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, Issue 6, June 2014

Mining User Profile Using Clustering From Search Engine Logs

DR.A.MUTHU KUMARAVEL[1]

MCA Department, Bharath Institute of Science and Technology, Bharath University , Chennai – 73[1]

ABSTRACT: Fundamental component of any personalization application is user profiling. The existing user profiling strategies are based on users interest (i.e. positive preferences).The main focus is on search engine personalization and to develop several concept-based user profiling methods. Concept-based user profiling methods deals with both positive and negative preferences. This user profiles can be integrated into the ranking algorithm of a search engine so that search result can be ranked according to individual users interest. The RSCF makes a search of data containing the item in the search results, the required data is been clicked by the user and this clicked data is given as the input and generates the rankers as the output.. The negative preference increases the separation between the similar and dissimilar queries. This separation provides a clear threshold for agglomerative clustering algorithm and improves the overall quality.

KEYWORDS: Negative preferences, search engine, user profiling.

I. INTRODUCTION

One criticism of search engines is that when queries are issued, most return the same results to users. Queries are submitted to the search engine short and ambiguous and different information needs and goals under the same query. For e.g. a biologist may use query “mouse” to get information about rodents, while programmers may use the same query to find information about computer peripherals.

Personalized search is an important research area that aims to resolve the ambiguity of query terms. Personalized search engines create user profiles to capture the user’s personal preferences. Given query, a personalized web search can provide different search results for different users based upon their interests, preferences and information needs. User profiling strategy is an essential and fundamental component in search engine personalization.

User profiling is a fundamental component of any personalization applications. Generate user profile based on their access patterns. Development of user profile implicit way (i.e.) can be automatically learnt from a user’s historical activities. User browsing histories are the most frequently used source of information about user interests. User profiling strategy can be either document based or concept based. Document based profiling methods try to estimate user’s document preferences. Concept based profiling methods aim to derive topics or concepts that user’s are highly interested.

Concept based user profiling strategies that are capable of deriving both users’ positive and negative preferences. Negative preferences improve the separation of similar and dissimilar queries. User profiling strategies are query-oriented. Profile is created for each user queries.

II. RELATED WORK

In order to carry out this project several references and white papers are referred, from which many valuable information are identified. The following sections provide those information.

Query Recommendation Using Query Logs in Search Engines

Given a query submitted to a search engine, suggests a list of related queries. The related queries are based in previously issued queries [6]. The method based on a query clustering process in which groups of semantically similar queries are identified. The clustering process uses the content of historical preferences of users registered in the query log of the search engine.

Personalized Concept-Based Clustering of Search Engine Queries

Concept based profiling method that captures the user's conceptual preferences in order to provide personalized query suggestions. [4] Two new strategies are used to achieve this goal. First develop online techniques that extract concepts from the web-snippets of the search result returned from a query. Second a new two phase personalized agglomerative clustering algorithm that is able to generate personalized query clusters.

Personalized Search Based on User Search Histories

User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles collect user information through proxy server or desktop bots [5]. Personalization is the process of presenting the right information to the right user at the right moment. Systems can learn about user's interests collecting personal information, analyzing the information, and storing the results in a user profile. Information can be captured from users in two ways. Explicitly, for example asking for feedback such as preferences or ratings; and implicitly, for example observing user behaviors such as the time spent reading an online document.

Personalized WebSearch

A new technique on Personalized Web search can provide different search results for different users, based upon their interests, preferences, and information needs [2]. User information can be specified by the user or can be automatically learnt from a user's historical activities. Personalized web search can be achieved by checking content similarity between web pages and user profiles. Personalized web search can improve performance of web search. Personalized web search can be implemented on either server side or client side. For server-side personalization, user profile are built, updated, and stored on the search engine side. User information is directly incorporated into the ranking process, or is used to help process initial search results. For client-side personalization, user information is collected and stored on the client side, usually by installing a client software or plug-in on a user's.

