International Journal of Hybrid Information Technology
Vol. 7, No. 6 (2014), pp. 43-56
http://dx.doi.org/10.14257/ijhit.2014.7.6.04

A Survey on Query Suggestion[1]

Lingling Meng

Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China,
llmeng@deit.ecnu.edu.cn

Abstract

Query suggestion attracts great concern recently. It is crucial for capturing frequently asked questions in question-answering system and most popular topics in search engine. Besides these, is also used in advertising retrieval systems, e-commerce system for advertising push to get more profits. The paper gives a general review of query suggestion methods. On the whole, all the methods can be grouped into two categories: session based methods and click-through based methods. Adjacency based query suggestion, co-occurrence based query suggestion, query-flow graph based query suggestion, clustering based query suggestion, and bipartite graph based query suggestion are presented respectively in detail. Furthermore, how to evaluate the performance of query suggestion is denoted. Finally the important related issues of the area in further research are discussed.

Keywords: *query suggestion, session based, URLs based, click through, bipartite graph, evaluation method*

1. Introduction

With the rapid growth of data in the Web, more and more people rely on search engines for exploiting extremely valuable information. Currently, most search engines using bag-of-words model to respond to a user's query, which matches keywords between the query and web documents [1]. However the limitations of this model become increasingly prominent. Firstly, the inherent ambiguity of natural language [2] [3, 4] makes the search engine can not find out the documents that meet the user's need. Secondly, the average length of queries submitted to search engines is only 2~3 words, which make it difficult to speculate the meaning of the queries [5-7]. Finally, in most cases, users have little knowledge about the topics that they searched, even they could be not certain about what to search for, which makes it hard to find the right words to construct queries to express their information requirements. All these lead to the search results can not make user satisfied. In order to find out the satisfied document, user will often revise or reformulate the query. As shown in Chinese search engine user behavior study data of 2009, 78.2% of the users will revise or reformulate queries, and only 19.7% of the users will give up [1]. However, even in the revision or reformulation the queries, how to construct a query is still a challenge.

An effective solution is to use query suggestion technology. Query suggestion, which enables the user to reformulate a query with a single click, has become one of the most

[1] The work in the paper was supported by Shanghai Industry-University Cooperation Foundation (Grant No. Shanghai CXY-2013-84).

ISSN: 1738-9968 IJHIT

fundamental features of Web search engines. It is based on the analysis of query logs, which register the history of queries that user submitted to search engine, and the pages selected after a search, among other data. Query suggestion assumes that in the same period of time, many users have the same or similar information requirements, but they usually use different queries. These queries are considered to be similar queries, which can express similar query intention [8-10]. When a user performs a query, the search engine will suggest a group of similar queries at the top or bottom of a page for users to choose to improve the effectiveness. The generation process of such queries, basically, exploits the expertise of skilled users to help inexperienced ones. The more the users that satisfied the same information need in the past, the more precise and effective the related suggestions provided by any query suggested technique [11].

Besides this, query suggestion is also used in advertising retrieval systems, e-commerce system for advertising push to get more profits [12-14]. In addition, the query suggestion can also be applied in answering system [15], personalized search [16] and other fields. In recent years, it has become a hot topic.

Query suggestion can date back to 1990s [12, 17]. Rutger University carries out a series of experiments to study man-machine interaction of information retrieval system. The results indicate that, compared to automatic query expansion, users prefer to use query suggestion techniques in information retrieval than query expansion and query suggestion can really help to improve the retrieval effectiveness and save search time [16, 18-19]. From a technical point of view, query suggestion can be viewed as an information retrieval problem that takes queries as processing target. However, due to the short queries and information sparse, traditional information retrieval methods cannot be used. Usually it is based on the unique characteristics of queries to find the relations and construct models. Various kinds of methods have been proposed for generating query suggestions. In spite of the specific methods are very different, yet they have in common about the exploitation of usage information recorded in query logs [20]. On the whole, all the methods can be grouped into two categories: session based methods and click-through based methods. This paper gives a review of different methods, discusses the features, the evaluation method, and the issues to be resolved of query suggestion.

The remainder of the paper is as follows. In Section 2 and Section 3, session based query suggestion and click-through based query suggestion are presented respectively. Section4 describes how to evaluate the performance of query suggestion. In Section 5 discussion and further research is described and a summary is given in Section 6.

2. Session Based Query Suggestion

Session based query suggestion used query sequences to model the user behavior for predicting queries that are likely to follow a given query [21-22]. It assumes that when a user submits a query, there may be continuous queries will be submitted for correcting the initial query. These query sequence constitutes query context information for each other, which will contribute to capture the user's intention. Besides this, query session has the following features [23]:

(1)Many queries in the same session in a short time are submitted by the same user.

(2)During the interaction of query session, users often try to change the queries, or use a new query to get the results. According to Lau and Horvitz’s study [24], after a failed search,

most users either use a new query or add more detailed description in the back of original query.

(3)In the same query session, most the queries that user submitted is based on the same topic. Only a small number of queries will switch topics or conduct multi-topics search. According to Ozmutlu's study [25], there are only 11.4% users that will conduct multi-topics research in the same session and other 88.6% users will search a single topic.

