Open access Journal International Journal of Emerging Trends in Science and Technology

Dynamic Query Clustering in Personalized Search to Improve Retrieval Relevance

Authors

Bethapudi Haritha[1], K Mohana Krishna[2]

[1]M.Tech.Programme Student of Vasireddy Venkatadri Institute of Technology
Nambur (v), Guntur, Andhra Pradesh.
Email: *haritha.feb11@gmail.com*
[2]Asst Professor of Vasireddy Venkatadri Institute of Technology
Nambur (v), Guntur, Andhra Pradesh.
Email: *Mohank1987@gmail.com*

Abstract

Today internet users are using web search engines for complex generic query processing to achieve the day to day activities like trip management, budget planning, shopping plan and text similarity etc. To avoid the complex generic query management many search engines are using query modulators, to break the main query in smaller sub queries to reduce the complexity and to extract the relevant data as results. Some search engines also maintain the user level search history customization to help the user by suggesting them. To achieve more efficiency in personalized search, in this paper we are introducing dynamic clustering in personalized search to assist the user search and to improve the precision of search relevance. This clustering is also useful to find result ranking, relevance ratio and result collaboration. Experimental results are showing that our approach is having the high precision and recall in terms of search relevance and scalable in terms of response time than other approaches.

Keywords: *web mining, personalized search, dynamic query clustering, search history analysis, query result processing.*

1. Introduction

In this decade, internet usage is dramatically increased due to the wide availability and adoption of customers. Almost every electronic device is allows surfing the web to extract the useful information by mining various web applications. Users are improved from using the simple web queries to complex web queries to extract the relevant information from web environment. Personalized search history is the innovative idea to assist web user to write queries feasibly with private search suggestions. Most of the current search engines like Google and Yahoo also following the same tradition, to achieve the more accuracy in retrieving result relevance. But today simple keyword search is unable to retrieve the all rigid web data by using simple queries and keyword search. As per the search navigation [5, 6] survey, today simple web search techniques can navigate only 60% of web data and this will fails to achieve the accuracy with image and video data. To accomplish the task achievement sake today web navigators are preparing more complex queries. For example to navigate the relevant information from web to search for complex task Health Management which includes check-up schedules, health records, disease information, hospital locations, medicine usage etc. User cannot keep all these relevant queries ina single search query to extract the relevant information. In this case some topic relevance search implementation is required at web search engine level to cluster the relevant data and to display the topic relevance information also as result to user. This approach will reduce the search burden at user level dramatically and helps him to know the more relevance information about the complex generic queries.

Personalized search is an emerging technique to improve the search result relevance at every user level by storing the user search data at web logs as private to that user. This new features are useful, the manual efforts involved can be disruptive and will be untenable as the search history gets longer over time. In this paper we are introducing dynamic clustering in personalized search to assist the user search and to improve the precision of search relevance. This clustering is also useful to find result ranking, relevance ratio and result collaboration. Initially search engines will analyze the topic relevance from various search topics to cluster all relevant topics and to display them as results to the user complex generic query. This process also may concentrates on public user search topic relevance to migrate with base clusters and private clusters of a user. For example a user given the complex generic query like health management to know the relevant information about the human health management, which may shows the relevant n results as output. Like that, if we consider the more users who are searching for the same they may go for some more relevant topics means the supportive results. By considering all these results for the same query we can create a public topic relevance result cluster. In another way, we are also targeting at a specific user level with personalized search and monitoring this user relevance sub topics for the same generic query and making this as a private topic result relevance cluster to a user. Oncea user given the query again our approach will retrieve the merge of private and public clusters with a predetermined high priority. This approach not only suggests the personal relevance, but also concentrates on public relevance. Experimental results are showing that our approach is having the high precision and recall in terms of search relevance and scalable in terms of response time than other approaches.

2. Related work

In this section we discuss the implementation and advantages of the personalized search, search history and query clusters in detail.

Personalized Search: Personalized search will concentrates at user level to extract the user relevant information by mining web applications. In this study we concentrated at online web mining related to user interest. This search will create the separate log file environment for every registered user. Initially this will result the public results to user search and monitors the user interests based on the selected results among the published results. Most of the search engines are implementing this feature at server side to reduce the burden of user. For a given query Q results sake first search engine will extract the all relevant data results from public perspective, than swaps the order of results displaying based on personalized search result priority. We can implement the personalized search at client side and server side to create private search logs for each user individually. Client side search [3, 4] is having the cookie acceptance problem; personalization is applicable from only one single machine, global search patterns etc. Server side search is also having security, storage and maintenance problems. Storing the user personal information at server side may create the data theft and misuse, Server side we need more secondary memory to keep the each user search log separately apart from public search logs.

