DESIGN AND IMPLEMENTATION OF QUERY CLUSTERING ON USER SEARCH HISTORY FOR EFFECTIVE INFORMATION RETRIEVAL

Sattari Ajay[1] Hema K L[2]

[1]M.Tech Scholar [2]Assistant Professor

[2]Department of Information Science

[1,2]RNS Institute of Technology Bangalore, Karnataka, India

Abstract— Query clustering is a concept of grouping the related searches based on a predefined set of rules which end up with a most appropriate grouping of the queries. In our day to day life we use search engines like google to search the information, but we end up with the searches which provide an inappropriate results. As the search engine are meant to be user friendly, the overhead of grouping of the queries should be the task of the search engines rather than, distracting the user experience by ending up with unrelated searches for their query. Clustering is the process of organizing the objects in a group which have some common similarities. Query clustering is an unambiguous process which groups the data item sets in a predefined way, where in a query header is given to each of the cluster which determines in which cluster each query should fall in.

Keywords: Query Clustering, Search Engine, click graph, query reformulation

I. INTRODUCTION

Clustering is a useful technique for the discovery of data distribution and patterns in the underlying data. The goal of clustering is to discover both the dense and the sparse regions in a data set. Data Clustering has been studied in the statistics, machine learning and database communities. Clustering is the process of organizing data into meaningful groups, and these groups are called clusters. Clustering can be seen as a generalization of classification. Classification has the knowledge about both the object and the characteristics the object is looking for. So the meaning of classification clearly implies that "Where to put the new object in?". The main advantage of clustering is that the clusters are created dynamically and is more suited for environment such as internet which dynamic in nature. This very nature enables the use of clustering than over the classification technique. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups. The instances are thereby organized into an efficient representation that characterizes the population being sampled.

Now a days it has been so often that all the information that the user needs through Online is performed initially by giving the *query*. Queries are the keywords given by a user when searching Online. Through the initial query given by the user, he/she may lead to enter multiple queries which tends to yield better results. The results yielded might be of three types i) Redundant ii) Relevant and iii) Irrelevant results. *Redundant* results are the results that are quite nearer to the search we are expecting i.e we can consider them as query suggestions. *Relevant* results are the exact results that the user is needed i.e these are the accurate results or information for the query or keyword given by the user. Irrelevant results are the results which are not at all nearer to the keyword the user has specified i.e these are the results that the user is not expecting.

From the survey we got to know that 25-30% of the results are relevant results. So now based on the history or session of the user search through queries we will form a Cluster. This type of clustering through the keywords or queries entered by the user is called as *Query Clustering.*

Clustering being one of main technique, used in the grouping of similar instances/objects by looking at the context in which it's being used. Hence some sort of measure is used to determine whether two objects are similar or dissimilar is required. There are two main type of measures used to estimate this relation: distance measures and similarity measures. The distance measure is calculated based on observing the distances between the observed variables and then applying one of the standard clustering algorithms to these distances .The main advantage of these observations is that, with more similar patterns of responses on the variables of interest are seen as closer to one another than are those with more disparate response patterns. Query Clustering plays a major role in the information retrieval process of searching online. Query Clustering is grouping the similar queries into groups i.e all the similar related queries should be made as a group forming a cluster. As the grouping is done mainly on the user history of the relevant results, the searching can become faster comparatively.

Query Clustering should be performed in an dynamic and automated fashion. Whenever the user enters the query, if the query group already exists related to that keyword, then the keyword entered should be placed in that group. If the query entered doesn't exist in the user history then it should automatically create a new query.

Query clustering is mostly based on the user clicks of the URLs related to the queries or keywords they searched i.e Sessionazination concept is used.

II. RELATED WORK

Many authors classified and observed that clustering can be considered the most important unsupervised learning problem as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way" [1].

