International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-5, April 2013

A Taxonomy of Web Search Using Search History Clustering Mechanism

Kala O. S, R.Premkumar

Abstract—The size and richness of information available on the web growing very rapidly. To this end the users are trying to accomplish more complex task through online. The users can break down the complex tasks into a few co-dependent tasks and issue as multiple queries around these tasks. Search engines are the primary means of accessing information through online.While searching, the search engine can keep their old queries and clicks. Grouping of related queries in the search history is useful for a variety of search engine applications.Query grouping allows the search engine to better understand a user's session and tailor that user's search experience according to their needs. Hence this system presents a mechanism that automatically identifies query groups in the search history.

Index Terms— search history, query group, search behavior graphs, query reformulation.

I. INTRODUCTION

As the increased demand for more precise information retrieval devices, a new generation of search engines have appeared on the internet.These new systems attempting to grasp the user's question so as to recommended similar queries that others have asked and that the system has the proper answers. The experiments on real web search informations indicates that the algorithms developed are effective in rising the search accuracy for each recent and continual queries. Additional accuracy is achieved when using click through data of past searches that are associated with the the present question. Users searching of information on the web is not only informational but also navigational.It may be transactional also. "Provide the URL of the website that I need to search" is an example of navigational query.The purpose of navigational queries is to provide the entry page of a particular website. The users are really interested in finding a document that offers the service described in a transactional query. Todays web search engines often complement the search results with a list of related search queries. The users are often provide very small queries with little or no context.However the related queries allow users to specify their information needs. The related searches are either presented at the bottom or top of the search results page or as a navigation bar on the left. If the related queries are grouped together the user will get a better search result. For example, the Bing search engines have a search history feature.It allows the users to track their online searches by recording their queries and clicks. The users can view this search history as well as they can manipulate it.

Manuscript published on 30 April 2013.
*****Correspondence Author(s)
Kala O.S, Student of M.E., Department of Computer Science & Engineering, Annai Mathammal Sheela College of Engineering, Namakkal, Tamilnadu, India
Mr. R.Premkumar, Asst. Professor, Department of Computer Science & Engineering, Annai Mathammal Sheela College of Engineering, Namakkal, Tamilnadu, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

A query group is an ordered collection of queries together with the corresponding set of clicked URLs. Query grouping allows the search engine to better understand the user's requirements and the user will get a better search result. When a new query comes the search engine can find the related query group. For example, if a search engine identifies that the financial statement and Bank of India are related queries, then the user will get a good search result instead of the Wikipedia article on financial statement or the pages related to other banks. Query grouping can also assist other users by promoting task-level collaborative search.

Organizing groups of related queries has applications beyond helping the users to make sense and keep track of queries and clicks in their search history. First, the query grouping allows the search engine to better understand a user's session and potentially tailor that user's search experience according to their needs.

After identifying the query groups, the search engines can have a good representation of the search result behind the current query using queries and clicks in the corresponding query group. It helps the search engines to improve the quality of key components such as query alterations,suggestions, result ranking and collaborative search.

II. LITERATURE SURVEY

The following are some of the papers reviewed to get an idea of the different systems existing in the relevant area. Each query group is a collection of queries issued by the same user around a common information. This paper studies the problem of arranging a user's search history into a set of query groups.An online mechanism is used to identify the related queries.These queries may have some relevance between each other. These query groups are dynamically updated as the user issues new queries. Then new query groups may be created over time. In the earlier systems the query grouping is based on text and time.

The paper, "A Web based Kernel Function for Measuring the Similarity of Short Text Snippets" finding the similarity of short text snippets,like search queries.There are only a small amount of similarity between two short snippets.So it works poorly with traditional document similarity measures . In this paper each snippet is employed as a query to a web search engine so as to seek out a variety of documents that contain the terms with in the original snippets. It then use these documents to make a context vector for the original snippet.The context vector contains several words that tend to occur in context with the original query terms. Such context vectors may be rather more robustly compare with a measure such as the cosine to determine the similarity between the original text snippets.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The users of web search engine can perform multitasking in two different ways. First, a user could begin their searching with multiple topics. The second begins with one topic and so develop further topics throughout the searching process. These two processes embody information task switching or switching back and forth between totally different topics throughout the sessions of search. For example, a user may switch between seeking financial information and new technology information as they think and work on multiple information problems concurrently.

To identify related queries, text similarity has been proposed in prior works. This method uses the overlap of terms of two queries to detect changes in the topics of the searches.It also identifies the refinement classes based on the keywords in queries, and attempted to predict these classes using a Bayesian classifier.

