www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 12 December 2014, Page No. 9576-9580

Survey of **various query suggestion system**

Prajakta Shinde[1], Pranjali Joshi[2]

[1]Pune University, PICT Pune,Dept. of Computer
Dhankwadi, Pune,India
Prajaktashinde5@gmail.com

[2]Pune University,PICT Pune, Dept. of Computer
Dhankwadi,Pune,India
pranjalianshu@gmail.com

Abstract: Now a days, Internet has become essential for banking, education, business and many more applications. Users search information on the web using search engines by clicking on hyperlinks or via keyword queries. So to develop user search intent application is challenging, satisfying increased expectations and diverse needs of user. In this paper, we will survey the various techniques used for query suggestion, personalization and FAQ identification. We will review query suggestion by considering query contents, document clicks, query frequency and semantic features. Thus, by automating the optimization process of searching on web; we can minimize user efforts; maximize user satisfaction for getting desired search.

Keywords: search history, clustering, suggestion and mining

1. Introduction

The effectiveness of information retrieval from the web depends on whether users can issue queries to search engines, which properly describe their information needs. Query logs are considered as rich source of knowledge on user behavior. Many different approaches have been proposed in order to discover essential features in query logs. Query clustering algorithms have become more popular to increase efficiency of search engine.

In this paper we survey query clustering techniques for query suggestion. Query clustering is useful for query log analysis, advertising, query suggestion; personalization and FAQ identification are surveyed. In recent years several query clustering have been proposed to accomplish these tasks. We review many query clustering methods with focus on query contents, document clicks, semantic features, latent term weight, time properties and query frequency.

In table 1 shows several query clustering methods with advantages, disadvantages and application.

The paper is organized as follows: Section 2 presents some literature review on the different query suggestion system; Section 3 contains a table of comparisons of different query suggestion techniques and their advantages, disadvantages and applications; and Section 4 finally concludes the paper.

2. Literature survey

The discussion addresses different query clustering techniques, trying to improve the query suggestion in different ways. The paper surveys different query suggestion techniques used in search engines like Google, Yahoo, Microsoft, Academia etc. This paper also goes through traditional query suggestion methods, showing their pros, cons and applications.

Query suggestion system: Query suggestions are based on users search history.

Query clustering methods based on query contents were used for grouping purpose but that methods are more reliable for short queries than long queries. The next method for query clustering is based on time feature; one can assume that a query is always followed by a related query. However, this may not be the case when the user is having more than one tab open in her browser, or digressing to an irrelevant topic and then resuming her searches.

Document selections are comparable to user relevance feedback in IR environment, except that document clicks denotes implicit and not always valid relevance judgments.

Various approaches have been proposed in recent years that use query logs for query suggestion.

Beeferrman and Berger [2] apply a hierarchical agglomerative clustering technique to click-through data to find clusters of similar queries and similar URLs in a Lycos log. A bipartite graph is constructed from queries and related URLs which iteratively clustered by choosing at each iteration the two pairs

of most similar queries and URLs; but with limitations of noise and small number of common clicks.

The query clustering approach in [16] uses K-Means clustering algorithm but which cannot work that much effectively in query clustering case due to the difficulty on specifying value of k.

Wen et al. [7][8] analyzed query contents as well as click through bipartite graph and applied a density-based algorithm DBSCAN [13] to form cluster of similar queries. Similar to agglomerative query clustering, DBSCAN algorithm requires high computation cost.

In [14], a graph representation of the interesting knowledge about latent querying behavior is done using query flow graph. In the query-flow graph a directed edge from query qi to query qj means that the two queries are likely to be part of the same "search mission". Time and textual properties are considered for grouping so it is not that much efficient. So In [5] for FAQ retrieval clustering of query log is done using latent term weights.

Jeonghee Y [6] introduced click through graph which consider query and clicked page relationship. Query clustering is done on the basis of Query and clicked page relationship, other features are not taken into account.

Ji-Rong Wen [10] introduced query clustering approach using content words and user feedback, combining content and feedback similarity approach so it is efficient but it's difficult to set parameters for linear combination of two similarity metrics.

Yuan Hung, Jaideep V [20] and Kajal Y Yyas [11] used search results for query clustering, similarity based on ranked url results return by search engine this approach is having better scalability.

Toru Onada, Takayaki Yumoto [17] introduced concept of query clustering based on history of query frequency, but its limitation is it is applicable for short terms only.

Lye Limam, David coquil [12] they applied semantic taxonomy to search log and perform grouping to extract user interest which is helpful for personalization application.

In [15] method of personalized query suggestion is introduced using hitting time; semantically consistent query suggestions with respect to current query are given. But how to generate personalized query suggestion is still problem. Solution to this problem is using incremental clustering algorithm [18] takes into account the individual history as streaming on-line sources and presents a personalized search model that can be updated by means of incremental clustering algorithm as new data arrive. It is efficient because two strategies for selecting cluster latest interest and content similar interest is taken into consideration but long term interests are not considered.

