ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 7, July 2015

User Search Goal Identification Using User Click Sequence Analysis

Miss. Radhika Rane[1], Mr. Sudip Tembhurne[2], Mr. Sanjeev Dwivedi[3]

M. E. Student, Computer Engineering, VIT Mumbai, India[1,2]

Associate Professor, Department of Computer Engineering, VIT Mumbai, India[3]

Abstract: Data Mining refers to extracting or mining knowledge from large amounts of data. It is also called as knowledge mining from data. Web mining is the application of data mining to extract knowledge from web data including web documents, hyperlinks between documents usage log of websites .Search engine is one of the most important applications in today's internet. For an ambiguous query, different users may have different search targets, so the search engine doesn't satisfy user information needs properly on the diverse aspects upon submission of same query. The computation and analysis of user search goals can be very useful in improving search engine relevance and user experience. Search history records have been clustered to discover different user search goals for a query. User click sequences are constructed from user click-through logs and can efficiently reflect the information needs of users. Virtual-documents are generated through user click sequences for clustering using clustering algorithm. We propose Cosine Similarity Algorithm to evaluate the performance of user search target computing based on restructuring web search results. Thus, we can determine the number of user search target for a query.

Keywords: Data mining, user search goals, clustering, cosine similarity algorithm.

I. Introduction

Identifying user goals from a given search query is a very difficult job as mostly search engines gives results in the form of simple keywords which may cover broad topics, to more technical precision ,or to proper nouns that can be used for guiding the search process to the meaningful collection of documents. For the need of information users usually fires a query to search engines and different users have different targets in their mind for a broad topic. Sometimes search engines query output may not match the user goals due to usage of less ambiguous keywords[1].

Sometimes user specific information needs may not be represented by queries since many ambiguous queries may cover a broad topic. Therefore, it is necessary to capture different user search goals. User search goals are information on different aspects of query that user want to obtain. Inference and analysis of user search goals have advantages such as restructure the web search results according to user search goals by grouping the search results with the same search goal, user search goals represented by some keywords can be utilized in query recommendation and distribution of user search goals. There are three classes representing user search goals:

1. Query classification
2. Search result reorganization
3. Session boundary detection.

In first class, some specific classes are predefined and query classification is performed accordingly. User goals are classified into navigational and informational. For navigational, user has particular web page in mind but for informational user's does not have particular page in mind or intends to visit multiple pages. Some other methods used for defining queries as product intent and job intent. Next method defined is tagging queries with some predefined contents to improve feature representation of queries. Disadvantages of this classification are finding suitable predefined search goal class is difficult because what user cares about varies a lot for different queries. In second class, people try to recognize search results. First method used is learning interesting aspects of queries by analysing the clicked URLs or text documents directly from user click through logs to organize search results. Limitation of this is number of clicked URL's or text documents may be small. Another method used is analysing the search results returned by a search engine when a query is submitted. But disadvantage of this method is feedback is not taken into account so noisy results that are not clicked by user may be analysed. In third class, aim is to detect session boundaries. This method predicts goal and mission boundaries to hierarchically segment queries logs. Limitation with this if it only identifies whether a pair of queries belong to same goal and does not care about the goal in detail. Here, aim is to discover the number of different kinds of user search goals for a query and describing each goal with some keywords. For this purpose first approach is to Cluster the feedback sessions to infer user search goals. Feedback session contains both clicked and unclicked URL's or text documents and ends with the last URL or text document that was clicked in a session. The distributions of different search goals can be obtained after feedback sessions are clustered. Then to reflect user information needs effectively map these feedback sessions to Virtual-documents. This is nothing but the optimization method to combine the enriched URL's or text documents in a feedback session. CAP (Classified average precision) is used to evaluate the performance of user search goal inference based on restructuring web search results. Using which we can determine number of user search goals for a

DOI 10.17148/IJARCCE.2015.4732

query. System can use URLS or text documents as URL's are very difficult to managed and also various search engines uses URLS to give output to the query. Text documents are easy to made and also gives clean output to the user. User finds the information to query so quickly and organised result is being displayed. Text documents are very fast to retrieve from system as compared to URL's.So based upon the need of user our proposed system gives output .

