Qualifier Mining for NTCIR-INTENT

Haitao Yu
Faculty of Engineering, The University of Tokushima
2-1 Minamijyousanjima-cho, Tokushima
+81-88-656-7304.JAPAN.770-8506
yuhaitao@is.tokushima-u.ac.jp

Fuji Ren
Faculty of Engineering, The University of Tokushima
2-1 Minamijyousanjima-cho, Tokushima
+81-88-656-7304.JAPAN.770-8506
ren@is.tokushima-u.ac.jp

Song Liu
Faculty of Engineering, The University of Tokushima
2-1 Minamijyousanjima-cho, Tokushima
+81-88-656-7304.JAPAN.770-8506
liusong@is.tokushima-u.ac.jp

ABSTRACT
We participated the Subtopic Mining subtask of NTCIR-9 INTENT task. Query Log is used as the primary resource to mine latent subtopics. Through analysis of query log, we observed that queries describing similar information needs will use a similar group of qualifiers, which may also frequently occur together within queries. We introduced the concept of qualifier graph for subtopic mining. To solve the sparseness problem, the search snippets returned by web search engines are used. The experiment results show that it is reasonable to make use of qualifier to mine latent subtopics.

Keywords
Subtopic Mining; Query Log; Qualifier Clustering

[TUTA1-Chinese Subtopic Mining][Google]

1. INTRODUCTION
Nowadays search engines are the primary ways of information access on the web. When an information-need is being formulated in users' mind, queries in the form of a sequence of words will be typed into the search box, ideally, the search engine should respond with a ranked list of snippet results that best meets the needs of users. Unfortunately, many queries are ambiguous and/or underspecified. An inherent ambiguity with respect to short queries occurs in many circumstances and users may seek for different information underlying the same query. For an ambiguous query, it may refer to different interpretations, e.g. "windows" may refer to Microsoft Windows software or house windows. For a query on a broad topic, users may seek for different aspects, i.e., profile, album, songs and concerts are all hot aspects for query "Michael Jackson". Despite this, retrieval models, in general, have not focused on explicitly representing users' intent, and query processing has just been limited to simple transformations such as stemming or spelling correction[1]. Therefore, it has been recognized as a crucial part of effective information retrieval to understand users' information need or intent that underlies the submitted query and diversify the results retrieved for ambiguous query maximizing the satisfaction of users with different intents.

Towards this direction, NTCIR-9 proposed the Intent task, which explores the above problem from the following two aspects:

(1) How to mine underlying intents/subtopics;

(2) How to selectively diversify search results;

This paper describes our work for the first subtask (Chinese). The remainder of this paper is organized as follows. We give a systematic review of the related work in section 2. Section 3 details the analysis of query log that inspired our approach. In section 4, we detail the overall experiments and the results we obtained. Finally we conclude our work in section 5.

2. Related Work
Query suggestion or query recommendation is a key technique for generating alternative queries to help users drill down to a subtopic of the original query[2-4]. Different from query suggestion or query recommendation, subtopic mining focuses more on the diversity of possible subtopics of the original query rather than merely inferring relevant queries. Jian Hu[5] integrated the knowledge contained in Wikipedia to predict the possible intents for a given query. A number of intent seeds are iteratively propagated through Wikipedia structure with Markov random walk. Filip Radlinski[6] proposed an approach for inferring query intents from reformulations and clicks. For an input query, the click and reformulation information are combined to identify a set of possibly related queries to construct an undirected graph. An edge is introduced between two queries if they were often clicked for the same documents. Finally, the random walk similarity is used to find intent cluster. Eldar Sadikov[7] integrated the session co-occurrence information to cluster the refinements of a query based on the underlying intents. The possibility of a drift in user intent to another topic was also considered.

Query log records the interactive activities of huge amounts of users, the submitted queries, session information and so on. In previous work, many approaches were proposed by researchers to mine the wealth of information hidden in the query log. In our work, query log is also used as the primary resource. The detailed analysis is introduced in next section.