A new approach to system-centered evaluation for query suggestion is introduced in [9]; in this a query search procedure constructs queries that rank the document high enough for user to see it; from this set of queries the suggestions is given .

In [3] proposed a novel approach to query suggestion using click-through and session data. Unlike previous methods, this approach considers not only the current query but also the recent queries in the same session to provide more meaningful suggestions. Moreover, for grouping similar queries into concepts and provide suggestions based on the concepts.

In [19] Yang Song proposed a novel query suggestion framework which leverages user re-query feedbacks from search engine logs. Specifically, mining user query reformulation activities where the user only modifies part of the query by add, delete and modify operation and then built term transition graph on mined data. Random walk model is used over term transition graph so it is efficient and scale upto large dataset; but random walk models only focus on query and clicks but ignore the rich information which is embedded in the entire user session.

In previous system we have observed that mining of search log is done for query suggestion but in [1] mining of log is done for optimization of web search results so in this way time user spends for seeking out required information from search result is reduced.

A new method [4] for query expansion based on user interactions recorded in user logs. The central idea is to extract correlations between query terms and document terms by analyzing user logs. These correlations are then used to select high-quality expansion terms for new queries; so aim is to establish correlation between query terms and document terms to narrow gap between query space and document space. Compared to previous query expansion methods, this method takes advantage of the user judgments implied in user logs but this method is effective for short queries than long queries.

In this way, we surveyed various query suggestion method. Query suggestion is very important feature to suggest better keywords to user while searching on web.

3. Table of comparison

Authors	Method	Pros	Cons	Application
DougBeeferman, AdamBerger [2]	Agglomerative clustering	1.Combined approach for similar queries and urls respectively; so it is efficient .	1.Content ignorant. 2.Quite expensive	Query suggestion
Ji-Rong Wen,Jian-Yun Nie,Hong-Jiang Zhang [8]	Query clustering using user log	1.In this , we want to find FAQs, so DBSAN algorithm has filtered out those queries with low frequencies. 2.DBSCAN will not require manual setting to form clusters.	1.Keyword similarity is applied for clustering but will be very inaccurate due to short length of queries.	FAQs identification
Paolo Boldi,Francesco Bonchi,Carlos castillo [14]	Query Flow graph	1.Improving query log analysis 2.Mining user behavior	1. Time and textual properties are considered for grouping so it is not that much efficient.	Query suggestion
Jeonghee Yi,Farzin Maghoul [6]	Click through graph	1. Query and clicked page relationship. 2.Syntactic and semantic features are considered	1.Due to the strict requirement of complete connectedness of the clusters by the algorithm, many potentially interesting query clusters are excluded if they slightly violate requirements.	Query suggestion
Ji-Rong Wen,Jian-Yun Nie,Hong-Jiang Zhang [10]	Using content words and user feedback	1. Content similarity. 2. User Feedback similarity. 3.Combining 1 and 2 so it is efficient	1. Semantic features are not considered.	Query suggestion
Toru Onada,Takayaki yumoto , Kazutoshi sumiya[17]	History of query frequency	1. Grouping based on frequency of query in search log.	1. Not applicable for short term.	Query suggestion
Harksoo Kim ,Jungyun seo [5]	FAQ retrieval based on latent term weights	1. Extracting and representing contextual usage meaning of words for clustering.	2.Too many calculations for calculating latent term weight of queries .	FAQs identification
Lyes Limam,David coquil,Harald kosch,Lionel Brunie[12]	Extracting user interests from search log	1.Semantic taxonomy of search log.	1.Clicks are not taken into account.	Query suggestion , Personalization

Yuan Hung,Jaideep V aidya,Haibing Lu [20]	Using Top –K search results	1. Similarity based on ranked url results return by search engine. 2.Better scalability	-	Query suggestion
Kajal Y. Vyas [11]	Search result rank optimization	1. Optimize search results. 2. It reduces users efforts and seeking time	1.Number of clusters should be specified.	Query suggestion , Search result optimization
Ranieri Baraglia, Carlos Castillo, Debora Donato[16]	Time based query flow graph	1.Recommendation model is always kept updated without reconstructing it from scratch every time, 2.there is no aging effect as incremental algorithm is used	-	Query recommendation
Qiaozhu Mei,Dengyong Zhou[15]	Personalized query suggestion	1.Semantically consistent query suggestion.	1.To generate personalized query suggestion is complex one.	Query suggestion
Shen Jiang, Sundra Zilles[9]	Query suggestion by query search	1.System centered query evaluation.	1. Not applicable for short term.	Query suggestion
Huanhuan Cao,Daxin Jiang[3]	Context aware Query suggestion	1. Not only the current query but also the recent queries in the same session to provide more meaningful suggestions.	-	Query suggestion
Yang Song, Li-Wei he[19]	Query suggestion using term transition graph	1. Random walk model is used over term transition graph so it is efficient and scale upto large dataset.	1. To be specific, random walk models only focus on query and clicks but ignore the rich information which is embedded in the entire user session	Query suggestion
A Sharma,Neelam Duhan[1]	Search result optimization by mining log	1.Seek time is reduced as search result optimization is done.	-	Search result optimization
Xiaochun Wang, Muyun Yung[18]	Incremental clustering for personalized search	1.It is efficient because two strategies for selecting cluster latest interest and content similar interest are taken into consideration.	1. Long term interest is not considered.	Personalization
Hang Cui, Wei-Ying Ma[4]	Query expansion by mining log	1.Aim to establish correlation between query terms and document terms to narrow gap between query and document space.	1.Effective for short queries than long queries.	Query expansion