Text similarity may work in some cases.It may fail to capture cases where there is "semantic" similarity between queries (e.g., "iPod" and "apple store") but no textual similarity. This paper first describes the initial data set and the methodology for transforming the data into a representation of user behavior with richer semantics. It reviews a set of definitions and informational goals that abstract queries into classes of query refinement. It then describes the construction of Bayesian network models that capture dependencies among variables of interest.

In the paper, Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs-automatically organize a user's search history into query groups. Each query group containing one or more related queries and their corresponding clicks.

Each query group corresponds to an specific information need that may require a small number of queries and clicks related to the same search goal. For navigational queries, a query group may contain as few as one query and one click. In the case of informational queries, a query group contain few queries and clicks. Most of the analysis of internet search relevance and performance measures takes only one query for the unit of the interaction of search engines.

In the article, Defining a session on web search engines, to obtain tasks at a variety of measures, users issues multiple queries to search engines. In most cases, the segmentation was based on a timeout threshold .Some of them are looked at the segmentation of a user's browsing activity, and not search activity. But time is not a good basis for identifying query groups.The users can perform multitasking when searching online thus resulting in interleaved query groups.

In the paper, Multitasking during web search sessions, the users of web search engine use information multitask in two ways. Initially, a user may begin their web search with multiple topics.The second begin with a single topic and then develop additional topics during the search process. Both processes include switching back and forth between different topics and information task switching during a search session.

In the paper, The query flow graphs: model and applications- search history in the search engines can record the queries and the actions of the users. They contain valuable data concerning the preferences, interests and therefore the behavior of the users, furthermore their implicit feedback to the results of search engines.

Getting the wealth of knowledge offered with in the search logs has several necessary applications including user profiling and personalization, query log analysis, query recommendation, and more. This paper provides query flow graph, in which it illustrates the interesting knowledge about latent querying behavior.

III. EXISTING SYSTEM

In the existing system, for identifying the query groups, every query in a user's history is treated as a singleton query group.Then merge these singleton query groups in an iterative fashion. It is important to have a suitable relevance measure between the current query singleton group and an existing query group.This is used to ensure that each query group contains closely related and relevant queries and clicks. Two commonly used query relevance measures are time and text.

A. Time Based Similarity

One may assume that two queries are relevant if they appear close to each other in time in the user's history. In other words, one may assume that users generally issue very similar queries and clicks within a short period of time. Then we can define a time-based relevance metric sim_{time}.

Definition: sim_{time} (sc,si) is defined as the inverse of the time interval, in seconds, between the times that qc and qi are issued.

$$1/\text{simtime}(sc,si) = 1/|\text{time}(qc) - \text{time}(qi)| \quad (1)$$

sc is the current query singleton group and si is an existing query group .The queries qc and qi are the most recent queries in sc and si. Higher sim_{time} values indicates that the queries are temporally closer.

B. Text Based Similarity

On a different note, we may assume that two query groups are similar if their queries are textually similar. Textual similarity between two sets of queries can be measured by metrics such as the fraction of overlapping words (Jaccard similarity) or characters (Levenshtein similarity).

Jaccard similarity : $Sim_{jaccard}$(sc,si) is defined as the fraction of common words between qc and qi .

$$\text{Simjaccard}(sc, si) = \frac{|\text{words}(qc) \cap \text{words}(qi)|}{|\text{words}(qc) \cup \text{words}(qi)|} \quad (2)$$

Levenshtein similarity is defined as follows.

Sim_{edit}(sc,si) is defined as 1 _ $dist_{edit}$(qc,qi). The edit distance $dist_{edit}$ is defined as the total number of character substitutions, insertions or deletions required to change one sequence of characters into another.It is normalized by the length of the longer character sequence.

C. Limitations of existing system

In some cases the above time-based and text-based relevance metrics may work well. They cannot capture certain aspects of query similarity. For example, sim_{time} assumes that

a query is always followed by a related query. This may not be the case when the user is multitasking .In multitasking, the users having more than one tabs open in their browser.The users can digressing to an irrelevant topic and then resuming their searches.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

In the text based similarity , $sim_{jaccard}$ and sim_{edit},is used as relevance metrics.They can capture the relevance between query groups around textually similar queries. But in some cases the related queries may not have similar words. So, this is impractical in this scenario for two reasons.

The first reason is that it may have the undesirable effect of changing a user's existing query groups.It potentially undoing the user's own manual efforts in organizing their history. Second, it involves a high computational cost, because we would have to repeat a large number of query group similarity computations for every new query. Therefore, it needs a relevance measure to identify similar query groups beyond the approaches that simply rely on the textual content of queries or time interval between them.