3. Exploration of Query Log
In this section, we explore the valuable knowledge contained in Query Log for subtopic mining. Take the dry-run topic "霸王别姬" in NTCIR-9[1] for example (Figure1). It consists of two parts, the query part (denoted as <query>) and the subtopic part (denoted as <subtopic>).

```
<query>霸王别姬</query>
<subtopic number="1">霸王别姬(电影)下载</subtopic>
<subtopic number="2">霸王别姬(电影)在线观看</subtopic>
<subtopic number="3">霸王别姬(屠洪纲)</subtopic>
<subtopic number="4">京剧霸王别姬</subtopic>
<subtopic number="5">霸王别姬 张国荣</subtopic>
<subtopic number="6">电影霸王别姬的一般信息</subtopic>
<subtopic number="7">二人转霸王别姬</subtopic>
<subtopic number="8">Others</subtopic>
```

[1] http://research.nii.ac.jp/ntcir/ntcir-9/index.html

Figure 1. An dry-run topic in NTCIR-9.

As shown in figure1, around the topic "霸王别姬", there are many aspects that users may be interested in, like "电影下载", "京剧", "二人转", etc. There is no doubt that search engines will be confused if only "霸王别姬" is used as the query without specifying the interested subtopic or aspect.

From SogouQ[2] we extract the queries that relates the topic "霸王别姬" to see what the users have done when they are searching information around the topic "霸王别姬". Two types of queries are extracted, the first type is the queries that contain the topic word "霸王别姬", the second type is the queries that occurred in the same session with the query of the first type. Table 1 shows the extracted queries, which are also manually segmented.

Table 1. User queries extracted from SogouQ

Queries	Manual Segmentation	User Number
霸王别姬	霸王别姬	137
霸王别姬 +李碧华	霸王别姬 李碧华	72
免费电影 +霸王别姬	免费 电影 霸王别姬	38
霸王别姬 下载/+下载	霸王别姬 下载	31
电影霸王别姬	电影 霸王别姬	21
…	…	…

From table 1, we observed that:

(1) A large proportion of users directly submitted underspecified queries, which poses challenges for search engines;

(2) Some users would like to indicate their interested aspects for an multi-aspect topic, like "李碧华", "下载" for "霸王别姬", we call this kind of words as *qualifier*, and the topic word "霸王别姬" as *subject-concept*;

(3) Though some qualifiers appear at different position of subject-concept, they means the same intent in deed, such as "下载" and "电影".

(4) There are semantic associations among qualifiers.① There are grouped qualifiers that have other qualifiers as sub-intents, such as "免费电影" and "电影". ② The same intent is represented using different qualifiers by different users.

To get a vivid understanding of the qualifiers, we illustrate the relationships from the view of graph as figure 2.

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

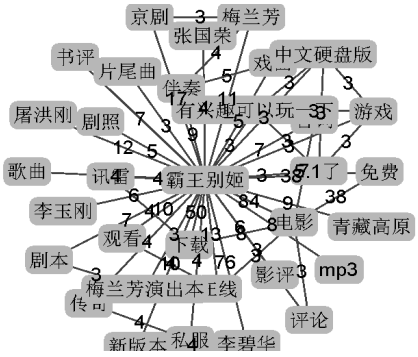

Figure 2. Co-occur graph of qualifier and subject-concept.

The set of nodes are the qualifiers plus the subject-concept. We add an edge between two nodes if the corresponding words co-occurred in one query, the weight of an edge corresponds to the co-occurred frequency. To simplify the graph in figure 2, we removed the node of subject-concept. The resulting graph composing of all the qualifiers is shown as figure 3.

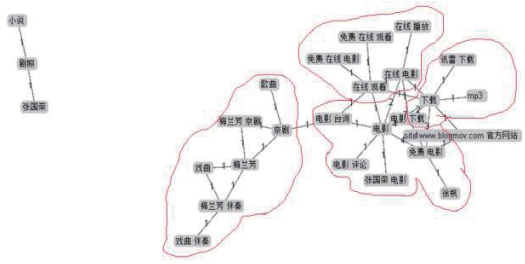

Figure 3. Qualifier graph for topic "霸王别姬"

Latent Dirichlet Allocation[8] (LDA) is a classic generative model in topic modeling that discovers latent semantic topics in collections of text documents. The key insight behind LDA is the premise that words in documents carry strong semantic information about the document's topic, and documents discussing similar topics will use a similar group of words. Thus, LDA posits that latent topics can be discovered by identifying groups of words that frequently co-occur within documents. Inspired by the LDA model, according to the analysis of query log stated above, it is reasonable to assume that queries describing similar information needs will use a similar group of qualifiers, which may also frequently occur together within queries. The region differentiated by red-circles in figure 3 also demonstrates the assumption. These regions of qualifier graph can be intuitively interpreted as subtopics of the subject "霸王别姬".