These form the basis of session based query suggestion. Generally speaking session based query suggestion can be divided into three groups: adjacency or query co-occurrence query suggestion, query flow graph based query suggestion, and other query suggestions, such as rules based query suggestion, user experience based query suggestion, machine learning based query suggestion and so on.

2.1. Adjacency based or Query Co-occurrence based Query Suggestion

This method analyzes the query sequence in sessions, exploits other queries co-occurrence in the same session with the initial query, and takes the adjacency queries or co-occurrence queries for suggestion. In adjacency based method, if many users submitted q_2 after q_1, the search engine will suggest q_2 to user when a user submitted q_1, and vice versa. In co-occurrence based method, if q_1 and q_2 often co-occur in the same session, then q_1 and q_2 can be suggested for each other.

Huang's study just uses the feature of sessions [26]. He mined co-occurring query pairs from session data and candidates relevant terms for a user query by drawing on terms that co-occurred in similar search processes then ranked the candidates based on their frequency of co-occurrence with the user input queries. Experience shown that single terms suggested based on a co-occurrence matrix mined from Chinese query logs had much higher precision than those suggested based on retrieved documents over 100 queries.

Jensen's study [27] considered not only the co-occurrence frequency of the candidates, but also their mutual information with the user input queries. It weighted with the logarithm of the co-occurrence frequency for scoring candidate suggestions and original query. Moreover, to further improve the coverage of the query suggestion method, the authors manually mapped query terms to topics and then aggregated the co-occurrence patterns at topic level.

In Zeng's study [28] a health information query assistant system was developed. The system suggests alternative queries related to the user's original query that can be used as building blocks to construct a better, more specific query. The suggested queries are selected according to their semantic distance from the original query, which is calculated on the basis of concept co-occurrences in medical literature and log data as well as semantic relations in medical vocabularies. When a suggested concept is selected by a user, its occurrence with the original query concept is increased by one.

It assumed in Zhang's study [29] that the degree of similarity of two queries depends on the adjacent degree of the queries, and he more adjacent, the more similar. For example, there is a query sequence q_1, q_2, q_3. When a user submitted q_1, then q_2 are suggested to user, not q_3. In 2008, Zhiyong Zhang and Olfa Nasraoui's improved the study. They combined the association or correlation-type information with the textual content between queries. A soft relation matrix is built to store the relation between consecutive queries that occur within the same session [30].

Zanon's [31] study takes the context, categories of user click through, and similarity of queries into account, and proposed an algorithm for query suggestion.

Jones' study [32] assumed that a candidate reformulation is a pair of successive queries issued by a single user on a single day and candidate reformulations will also be referred to as query pairs. It extracted frequently adjacent query pairs in the same search processes.

Shuo-En Tsai and Yi-Shin Chen [33] refer to session data which implicitly embedded users' consecutive queries as crowd wisdom and proposed a Pattern Recognition Query (PRQs) method, which aims to achieve a query suggestion method. The study applies crowd wisdom in two ways, specialization and association. And a co-occurrence query index was built to collect co-occurrence click queries for each query.

He [21] proposed a query suggestion algorithm based on the resemblance between the user's query sequence and query sequence history.

Silviu Cucerzan and Eric Brill [34] collect co-occurrence statistics for all queries submitted by users over a long period of time. Use maximum likelihood estimation to approximate the probability that a query q_j follows immediately another query q_i in a user search session and the probability of a query q_i to be sent by a user to a search engine to generate semantically similar queries to a target query q_i.

2.2. Query Flow Graph based Query Suggestion

Query-flow graph is a usage oriented, actionable, compact representation of the information contained in a query log, and it is aimed at facilitating the analysis of user behavior. Query flow graph based query suggestion is a more structured processing of the sessions, which used a graph to represent the interesting knowledge about latent querying behavior. It is an outcome of query-log mining. In the query-flow graph each node represents a distinct query, and a directed edge from query q_i to query q_j means that at least one user submitted query q_j immediately after submitting q_i in the same session. Any path over the query-flow graph may be seen as a searching behavior, whose likelihood is given by the strength of the edges along the path [35].

In 2008 Boldi proposed a method for building a query-flow graph [35]. In the method nodes are queries and an edge from q_i to q_j is also associated with some weight to indicate how likely a user moves from q_i to q_j within a session. The edge weights were measured by the frequency of observed transitions from q_i to q_j in search logs. Then, neighbors with the largest edge weights are selected as suggestions for an input query. In 2009 Boldi *et al.* suggested labeling the edges in a query-flow graph into four categories, namely, generalization, specialization, error correction, and parallel move, and only using the edges labeled as specialization for query suggestion [35].

Recent years, some further studies extended Boldi's work along different directions. Anagnostopoulos [20, 36] argued that providing query suggestions to users may change user behavior. They thus modeled query suggestions as shortcut links on a query-flow graph and considered the resulted graph as a perturbed version of the original one. Then the problem of query suggestion was formalized as maximizing the utility function of the paths on the perturbed query-flow graph.

Sadikov [37] extended the query-flow graph by introducing the clicked documents for each query. The queries q_j following a given query q_i are clustered together if they share many clicked documents.

Zhen Liao *et al.* summarize similar queries into concepts and use concepts in both context modeling and suggestion generation, which is more effective to address the sparseness of queries [38].

2.3. Other Methods based on Sessions for Query Suggestion

Besides the methods above, there are some other query suggestion methods.

