Suggestions for Fresh Search Queries by Mining Mircoblog Topics

Lin Li[1], **Xing Chen**[2], **Guandong Xu**[3]
[1,2]Wuhan University of Technology, China
{ [1]cathylilin, [2]rebecca_lymx}@whut.edu.cn
[3]University of Technology, Sydney, Australia
[3]Guandong.Xu@uts.edu.au

Abstract

Query suggestion of Web search has been an effective approach to help users quickly express their information need and more accurately get the information they need. All major web-search engines and most proposed methods that suggest queries rely on query logs of search engine to determine possible query suggestions. However, for search systems, it is much more difficult to effectively suggest relevant queries to a fresh search query which has no or few historical evidences in query logs. In this paper, we propose a suggestion approach for fresh queries by mining the new social network media, i.e, mircoblog topics. We leverage the comment information in the microblog topics to mine potential suggestions. We utilize word frequency statistics to extract a set of ordered candidate words. As soon as a user starts typing a query word, words that match with the partial user query word are selected as completions of the partial query word and are offered as query suggestions. We collect a dataset from Sina microblog topics and compare the final results by selecting different suggestion context source. The experimental results clearly demonstrate the effectiveness of our approach in suggesting queries with high quality. Our conclusion is that the suggestion context source of a topic consists of the tweets from authenticated Sina users is more effective than the tweets from all Sina users.

1 Introduction

Web search engines have greatly changed the way that people acquire information during the last ten years. As an end-user starts typing a query in a search engine's query box, most search engines assist users by providing a list of queries that have been proven to be effective in the past [19]. The user can quickly choose one of the suggested completions (in some cases, alternatives) and thus, does not have to type the whole query herself. Feuer et al. [10] analyzed approximately 1.5 million queries from the search logs of a commercial search engine and found that query suggestions represented nearly 30% of the total queries and the engine with phrase suggestions performs better in terms of precision and recall than the same search engine without suggestions. Furthermore, Kelly et al. [13] observed that the use of offered query suggestions is more for difficult topics, i.e., topics on which users have little knowledge to formulate good queries. Yang et al. [20] presented an optimal rare query suggestion framework by leveraging implicit feedbacks from users in the query logs. Sumit et al. [3] put forward a probabilistic mechanism for generating query suggestions from a corpus without using query logs and utilized the document corpus to extract a set of candidate words.

Traditional methods rely on some other users who searched for the same information before, and then utilize these large amounts of past usage data to offer possible query suggestions. Although there are many works using query logs to suggest queries [1; 2; 4; 6; 8; 12; 16; 17; 20], there still exist some difficulties.

First, query logs may not always be accessible in some applications due to privacy and legal constraints. Second, even in the case of general-purpose web search engines, end-users sometimes pose queries that are not in query logs or are not very frequent. Third, with the rise of social network, there has been emerging a group of new network vocabulary. When these newly appeared words formulate search queries, they always have few search history in query logs. Thus they are insufficient in context. Both the queries that in the absence of query logs and the newly appeared queries together constitute a kind of search queries, so called fresh queries. They may cause a great amount of search traffic potentially affecting the performance of search engines significantly. Therefore, how to offer an effective query suggestion for fresh search queries is a challenging research problem, which we discuss in this paper.

At present, as a widely used medium platform, microblog's diverse features meet the people's information, interpersonal information and other aspects of the new requirements. Compared with the traditional media, microblogging as a new service has the following characteristics and advantages.

(1) Its information propagation is convenient and rapid.

(2) Its information dissemination is of high efficiency.

(3) It has great potential business value.

Among the three features, the second one motivates our work.

Nowadays the speed of information propagation through the microblog service is faster than most of media products, and more people pay attention to it. The intuitive, convenient, and efficient communication makes microblog popular and the micoblog information updated quickly, which is the reason that we choose the microblog topics as our study background. The key idea of our work is that extracting and analyzing fresh queries by mining microblog topics in order to give query suggestions to web search users. Our main contribution is that the suggestion context source of a topic consists of the tweets from authenticated Sina users is more effective than the tweets from all Sina users.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the prior work on query suggestion along with explaining how our approach differs from the previous methods. In Section 3, we give a suggestion flowchart to describe our whole query suggestion process. In Section 4, we further describe the specific approach of offline processing in details. Experiments and results are presented in Section 5. Section 6 concludes the paper and outlines future research directions.

