Mining Search Subtopics from Query Logs

Dan Zhu, Jianwei Cui, Jun He, Xiaoyong Du
Key Labs of Data Engineering and Knowledge Engineering
Ministry of Education, China
School of Information, Renmin University of China
Beijing 100872, China
{zhudan, cjw, hejun, duyong}@ruc.edu.cn

Hongyan Liu
School of Economics and Management
Tsinghua University
Qinghuayuan, Haidian District
Beijing 100084, China
hyliu@tsinghua.edu.cn

ABSTRACT

Web queries are usually short and ambiguous. Subtopic mining plays an important role in understanding user's search intent and has attracted many researchers' attention. In this paper, we describe our approach to identify users' intents from query logs, which is a subtopic mining subtask of the NTCIR-9 Intent task for Chinese. We extract queries that are semantically related to the original query from query log, measure their similarities based on their relationship with urls, and cluster them into groups which represent different subtopics. In the experiment section, we show the results of our method evaluated by the organizers and a case study for one of the NTCIR-9 intent queries. The results shows that most found intents are different. The proposed method is easy to implement in real applications and can be computed quickly.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Subtopic mining, Query intent, Query logs

Team Name [DBIIR]

Subtasks/Languages [Chinese Subtopic Mining]

External Resources Used [Google][Baidu]

1. INTRODUCTION

Web queries are usually short and ambiguous. The query word might be polysemous and contain several subtopics. Therefore, different users often have different intents for the same query, which corresponds to the query's different subtopics. Thus, understanding different intents of users is a significant problem in information retrieval, which is critical for search engines to provide correct information for users and to enhance user's search experience.

We participate in the subtopic mining subtasks of the NTCIR-9 intent task for Chinese, and propose a method to mine subtopics of each query issued by users. Given an original query, queries that are semantically related to it are selected. Each of them is regarded as an interpretation of the original query and belongs to a subtopic. To evaluate the similarities between selected queries, we build query-url bipartite graph to model the relationship between them, and compute similarities based on link information among them. Similar queries may belong to the same intent, so we cluster selected queries and get different intents. We conducted evaluations of our measure based on SogouQ log data [1]. We also present a case study for one NTCIR-9 intent query. The results show that our measure is effective. Most found intents are different and they are ranked according to the possibility that each intent is related to the original query.

2. OUR APPROACH

We define related queries as queries that are semantically related to the original query, and may suggest different aspects of the original query. Then the query intent mining problem can be decomposed into the following problems. First, given a query, how can we find those related queries? Because different related queries are not orthogonal, some of them might represent the same intent. Consequently, candidates can be divided into several groups which represent different intents. Then, the second problem is how to measure the relationships between different related queries and find users' intents?

As mentioned above, query intents of the same query issued by different users may vary widely. Then what can we use to learn users' real intent? Fortunately the URL a user clicks indicates her intent to some extent. For example, if a user issued the query "*mars"* and the URL *www.mars.com* is clicked, we can infer that *mars* stands for the candy company. However, if the user click a URL that introduces solar system for the same query, we may know the user prefers the planet Mars.

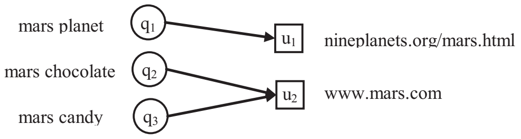

Figure 1 An example of query-url graph

Query logs records queries and corresponding URLs users clicked. Therefore, we propose a straightforward method to mine subtopics of web queries mainly based on query logs. Different users having same search intent may use different queries. For instance, when another user wants to know something about the planet, she may use the query *mars planet* instead of *mars*. Although the two queries are different, since they have the same intent, it is highly possible that the two users click similar documents or even the same one. Based on the above analysis, we build a query-url bipartite graph from query logs to model the click-through relationship between queries and urls. Figure 1 shows an example of query-url bipartite graph. In the graph, q_1, q_2, q_3 are queries, and u_1, u_2 are urls.