Bruno M. Fonseca uses association rule to measure the similarity of queries [39]. The method is based on two steps. Firstly, it extracts the user sessions. Secondly, similarity queries are determined by using association rules among the collection of user sessions. In his research query is taken as item and sessions is taken as a set of transactions of association rule mining. Each transaction represents a session in which a single user submits a sequence of related queries in a time interval. If is found that when users manually expanded 153 queries with concepts mined from associated queries in a session log a 32-52% relative improvement in retrieval average precision was obtained

Umut Ozertem [40] considered the task of suggesting related queries to users after they issue their initial query to a web search engine and proposed a machine learning method to learn the probability that a user may find a follow-up query both useful and relevant, given his initial query. The method is based on a machine learning model which enables the system to generalize queries that have never occurred in the logs as well. The model is trained on co-occurrences mined from the search logs, with novel utility and relevance models, and the machine learning step is done without any labeled data by human judges. The learning step allows system to generalize from the past observations and generate query suggestions that are beyond the past co-occurred queries.

Cucerzan and White [41] use previous users' experience for query suggestion. In the previous search process, if a user with a query finds out the documents that he satisfied, then you can take advantage of the user's experience and directly put forward the query to other users. Based on the idea, Cucerzan and White proposed a set of rules to determine the final returned documents with a query in a session for suggestion.

Daniele Broccolo [42] exploited a weak function for assessing the similarity between the current query and the knowledge base built from historical users' sessions.

Daniel [43] argue that relative to a single query, session provides more information to help users clearing query intention and query suggestion based on the entire session will be more accurate.

3. Click-through Based on Query Suggestions

The click-through based method focus on mining similar queries from a click-through in search logs. When user conducts a query, the log records the click URLs each time. The URLs can be used to exploit the relationship of different queries. It argues that two queries are similar to each other if they share a large number of clicked URLs. Based on the idea, some query suggestion methods are presented. On the whole, all the methods are categorized into two groups: clustering based query suggestion and bipartite graph based query suggestion.

3.1. Clustering based Query Suggestion

It is a popular method that clustering queries based on the clicked URLs. After the clustering process, for given a query q_i, it can be identified that cluster C which q_i belongs to. The other queries of cluster C can be presented as query suggestions [44-45]. That is to say, the queries in the same cluster indicate the same or similar topics and the queries within the same cluster are used as suggestions for each other. Ricardo Baeza-Yates presented an improved method. In his study [46], a rank score for each query in cluster C is computed. The rank score of each query measures the interest of the query to users that submitted the

input query. And the similar queries are returned ordered according to their rank score for suggestion. The rank score of a query is based on notions of similarity to the initial query and support of the query in the cluster. Besides this, by analyzing the logs in his experiments it is found that popular queries whose answers are of little interest to users.

Beeferman and Berger [47] incorporated the common clicked URLs. One can apply an agglomerative clustering algorithm to identify related queries and URLs for deriving group of queries that are similar in an iterative way. The queries in the same cluster are used as suggestions for one another. The quality of the query suggestions was evaluated by the click-through rate on the live Lycos search engine. However, this method has high computational cost and cannot scale up to large data. Wing Shun Chan [48] pointed out a weakness of Beeferman's method and proposed an improved clustering algorithm which ignored the noise relationships from the search engine log.

Ji-Rong and Jian-Yun [49] propose a similar queries clustering algorithm to recommend URLs to frequently asked queries of a search engine. It assumed that,

(1)If two queries contain the same or similar terms, they denote the same or similar information requirements.

(2)Two queries are similar if they lead to the selection of the same or similar document.

The function of similar queries is defined by combining both assumptions linearly.

It is noticed that the method combined query content information and click-through information. Next, a density-based method, DBSCAN [50] is applied to cluster queries. Unfortunately, it is expensive too.

In Ji-Min Wang's research [6], a new method for discovering related web queries was presented. First, some statistical characteristics of a candidate query for a given query were extracted from the log files, such as the numbers of different users submitted, the numbers of the candidate query submitted as well as the returned result clicked, the numbers of common terms and common URLs clicked between the candidate query and the given query. Then these candidate queries were ranked with a linear regression model learned from human labeled training data.

Zaiane and Strilets [51] present a method to recommend queries based on seven different notions of query similarity. The method is intended for a meta-search engine. It not only takes the keywords, phrases of the query or common clicked URLs into considered, but also takes the content and title of the URL's in the result of a query into considered. However none of their similarity measures consider user preferences in form of clicks stored in query logs.

Baeza-Yates first builds a term-weight vector for each query [21, 44]. Each term is weighted according to the number of occurrences and the number of clicks of the documents in which the term appears. He measures the similarity of two queries as the similarity of their trace vectors using the cosine function. Then an efficient k-means algorithm is used to group similar queries. The k-means algorithm requires a user to specify the number of clusters ahead of time, which is difficult for clustering search logs.

Larry Fitzpatrick and Mei Dent's study [52] calculated the similarity of queries according to the relationships of URLs. The study computed similarity based on the number of common URLs in click-through. It assumed that the more common URLs, the more similar. Furthermore, the method takes into account the position of document in results list, and established a corresponding weight function depending on the position. The weight was

determined by the probability that related document can be found in the position. If a document was not contained in the results list, its weight is 0.