2 Related Work

2.1 Query Suggestion

There are a variety of research works about query suggestion. Initial works focus on identifying past queries similar to a current user query. Baeza-Yates et al. [1] cluster queries presented in search logs. Given an initial query, similar queries from its cluster are identified based on vector similarity metrics and are then suggested to a user. Barouni-Ebrahimi and Ghorbani utilize words frequently occurring in queries submitted by past users as suggestions [2]. Gao et al. describe a query suggestion mechanism for cross lingual information retrieval where for queries issued in one language, queries in other languages can also be suggested [11]. By utilizing clickthrough data and session information, Cao et al. propose a context aware query suggestion approach [6]. In order to deal with the data sparseness problem, they use concept based query suggestions where a concept is defined as a set of similar queries mined from the query-URL bi-partite graph.

Lately, Broder et. al propose an online expansion of rare queries in [5]. Their framework starts by training an offline model that is able to suggest a ranked list of related queries to an incoming rare query. The rare query is then expanded by a weighted linear combination of the original query and the related queries according to their similarity. Yang et al. [20] also work on rare query suggestion by using implicit feedbacks, while Sumit et al. [3] make use of a corpus instead of query logs. To the best of our knowledge, our work makes the fist to study the query suggestion of fresh queries (both newly appeared queries and queries absent from query logs).

2.2 Social Media

The rising popularity of online social networking services has spurred research into microblogs and their characteristics. There are a number of research works which explore and study microblog., especially English microblogging, i.e., twitter. Newman et al. [18] make the first quantitative study on the entire Twitter sphere and information diffusion on it. They study the topological characteristics of Twitter and its power as a new medium of information sharing and have found a non-powerlaw follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. In 2010, the work in [14] further discusses the topological characteristics of Twitter and its power as a new medium of information sharing. Chen et al. [7] compare two kinds of approaches, traditional cosine-based approach and WordNet-based semantic approach, when computing similarities between microblogs to recommend top related ones to users.

With the prevalence of Sina microblogging, some researchers begin to study the new Chinese microblog media. Liu et al. [15] combine a translation-based method with a frequency-based method for keyword extraction. They extract keywords for microblog users from the largest microblogging website in China, Sina Weibo. Different from them, we present how to extract and analyze microblog topics to produce effective suggestions to fresh queries, and experimentally discuss the selection of suggestion context sources in terms of precision and efficiency.

3 Suggestion Flowchart

As shown in Figure 1, the whole suggestion process is divided into two modules: offline processing and online processing.

First, let us look at the offline module. It consists of the following four steps. Step 1 extracts microblog information from a topic. Here we not only get the tweets from all Sina users for each topic, but also extract the microblog tweets from authenticated Sina users. Then we can take the two types of data as two different suggestion context sources. Step 2, step 3 and step 4 make a text preprocessing for our selected suggestion context sources, including removing stopwords, Chinese participle preprocessing and word frequency statistics. After completing these four steps, we extract the top 10 representative nouns or verb-nouns in the sequence according to the results of word frequency statistics. Nouns and verb-nouns can represent the meaning of a topic. Last, the produced words are put into the suggestion list in turn.

We give an example to describe our online processing. When a user has an information need, she will transform the information need into a query and start typing the query in the query box of a search engine. The user has some information need but is not sure which words to use to formulate a query because traditional method that documents indexed by the search engine are not visible to the user. The terms selected by the user to formulate the queries often do not lead to a good retrieval performance due to the gap between query-term space and document-term space [9]. This problem is especially difficult for the fresh search queries because of lacking context in query logs. To help the user formulate good search queries, our suggestion list may give the useful query suggestion to the user. When the query exists in the suggestion list of a certain topic, we can recommend other words in this suggestion list to the user.

