ISCAS at Subtopic Mining Task in NTCIR9

Xue Jiang, Xianpei Han, Le Sun
Institute of Software, Chinese Academy of Sciences
P.O.Box 8718, Beijing, 100190, P.R. China
{jiangxue, xianpei, sunle}@nfs.iscas.ac.cn

ABSTRACT

In this paper, we describe our work at subtopic mining subtask in NTCIR-9 in simplified Chinese. To find possible subtopics of a specific query, we select related queries recorded by query log, or titles of searching results provided by Google and Baidu, or the catalog of corresponding entry in Baidu encyclopedia, which are lexically similar as the original query, then we apply k-means algorithm to cluster these candidate queries with different k (k=5, 10), and rank these queries with consideration of similarities and clusters.

Keywords

NTCIR, subtopic mining, query log

1. INTRODUCTION

In our entry to NTCIR-9 INTENT task [6], we focus on the Subtopic Mining subtask, and trying to develop a system to automatically find all possible subtopics of a given query and organize these subtopics for diversity.

In the Subtopic Mining subtask, we should return a ranked list of subtopics in response to a given query. According to the task, a subtopic is a specific interpretation of an ambiguous query, or an aspect of a faceted query. For example, “apple ipod” or “apple fruit” in response to “apple”, or “windows 7 update” in response to “windows 7”.

Finding possible subtopics of a query may be beneficial for diversifying query suggestions and searching results. As we know, many queries constructed by users are ambiguous and difficult to find precise documents. If we give all possible subtopics of this query, user would select what he/she exactly wants from these query suggestions, or we can offer documents in different subtopics to improve user experience.

For this subtopic mining task, we have submitted 4 different runs on different corpus or different numbers of clusters. There are two basic steps for all of these 4 runs. Firstly, we find the related queries in the corpus, such as query log, searching results, and assign each subtopic a score according to the similarity. Then, we utilize K-means algorithm to cluster these subtopics and rank them with consideration of the score assigned at first step and the results of clustering.

We submitted 2 standard runs which are developed on the Sogou query log [1] only, and the other 2 nonstandard runs on opened data set, including the titles of retrieved documents by Google and Baidu, and the catalogue of corresponding entries in Baidu encyclopedia. The experiment results shows that the opened data set would improve the effectiveness greatly, especially the user generated data, i.e., the catalog of corresponding entries in Baidu encyclopedia.

The reminder of this paper will introduce some related work in this field and then describe the details of our work and compare our results with other teams and runs. At last, we list some difficulties we have found.

2. RELATED WORK

The INTENT task is a new NTCIR task, and there is little work on mining subtopic before. Uluhan and Badur try to mine subtopics from top-ranked web pages returned by a search engine, then extract key phrases and apply data mining algorithm to find candidate subtopics [7]. Besides this, we can refer to some work which has the similar goal. Many researchers focus on finding words related to the user query to expand the original query, we may get some inspiration from these researches.

There are many researches focus on modeling the dependencies between words. Metzler and Croft model dependencies between query terms and latent concepts through Markov random fields, and select top k concepts to expand original queries [4]. Lang et al improved this work by using hierarchical Markov random fields to model term dependencies [3]. Wang and Zhai proposed a contextual model by investigating the context similarity of terms in history queries [9]. Two terms with similar context are used to substitute each other in candidate query generation. Then a context based translation model is employed to score the candidate queries. Jones et al. employed hypothesis likelihood ratio to identify those highly related query phrases or term pairs in user sessions [2].

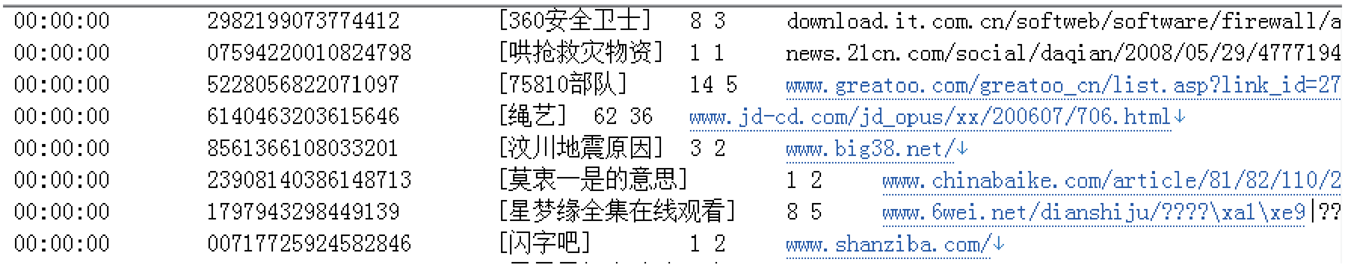
Diversity is also a hot topic in recent years. Researchers proposed kinds of approaches to satisfy the requirement of diversity. In [1, 5], the authors used Scatter/Gather algorithm to cluster the top documents returned from a traditional information retrieval system. In [10], supervised learning algorithms are studied to extract meaningful phrases from the search result snippets and these phrases are then used to group search results. Wang and Zhai apply star clustering algorithm to find the central concept of each cluster [8].