Bordogna's study [53] is based on an iterative query disambiguation mechanism for query suggestion. The method is divided into three steps. Firstly, the retrieved documents are clustered, on the basis of words extracted from their titles and snippets. Secondly, for each cluster, a personalized rank is computed, based on the aggregation of two criteria: the novelty of contents of the cluster with respect to past results, and the overall content similarity of clusters with respect to the original query. Finally, from each cluster's representation, a disambiguated query is generated and suggested to the user to deepen the search. The disambiguated queries are computed based on terms in titles and snippets of documents in the clusters. They should be able to highlight the main contents of the cluster and potentially may retrieve further relevant documents.

Saurabh Sharma [54] created user profiles to capture the user's personal preference and identified the actual goal of the input query. Furthermore agglomerative clustering algorithm was used to find the queries that are close to each other conceptually. In the method relationship between users, queries and concepts were taken into account to obtain accurate and more personalized query suggestions for the user.

Kenneth [55] introduced an approach that captured the user's conceptual preferences in order to provide personalized query suggestions. They achieved this goal with two new strategies. First, online techniques that extract concepts from the web-snippets of the search result returned from a query are developed and concepts are used to identify related queries for that query. Second, a new two phase personalized agglomerative clustering algorithm based on click-through was proposed that was able to generate personalized query clusters. To the best of the authors' knowledge, no previous work had addressed personalization for query suggestions.

3.2. Bipartite Graph based Query Suggestion

In the user query log, queries and URLs are represented as nodes. Each record contains one pair <query, URL>. Merging these pairs, you can create a bipartite graph with the vertices on one side corresponding to queries and on the other side to URLs, which join the collections of queries and the collections of click-through. It represents an implicit judgment of the relationships between queries, relationships between click-through and relationships between query and click-through. The method attempts to find such two sets: (1) a disjoint similar query sets. The elements in the set represent different expressions with the same or similar information requirements; (2) a disjoint URLs. The elements in the set represent different pages with the same or similar information requirements.

H. Ma [56] established a user-query bipartite graph and a query-URL bipartite graph based on click-through. In his study, a two-tier query suggestion model was presented and a similar query model was proposed based on features.

Mei *et al.* [57] performed a random walk starting from a given query q_i on the click-through bipartite to find queries similar to q_i. Each similar query q_j is labeled with a "hitting time," which is essentially the expected number of random walk steps to reach q_j starting from q_i. The queries with the smallest hitting time were selected as the query suggestions.

Yang Song and Li-wei He [58] took the clicked URLs and skipped URLs into account and proposed an optimal rare query suggestion framework by leveraging implicit feedbacks from users in the query logs. The proposed model is based on the pseudo-relevance feedback. It assumes that clicked and skipped URLs contain different level of information, and thus, they

should be treated differently. Therefore, the framework optimally combines both clicked and skipped information from users, based on which query-URL bipartite was established respectively for determining similar queries. And a random walk model was used to optimize (1) the restarting rate of the random walk, and (2) the combination ratio of click and skip information. Experimental results on a log from a commercial search engine show the superiority of the proposed method over the traditional random walk models and pseudo-relevance feedback models.

H. Tong [59], Craswell [60] and Baluja [61] used a random walk-based method for query suggestion. The basic idea behind random walk models is quite straightforward. Queries and URLs are represented as nodes in a bipartite graph where each edge connects one query with one URL, which indicates a click. The model calculates the stable transition probability from one node to another and uses the probability to estimate the closeness between two nodes.

Yan Chen [62] constructed personalized query suggestion agent based on query-concept bipartite graphs and concept relation trees. There are three steps. First of all, the personalized query suggestion agent uses both concepts' semantic relations and concepts' co-occurrence for concept clustering. Furthermore, the agent constructs concept relation trees that can provide more suggested queries than a query-concept based method. Finally, the agent dynamically updates weights between query-concept and concept-concept to personalize suggestions.

Besides these methods mentioned above that focused on providing new queries, Bai Lv [63] constructed long tail query suggestion model based on query intent. Wei Gao [64] proposed a new method for cross-lingual query suggestion by exploiting, in addition to the translation information, a wide spectrum of bilingual and monolingual information, such as term co-occurrences, query logs with click-through data, and so on. In Jiang-Ming Yang's study [65] proposed a unified strategy to combine query log and search results for query suggestion was proposed. In this way, the study leveraged both the users' search intentions for popular queries and the power of search engines for unpopular queries. The suggested queries are also ranked according to their relevance and qualities; and each suggestion is described with a rich snippet including a photo and related description. Markus Strohmaier [66] introduced an intentional query suggestion as a novel idea that is attempting to make users' intent more explicit during search and presented a prototypical algorithm for intentional query suggestion. Yang Song [67] from the search engine session logs mined a large amount of user preference data and proposed a query suggestion method by constructing term-transition graphs. In the method it was considered the following tuple $\{q_1,q_2,u\}$ where a user abandoned a query q_1 and immediately reformulated it into q_2 then made a click on URL u, during the same session. These activities strongly indicate a user's preference on query q_2 over q_1, which often differs by only a few terms. Then a term-preference graph was constructed from the above data where each node is a term in the query and each directed edge a preference. And a topic-biased Pagerank model was trained for each of the query topics by extracting topics from clicked URLs Given a query, this model guides the decision of (1) expanding relevant terms to the original query, (2) removing terms from the original query, or (3) replacing existing terms with relevant terms. Yiqun Liu [68] analyze the nature of query suggestion process from user's perspective and propose a query suggestion framework in which keywords are suggested because of their appearance in clicked snippets instead of similarity with previous queries. Two snippet click models and corresponding suggestion algorithms were presented based this analysis. Some studies tackled the query suggestion problem by merely substituting or stemming terms [69, 32, 70-71]. A few recent studies diversifying the top-returned search

