned clustering algorithm is used to group users with similar navigation pattern.

Ji-Rong Wen, Jia-Yun Nie and Hong-Jiang Zhang tried to cluster similar queries according to their contents as well as user logs[6]. They used density based clustering method DBSCAN and its Incremental DBSCAN to build a comprehensive query clustering tool.

Previous System

In a typical information retrieval setting, the user describes their information need with a query Q, in response the retrieval system returns a ranked list of documents D1, D2, D3,... as results, as shown in Figure 2. If the initial ranking is poor, one way for improvement is to ask the user to provide feedback, i.e., to evaluate the relevance of some top-ranked documents. According to the cluster hypothesis, which states that relevant documents tend to be more similar to each other than to irrelevant ones, if some example relevant documents are identified from user feedback, one may find more relevant documents by seeking similar ones**.**

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

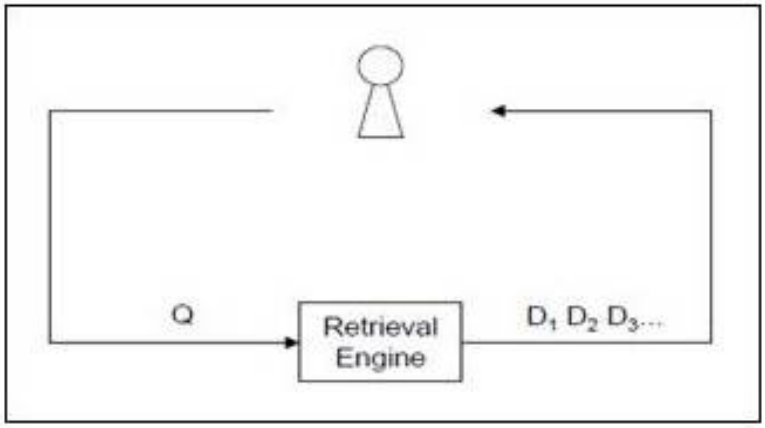

Figure 2. Relevance Feedback

III. SYSTEM ARCHITECTURE

Proposed system takes the user query dataset as input and generates original search results and performs various steps on these original results to generate output. System performs preprocessing on generated original search results then search result clustering and ranking is applied to provide restructured search results. Figure 3 shows the architecture of the system.

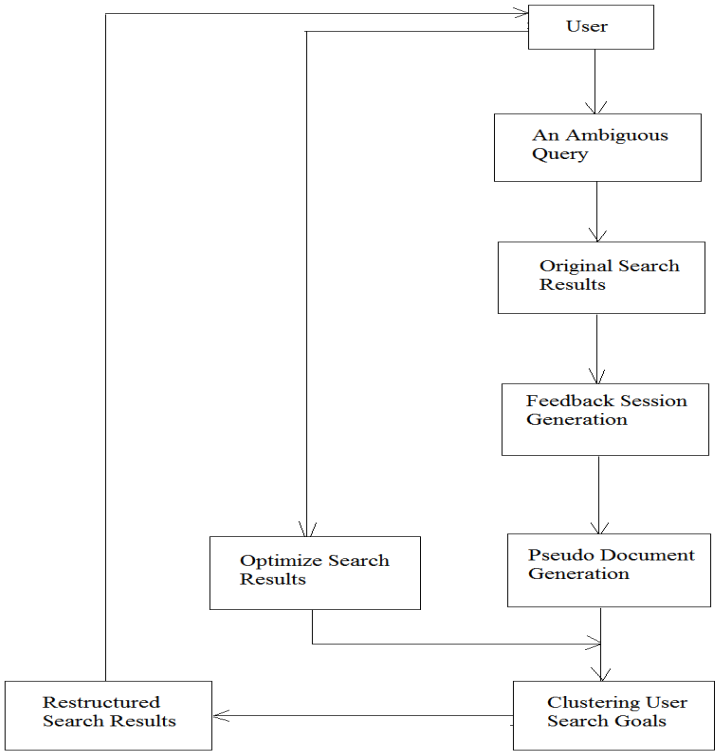

Figure 3. System Architecture

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

List of Modules

- User Login
- Static Search
- Dynamic Search
- Evaluation

Module 1 (User Login)
Input: User name and password
Output: Successful login
The user has to first login to the system get access to the system. A new user registers itself by providing first name, last name, email-id, password, phone number. After successful registration user login to the system using username and password. When user no longer needed access he may logout from the system.

Module 2 (Static search)
Input: User query
output: Reranked results
In static search module, user first enters the query to the search engine. Search engine provides the results in the form of URLs from the database. Based on user's choice these fetched results are grouped and reranked using clustering and ranking algorithm. Administrator indexes the query, category and URLs into the database. Manually adding these input set has the advantage of getting more accurate results.

Module 3 (Dynamic search)
Input: User query
Output: Reranked results
In dynamic search module, query is issued by the user to the search engine. Search engine first look for the results in local database. If available those results will be displayed to the user and user can proceed for further searching. Otherwise results will be fetched online and these raw results will be provided to the user. User selects URLs which he finds relevant to him. Titles and descriptors will be extracted from all these returned URLs and feature vector for each URL is constructed based on occurrence of terms in title and descriptor. Input Query is matched with these URLs and only those URLs which have threshold greater than 05. are selected as input to clustering algorithm. The proposed cosine similarity clustering algorithm is used to cluster these selected URLs which infers user search goals. These clustered search results are ranked in descending order of click frequency.