As query log records the activities of huge amounts of users and the interested aspects of one topic is consistent over time. The qualifier words submitted by users for a specific topic would be invaluable for subtopic discovery. Going further, we can take advantage of the wisdom of crowds to mine the possible subtopics. Thus, the subtopic mining problem can be viewed as a qualifier clustering problem.

The first challenge is how to properly segment a query to identify the subject-concept and qualifiers. Through prior works[9-10], it would be error prone if we directly project natural language structures onto user queries. The well-known bag-of-words model[11] simply assumes that a text such as a sentence or a document can be represented as an unordered collection of words,

disregarding grammar and even word order. Inspired by bag-of-words model, we use bag-of-units model to simplify the unique structural properties of queries. In bag-of-units model, each segment refers to a semantic unit, which can be a single word, an idiom, named entities or multi-word expressions. A query is represented as an unordered collection of unit, disregarding grammar and segment order. Named entities and noun phrase are proven to be reliable for key concept discovery in past works on information retrieval and natural language processing. We use named entities or noun phrases extracted from the given query as subject-concepts. For instance, the query "霸王别姬下载" presented in figure 1 can be split into "霸王别姬" and "下载", "霸王别姬" is taken as subject-concept, and "下载" is taken as qualifier.

The second challenge is how to discover the latent subtopic clusters. We introduce the concept of qualifier graph. For a given query, when the subject-concepts and qualifiers are identified, we can extract a set of queries from query log. Each query either co-occurred in the same session with the given query or has the same subject-concepts and all the co-session queries. Then each query will be segmented, and all the qualifiers are used to construct the qualifier graph. An edge is introduced between two qualifiers if the two qualifiers appeared together in one query. The weight of an edge corresponds to the co-occurred frequency. Then we formulate the subtopic mining problem into an overlapping community discovery problem. Our approach is based on an modified version of the star clustering algorithm[12] (co-occurred frequency is used as the parameter to control granularity), each star-shaped cluster consists of a center qualifier and several satellite qualifiers. Based on these star-shaped clusters, we use Width First Traversing to construct subtopic community. To solve the sparseness problem, we firstly submit the topic to a web search engine, the top-n (n is an experienced value) returned snippets are collected as snippets corpus. Secondly, we extract NP and VP segments from the snippet corpus using Stanford Parser[3]. Each snippet is viewed as a user query simulating queries submitted by users. The extracted NP and VP segments of each snippet are view as co-occurred qualifiers. We add an edge between two co-occurred qualifiers to construct the qualifier graph.

4. Experiment Results in NTCIR-9

4.1 Dataset

For the Chinese subtask, the SogouT[4] corpus is provided as document collection, the Chinese query log called SogouQ is provided as additional resource. We take the SogouQ as the query log instance in later sections.

4.2 Results and Discussion

For the formal run of Subtopic Mining subtask, a set of 100 Chinese topics are provided, which were selected from the June 2008 query log of Sogou. Participants were required to submit ranked list of subtopics for each topic[13]. I-rec, D-nDCG and D#-nDCG[14] are used as evaluation metrics, I-rec measures diversity, D-nDCG measures overall relevance across intents, D#-nDCG is a linear combination of I-rec and D-nDCG, which is used as the primary evaluation metric.

For each formal topic, we first identify the subject-concept, then construct the corresponding qualifier graph. When mining the subtopic communities, the co-occur frequency threshold is 3. Each obtained subtopic community is ranked by the sum of frequency of each member qualifier. For fresh topic that query log contains little information, the top-50 returned snippets by Google is used to extract NP and VP segments to construct qualifier graph. Table 3-5 show the mean intent recall, D-nDCG and D#-nDCG values for different measurement depths (i.e. number of top ranked items to be evaluated) of l=10,20 and 30. The runs of all participants are sorted by D#-nDCG.