IV. PROPOSED SYSTEM

In the proposed system the query relevance measure is based on web search logs. The query relevance is measured by considering two important properties of relevant queries:

1) Frequently appeared queries are reformulations of each other.
2) The queries that have elicited by the users to click on similar sets of pages.

Therefore the relevance is measured from search behavior graph.The proposed system approaching a new way of search strategies .A new mechanism called"online query grouping" is used in the proposed system.

A. Search Behavior Graphs

From the search logs of a commercial search engine, we can create three types of graphs. These are query reformulation graph,query click graph and the query fusion graph.The query reformulation graph, QRG, represents the relationship between a pair of queries that are reformulations of each other. The query click graph, QCG, represents the relationship between two queries that frequently lead to clicks on similar URLs. The information in the query reformulation graph and query click graph is merged together to form a query fusion graph(QFG). All three graphs are defined over the same set of vertices. All these graphs consisting of queries which appear in at least one of the graphs.The edges in the graph are defined differently.

So the proposed system investigates how signals from search logs such as query reformulations and clicks can be used together to determine the relevance among query groups. In order to enhance this process we study two ways of using clicks.The first method fusing the query reformulation graph and the query click graph into a single graph that we refer to as the query fusion graph.The second method expanding the query set when computing relevance to other queries with similar clicked URLs.

B. Advantages of the Proposed System

The proposed system uses the signals from search logs such as query reformulations and query clicks can be used together to determine the relevance among query groups. So Internet users easily can able to find out their requirements as per entering their queries. The proposed system has another advantage that it highly support users in their long-run data quests on the web. So the search engines can keep track of their recent queries and clicks whereas searching online. The proposed system uses query grouping mechanism. So the searching manner will become more flexible to users.

V. SYSTEM IMPLEMENTATION

The following sections depict several implementation modules in building the new system. This section also presents an analysis of the paper in terms of the module elaboration. The system is analyzed by dividing it into separate sections. The main sections of the system are the query grouping and the query relevance and search logs.

A. Query Grouping

This section again divided into two sub sections. They are creating search history and query clustering. In the Creating Search history, any personal documents such as browsing history and emails on a user's computer could be the data source for user profiles. This focus on frequent terms limits the dimensionality of the document set, which further provides a clear description of users' interest. Fig(1) shows the search logs created by the user. This section allows the search engine to better understand a user's search session and tailor the user's search experience according to their needs.

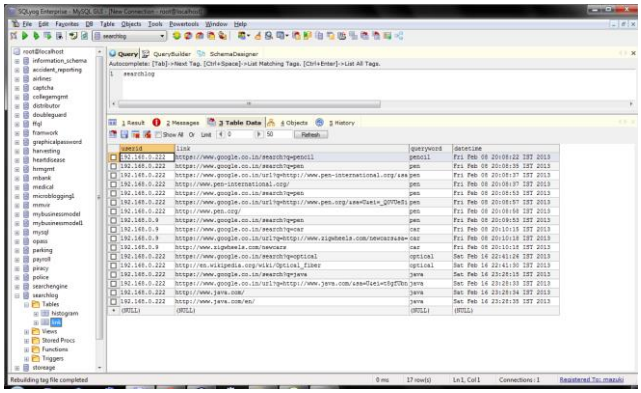

Fig. 1 search history created by the user

In the query clustering module, the queries that is issued by the user can be classified into different query clusters. To achieve personalization effect, concept based user profiles are employed in the clustering process. Merge the most similar pair of concept nodes and the most similar pair of query nodes, and so on.Each individual query submitted by each user is treated as an individual node and each query with a user identifier. Fig(2) shows the clustering of queries in the search history created by the user. After identifying the query groups the search engines can have a good representation of the search result in the corresponding query group.

For identifying similar query groups we need a relevance measure that is beyond the approaches that simply rely on the textual content of queries or time interval between them. The proposed system makes use of search logs in order to determine the relevance between query groups more effectively.

In this section the items in the search logs of a user is grouped.For identifying groups,it is first treat every item in a user's history as a query group.Then merge these query groups in an iterative fashion .For this grouping k-means algorithm is used. k-means clustering is a method of cluster analysis which partition the search history into *k* clusters in which each item belongs to the cluster with the nearest mean. The algorithm is iterative in nature. This results into a partitioning of the data space into seperate cells.The grouping can be performed in a dynamic fashion.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

For grouping we first place the current query and clicks into a query group. The k-means clustering procedure is implemented by taking a single query in the search history as the initial group.When a new query comes ,it identifies the most relevant group .

The query relevance is measured from the search behavior graphs.In one iteration, consider all of the cases in relation to each of the clusters.From the search behavior graphs a context vector is created for each query groups.Each value in the context vector reflects the relevance of other query groups to this query group.