Search History: Tracking the search history is an important aspect for search engine customization and personalization. Today all search engines are tracking the search history[6, 9] for introducing new search mechanisms and improving the scalability in existing approaches. In this case we are using the search engine logs to store the information at search engine level and these semi structured log data will be used to find the interest of user by monitoring the search queries. Some search engines are supporting the unstructured and structured queries both for keyword based search at search engine level. Unstructured queries example is Google web search and structured web search example is Google news in present. In search history logs every query, retrieved results, results order and selected results are stored in a updatable manner. This information is useful in terms of calculating result efficiency, relevance factor, result ranking, result clustering and interest mining etc.

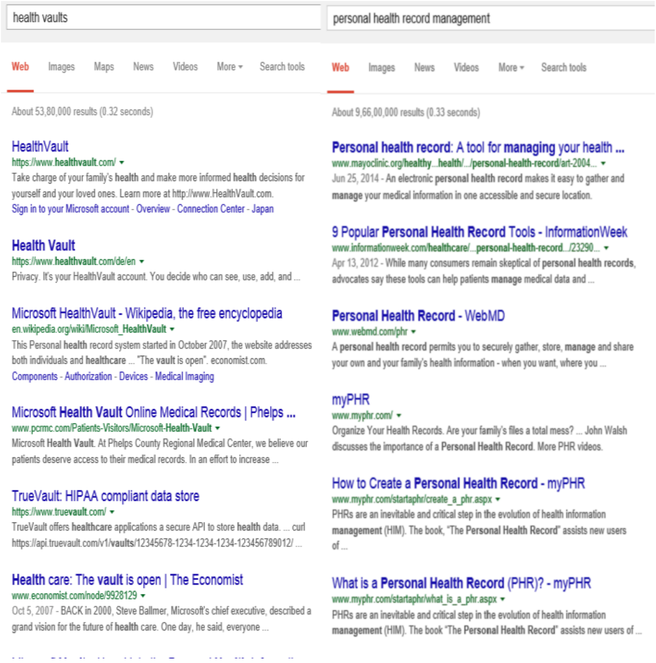
Query Clusters: In order to improve the result efficiency for a given query, present search engines are using the query grouping or clustering technique. This clustering will done automatically by the time of web search is happening is called as dynamic clustering. Dynamic clustering is a process of creating the clusters based on topic relevance in terms of user search criteria at every user level. Each cluster is having a set of interrelated queries from the same user and other web users to mine the interest of individual web user. These clusters will be updated automatically as per the user keyword based query search, based on specified time periods. There are many problems we may encounter by the time of clustering are topic relevance, query similarity, subject relevance and time line considerations. in this paper we are introducing private user search history to address all the above problems while query clustering.

3. Dynamic Query Clustering in Personalized Search

In this paper we are introducing a new mechanism to improve the relevance in keyword based web search [1, 2] based on web search histories. In our terms the relevance means co-existence and topic relevance which is an important criterion from user interest mining. To achieve this relevance in searching we are implemented the given below techniques in a passionate manner. Our main goal is dynamically clustering the user search queries and results based on topic relevance to mine user interest.

Search Relevance Graphs: these graphs are the back bones for personalized search to generate the user query clustering based on topic relevance. In order to get the topic relevance they use to find the query reformulation graphs [9] by removing the stem words [11] from the given query. If two users are given the similar query, our search relevance graphs identify the same relevant clicks for the given query among multiple users. For example, two users are given the two queries like "Health Data Management" and "Health Vaults" with the same intention to know about how to store and manage the personal health records online for remote access. In this case search engine will display the relevance results for each query and the results are shown in fig.1.

From the above results of two queries we noticed that although both queries are same, search engine is retrieving the different results for each query at least not more than 30% relevance. In order to avoid these differences in results we first have to find the associate among the given two queries with help of search relevance graphs. This search relevance graphs will monitor the user clicks for given query results at every query level and traces them with keyword relevance also. After these clicks based information is used to find the relevance among the topics at every query of a user level. At the end of this process after sometime relevant topics will be ready at each user level and query level.



About 53,80,000 results (0.32 seconds)

HealthVault
https://www.healthvault.com/
Take charge of your family's **health** and make more informed **health** decisions for yourself and your loved ones. Learn more at http://www.HealthVault.com.
Sign in to your Microsoft account - Overview - Connection Center - Japan

Health Vault
https://www.**healthvault**.com/de/en
Privacy. It's your HealthVault account. You decide who can see, use, add, and ...