Table 2. Runs ranked by D#-nDCG@10

Run Name	I-rec @10	D-nDCG @10	D#-nDCG @10
THU-S-C-2	0.4801	0.7186	0.5993
MSINT-S-C-2	0.5130	0.6806	0.5968
THU-S-C-3	0.4828	0.7107	0.5967
THU-S-C-1	0.4946	0.6896	0.5921
ICTIR-S-C-1	0.5161	0.6434	0.5797
uogTr-S-C-5	0.4947	0.6598	0.5772
MSINT-S-C-4	0.4864	0.6604	0.5734
ICTIR-S-C-4	0.5035	0.6417	0.5726
ICTIR-S-C-2	0.4826	0.6576	0.5701
HITIR-S-C-5	0.4936	0.6449	0.5693
ISCAS-S-C-1	0.5022	0.6336	0.5679
ICTIR-S-C-3	0.4808	0.6530	0.5669
HITIR-S-C-1	0.4854	0.6453	0.5653
ISCAS-S-C-3	0.4910	0.6386	0.5648
MSINT-S-C-1	0.5002	0.6240	0.5621
NTU-S-C-2	0.4683	0.6546	0.5615
MSINT-S-C-5	0.4578	0.6543	0.5560
NTU-S-C-3	0.4807	0.6308	0.5558
HITIR-S-C-4	0.4738	0.6291	0.5514
HITIR-S-C-3	0.4738	0.6291	0.5514
HIT2jointNLP Lab-S-C-2	0.4596	0.6407	0.5501
MSINT-S-C-3	0.4587	0.6256	0.5422
ICTIR-S-C-5	0.4714	0.5832	0.5273
DBIIR-S-C-1	0.4694	0.5620	0.5157
HIT2jointNLP Lab-S-C-1	0.4240	0.5946	0.5093
TUTA1-S-C-1	***0.3405***	***0.6762***	***0.5084***
NTU-S-C-1	0.4335	0.4836	0.4586
ISCAS-S-C-4	0.3062	0.4810	0.3936
KLE-S-C-3	0.3185	0.4461	0.3823
KLE-S-C-1	0.3162	0.4466	0.3814
KLE-S-C-2	0.3162	0.4464	0.3813
ISCAS-S-C-2	0.3019	0.4491	0.3755
THU-S-C-5	0.2888	0.4455	0.3672
III_CYUT_NTHU-S-C-1	0.3085	0.4099	0.3592
THU-S-C-4	0.2654	0.4040	0.3347
UWat-S-C-2	0.3324	0.3355	0.3340
uogTr-S-C-1	0.3210	0.3385	0.3297
uogTr-S-C-4	0.3176	0.3364	0.3270
UWat-S-C-1	0.2388	0.2492	0.2440
uogTr-S-C-2	0.1753	0.1772	0.1763