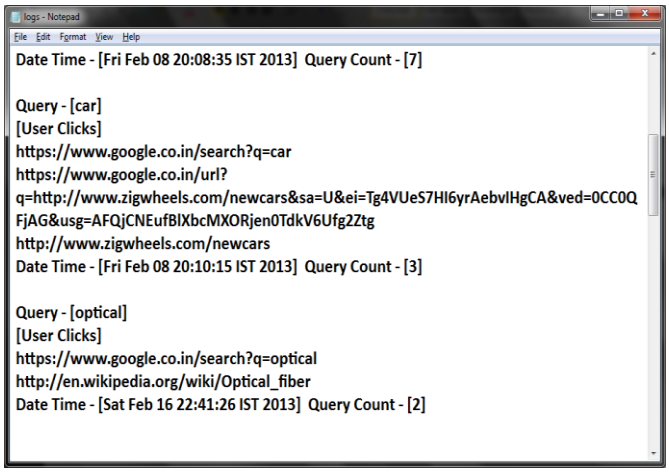

Fig.2 clustering of queries in the search history

The search history of a large number of users contains signals regarding query relevance,such as query reformulations and query clicks. Query reformulations means which queries tend to be issued closely together, and query clicks means which queries tend to lead to clicks on similar URLs. Such signals are user generated and are likely to be more robust. For measuring the relevance between query groups we exploiting the query logs and the click logs simultaneously.

B. Query Relevance and Search Logs

This module explains how to identify the relevance measure that is robust enough to identify similar query groups beyond the approaches that simply rely on the textual content of queries or time interval between them. Based on Web search logs the query relevance is measured.

In order to determine the relevance between query groups more effectively, the search logs are used. The search history of a large number of users contains signals regarding query relevance, such as query reformulations and query clicks. The aforementioned properties are captured by introducing three search behavior graphs.The search behavior graphs used for determining the relevance are query reformulation graph,query click graph and the query fusion graph.

In the query reformulation module, it is important to have a suitable relevance between the current query groups to determine that each query group contains closely related and relevant queries and clicks. One may assume that users generally issue very similar queries and clicks within a short period of time. The search history of a large number of users contains signals regarding query relevance. This captures the relationship between the frequently issued queries and clicks on similar URLs. The query reformulation graph and the query click graph from search logs are used to determine the relevance between queries or query groups within a user's history.

The web is not a well-organized information source where in numerous "authors" create their websites independently. Those authors "vocabularies" vary greatly. Moreover, most words in the natural language have inherent ambiguity. These reasons make it rather difficult for the web users to formulate queries with appropriate words. Many web search engines, such as AltaVista (http://www.altavista.com), Excite (www.excite.com), Lycos (www.lycos.com), etc., attempt to identify some of the users' intentions and suggest a list of alternate terms for the user to reformulate their query.

Therefore this term suggestion mechanism needs to track the users emerging topics of interest. For instance events in the news or newly publicized web sites. Query clustering, obviously which intends to find related queries from users' daily logs, could be one of the critical techniques to drive this term suggestion process. For any clustering problem, researches have been concerned mainly with two aspects — similarity functions and algorithms for the clustering process.

VI. CONCLUSION

The query reformulation and click graphs contain useful information on user behavior when searching online. This paper shows how such information can be used effectively for the task of organizing user search histories into query groups. More specifically, this proposes combining the two graphs into a query fusion graph. Further it shows that this approach that is based on probabilistic random walks over the query fusion graph outperforms time-based and keyword similarity-based approaches. It also finds value in combining our method with keyword similarity-based methods, especially when there is insufficient usage of information about the queries.

The query clustering process is also a batch process that can be accomplished offline. As a future work it is used to intend for the investigation of the usefulness of the knowledge gained from these query groups in various applications such as providing query suggestions and biasing the ranking of search results. Another enhancement of this paper is that it can apply the idea to increase the performance of cache memories.

REFERENCES

1. M. Sahami and T.D. Heilman, "A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. the 15th Int'l Conf. World Wide Web 2006."
2. 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, 2008.
3. B.J. Jansen, A. Spink, C. Blakely, and S. Koshman, "Defining a Session on Web Search Engines: Research Articles," J. the Am. Soc. for Information Science and Technology, 2007.
4. A. Spink, M. Park, B.J. Jansen, and J. Pedersen, "Multitasking during Web Search Sessions," Information Processing and Management,2006.
5. P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, "The Query-Flow Graph: Model and Applications," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM), 2008.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)
© Copyright: All rights reserved.