Microsoft HealthVault - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Microsoft_**HealthVault**
This Personal **health** record system started in October 2007, the website addresses both individuals and **healthcare** ... "The **vault** is open". economist.com.
Components - Authorization - Devices - Medical Imaging

Microsoft **Health Vault** Online Medical Records | Phelps ...
www.pcrmc.com/Patients-Visitors/Microsoft-**Health-Vault**
Microsoft **Health Vault**. At Phelps County Regional Medical Center, we believe our patients deserve access to their medical records. In an effort to increase ...

TrueVault: HIPAA compliant data store
https://www.truevault.com/
TrueVault offers **healthcare** applications a secure API to store **health** data. ... curl https://api.truevault.com/v1/**vaults**/12345678-1234-1234-1234-123456789012/ ...

Health care: The **vault** is open | The Economist
www.economist.com/node/9928129
Oct 5, 2007 - BACK in 2000, Steve Ballmer, Microsoft's chief executive, described a grand vision for the future of **health** care. One day, he said, everyone ...

About 9,66,00,000 results (0.33 seconds)

Personal health record: A tool for **managing** your health ...
www.mayoclinic.org/healthy...health/.../personal-health-record/art-2004...
Jun 25, 2014 - An electronic **personal health record** makes it easy to gather and **manage** your medical information in one accessible and secure location.

9 Popular **Personal Health Record** Tools - InformationWeek
www.informationweek.com/healthcare/...personal-health-record.../23290...
Apr 13, 2012 - While many consumers remain skeptical of **personal health records**, advocates say these tools can help patients **manage** medical data and ...

Personal Health Record - WebMD
www.webmd.com/phr
A **personal health record** permits you to securely gather, store, **manage** and share your own and your family's health information - when you want, where you ...

myPHR
www.myphr.com/
Organize Your Health Records. Are your family's files a total mess? ... John Walsh discusses the importance of a **Personal Health Record**. More PHR videos.

How to Create a **Personal Health Record** - myPHR
www.myphr.com/startaphr/create_a_phr.aspx
PHRs are an inevitable and critical step in the evolution of health information **management** (HIM). The book, "The **Personal Health Record**" assists new users of ...

What is a **Personal Health Record** (PHR)? - myPHR
www.myphr.com/startaphr/what_is_a_phr.aspx
PHRs are an inevitable and critical step in the evolution of health information **management** (HIM). The book "The **Personal Health Record**" assists new users of ...

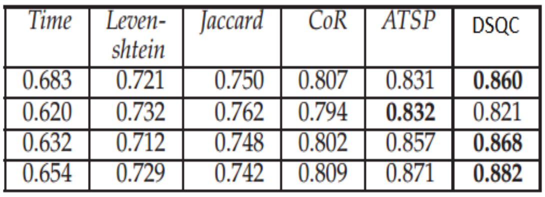
Fig.1. Topic relevance for given queries.

Dynamic Query Clustering: After finding the topic relevance at user level we have to improve the query clustering process based on relevance results. To achieve this we create an image for each query in the search criteria to represent the relevance with other queries also. Query similarity vector is used to integrate the all the relevant images at a single location for public relevance mining. This vector stores the query relevance percentage about each query and supports to find the similarity at end. This vector will add the user clicks to search relevance to find the relevance percentage while searching the data by user. For example the given query q1 to create the relevance group g1 based on the topic relevance factor r1 with number of clicks cn. To cluster this query in real implementation as cluster c1 = {((q1,g1)*r1),((q2,g2)*r2),. . . ((qn,gn*rn)} with R> = 1, is the way to create the cluster. After this for every click on search result will improve the relevance percentage, which is proportional to number of clicks by user. This process will continue with every query dynamically to create multiple query relevance clusters to hold the top relevant results from user clicks.

After creation of clusters from v1 to vn we have to insert the recorded queries as optimal values to that cluster group and sort them by relevance percentage, which is obtained from search relevance graphs [7]. We set the minimum relevance threshold as 78% based on the relevance ration calculations from the above calculations. Increasing the threshold value is caused to add more irrelevant information to results and decreasing the threshold value is caused to remove the relevant results from cluster. In this case our approach can finds the very much associate with respect to relevance factor of clusters and finds the matches among all user's query cluster relevance ratio with all other user's clusters. Finally this is an integrated environment which finds the topic similarity [8], results retrieval relevance at a single scoped location of search engine. This process does not require any external hard ware except to maintain user level search logs, which are a part of every search engine today like Google and Yahoo.

4. Experiments

In this section, we explore the experiments of our work and performance of our results in an experimental manner. For this operation, we considered some complex queries like trip planning, personal health data management and budget planning etc. we created some users and their private accounts to monitor the user search personalization at every user account level. We continued this process for three weeks to analyze the search relevance and dynamic clustering mechanism. In this period of testing time we created query clusters for each user query and monitors the relevance updates in every possible level. In this case, finally we created 4 groups to implement dynamic cluster mechanism to mine the user interest as shown in Table 1. To test this approach we used a Linux based intranet university server for three months of time. Total 500 of college students were participated in this experimental research and every search query is addressed through proxy server.