[3] http://nlp.stanford.edu/software/lex-parser.shtml

[4] http://www.sogou.com/labs/dl/t.html

uogTr-S-C-3	0.1682	0.1698	0.1690
HITIR-S-C-2	0.0393	0.0242	0.0317

Table 3. Runs ranked by D#-nDCG@20

Run Name	I-rec @20	D-nDCG @20	D#-nDCG @20
ICTIR-S-C-1	0.6997	0.6162	0.6579
THU-S-C-3	0.6311	0.6727	0.6519
THU-S-C-2	0.6227	0.6739	0.6483
ICTIR-S-C-2	0.6444	0.6460	0.6452
ISCAS-S-C-3	0.6478	0.6370	0.6424
THU-S-C-1	0.6311	0.6508	0.6409
ISCAS-S-C-1	0.6406	0.6387	0.6397
HITIR-S-C-5	0.6421	0.6180	0.6300
MSINT-S-C-2	0.6066	0.6462	0.6264
HITIR-S-C-1	0.6316	0.6213	0.6264
MSINT-S-C-4	0.6293	0.6008	0.6150
HITIR-S-C-4	0.6235	0.6027	0.6131
HITIR-S-C-3	0.6235	0.6027	0.6131
MSINT-S-C-5	0.6069	0.6122	0.6096
MSINT-S-C-3	0.5962	0.5852	0.5907
ICTIR-S-C-4	0.6206	0.5579	0.5893
ICTIR-S-C-3	0.5849	0.5913	0.5881
MSINT-S-C-1	0.6187	0.5506	0.5846
TUTA1-S-C-1	***0.4794***	***0.6677***	***0.5735***
HIT2jointNLP Lab-S-C-2	0.5583	0.5736	0.5660
ICTIR-S-C-5	0.5803	0.5427	0.5615
HIT2jointNLP Lab-S-C-1	0.5116	0.5498	0.5307
uogTr-S-C-5	0.4947	0.4278	0.4612
NTU-S-C-2	0.4683	0.4242	0.4463
NTU-S-C-3	0.4807	0.4090	0.4449
DBIIR-S-C-1	0.4926	0.3948	0.4437
UWat-S-C-2	0.4945	0.3893	0.4419
KLE-S-C-3	0.4482	0.4344	0.4413
KLE-S-C-2	0.4443	0.4329	0.4386
KLE-S-C-1	0.4443	0.4326	0.4385
ISCAS-S-C-4	0.4053	0.4626	0.4340
ISCAS-S-C-2	0.3922	0.4434	0.4178
uogTr-S-C-1	0.4187	0.3670	0.3929
THU-S-C-5	0.3567	0.4286	0.3926
III_CYUT_NTHU-S-C-1	0.3890	0.3946	0.3918
uogTr-S-C-4	0.4170	0.3662	0.3916
THU-S-C-4	0.3568	0.3967	0.3767
NTU-S-C-1	0.4335	0.3140	0.3738
UWat-S-C-1	0.3236	0.2459	0.2847
uogTr-S-C-2	0.2407	0.1691	0.2049
uogTr-S-C-3	0.2245	0.1796	0.2020
HITIR-S-C-2	0.0416	0.0226	0.0321

Table 4. Runs ranked by D#-nDCG@30

Run Name	I-rec @30	D-nDCG @30	D#-nDCG @30
ICTIR-S-C-2	0.7070	0.5895	0.6482
THU-S-C-3	0.6844	0.6101	0.6473
ISCAS-S-C-1	0.6861	0.5783	0.6322
ICTIR-S-C-1	0.7224	0.5299	0.6261
THU-S-C-2	0.6663	0.5750	0.6206
THU-S-C-1	0.6686	0.5667	0.6176
ISCAS-S-C-3	0.6884	0.5419	0.6152
HITIR-S-C-5	0.6730	0.5529	0.6130
HITIR-S-C-1	0.6634	0.5531	0.6083
MSINT-S-C-5	0.6500	0.5412	0.5956
MSINT-S-C-4	0.6638	0.5150	0.5894
MSINT-S-C-2	0.6275	0.5390	0.5832
HITIR-S-C-4	0.6479	0.5182	0.5830
HITIR-S-C-3	0.6479	0.5182	0.5830
MSINT-S-C-3	0.6218	0.5022	0.5620
MSINT-S-C-1	0.6432	0.4662	0.5547
ICTIR-S-C-3	0.6062	0.4867	0.5464
ICTIR-S-C-4	0.6340	0.4441	0.5390
TUTA1-S-C-1	***0.4982***	***0.5602***	***0.5292***
HIT2jointNLP Lab-S-C-2	0.5775	0.4681	0.5228
ICTIR-S-C-5	0.5924	0.4407	0.5165
HIT2jointNLP Lab-S-C-1	0.5297	0.4566	0.4931
KLE-S-C-2	0.4769	0.3712	0.4241
KLE-S-C-1	0.4769	0.3709	0.4239
ISCAS-S-C-4	0.4394	0.4066	0.4230
KLE-S-C-3	0.4776	0.3677	0.4226
ISCAS-S-C-2	0.4320	0.4059	0.4189
UWat-S-C-2	0.5110	0.3154	0.4132
uogTr-S-C-5	0.4947	0.3309	0.4128
DBIIR-S-C-1	0.4953	0.3068	0.4010
NTU-S-C-3	0.4807	0.3163	0.3985
NTU-S-C-2	0.4683	0.3278	0.3980
uogTr-S-C-4	0.4534	0.3386	0.3960
uogTr-S-C-1	0.4534	0.3386	0.3960
THU-S-C-4	0.3917	0.3507	0.3712
THU-S-C-5	0.3803	0.3598	0.3700
III_CYUT_NTHU-S-C-1	0.3890	0.3042	0.3466
NTU-S-C-1	0.4335	0.2432	0.3384
UWat-S-C-1	0.3270	0.1926	0.2598
uogTr-S-C-2	0.2640	0.1470	0.2055
uogTr-S-C-3	0.2422	0.1650	0.2036
HITIR-S-C-2	0.0432	0.0197	0.0315