The detail of our method is as follows. First, we find related queries of the original one from the query log. Since SogouQ log data lacks the session information, we just choose queries that completely contain the original query string as substring as related queries. Considering the particular nature of Chinese, we use the segmentation techniques to eliminate noises. Noises mean queries that contain the original query, but have totally different meanings and are not related to original query actually. For instance, the query *灯红酒绿* (feasting and revelry) is regarded as a noise of *红酒*(red wine). The number of related queries may be huge. To cope with the efficiency problem, only a certain number of queries that occurs frequently are considered. Then, we build the bipartite graph based on related queries. Specifically, we treat queries and urls as nodes, and add edges between chosen related queries and corresponding urls if the url is once clicked for the query. The intuition behind our method is that users with similar intents will click similar urls. So we use a well-known link-based similarity measure, SimRank [2], to compute the similarity of related queries. After this procedure, queries with similar intents will have relative higher similarity scores. The next step is to cluster related queries according to the scores. Each cluster is considered to share the same intent. And we choose the query that occurs the most frequently in query log as a representative of certain intent.

Meanwhile, there are many queries that we can hardly find information from the log. For those queries, we put them in the mainstream search engines like Google and Baidu, and use their returned related queries. Then we use edit distance to measure the differences of those queries. And similar related queries are clustered together. For each cluster the proximal query to the original query is selected to express the intent. After these processes, we can get a list of possible intents for each query.

3. EVALUATION AND DISCUSSION

Paper [3] illustrates the mean I-rec, D-nDCG and D#-nDCG values for all participated runs. Table 1 shows the values for our run (DBIIR-S-C-1) while evaluating different number of top ranked items. According to the paper [3], D#-nDCG is a linear combination of intent recall (or "I-rec", which measures diversity) and D-nDCG (which measures overall relevance across intents).

Table 1. Mean I-rec, D-nDCG and D#-nDCG

I-rec@10	D-nDCG@10	D#-nDCG@10
0.4694	0.5620	0.5157
I-rec@20	D-nDCG@20	D#-nDCG@30
0.4926	0.3948	0.4437
I-rec@30	D-nDCG@30	D#-nDCG@30
0.4937	0.3068	0.4010

Table 2 presents results of our run for query *王子变青蛙* (The Prince changed to a frog). From the result, we can see the majority of them represent different intents of users. Meanwhile, the intents of most frequently searched queries rank high like watching and downloading the play. While space is limited in real applications, ranking like this can help to enhance users' search experiences. But some of the results are still duplicate and we will focus on how to generate more accurate intents with less redundancy in future work.

Table 2. Top 20 intents produced by our measure for query "王子变青蛙" (The Prince changed to a frog)

Intents	English translation
免费看《王子变青蛙》	Watch for free
《王子变青蛙》下载	Download
观看《王子变青蛙》	Watch
免费下载《王子变青蛙》	Download for free
《王子变青蛙》结局	Ending
电视剧：《王子变青蛙》	TV play
《王子变青蛙》16-20集	Episode 16 to 20
在线播放《王子变青蛙》	Watch online
《王子变青蛙》花絮	Tidbits
《王子变青蛙》分集介绍	Introduction to each episode
《王子变青蛙》音乐	Music
《王子变青蛙》的剧本	Script
《王子变青蛙》免费下载	Download for free
《王子变青蛙》电视剧	TV play
《王子变青蛙》最新消息	Latest news
《王子变青蛙》主题曲	Theme song
《王子变青蛙》的歌	Songs
《王子变青蛙》图	Pictures
《王子变青蛙》28 集	Episode 28
site:ent.sina.com.cn+<王子变青蛙>	Search in the specified site

4. CONCLUSIONS

We participate in the subtopic mining subtasks of the NTCIR-9 intent task for Chinese, and study the problem of subtopic mining in this paper. We find queries that are related to the original query, compute the similarities between different queries using link-based similarity measure, and then cluster queries into different intents. Our approach is easy to be computed, and experiments show that it is effective.

5. ACKNOWLEDGEMENT

The work was supported in part by the National Nature Science Foundation of China under Grant No. 70871068 and 70890083 and HGJ project 2010ZX01042-002-002-03.

6. REFERENCES

[1] http://www.sogou.com/labs/dl/q.html

[2] G. Jeh , J. Widom. SimRank: A Measure of Structural-Context Similarity. pp.538-543, SIGKDD 2002.

[3] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 Intent task. In Proc. of NTCIR, 2011