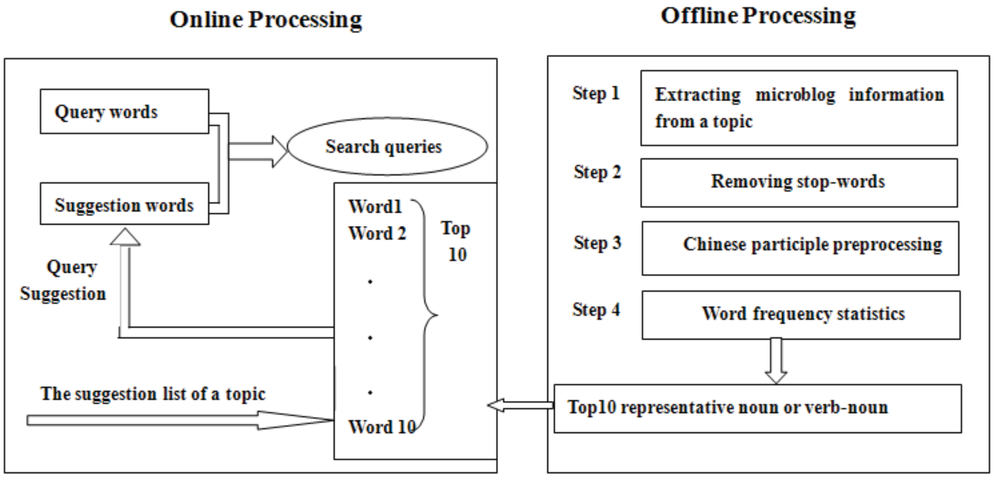
Online Processing
Offline Processing
Query words
Search queries
Suggestion words
Word1
Word 2
Top 10
Query Suggestion
The suggestion list of a topic
Word 10
Step 1
Extracting microblog information from a topic
Step 2
Removing stop-words
Step 3
Chinese participle preprocessing
Step 4
Word frequency statistics
Top10 representative noun or verb-noun

Figure 1: Suggestion Flowchart for A Search Query

4 Offline Processing

In this section, we first introduce how to extract microblog data by crawling. Then we describe our approach for text processing. Finally, we make some discussions.

4.1 Extracting Microblog Data by Crawling

For benefiting from our professional point of view of computer science technology, we select the technology/IT Internet as our data source. The chosen 14 topics all are the most popular topics at the crawling time as our experimental data. We extract the tweets of these 14 topics. Sina microblog only gives 10 pages space capacity to display the tweets of each micro-topic and each page only exists 20 tweets. From the end of March 2012, we start collecting micro-topic data. In order to ensure nonduplication of data, nearly every two days, we download the html webpages of each Micro-topic, and then save them in the local disk folder as .txt file format. By the end of June 2012, we obtain almost 3750 web pages. But how to extract our needed information from these html files? Here we use HtmlParser [1]. HtmlParser is an open-source project used to parse the HTML document. It is small, fast, simple and has a powerful function.

4.2 Text Processing

First, we need to point out that so-called the authenticated users mainly includes the users of Sina microbog VIP, Sina approved Sina agencies and authenticated Sina individuals. These users are certified by Sina microblog with a certain authenticity and authority. Then we have collected profiles of users who mentioned about 14 Sina microblog trending topics from March 25th to June 17th, 2012 by crawling. We separately extracted tweets information from all users and from authenticated users for each topic among 14 topics. Under our preliminary statistics, there are 63,354 tweets form all users. The number of tweets by the authenticated users is about 22,724, accounting for about 35.9% of total users' tweets.

To complete text processing, two NLP tools (i.e., $MyTxtSegTag$ and $MyZiCiFreq$) is used [2]. The next step is to remove all stopwords. We remove the words in our tweets for each topic that appeared in a stopwords list. After making stopwords processing for authenticated tweets and the whole tweets of each topic, we make Chinese participle preprocessing for the filtered tweets by a set of word segmentation and POS tagging tool which is named $MyTxtSegTag$. The big advantage of this software is that it can identify proper and newly appeared nouns and minimize the word granularity, such as the new word of Mirco-Letters, weixin(a mobile phone chat software) which is a new application launched by Tencent company in 2011. If the option of starting proper nouns is not selected, then after making participle processing,the word of weixin will be divided into two words that ”Micro” and ”letters”. So using this software can improve the precision and accuracy of participle processing results.

Last, we save these produced words in a .txt formatted file, making a preparation for word frequency statistics and analysis. We adopt a word frequency statistic tool named MyZiCiFreq. This software can not only make character frequency statistics but also make word frequency statistics. Furthermore, we also observe that the processing times for the same topic from two different sets of users' tweets are significantly different. Averagely, it takes about 4 hours for all users' tweets of a certain topic, but for processing authenticated users' tweets, it just take about 30 minutes less than an hour.

4.3 Discussions

What we will discuss is about why we collected microblog data by crawling not using Sina API (Application Programming Interface). As we all know it will greatly shorten the

[1] http://htmlparser.sourceforge.net/

[2] They are recommended by the website of http://www.china-language.gov.cn/index.htm

time so as to improve the efficiency if we collect data using API. But there are many limiting factors, such as only a part of API, not all API is provided. Moreover, some API just can be used by senior member users, making that we cannot crawl and collect data in time and completely.