From the obtained results in above tables, we can draw the following conclusions for the proposed approach: (1) The I-rec value is low compared with the top results. One reason is that some subtopics have a low frequency in query log and it is hard to capture by qualifier graph. Another reason is that we rely mainly on query log. Other resources can also be integrated. (2) The D-nDCG value is acceptable. The reason is that all the obtained subtopics are mainly formulated by users, which is compatible with the structural property of query itself.

5. Conclusions and Future Work

In our work of NTCIR-9, we tested the proposed technique based on qualifier graph mined from query log in order to take advantage of the wisdom of crowds to organize subtopics indicated by users. Experimental results show that it is reasonable to make use of qualifier for subtopic mining.

In the future, we will make further efforts to study the query that has several subject-concepts, integrate more resources to compare performances. The sparseness problem of fresh topic is another challenge when using query log to mine information.

6. REFERENCES

[1] Hang Li, Gu Xu, Bruce Croft, etc. Query Representation and Understanding. Proceeding of the Second Workshop on Query Representation and Understanding.

[2] Mehran Sahami, Timothy D. Heilman. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland, 2006.

[3] Zhiyong Zhang , Olfa Nasraoui. Mining Search Engine Query Logs for Query Recommendation. Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland, 2006.

[4] Qiaozhu Mei, Dengyong Zhou and Kenneth Church. Query Suggestion Using Hitting Time. Proceeding of the 17th ACM conference on Information and knowledge management, California, USA, 2008.

[5] Jian Hu, Gang Wang, Fred Lochovsky, etc. Understanding User's Query Intent with Wikipedia. Proceedings of the 18th international conference on World Wide Web, Madrid, Spain, 2009.

[6] Filip Radlinski, Martin Szummer and Nick Craswell. Inferring Query Intent from Reformulations and Clicks. Proceedings of the 19th international conference on World Wide Web, North Carolina, USA, 2010.

[7] Eldar Sadikov, Jayant Madhavan, Lu Wang, etc. Clustering Query Refinements by User Intent. Proceedings of the 19th international conference on World Wide Web, North Carolina, USA, 2010.

[8] David M. Blei, Andrew Y. Ng, Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, Vol. 3, No. 5, pp. 993-1022, 2003.

[9] Nikita Mishra, Rishiraj Saha Roy, Niloy Ganguly, etc. Unsupervised Query Segmentation Using Only Query Logs. Proceedings of the 20th international conference companion on World Wide Web, Hyderabad, India, 2011.

[10] Matthias Hagen, Martin Potthast, Benno Stein, etc. Query Segmentation Revisited. Proceedings of the 20th international conference companion on World Wide Web, Hyderabad, India, 2011.

[11] David D. Lewis. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 4-15, 1998.

[12] J. A. Aslam, E. Pelekhov, D.Rus. The Star Clustering Algorithm For Static And Dynamic Information Organization. Journal of Graph Algorithms and Applications, Vol. 8, No. 1. pp. 95-129, 2004.

[13] R. Song, T. Sakai, M. Zhang, etc. Overview of the NTCIR-9 INTENT Task. NTCIR-9 Workshop Meeting, Tokyo, Japan, 2011.

[14] T. Sakai, R. Song. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, Beijing, China, pp. 1043-1052, 2011.