Query Clustering Using User Logs

Query clustering is a process used to discover frequently asked questions or most popular topics on a search engine. This process is critical for search engines based on question-answering. Because of the short lengths of queries, keywords are not suitable for query clustering. [1] This paper describes a new query clustering method that makes use of user logs which allow us to identify the documents the users have selected for a query. The similarity between two queries may be deduced from the common documents the users selected for them.

Deriving Concept-Based User Profiles from search Engine Logs

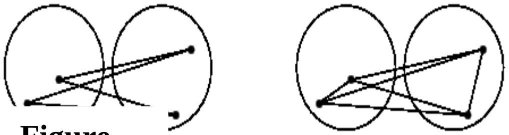
Figure.

Fundamental component of any personalization application is user profiling. The existing user profiling strategies are based on users are interested (i.e. positive preferences).The main focus is on search engine personalization and to develop several concept-based user profiling methods. Concept-based user profiling methods deals with both positive and negative preferences. The concept-based user profiles can be integrated into the ranking algorithm of a search engine so that search result can be ranked according to individual user's interest. To terminate and improve the overall quality of resulting query cluster the agglomerative cluster algorithm being used. User profiling strategy can be either document based or concept based. Concept based method provides personalized query suggestions based on a personalized concept based clustering technique. When a user submits a query, concepts and their relations are mined online from web-snippets to build a concept relation graph.

III. ALGORITHM

HAC (Hierarchical agglomerative clustering) Algorithm

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC . Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached. HAC is more frequently used in IR than top-down clustering and is the main subject.

Personalized Agglomerative Clustering

Personalized Agglomerative Clustering is divided into two steps: Initial clustering and Community merging.
Initial Clustering:

1. Obtain the similarity scores for all possible pairs of query node.
2. Merge the pair of most similarity query nodes that does not contain the same query from different users. Concept node c is connected to both query nodes.
3. Obtain the similarity scores for all possible pairs of concept node.
4. Merge the pair of concept nodes.

Community Merging:

1. Obtain the similarity scores for all possible pairs of query node.
2. Merge the pair of most similarity query nodes that contains the same query from different users. Concept node c is connected to both query nodes.

IV. DESIGN

In the existing system, user profiling strategies are based on objects that users are interested. User profiling strategy can be either document based or concept based. Here in this search, query is given by the user and the entire related search results are displayed. Here the users have to search the entire data and select the needed data. This increases the time to search. Here is this type of search the query is given by the user, and result for the query is based on the search history. Based on the search history the preference of the user query is checked and the result is given according the search history. Time to search the data is reduced.

In the proposed system, the main focus is on search engine personalization and to develop several concept-based user profiling methods. Concept-based user profiling methods deals with both positive and negative preferences. The concept-based user profiles can be integrated into the ranking algorithm of a search engine so that search result can be ranked according to individual user's interest. To terminate and improve the overall quality of resulting query cluster the agglomerative cluster algorithm being used.

Concept based method provides personalized query suggestions based on a personalized concept based clustering technique. Existing method that provide the same suggestions to all user’s, our approach uses click through data to estimates user’s conceptual preferences and then provides personalized query suggestions for individual user according to user’s conceptual needs. The main aim of this concept based method is that queries submitted to a

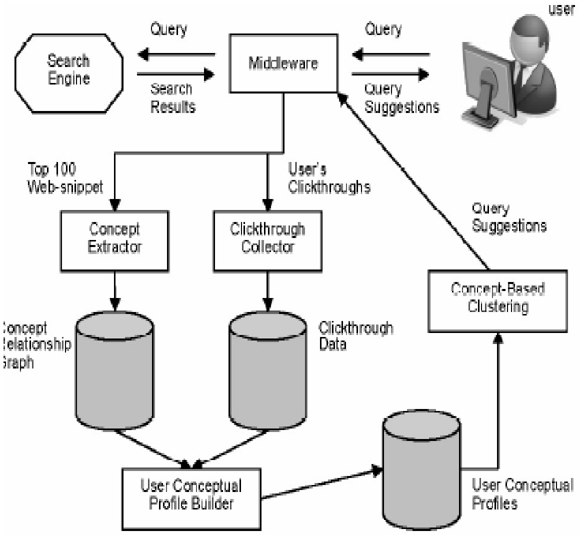

Figure. 2. Concept based clustering

search engine may have multiple meanings. For e.g. depending on the user, the query “apple” may refer to a fruit, the company Apple Computer or the name of a person, and so forth.