results [72-73] for query suggestion. Sadikov [37] combined the query-flow graphs with click URLs information to find query suggestions. Hotho [74] improved bipartite and proposed a method in the context of modeling folksonomies, which can be represented as a tri-partite document-user-tag graph. It has been successfully applied in web ranking tasks. Chien and Immorlica [75] used pearson correlation of query time distributed vector to characterize similar queries. Based on Chien's work, Zhang [76] considered the factor of significant time interval for query suggestion.

4. Evaluation

Another problem is how to evaluate the performance of different method.

On the one hand, nowadays there is no public corpus for query suggestion. Because user query logs involves privacy and some commercial factors, most universities and research institutions' labs are difficult to obtain real query logs from search engine companies, such as Microsoft, Yahoo, Google and so on. Only the English logs of three companies that are Excite, AlltheWeb, AltaVista are available. The latest version is AltaVista_2003. Unfortunately most pages in AltaVista_2003 are not existed. Part of Sougou company's query logs in Chinese of 2008 is available in which userID, queries, click-through, rank, timestamp and others are included.

On the other hand, there is unified standard for evaluating the performance. It makes great difference in different literatures. Some frequently used evaluations are as follows.

(1) P@N(precision @ N). It is the precision in top N suggested queries.

Precision is the ratio of the correct suggestions' number to all suggested queries' number.

$$percision \quad (\%) \quad = \frac{A}{A + C} * 100 \ \% \tag{1}$$

Where A is the number of correct suggestions, C is the number of not correct suggestions. That is to say, A+C is the number of all suggested queries.

(2) Coverage. Coverage is the ratio of correct suggestions' number to the true query set.

$$\text{cov} \ erage \quad (\%) \quad = \frac{A}{A + B} * 100 \ \% \tag{2}$$

Where A is the number of similar queries that are suggested, B is the number of similar queries that are not suggested. And A+B is the true query set.

(3)DCG(Discounted Cumulative Gain). Set $<v_1,v_2,\ldots\ldots, v_n>$ is the results of query q, and R(k) is the score of v_k, then DCG in the k position is defined as:

$$DCG \quad @ \ k = \sum_{i=1}^{k} \frac{1}{\log_2(1 + i)} (2^{R(i)} - 1) \tag{3}$$

(4)NDCG (Normalized DCG). NDCG is normalized DCG. For any query, an ideal DCG is calculated. NDCG is the ratio of DCG@k to IDCG@k, formally:

$$NDCG \quad @ \ k = \frac{DCG \quad @ \ k}{IDCG \quad @ \ k} \tag{4}$$

(5)QSCTR. Query suggestion click-through rate (QSCTR) is another metric used to evaluate the quality of query suggestions. In literature [77], it defined QSCTR of a <query, suggestion> pair as:

$$QSCTR \quad (\%) \quad = \frac{its \quad clicked \quad count}{its \quad impression \quad count} * 100 \ \% \tag{5}$$

Therefore, QSCTR can be interpreted as the probability that a user clicks on a query suggestion given in response to a query.

5. Discussion and Future Research

Query suggestion is an emerging field of web research. Some performance have been achieved, however some issues are needed to be solved.

In session based query suggestion how to properly divide session is a problem, which will directly affect the accuracy of suggestion. A traditional method is to divide sessions according to the time intervals of two adjacent queries. If the interval is greater than a threshold, the two adjacent queries will be divided into two sessions. For example, Ryen‘s study [78] discussed the problem. In recent years, some new session division methods were proposed [43]. Another reason that will cause inaccuracy of query suggestion is that in a specific period of session data, search interests might even change over time. That is to say query intention will drift.

In click-through based query suggestion, there is clicked data sparse problem, because only a few URLs are clicked during an information retrieval process. Many similar queries have no common URLs. Even sometimes maybe there are no clicked URLs for some queries. Click-through information is not enough for deriving similar queries. Furthermore, there is noise in user click-through because different users may have different clicked behavior. Some users may click their interested URLs, others may clicked URLs that they are not interested in a random way, which noise data is resulted. Beside this, rare queries possess much less information than popular queries in query logs, which results it more difficult to suggest similar queries to a rare query.

6. Summary

This paper gives a survey on query suggestion methods, including session based query suggestion and click-through based query suggestion. Adjacency based query suggestion, co-occurrence based query suggestion, query-flow graph based query suggestion, clustering based query suggestion, and bipartite graph based query suggestion are presented in detail. How to evaluate the performance of query suggestion is denoted. Furthermore, the important related issues are discussed. Finally the paper gives some suggestions of the area in further research.