To help the user formulate effective queries, a suggestion list is produced after text processing. When the query exists in the suggestion list of a certain topic, we can recommend other words in this suggestion list to the user. In this step, we can adopt computing semantic similarity between query word and other words that appeared in the suggestion list based on the path length similarity, in which we treat taxonomy as an undirected graph and measure the distance between them in HowNet which serves as a base of research in knowledge processing and multilingual NLP [3]. This method will be used in our future work.

5 Experiment

In this section, we first introduce the data sets and evaluation method. Then we present the experimental results. Finally, a dicussion is given.

5.1 Data Set and Evaluation Method

We collected a sample of almost three months tweets between March 25th and June 17th from Sina microblogging platform. We got 22,724 tweets from authenticated users and 63,354 tweets from total users. For each of topics and each of tweets, we conduct the preprocessing of removing stopwords and chinese participle preprocessing. Then word frequency statistics are done and top 10 representative nouns or verb-nouns are extracted.

For a given query, the precision of a query suggestion method is defined as the fraction of suggestions generated that are meaningful. Note that since an exhaustive set of all possible suggestions for a given query is not available, recall cannot be computed. Also, for the query suggestion task, precision is a much more important metric than recall as the number of suggestions that can be offered is limited by the screen space. Precision is defined as

$$Precision@N = \frac{\#related\ words\ in\ a\ suggestion\ list}{N}. \tag{1}$$

In Equation 1, we take the extracted top 10 representative nouns or verb-nouns as the words that can represent a certain topic. We manually judge whether these words can be considered to accurately reflect the topic. The precision value of each topic is computed.

5.2 Experimental Results

Here, in Table 1, we have made a specific explanation for all abbreviations appeared in Table 2 which show the results of each microblog topic. Using the tweets from all users as suggestion context source is our baseline. The average results of all the 14 topics are listed in Table 3.

From the results that presented in Table 2, it seems that the differences between Atop10 precision and Ttop10 precision are not particularly obvious in terms of precision. For further observation, we decide to make an average for the precision values of 14 topics. To our surprise, the precision value of top 10 words that produced by the tweets that from authenticated users is higher than that produced by the tweets from all of users involved in a topic on average.

Table 3: The average precision values of all the 14 topics

	Average
Authenticated/Total	0.3575
Atop10 precision	0.4357
Ttop10 precision	0.4286
Atop5 precision	0.3286
Ttop5 precision	0.3214

Before conducting experiments, we think that the number of the total tweets for one topic actually not only contains the tweets from authenticated users, but also contain others tweets from common users. In comparison, it has the larger suggestion context source and the richer content information. Thus, it should output higher precision scores. However, the results run adversely to what we might intuitively expect the average precision value of top 10 words that produced by the tweets that authenticated users, i.e., slightly higher.

So what does this show? It illustrates that we do not need to select all tweets of a topic as our suggestion context source. Considering the final result, the tweets that from authenticated users could be on behalf of the entire tweets under a topic. During the preprocessing, we observed that under the background of computer configuration with a 32-bit operating system, dual-core CPU and 3.00GB memory, it takes about 4 hours for all users' tweets of a certain topic, but for processing authenticated users' tweets, it just takes about 30 minutes. Taking tweets that from authenticated users as our suggestion context source saves not only the processing time, but also the storage space. How much storage space does it save at all? From experimental data, we can see that average Authenticated/Total is about 0.3575. In other words, the authenticated context accounts for around 1/3 in total tweets and almost saves 2/3 storage space.

5.3 Discussions

There are some limitations in our approach. That is, for the words of a query that do not appear in the suggestion list of a topic, we cannot give a suggestion. In other words, our approach is based on the query that has already appeared in microblog topics, but they are little or even no in the history record of search engine. That is the so-called fresh query. The characteristic of high efficiency of information dissemination in microblog is our motivation to do this research.

6 Conclusions and Future Work

In this paper, we have introduced our approach for the suggestion of fresh search queries by mining microblog topics. We gave out the whole process that how to be access to microblog topics data and how to do text processing for these tweets until the words produced. It is worth mentioning that

[3] http://www.keenage.com

Table 1: Explanation for the abbreviations in Table 2

	Meaning
Authenticated tweets	The tweets that from authenticated users
Total tweets	The tweets that from all of users involved in a topic
Authenticated/Total	Authenticated tweets/ Total tweets
Atop10 precision	The precision of top 10 words produced by authenticated tweets
Ttop10 precision	The precision of top 10 words produced by total tweets