Kenneth Wai-Ting Leung said that Clustering is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data [2]. Clustering algorithms can be applied in many fields, for instance:

- Marketing*:* finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.
- Biology*:* classification of plants and animals given their features.
- Libraries*:* book ordering.
- Insurance*:* identifying groups of motor insurance policy holders with a high average claim cost, identifying frauds.
- City-planning*:* identifying groups of houses according to their house type, value and geographical location.
- Earthquake studies*:* clustering observed earthquake epicenters to identify dangerous zones.
- WWW*:* document classification, clustering weblog data to discover groups of similar access patterns.

The main requirements that a clustering algorithm should satisfy are:

- Scalability.
- Dealing with different types of attributes.
- Discovering clusters with arbitrary shape.
- Minimal requirements for domain knowledge to determine input parameters.
- Ability to deal with noise and outliers.
- Insensitivity to order of input records.
- High dimensionality.
- Interpretability and usability.

There are a number of problems with clustering. Among them:

- Current clustering techniques do not address all the requirements adequately (and concurrently).
- Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
- The effectiveness of the method depends on the definition of "distance" (for distance-based clustering).
- If an obvious distance measure doesn't exist it must "define" it, which is not always easy, especially in multi-dimensional spaces;
- Result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Authors also said that many search engine applications such as query recommendation require query clustering as a pre-requisite to function properly and clustering is necessary to unlock the true value of query logs [2]. Clustering search queries effectively is quite challenging, due to the high diversity and arbitrary input by users. Search queries are usually short and ambiguous in terms of user requirements [3]. Many different queries may refer to a single concept, while a single query may cover many concepts. Existing prevalent clustering methods, such as K-Means or DBSCAN cannot assure good results in such a diverse environment [4]. Agglomerative clustering gives good results but is computationally quite expensive [5]. In this paper author presented a novel clustering approach for diverse queries based on the ranked URL results returned by a search engine for queries. Experimental results demonstrated that more accurate clustering performance, better scalability and robustness of the approach against known baselines [6]. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users [7].

III. PROPOSED WORK

Organizing the query groups within a user's history is challenging for a number of reasons. First, related queries may not appear close to one another, as a search task may span days or even weeks. This is further complicated by the interleaving of queries and clicks from different search tasks due to users' multitasking, opening multiple browser tabs, and frequently changing search topics.

This type of organizing user search histories into groups is impractical for two reasons. First, it may have the undesirable effect of changing a user's existing query groups, potentially undoing the user's own manual efforts in organizing the history. Second, it involves a high computational cost, since it have to repeat a large number of query group similarity computations for every new query.

So to avoid these reasons and to overcome manual query grouping methods we are proposing a dynamic query grouping methods in this paper. The architecture of the proposed work is as shown below.

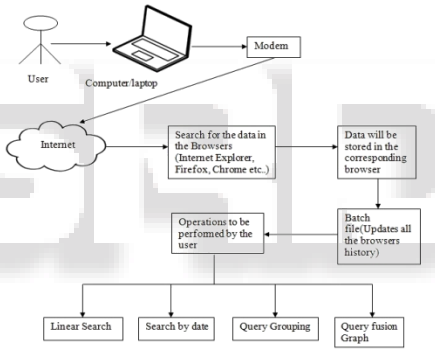

Fig. 1: System Architecture

The above figure shows the whole system architecture of the paper. It shows that the user is connected to the internet via a modem. When the user is connected to the internet, the user performs the browsing of data in different browsers. The data searched by the users will be stored in the histories of the corresponding browsers. In this paper, the entire browser's data will be fetched by using a batch file. When the batch file is clicked it automatically updates the browser's data required for the paper.

This paper has four modules like Linear Searching, Search by Date, Query grouping and Query click graph. When the browsers data is updated, the user can start performing the above operations.

In the Linear Search module user will fetch the history from all the browsers which is arranged in a sequential (Linear) way. It contains the following details like URL, Title searched Visit time, Visit from, Browser name.