References

[1] J. Yang, J. Yu-Gang, A. G. Hauptmann and N. Chong-Wah, “Evaluating bag-of-visual-words representations in scene classification”, Proceedings of the international workshop on Workshop on multimedia information retrieval, University of Augsburg, Germany, **(2007)** September 28-29.
[2] http://www.cnnic.net.cn/uploadfiles/2009/9/21/104149.doc
[3] H. Cui, J.-R. Wen, J.-Y. Nie and W.-Y. Ma, “Probabilistic Query Expansion Using Query Logs”, Proceedings of the 11th International Conference on World Wide Web, Honolulu, Hawaii, USA, **(2002),** May 7-11.

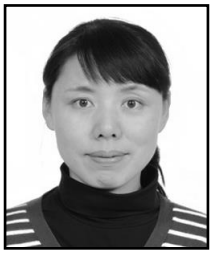
[4] R. Song, Z. Luo, J. R. Wen, Y. Yu, and H. W. Hon, "Identifying Ambiguous Queries in Web Search", Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, **(2007)** May 8- 12.
[5] B. J. Jansen, A. Spink, J. Bateman and T. Saracevic, "Real life Information Retrieval: A Study of User Queries on the Web", ACM SIGIR Forum, vol. 32, no. 1, **(1998)**.
[6] J. Wang, C. Chen and B. Peng, "Analysis of the User Log for a Large Scale Chinese Search Engine", Journal of South China University of Technology (Natural Science), vol. 32, SUPPL, **(2004)**.
[7] A. Spink and B. J. Jansen, "A Study of Web Search Trends", Webology, vol. 2, no.1, **(2004)**.
[8] M. Agosti, F. Crivellari and G. Maria Di Nunzio, "Web Log Analysis: A Review of a Decade of Studies About Information Acquisition, Inspection and Interpretation of User Interaction", Data Mining and Knowledge Discovery, vol. 24, no. 3, **(2012)**.
[9] Z. Bar-Yossef and M. Gurevich, "Mining Search Engine Query Logs Via Suggestion Sampling", Proceedings of PVLDB, vol.1, no.1, **(2008)**.
[10] W. Wu, H. Li and J. Xu, "Learning Query and Document Similarities From Click-through Bipartite Graph with Metadata", Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy, **(2013)** February 6-8.
[11] D. Broccolo, L. Marcon, F. M. Nardini, R. Perego and F. Silvestri, "Generating suggestions for queries in the long tail with an inverted index", Information Processing & Management, vol. 48, no. 2, **(2012)**.
[12] Y. Li, B. Wang and J. Li, "A Survey of Query Suggestion in Search Engine", Journal of Chinese Information Processing, vol. 24, no. 6, **(2010)**.
[13] https: //adwords.google.cn /
[14] http: // e.baidu.com/pro/
[15] J. Jeon, W. B. Croft, and J. H. Lee, "Finding Similar Questions in Large Question and Answer Archives", Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, **(2005)** October 31 - November 05.
[16] P. A. Chirita, C. S. Firan and W. Nejdl, "Personalized Query Expansion for the Web", Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, **(2007)** July 23-27.
[17] A. Lee and M. Chau, "The Impact of Query Suggestion in E-commerce Websites", The 10th Annual Workshop on E-Business, Shanghai, China, **(2011)** December 4.
[18] . Al Hasan, N. Parikh, G. Singh and N. Sundaresan, "Query Suggestion for E-commerce Sites", Proceedings of the 4th ACM International Conference on Web Search and Data Mining, Hong Kong, China, **(2011)** February 9-12,
[19] J. Jeon, W. B. Croft and J. H. Lee, "Finding Similar Questions in Large Question and Answer Archives", Proceedings of the 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, **(2005)** October 31-November 5.
[20] F. Silvestri, "Mining query logs: Turning search usage data into knowledge", Foundations and Trends in Information Retrieval, vol. 4, no.1, **(2010)**.
[21] Q. He, D. Jiang, Z. Liao, S. Hoi, K. Chang, E. Lim and H. Li, "Web query recommendation via sequential query prediction", Proceedings of the 25th International Conference on Data Engineering, Shanghai, China, **(2009)** March 29 - April 2.
[22] F. Radlinski and T. Joachims, "Query chains: Learning to rank from implicit feedback", Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, **(2005)** August 21-24.
[23] C. Liu, "Query Recommendation using Random Walk Model", Master thesis, Nankai University, **(2011)**.
[24] T. Lau and E. Horvitz, "Patterns of search: analyzing and modeling web query refinement", Proceedings of the 7th International Conference on User Modeling, Banff, Canada, **(1999)** June 20-24.
[25] S. Ozmutlu, H. C. Ozmutlu and A. Spink, "Multitasking web searching and implications for design", Journal of the American Society for Information Science and Technology, vol. 40, no.1, **(2003)**.
[26] C.-K.Huang, L.-F.Chien, and Y.-J. Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs", Journal of the American Society for Information Science and Technology, vol. 54, no.7, (**2003**).
[27] E. C. Jensen, S. Beitzel, A. Chowdhury and O. Frider, "Query phrase suggestion from topically tagged session logs", Proceedings of the 7th International Conference on Flexible Query Answering Systems, Milan, Italy, (**2006**) June 7-10.
[28] Q. T. Zeng, J. Crowell, R. M. Plovnick, E. Kim, L. Ngo and E. Dibble, "Assisting Consumer Health Information Retrieval with Query Recommendations", Journal of the American Medical Informatics Association, vol. 13, no. 1, **(2006)**.