Table 2: The precision values of each topics

Topics	New ipad sale	Iphone news	Ipad show	Apple ceo salary	App Store	CES2012	HTC	Tablet pc	Kodak bankrupt	Huawei for new life	Iphone 4s sale	Windows 8	iOS jailbreak	Facebook
Authenticated tweets	2079	471	2420	1831	3005	1866	487	443	2472	340	2491	2454	604	1761
Total tweets	6043	1242	6889	5600	6760	6126	1237	1283	7137	1116	6714	6516	1645	5046
Authenticated/Total	0.344	0.3792	0.3513	0.327	0.4445	0.3046	0.3937	0.3453	0.3464	0.3047	0.371	0.3766	0.3672	0.3489
ATop10 precision	0.4	0.6	0.4	0.5	0.4	0.4	0.5	0.4	0.4	0.4	0.3	0.5	0.6	0.3
TTop10 precision	0.5	0.6	0.4	0.5	0.4	0.3	0.5	0.4	0.4	0.4	0.3	0.4	0.6	0.3
ATop5 precision	0.3	0.4	0.3	0.2	0.3	0.3	0.4	0.4	0.4	0.4	0.3	0.4	0.6	0.3
TTop5 precision	0.3	0.3	0.3	0.4	0.4	0.2	0.5	0.4	0.2	0.2	0.3	0.3	0.5	0.2

we not only extracted the tweets that from all of users involved in a topic, but also extracted the tweets that from authenticated users. Through the final experimental results, we can see that the average precision value of the top 10 words that produced by the tweets that from authenticated users is actually slightly higher. In addition, taking tweets that from authenticated users as our suggestion context source saves both the processing time and the storage space. In the future, an interesting topic is how to combine other social evidence to enhance query suggestion quality.

Acknowledgments

This research was undertaken as part of Project 61003130 funded by National Natural Science Foundation of China and Project 2011CDB254 founded by Ministry of Education of China.

References

[1] R. A. Baeza-Yates, C. A. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. In *Proceedings of Current Trends in Database Technology - EDBT 2004 Workshops*, pages 588–596. Springer, 2004.

[2] M. Barouni-Ebrahimi and A. A. Ghorbani. A novel approach for frequent phrase mining in web search engine query streams. In *Proceedings of Fifth Annual Conference on Communication Networks and Services Research (CNSR 2007)*, pages 125–132. IEEE Computer Society, 2007.

[3] S. Bhatia, D. Majumdar, and P. Mitra. Query suggestions in the absence of query logs. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, pages 795–804. ACM, 2011.

[4] P. Boldi, F. Bonchi, C. Castillo, D. Donato, and S. Vigna. Query suggestions using query-flow graphs. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, pages 56–63, 2009. ACM.

[5] A. Z. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion

of rare queries for sponsored search. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009*, pages 511–520. ACM, 2009.

[6] 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 the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 875–883. ACM, 2008.

[7] X. Chen, L. Li, G. Xu, Z. Yang, and M. Kitsuregawa. Recommending related microblogs: A comparison between topic and wordnet based approaches. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2012*. AAAI Press, 2012.

[8] S. Cucerzan and R. W. White. Query suggestion based on user landing pages. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 875–876. ACM, 2007.

[9] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the Eleventh International World Wide Web Conference, WWW2002*, pages 325–332. ACM, 2002.

[10] A. Feuer, S. Savev, and J. A. Aslam. Evaluation of phrasal query suggestions. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007*, pages 841–848. ACM, 2007.

[11] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon. Cross-lingual query suggestion using query logs of different languages. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 463–470. ACM, 2007.

[12] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006*, pages 387–396. ACM, 2006.

[13] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*, pages 371–378. ACM, 2009.

[14] H. Kwak, C. Lee, H. Park, and S. B. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 591–600. ACM, 2010.

[15] Z. Liu, X. Chen, and M. Sun. Mining the interests of chinese microbloggers via keyword extraction. *Frontiers of Computer Science in China*, 6(1):76–87, 2012.

[16] H. Ma, H. Yang, I. King, and M. R. Lyu. Learning latent semantic relations from clickthrough data for query suggestion. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 709–718. ACM, 2008.

[17] Q. Mei, D. Zhou, and K. W. Church. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 469–478. ACM, 2008.

[18] M. E. J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, Sep 2003.

[19] F. Silvestri. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval*, 4(1-2):1–174, 2010.

[20] Y. Song and L. wei He. Optimal rare query suggestion with implicit user feedback. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010*, pages 901–910. ACM, 2010.