II. Related work

A. Automatic identification of user goals:

U. Lee, Z. Liu, and J. Cho[2], proposed automatic identification of user search goals. They stated that majority of queries have a predictable goal. Taxonomy of query goals based on two types:

A.1. Navigational

In this type, user has a particular web page in mind and is primarily interested in visiting that web page. User may either have visited that site before, or just assumes such a site exists. Here, user's will only visit the correct sites.

A.2. Informational queries

These are the queries where user does not have a particular page in mind or intends to visit multiple pages to learn about the topic. User is exploring Webpages that provide background knowledge about a particular query topic. Users click on multiple results because they do not assume a particular website to be single correct answer. Here, two features are used for the prediction of user goal:

1. Past user-click behavior: If a query is navigational, users will primarily click on the result that the user has in mind. Therefore, by Observing the past user-click behavior on the query, we can identify the goal.

2. Anchor-link distribution: If users associate particular query with a particular website then most of the links that contain the anchor will point to that particular website. Hence by observing the destinations of the links with the query keyword as the anchor, we can identify the potential goal of the query.

Limitations:

User queries are taken from the CS department that may show technical bias and are well crafted. In short, queries given by CS students are potentially work related. So, if we consider user queries by general people characteristics observed may not be true.

B. Web query classification

D. Shen, J. Sun, Q. Yang, and Z. Chen[3], published a work on classifying web queries into a set of target categories where the queries are very short and there are no training data. Here, intermediate taxonomy is used to train classifiers bridging and target categories so that there is no need to collect training data. Classifier bridging is used to map user queries to target categories. Classification approaches:

B.1. Classification by exact matching Two categories defined here are intermediate taxonomy and target taxonomy. One or more terms in each node along the path in the target category appear along the path corresponding to the matched intermediate category. For example, the intermediate category contains "Computers\Hardware\Storage" and target category contains "Computers\Hardware". We can directly map intermediate category to target category since both appears along the path "Computers\Hardware\Storage". In this approach, for each intermediate category we can detect whether it is mapped to target categories according to the matching approaches. It produces low recall because many search result pages no intermediate categories. B.2. Classification by SVM In this technique, it first constructs training data for target queries based on mapping functions between categories. If an intermediate category is mapped to a target category then the web pages are mapped into train SVM classifiers for the target categories. For each web query classify the query using SVM classifiers. This can improve the recall of classification result.

B.3. Classifiers by bridges It connects the target taxonomy and queries by taking an intermediate taxonomy as bridge. The intermediate taxonomy may contain enormous categories and some of them are irrelevant to the query classification task corresponding with the predefined target taxonomy. Therefore, to reduce the computation complexity, we should perform "Category Selection".

C. Reorganizing search results

X. Wang and C.-X Zhai[4], proposed clustering of search results which organizes it and allows a user to navigate into relevant documents quickly. This approach organizes search results learned from search engine logs. Steps of this approach are as follows:

Given a query,

1. Get its related information from search engine logs. Working set is formed by using this information.

2. Learn the aspects from information in the working set. These aspects correspond to users interests.

3. Each aspect is labeled with representative query.

4. Categorize and organize the search results of the input query according to the aspects.

First we will find related past queries in our preprocessed history data collection. Next learn the aspects by clustering. And finally categorize the search results using categorization algorithm.

D. Clustering web search results

H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma[5], researched on reformalizing the clustering problem. This approach consists of four steps:

1. Search result fetching

2. Document parsing and phrase property calculation

3. Salient phrase ranking

4. Post-processing.

Given a query and ranked list of search results. Firstly, the whole list of titles and snippets is parsed, extracts all possible phrases from the contents and calculates several properties for each phrase such as document frequencies, phrase frequencies. Then the regression model is applied to combine these properties into a single salience score. Phrases are ranked according to salience score and the top ranked phrases are taken as salient phrases. In post processing, filter out the pure stop words Disadvantages:

DOI 10.17148/IJARCCE.2015.4732

Feedbacks are not considered. So, noisy results that are not clicked by user may be analyzed.