TABLE 1.A survey of different clustering techniques for clustering query log

Conclusion

In this paper we surveyed different query clustering methods used for query suggestion, personalization and FAQ

identification. The main focus on query suggestion by considering query contents, document clicks, query frequency and semantic features. These systems use a variety of techniques to help users identify the items that best fit their needs. In this way query suggestion is very important feature to suggest full queries that have been formulated by previous users so that query integrity and coherence are preserved in suggested queries.

References

[1] A.Sharma, N.Duhan , "Web search result optimization by mining search engine query log",International conference on methods and models in Computer science,pp.40-45,2010

[2] D. Beeferman and A. Berger. "Agglomerative clustering of a search engine query log" In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, pages 407 – 416. ACM Press, August 2000

[3] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li,"Context-Aware Query Suggestion by Mining Click-Through and Session Data," KDD '08: Proc. 14th ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, pp. 875-883, 2008.

[4] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query Expansion byMining User Logs," IEEE Trans. Knowledge Data Eng., vol. 15, no. 4,pp. 829-839, July/Aug. 2003.

[5] Harksoo Kim,Jungyun Seo, "Cluster-based FAQ retrieval using Latent term weights",Journal of Natural Language Processing ,2008 IEEE 15 41-1672/08

[6] J. Yi and F. Maghoul, "Query Clustering Using Click-through Graph," Proc. the 18th Int'l Conf. World Wide Web (WWW '09), 2009

[7] J.Wen, J.Nie, and H.Zhang. "Clustering user queries of a search engine". In Proceedings of the Tenth International World Wide Web Conference, Hong-Kong, China, May 1-5, pages 162–168, 2001.

[8] J.Wen, J.Nie, and H.Zhang. "Query clustering using user logs". ACM Transactions on Information Systems, (1):59–81, 2002

[9] Jiang, Shen ; Zilles, S. ; Holte, Robert "Query suggestion by query search : a new approach to query suggestion",Web Intelligence and Intelligent Agent Technologies WI-IAT '09. IEEE/WIC/ACM International Joint Conferences on (Volume:1), pp.679-684,2009.

[10] Ji-Rong Wen, Jian-Yun Nie , "Query Clustering Using Content Words and User Feedback"ACM conf .2001

[11] Kajal Y.VYAS, "Improved web search result rank optimization using search engine query log" Journal of information knowledge and research in Computer Engineering ISSN:0975-6760 2012 volume-02,ISSUE-02

[12] Lyes Limam ,David Coquil, Harald Kosch ,Lionel Brunie "Extracting user interests from search log: A clustering approach"2000

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density based algorithm for discovering clusters in large spatial databases with noise in *KDD*, pp. 226–231, 1996.

[14] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna,"The Query-Flow Graph: Model and Applications," Proc. 17thACM Conf. Information and Knowledge Management (CIKM), 2008

[15] Q. Mei, D. Zhou, and K. Church, "Query Suggestion Using Hitting Time," CIKM '08: Proc. 17th ACM Conf. Information and Knowledge Management, pp. 469-477, 2008

[16] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendation using query logs in search engines," in *EDBT*,2004

[17] Toru Onoda, Takayuki Yumoto, "Extracting and Clustering Related Keywords based on History of Query Frequency", 2008 Second International Symposium on Universal Communication

[18] X.Wang,M.Yang,S.Li, "Incremental clustering of search history in Personalized search", Journal of Computational Systems , pp.2285-2292, 2013

[19] Y.Song, D.Zhou,Li-He, "Query suggestion by constructing term-transition graphs" Proceeding of the fifth ACM international conference on web search and data mining, WSDM 12,pp. 353-362,2012

[20] Yuan Hong, Jaideep, Vaidya and Haibing Lu Rutgers University "Search engine query clustering using Top-K Search Results", 2011 IEEE/WIC/ACM International Conferences on Web intelligence and intelligent Agent technology,DOI10.1109/WI-IAT.2011.224

.