Module 4 (Evaluation)
Input: Reranked results
Output: Evaluated results
The final restructured results are evaluated by using CAP (Classified Average Precision) as an evaluation metric. It takes into account risk of misclassifying the restructured search results.

Proposed Cosine Similarity based Clustering Algorithm
Input: Two strings V1 and V2
Output: Similarity score s
Step 1: Set Input V1 as user query and V2 as pseudo-document for web URL.
Step 2: for each(I in V1)
for each(J in V2)
Map each i^{th} and j^{th} word or character.
end for
end for
step 3: Verify similar count
Step 4: Calculate similar count using similarity vector as s
Step 5: Return s

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

IV. EVALUATION OF COSINE SIMILARITY BASED CLUSTERING ALGORITHM

To evaluate the performance of Cosine Similarity based Clustering Algorithm, we adopted CAP (Classified Average Precision) as the evaluation metric. Given a ranked results of query, CAP is calculated as

$$CAP=VAP * (1-Risk)^{r}$$

where VAP (Voted Average Precision) is AP (Average Precision) of class having more clicks. AP (Average Precision) is calculated as

$$AP=1/N^{+}\sum^{n}_{r=1} R_r / r$$

where N^{+} is the number of clicked documents in the retrieved ones and, 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 R_r is the number of relevant retrieved documents of rank r or less. Risk is declared as follows

$$Risk=\sum^{m}_{i,j=1(i>j)} d_{i,j} / (C_m)^2$$

It calculates the normalized number of clicked URL pairs that are not in the same class.
In order to evaluate the effectiveness of proposed clustering algorithm, it is compared with k-means clustering based algorithm.
Figure 4 shows the result comparison of k-means based method with proposed cosine similarity based method.

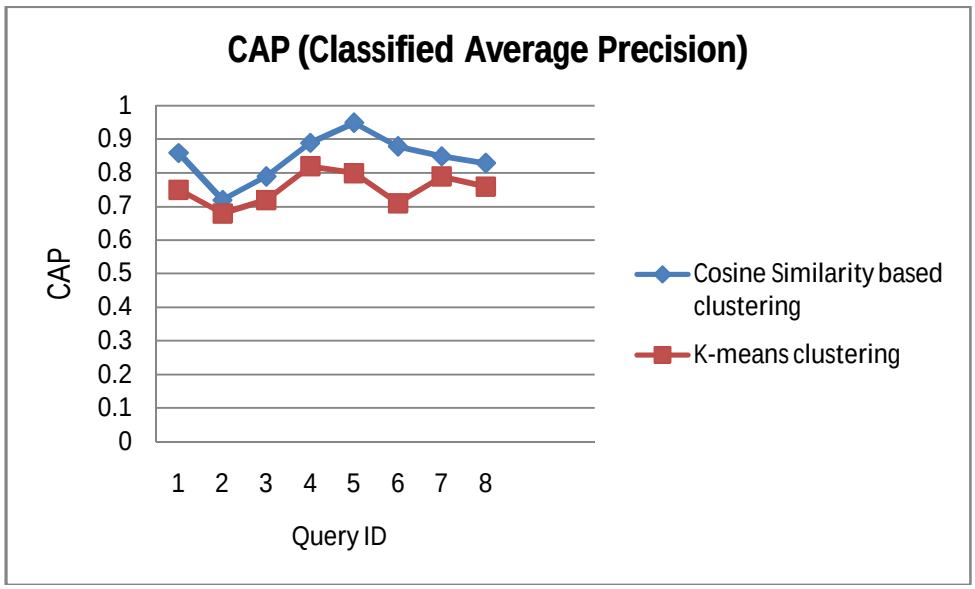

Figure 4. Performance of cosine similarity based clustering method in terms of CAP.

Here, from the figure we can speculate that the proposed approach can effectively infer user search goals from user given query by simultaneously exploiting feature vector and click preference given by the user. The k-means clustering based method cannot infer user search goals precisely because it requires the number of clusters to be formed in advance. Although it infers user search goals, it cannot find it effectively. The proposed cosine similarity based method finds user search goals efficiently.

V. CONCLUSION

The proposed system is used to infer the user search goals from the user given query. System includes four main components: user login, static search, dynamic search, and evaluation. First, system performs preprocessing on original search results. Then, the feature vector of URL's which has threshold greater than 0.5 are selected as input to clustering algorithm. The cosine similarity based clustering algorithm is applied to obtain the clusters of user search goals and

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

ranking is performed in descending order of click numbers. Proposed method shows the performance improvement above the existing static based system in terms of CAP by 9%.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin and 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.
[2] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search", Proc. 14th Intl Conf. World Wide Web (WWW 05),pp. 391-400, 2005.
[3] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log", Proc. SixthACM SIGKDD Intl Conf. Knowledge Discovery and Data Mining (SIGKDD 00), pp. 407-416, 2000.
[4] V. SUJATHA, PUNITHAVALLI, " Improved User Navigation Pattern Prediction Technique from Web Log Data", International Conference on Communication Technology and System Design 2011.
[5] Rakesh Kumar, Aditi Sharan, "Personalized Web Search Using Browsing History and Domain Knowledge", 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).
[6] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User Queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web (WWW '01),pp. 162-168, 2001.
[7] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines", Proc. Intl Conf. Current Trends in Database Technology (EDBT 04), pp. 588-596,2004.
[8] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification", Proc. 30th Ann. Intl ACM SIGIR Conf. Research and Development (SIGIR 07),pp.783-784, 2007.
[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 Discovery and Data Mining(SIGKDD 08), pp. 875-883, 2008.