E. Session boundaries

R. Jones and K.L. Klinkner [6], defined session boundaries and automatic hierarchical segmentation of search topics. In this approach, analysis of typical timeouts used to divide query streams into sessions and the hierarchical analysis of user search tasks into shorter goal and long-term missions is done.

Timeout is nothing but elapsed time of 30 minutes between queries which signifies that the user has discontinued searching. Here, combination of diverse set of syntactic, temporal, query log and web search features can predict mission boundaries and goals. Hence, best approach to clustering queries within the same goal may build on first identifying the boundaries then matching subsequent queries to existing segments.

Disadvantages: It only identifies whether a pair of queries belong to the same goal or mission but does not care about what the goal is in detail.

III. SYSTEM ARCHITECHURE

3.1 Proposed Architecture is developed for overcoming the drawbacks of traditional Kmeans algorithm .Proposed system mainly consists of 2 modules i.e.Admin and User Module.

Admin Module first logged into the system and prepares text documents. This documents consists of all the information that user will search in future while surfing. Admin prepares this document and stores it in database or either in cloud. This document contains all the gathered information of various topic or keywords. From various sources, information is collected and organised into the text document. This text document is well organised according to the query or keywords that users may use for searching. Admin will also keep the track of users that logged in to system means how much time the user fires the same query, according to that result is displayed. So whenever the user logged in again to the system result is again filtered. Different output is displayed whenever the user search for the same keyword. Hence, in this way refined and sorted search results can be obtained from the proposed system.

User module also first had to login to the system. If the user is authorised then only system will allow user to search the information. User fires a query or a keyword, result is obtained in the form of text document that can be downloaded and viewed to the user. If a user search for apple keyword so apple can be mobile phone or fruit or nursery rhyme.so different user has different goals in their mind. so he initially get all result i.e. all results related to the apple is displayed .user will select document according to the need and the click sequence is recorded in the admin side . Results is well defined and filtered if the same user next time logged into system and searches for the same keyword .

Click sequence is calculated from the frequency of downloaded document file. IF-IDF Analysis is done on the click sequence which gives weight to the document this is given as input to Cosine sine similarity algorithm and output is the well organised clustered set of data.

There are four parts in this system like user authentication, Keyword search mechanism, feedback session creation, search result restricting based on the history of users.

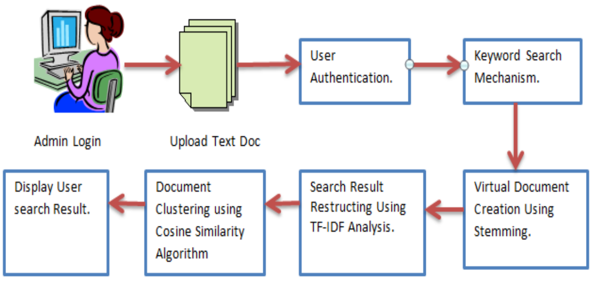
Admin Login
Upload Text Doc
User Authentication.
Keyword Search Mechanism.
Virtual Document Creation Using Stemming.
Search Result Restructing Using TF-IDF Analysis.
Document Clustering using Cosine Similarity Algorithm
Display User search Result.

Fig. 3. 1. System Architecture

3.2 Search History Records

A session for web search is a series of successive queries to satisfy a single information need and some clicked search results. We focus on computing user search goals for a particular query. Therefore, the single session containing only one query is introduced, which distinguishes from the conventional session. Meanwhile, the search history records in this paper are based on a single session, although it can be extended to the whole session. The proposed search history records consist of both clicked and unclicked URLs or text documents and ends with the last URL or text document that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks.