The main idea of this proposed technique is based on concepts and their relations extracted from the submitted user queries, the web-snippets, and the click through data.”Web-snippet” denotes the title, summary, and url of a web page returned by search engines. A new two phase personalized agglomerative clustering algorithm that is able to generate personalized query cluster.

Proposed system consists of the following major steps. First, when a user submits a query, concepts and their relations are mined online from web-snippets to build a concept relation graph. Second, click through are collected to predict user’s conceptual preferences. Third, the concept relation graph together with the user’s conceptual preferences is used as input to a concept-based clustering algorithm that finds conceptually close queries. Finally, the most similar queries are suggested to the user for search refinement.

Concept Extraction

Concept extraction method is used for finding frequent item sets in data mining. User submits a query to the search engine, a set of web-snippets are returned to the user for identifying the relevant items. Support formula is used for measuring particular keyword/phrase c_i with respect to returned web-snippets arising from a query q.

Support Formula : -

$$\text{Support } (c_i) = \frac{sf(c_i)}{n} \cdot |c_i|$$

Where n → Total no of web-snippets returned
Sf(c_i) → snippet frequency of the keyword/phrase c_i
|c_i| → no of terms in the keyword/phrase c_i

Query clustering

Query clustering is a technique for discovering similar queries on a search engine. Query clustering method based on the agglomerative clustering algorithm. Agglomerative clustering algorithm can cluster similar queries effectively .Agglomerative clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster.

User profiling

User profiling strategy can be either document based or concept based. SPYNB-C Method is one method to implement the concept based method. This is based on strong assumption that a page scanned but not clicked by user is considered uninteresting to the user. Generate user profiles based on their access patterns.

V. METHODOLOGY

The RSCF (Ranking SVM in Co-Training Framework) algorithm takes the click through data containing the items in the search result that have been clicked on by a user as an input, and generates adaptive rankers as an output. The click through data, RSCF first categories the data as the labeled data set, which contains the items that have been scanned already, and the unlabelled dataset, which contains the items that have not yet been scanned. The labeled data is then augmented with unlabelled data to obtain a larger data set for training the rankers.

VI. CONCLUSION

The user profile improves the search engine's performance by identifying the information needs for individual users. The user's positive preferences were inferred using the mining rules and utilized the preferences in deriving user's profiles. The user profiling strategies were evaluated and compared with the personalized query clustering method. The agglomerative clustering algorithm is employed for finding queries that are conceptually close to one another. The user profiles capturing both the user's positive and negative preferences perform the best among the user profiling strategies. The RSCF makes a search of data containing the item in the search results, the required data is been clicked by the user and this clicked data is given as the input and generates the rankers as the output.

REFERENCES

[1] J.-R.Wen, J.-Y.Nie, and H.-J. Zhang,"Query Clustering Using User Logs,"ACM Trans. Information systems, vol.20, no. 1, pp.59-81, 2002.

[2] Kenneth Wai-Ting Leung and Dik Lun Lee"Deriving Concept-Based User Profiles from search Engine Logs," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 7, July 2010.

[3] K.W.-T. Leung, W. Ng, and D.L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 11, pp.1505-1518, Nov. 2008.

[4] M.speretta and S .Gauch, "Personalized Search Based on User Search Histories," proc. IEEE/WIC/ACM Int'l Conf .Web Intelligence, 2005.

[5] R.Baeze-yates, C.Hurtado, and M.Mendoza, "Query Recommendation Using Query Logs in Search Engines," proc.Int'l Workshop Current Trends in Database Technology, pp.588-596, 2004.

[6] T. Joachims,"Optimizing Search Engines Using Clickthroygh Data,"Proc. ACM SIGKDD, 2002.

[7] Y.Xu,K. Wang, B.Zhang, and Z.Chen, "Privacy-Enhancing Personalized Web search," Proc. World Wide Web(WWW) Conf., 2007.