3. OUR APPROACH

The subtopic mining task aims to find all possible subtopics of a given query, or aspects of a faceted query. This section will introduce our approaches to find subtopics of a queriy and how to rank them.

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

NTCIR-9 Workshop Meeting, 2011, Tokyo, Japan.



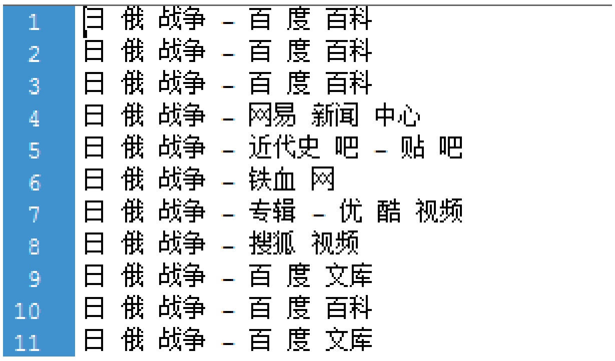
```
00:00:00        2982199073774412        [360安全卫士]   8 3     download.it.com.cn/softweb/software/firewall/a
00:00:00        07594220010824798       [哄抢救灾物资]  1 1     news.21cn.com/social/daqian/2008/05/29/4777194
00:00:00        5228056822071097        [75810部队]     14 5    www.greatoo.com/greatoo_cn/list.asp?link_id=27
00:00:00        6140463203615646        [绳艺]  62 36   www.jd-cd.com/jd_opus/xx/200607/706.html↓
00:00:00        8561366108033201        [汶川地震原因]  3 2     www.big38.net/↓
00:00:00        23908140386148713       [莫衷一是的意思]        1 2     www.chinabaike.com/article/81/82/110/2
00:00:00        1797943298449139        [星梦缘全集在线观看]    8 5     www.6wei.net/dianshiju/????\xa1\xe9|??
00:00:00        00717725924582846       [闪字吧]        1 2     www.shanziba.com/↓
```

Figure 1, Samples of Sogou query log (SogouQ).

3.1 Data

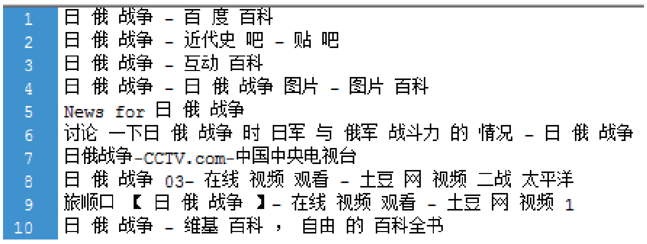
For the standard run, we are allowed to use the query log and document sets released by Sogou. This query log contains a large number of queries issued by users in June, 2008. It records some useful information, including time, query string, clicked url, document rank, and order of clicks. Figure 1 shows some samples of this query log.

Observing the experiment results of standard runs, we found that there are no proper subtopics for some queries in the evaluation set, such as the first query "日俄战争". Based on this observation, we crawled several kinds of data from the web to supplement the query log, including the searching results of Baidu and Google, and the catalogue of Baidu encyclopedia. We submitted these queries to the search engine, i.e., Baidu and Google, then we crawled the top 100 documents and extracted the titles to replace corresponding documents. Similarly, we searched these queries in Baidu encyclopedia and extracted the catalogue of corresponding entries if existed. Figure 2,3,4 shows the samples of these data respectively.



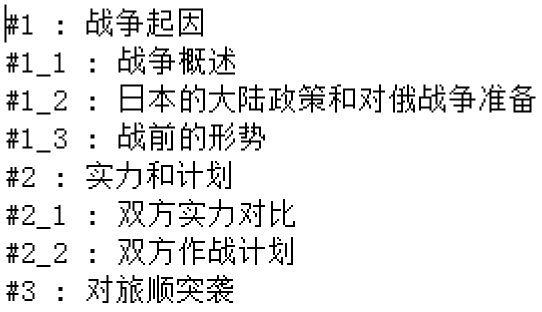
```
日 俄 战争 - 百 度 百科
日 俄 战争 - 百 度 百科
日 俄 战争 - 百 度 百科
日 俄 战争 - 网易 新闻 中心
日 俄 战争 - 近代史 吧 - 贴 吧
日 俄 战争 - 铁血 网
日 俄 战争 - 专辑 - 优 酷 视频
日 俄 战争 - 搜狐 视频
日 俄 战争 - 百 度 文库
日 俄 战争 - 百 度 百科
日 俄 战争 - 百 度 文库
```