In the Search by Date module groups are created based on the particular day. In this module we are fetching the history based on the date. For a Particular day we can fetch the history from all the browsers. It contains the details like GROUP STARTED ON= "Date", URL, Title searched Visit time, Visit from, Browser name.

In the Search by Group i.e Query Clustering module groups are created based on the particular similar data. In this module we are fetching the groups based on the similar data i.e all the similar data are made as one group and it will be dynamically updated as the user history changes. For example all the data regarding to gmail are made as a group. It contains the details like GROUP STARTED ON= "Query", URL, Title searched Visit time, Visit from, Browser name.

In the Query click graph module it generates a graph based on the queries given by the user in the browser. On the X-axis query groups are shown whereas on the Y-axis represents the count of the query group.

IV. EXPERIMENTAL RESULTS

This section briefly shows the presentation of the results and discussions of those results. This section analyses the results of the experiment went as expected with no unusual events that could have introduced error. The following figure shows the query click graph.

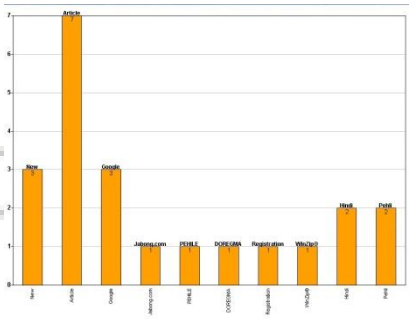

Fig. 2: Query click graph

V. CONCLUSIONS

In this paper we have seen that from the Linear Search Module, browsing history can be fetched from all the browsers at a time in a sequential fashion. By this the time complexity is reduced in opening each browser for their corresponding history. At a time sequentially the data from all the browsers will be displayed. From this paper it is observed that by the Search by Date Module, browsing history is fetched from all the browsers at a time and they automatically create a group based on the particular day. For a particular day the user can get to know the browser history details from all the browsers. From the Search by Group Module, browsing history is fetched from all the browsers at a time and query groups are created dynamically. Finally the query click graph clearly shows the users keen of interest by query count specified.

ACKNOWLEDGMENT

I take this Opportunity to express my profound gratitude and deep regards to my guide Ms. Hema K L, Assistant Professor, RNS Institute of technology, Bangalore, for her exempla nary guidance, and constant encouragement throughout.

I would also like to thank Director Dr. H N Shivashankar, Principal Dr. M K Venkatesha and Dr. M V Sudhamani, professor and Head, Dept of Information Science and engineering, RNSIT, for constant encouragement in implementing this paper and pursuing this paper.

REFERENCES

[1] Yuan Hong, Jaideep Vaidya and Haibing Lu, Search Engine Query Clustering using Top-k Search Results, MSIS Department and CIMIC, Rutgers University One Washington Park, Newark, NJ 07102, USA, 1998.
[2] Adam L. Kaczmarek, Interactive Query Expansion With the Use of Clustering-by-Directions Algorithm, IEEE Transactions on Industrial Electronics, vol. 58, no. 8, Aug 2000.
[3] Kenneth Wai-Ting Leung, Wilfred Ng, and Dik Lun Lee, Personalized Concept-Based Clustering of Search Engine Queries, IEEE Transactions on knowledge and data engineering, vol. 20, no. 11, Nov 2000.
[4] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE,Weiyao Lin, Member, IEEE, 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, Mar 2004.
[5] Doug Beeferman, Adam Berger, Agglomerative clustering of a search engine query log, IEEE Transactions on knowledge and data engineering, vol. 16, no. 7, Apr 2004.
[6] Jeonghee Yi Farzin Maghoul, Query Clustering using Click-Through Graph, Yahoo ! Inc. Mission College Blvd. Santa Clara, CA 95054 USA, Jul 2005.
[7] Eldar Sadikov Jayant Madhavan Lu Wang Alon Halevy, Clustering Query Refinements by User Intent, ACM Comput. Surv., 31(3), 2006.