[29] Z. Zhanga nd O. Nasraoui, "Mining Search Engine Query Logs for Query Recommendation", Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, **(2006)** May 23-26.
[30] Z. Zhang and O. Nasraoui, "Mining search engine query logs for social filtering-based query recommendation", Soft Computing for Dynamic Data Mining, vol. 8, no. 4, **(2008)**.
[31] R. Zanon, S. Albertini, M. Carullo and I. Gallo, "A New Query Suggestion Algorithm for Taxonomy-based Search Engines", Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval, Barcelona, Spain, **(2012)** October 4-7.
[32] R. Jones, B. Rey, O. Madani and W. Greiner, "Generating Query Substitutions", Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland, **(2006)** May 23-26.
[33] S.-E. Tsai and Y.-S. Chen, "Improving Query Suggestion by Utilizing User Intent", Proceedings of the 11th IEEE International Conference on Information Reuse and Integration, **(2010)** August 4-6.
[34] S. Cucerzan and E. Bril, "Extracting Semantically Related Queries By Exploiting User Session Information", Proceedings of the 5th ACM Workshop on Exploiting Semantic Annotations in Information Retrieval, Maui, HI, United states, **(2012)** November 2.
[35] P. Boldi, F. Bonchi, C. Castillo, D. Donato and S. Vigna, "Query suggestions using query-flow graphs", Proceedings of the 2009 workshop on Web Search Click Data, Barcelona, Spain, **(2009)** February 9.
[36] A. Anagnostopoulos, L. Becchetti, C. Castillo and A. Gionis, "An optimization framework for query recommendation", Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York City, NY, United states, **(2010)** February 3-6.
[37] E. Sadikov, J. Madhavan, L. Wang, and A. Halevy, "Clustering query refinements by user intent", Proceedings of the 19th international conference on World wide web, Raleigh, NC, United states, **(2010)**, April 26-30.
[38] L. Zhen, J. Daxin, C. Enhong, P. Jian, C. Huanhuan and L. Hang, "Mining Concept Sequences from Large-Scale Search Logs for Context-Aware Query Suggestion", ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 1, **(2011)**.
[39] B. M. Fonseca, P. B. Golgher, E. S. de Moura and N. Ziviani, "Using Association Rules to Discovery Search Engines related Queries", Proceedings of the 1st Conference on Latin American Web Congress, Santiago, Chile,**(2003)** November 10-12.
[40] U. Ozertem and O. Chapelle, "Learning to Suggest: A Machine Learning Framework for Ranking Query Suggestions", Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, Portland, Oregon, USA, **(2012)** August 12–16.
[41] S. Cucerzana and R. W. White, "Query Suggestion Based on User Landing Pages", Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in Information Retrieval, Amsterdam, Netherlands, **(2007)** July 23-27.
[42] D. Broccolo, L. Marcon, F. M. Nardini, R. Peregoa and F. Silvestri, "Generating suggestions for queries in the long tail with an inverted index", Information Processing & Management, vol. 48, no. 2, **(2012)**.
[43] G.-A. Daniel, "A survey on session detection methods in query logs and a proposal for future evaluation", Information Science: an International Journal, vol. 179, no. 12, **(2009)**.
[44] R. Baeza-Yates, C. Hurtadoa and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines", Proceedings of the 2004 International Conference on Current Trends in Database Technology, Heraklion, Crete, Greece, (**2004**) March 14-18.
[45] 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", In Proceedings of 14th International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, United states, **(2008)** August 24-27.
[46] R. Baeza-Yates, "Applications of Web Query Mining", Proceedings of 27th European Conference on IR Research, Santiago de Compostella, Spain, **(2005)** March 21.
[47] D. Beeferman and A. Berger, "Agglomerative Clustering of Search Engine Query Log", Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, United states,**(2000)** August 20-23.
[48] W. S. Chan, W. T. Leunga and D. L. Lee, "Clustering Search Engine Query Log Containing Noisy Clickthroughs", Proceedings of the 2004 International Symposium on Applications and the Internet, Tokyo, Japan, **(2004)** January 26-30.
[49] J.-R. Wen, J.-Y. Nie and H.-J. Zhang, "Query Clustering Using User Logs", ACM Transactions on Information Systems, vol. 20, no. 1, **(2002)**.
[50] M. Ester, H. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise", Information Systems, vol. 32, no. 7, **(2007)**.