Time	Levenshtein	Jaccard	CoR	ATSP	DSQC
0.683	0.721	0.750	0.807	0.831	**0.860**
0.620	0.732	0.762	0.794	**0.832**	0.821
0.632	0.712	0.748	0.802	0.857	**0.868**
0.654	0.729	0.742	0.809	0.871	**0.882**

Table1. Precision values comparison for four clusters

In order to evaluate the relevance in web search results based on dynamic clustering mechanism, we had taken precision [10] and recall [11] as the measurements. Precision will gives the relevance result ratio based on retrieved number of results and relevant results as shown below.

Relevant results = (Retrieved results ∩ Relevant results)/ Retrieved results

Unless the textual comparison of results we find the result relevance based on user clicks for the selected result option. We deployed a monitoring program to count the user clicks for a given query and to bind the query with relevant results. This process also concentrates on all users search history to find similarity in query result selection to improve the topic relevance and to construct more reliable clusters as shown in below fig.2.

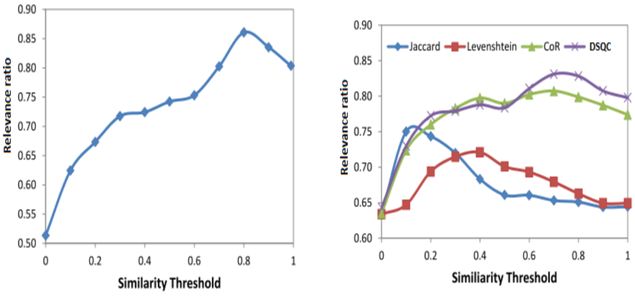


Fig.2. Search query relevance comparison with similarity threshold and other approaches

From the above diagram it is clear that our approach achieved the more similarity and high relevance ration (precision) than other approaches like Jaccrd [11], Levenshtein[12] and CoR [13]. Our approach Dynamic Search Query Clustering [DSQC] is having high relevance ration under all circumstances with other approach threshold levels.

5. Conclusion

Today search engines are moving towards personalized search at user level to mine user interest in a smart way and in short time. Query clustering is an important concern to improve the result relevance in personalized search. Up to now some researches were concentrated on query clustering in personalized search criteria. Topic relevance is a primary concern to success in query clustering which is not separately measured yet. In this paper we are introducing dynamic clustering in personalized search to assist the user search and to improve the precision of search relevance. To find the topic relevance cluster we are using search relevance graphs in this research area to assist the user search intention mining. This clustering is also useful to find result ranking, relevance ratio and result collaboration. Experimental results are showing that our approach is having the high precision and recall in terms of search relevance and scalable in terms of response time than other approaches.

References

1. J. Yi and F. Maghoul, “Query clustering using click-through graph,” in WWW, 2009.
2. E. Sadikov, J. Madhavan, L. Wang, and A. Halevy, “Clustering query refinements by user intent,” in WWW, 2010.
3. A. Z. Broder, S. C. Glassman, M. S. Manasse and G. Zweig, Syntactic clustering of the Web, in: Proceedings of the Sixth International Web Wide World Conference (WWW6), 1997.
4. D. R. Cutting, D. R. Karger and J. O. Pedersen, Constant interaction-time Scatter/Gather browsing of large document collections, in: (SIGIR’93), 1993, pp 126-135.
5. A. Spink, M. Park, B. J. Jansen, and J. Pedersen, “Multitasking during Web search sessions,” Information Processing and Management, vol. 42, no. 1, pp. 264–275, 2006.

6. O. Zamir, Visualization of search results in document retrieval systems, General Examination Report, University of Washington, 1998.
7. M. A. Hearst, The use of categories and clusters in information access interfaces, in: T. Strzalkowski (Ed.), Natural Language Information Retrieval, Kluwer Academic Publishers, 1998.
8. R. Baeza-Yates and A. Tiberi, "Extracting semantic relations from query logs," in KDD, 2007.
9. A. Spoerri, InfoCrystal: A visual tool for information retrieval and management, in: Proceedings of Information Knowledge and Management (CIKM'03), 2003, pp 150-157.
10. T. Radecki, "Output ranking methodology for documentclustering-based boolean retrieval systems," in SIGIR. New York, NY, USA: ACM, 1985, pp. 70–76.
11. F. Radlinski and T. Joachims, "Query chains: Learning to rank from implicit feedback," in KDD, 2005.
12. L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," in Technical report, Stanford University, 1998.
13. P. Boldi, M. Santini, and S. Vigna, "Pagerank as a function of the damping factor," in WWW, 2005.