3.3 Search History Records to Virtual Document

Representing the URLs or text documents in the search history records. In the first step, we first enrich the URLs or text document with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the search history records. In this way, each URL or text document in a search history records is represented by a small text paragraph that consists of its title and snippet. Then, some textual processes are implemented to those text paragraphs, such as transforming all the letters to lower cases, stemming and removing stop words. Finally, each URLs or text document title and snippet are represented by a Term Frequency-Inverse Document Frequency (TF-IDF) vector.

3.4 Computing User Search Targets by Clustering Virtual-Documents

With the proposed pseudo-documents, we can compute user search targets. In this section, we will describe how to compute user search targets and depict them with some meaningful keywords.

3.4.1 Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).A set of cluster resulting from a cluster analysis

DOI 10.17148/IJARCCE.2015.4732

can be referred to as a clustering. In Web search application, a keyword search may often return a very large number of hits (ie. Page relevant to the search) due to the extremely large no of pages. Clustering can be used to organized search results in the groups and present the result in concise and easily accessible way and it is IR process. Clustering is some time called automatic classification, clustering can automatically find grouping .This is distinct advantages of cluster analysis. Cluster in is also called data segmentation in some application because clustering partitions large data set into groups according to their similarity

3.4.2 Cosine Similarity [11]

A document can be represented by thousands of attributes, each recording the frequency of particular word (such as a keyword) or phrase in the document. Each document is the object presented by what is called a term frequency vector. Term frequency vector is typically long and sparse (i.e. the may 0 values) applications using such structures includes information retrieval, text document clustering, biological taxonomy and gene feature mapping. Cosine similarity is a measure of similarity that can be used to compare documents or say, give a ranking of documents with respect to a given vector of query words.

3.5 Restructuring Web Search Result

The evaluation of user search goal inference is a big problem, since user search goals are not predefined and there is no ground truth. Previous work has not proposed a suitable approach on this task. Furthermore, since the optimal number of clusters is still not determined when inferring user search targets, a search history record information is needed to finally determine the best cluster number. Therefore, it is necessary to develop a metric to evaluate the performance of user search goal inference objectively. Considering that if user search targets are computed properly, the search results can also be restructured properly, since restructuring web search results is one application of computing user search targets. Therefore, we propose an evaluation method based on restructuring web search results to evaluate whether user search targets are inferred properly or not. In this section, we propose this novel criterion "Classified Average Precision" to evaluate the restructure results. Based on the proposed criterion, we also describe the method to select the best cluster number. 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 computing user search targets. We will introduce how to restructure web search results by inferred user search targets at first. Then, the evaluation based on restructuring web search results will be described.

IV. CONCLUSION

This system first introduces search history records to be analyzed to compute user search targets rather than search results or clicked URLs or text documents. Both the clicked URLs or text documents and the unclicked ones before the last click are considered as user implicit search history and taken into account to construct search history records. Therefore, search history records will reflect user information needs more efficiently. Second, Map search history records to pseudo documents to approximate goal texts in user minds. The pseudo documents can enrich the URLs or text documents 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.Text documents are created as compared to URL'S . As URL'S are very difficut to manage. Finally, a new criterion CAP will formulated to evaluate the performance of user search targets computance.

REFERENCES

[1] 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

[2] 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.

[3] D. Shen, J. Sun, Q. Yang, and Z. Chen, "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, 2006.

[4] 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.

[5] 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.

[6] 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.

[7] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann.Int'l ACM SIGIR Conf. Research and Development (SIGIR '07),pp. 783-784, 2007.

[8] 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.

[9] 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.

[10] 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.

[11] www.bionicspirit.com/blog/2012/01/16/cosine-similarity-euclidean-distance.html.

BIOGRAPHIES

Radhika Rane, M.E. student in Vidyalankar Institute of Technology, Department of Computer Engineering, Mumbai

Sudip Tembhurne, M.E. student in Vidyalankar Institute of Technology, Department of Computer Engineering, Mumbai

Prof. Sanjeev Dwivedi, Associate Professor in Vidyalankar Institute of Technology, Department of Computer Engineering, Mumbai

DOI 10.17148/IJARCCE.2015.4732