[51] O. R. Zaïane and A. Strilets, "Finding Similar Queries to Satisfy Searches Based on Query Traces", Proceedings of the Workshops on Advances in Object-Oriented Information Systems, Montpellier, France, **(2002)** September 2.
[52] L. Fitzpatricka and M. Dent, "Automatic Feedback Using Past Queries: Social Searching", Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, **(1997)** July 27-31.
[53] G. Bordogna, A. Campi, G. Psailaa and S. Ronchi, "Disambiguated Query Suggestions and Personalized Content-similarity and Novelty Ranking of Clustered Results to Optimize Web Searches", Information Processing & Management, vol. 48, no. 3, **(2012)**.
[54] S. Sharma and N. Mangla, "Obtaining Personalized and Accurate Query Suggestion by Using Agglomerative Clustering Algorithm and P-QC Method, International Journal of Engineering Research & Technology, vol. 1, no. 5, **(2012)**.
[55] K. Wai-Ting Leung, W. Ng, and D. L. Lee, "Personalized Concept-Based Clustering of Search Engine Queries", IEEE Transactions on knowledge and data engineering, vol. 20, no.11, **(2008)**.
[56] H. Ma, H. Yang, I. King and M. R. Lyu, "Learning latent semantic relations from clickthrough data for query suggestion", Proceedings of the CIKM 2008, Napa Valley, CA, United states, **(2008)** October 26-30.
[57] Q. Mei, D. Zhou and K. Church,"Query suggestion using hitting time", Proceedings of 17th ACM Conference on Information and Knowledge Management , Napa Valley, CA, United states, **(2008)** October 26-30.
[58] Y. Song and L. He, "Optimal rare query suggestion with implicit user feedback", Proceedings of the WWW 2010, Raleigh, NC, United states, **(2010)** April 26-30.
[59] H. Tong, C. Faloutsos and J.-Y. Pan, "Random walk with restart: fast solutions and applications", Knowledge Information System, vol. 14, no. 3, **(2008)**.
[60] N. Craswell and M. Szummer, "Random walks on the click graph", Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, Netherlands, **(2007)** July 23-27.
[61] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran and M. Aly, "Video suggestion and discovery for Youtube: Taking random walks through the view graph", Proceedings of the 17th international conference on world wide web, Beijing, China, **(2008)** April 21 - 25.
[62] Y. Chen and Y.-Q. Zhang, "A Personalized Query Suggestion Agent based on Query-Concept Bipartite Graphs and Concept Relation Trees", International Journal of Advanced Intelligence Paradigms, vol. 4, no.1, **(2009)**.
[63] B. Lv, G. Jia-feng, C. Lei and C. Xue-qi, "Long Tail Query Recommendation Based on Query Intent", Chinese Journal of Computers, vol. 36, no. 3, **(2013)**.
[64] W. Gao, N. Cheng, J.-Y. Nie, M. Zhou, K.-F. Wong and H.-W. Hon, "Exploiting Query Logs for Cross-Lingual Query Suggestion", ACM Transactions on Information Systems, vol. 28, no. 2, **(2010)**.
[65] J.-M. Yang, R. Cai, F. Jingz, S. Wang, L. Zhang and W.-Y. Ma, "Search-based Query Suggestion", Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, California, USA, **(2008)** October 26-30.
[66] M. Strohmaier, M. Kröll and C. Körner, "Making User Goals More Explicit During Search", Second ACM International Conference on Web Search and Data Mining, Barcelona, Spain, **(2009)** February 9-12.
[67] Y. Song, D. Zhou and L.-w. He, "Query Suggestion by Constructing Term-Transition Graphs", WSDM'12, Seattle, Washington, USA, **(2012)** February 8-12.
[68] Y. Liu, J. Miao, M. Zhang, S. Maa and L. Ru, "How do users describe their information need: Query recommendation based on snippet click model", Expert Systems with Applications, vol. 38, no.11, (**2011**).
[69] V. Dang B. W. Croft, "Query reformulation using anchor text", Proceedings of the Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, New York City, NY, United states, (**2010**) February 3-6.
[70] R. Kraft and J. Zien, "Mining anchor text for query refinement", Proceedings of the 13th international conference on World Wide Web, New York, NY, USA, **(2004)** May 17-20.
[71] X. Wang and C. X. Zhai, "Mining term association patterns from search logs for effective query reformulation", Proceedings of ACM 17th Conference on Information and Knowledge Management, Napa Valley, California, **(2008)** October 26-30.
[72] H. Ma, M. R. Lyu and I. King, "Diversifying query suggestion results", Proceedings of the 24th AAAI Conference on Artificial Intelligence, Westin Peachtree Plaza in Atlanta, Georgia, USA, **(2010)** July 11-15.
[73] Y. Song, D. Zhou and L. He, "Post-ranking query suggestion by diversifying search results", Proceedings of Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, **(2011)** July 24-28.

[74] A. Hotho, R. Ja¨schke, C. Schmitz and G. Stumme, "Information retrieval in folksonomies: Search and ranking", Proceedings of the 3rd European Semantic Web Conference, Budva, Yugoslavia, **(2006)** June 11-14.
[75] S. Chien and N. Immorlica, "Semantic similarity between search engine queries using temporal correlation", Proceeding s of WWW05, International Conference on the World Wide Web, Chiba, Japan, **(2005)** May 10-14.
[76] W. Zhang, J. Yan, Sh.-Ch. Yan, N. Liu and Zh. Chen, "Temporal query substitution for ad search", Boston, Massachusetts, **(2009)** July 19-23.
[77] M. P. Kato, T. Sakai and K. Tanaka, "Query Session Data vs. Clickthrough Data as Query Suggestion Resources", ECIR 2011 Workshop on Session Information Retrieval, Dublin, Ireland, **(2011)** April 18.
[78] R. W. White, M. Bilenko and S. Cucerzan, "Studying the Use of Popular Destinations to Enhance Web Search interaction", Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, Netherlands, **(2007)** July 23-27.

Author



Lingling Meng, she is an associate professor of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