Figure 2, Samples of titles of searching results from Baidu



```
日 俄 战争 - 百 度 百科
日 俄 战争 - 近代史 吧 - 贴 吧
日 俄 战争 - 互动 百科
日 俄 战争 - 日 俄 战争 图片 - 图片 百科
News for 日 俄 战争
讨论 一下日 俄 战争 时 日军 与 俄军 战斗力 的 情况 - 日 俄 战争
日俄战争-CCTV.com-中国中央电视台
日 俄 战争 03- 在线 视频 观看 - 土豆 网 视频 二战 太平洋
旅顺口 【 日 俄 战争 】- 在线 视频 观看 - 土豆 网 视频 1
日 俄 战争 - 维基 百科 ， 自由 的 百科全书
```

Figure 3, Samples of titles of searching results from Google



```
#1 ： 战争起因
#1_1 ： 战争概述
#1_2 ： 日本的大陆政策和对俄战争准备
#1_3 ： 战前的形势
#2 ： 实力和计划
#2_1 ： 双方实力对比
#2_2 ： 双方作战计划
#3 ： 对旅顺突袭
```

Figure 4, samples of catalogue of entry in Baidu encyclopedia, the number before the colon indicates the structure of a catalogue.

3.2 Find Possible Subtopics

Recall the definition of subtopics and the examples illustrated above (i.e., "apple ipod" or "apple fruit" are subtopics of "apple", and "windows 7 update" is a subtopic of "windows 7".), we can find that the query is a substring of its subtopics.

Most of the queries in this task contain more than one words, so we use $Q = q_1 q_2 \cdots q_n$ and $S = s_1 s_2 \cdots s_n$ to denote a query and a corresponding subtopic respectively, where q_i and s_i denotes one word. We observed that if S is a subtopic of query Q, it may contain all words in Q, all some words of query Q. We define the common words between query Q and subtopic S as $C = c_1 c_2 \cdots c_n$, where c_i is the word which appears both in query Q and subtopic S. According to our observation, the string C may be not at the same position in Q and S, and sometimes these words in C are not even of the same order in Q and S. So it is not appreciate to judge whether a query string S is a subtopic of another query Q by exact matching. We assume that these words in Q and S are independent and use bag of words to calculate the similarity between a given query Q and any other query S. If the similarity is larger than a pre-defined threshold δ, we regard this query S is a subtopic of query Q. We can calculate the similarity with equation (1) as follows

$$Sim(Q,S) = \frac{\sum_{w \in Q \cap S} c_Q(w) c_S(w)}{|Q| \bullet |S|} \qquad (1)$$

Where $c_Q(w)$ means the count of word w in query Q, and $|Q|$ means the length of query Q, the definition of $c_S(w)$ and $|S|$ are similar.

3.3 Cluster and rank

We can find many possible subtopics through the first step described in the last section. But we must face another question, that many subtopics found in the first step are duplicated, or of the same aspect. To satisfy the requirement of diversity, we need to cluster these subtopics and select some representative subtopics for further application, such as query suggestion, clustering the searching results.

We utilize K-means algorithm to cluster these subtopics, where K is different in different runs, because we want to examine how the granularity of cluster will influence the results. In our experiments, we set K = 5 and 10.

We assign a score for each cluster to rank these clusters. If a cluster $G = \{S_1, S_2 \cdots S_n\}$, the score of G equals to the accumulation of the similarities of all subtopics in cluster G (i.e., equation (2)).

$$Score(G) = \sum_{i=1}^{n} Sim(Q, S_i) \qquad (2)$$

In each cluster, we rank the subtopics with respect to their similarities calculated before. To rank all of these subtopics, we rank the clusters with corresponding score first, and then iteratively select the top subtopics in each cluster.

4. RESULTS

We submitted 4 different runs for this task, 2 of them are based on the standard data set (Sogou query log, i.e., SogouQ), we call them "standard runs", and another 2 runs are called as "non-standard runs", because the subtopics are found not only from SogouQ, but also from the searching results of Baidu, Google and the catalog of corresponding entry in Baidu encyclopedia.

As mentioned above, the query S will be a subtopic of a given query Q once their similarity is larger than δ. In our experiments we set $\delta = 0.8$. For K-means algorithm, we set k =5 and 10 respectively.

Table 1 describes the basic information of each run we submitted, including name, data, and number of clusters.

4.1 Evaluation

The primary evaluation metric used in this task is D#-nDCG, which is a linear combination of intent recall and D-nDCG. In this task, the assessors select the top 10, 20, 30 subtopics of each run for evaluation. Table 2 shows the evaluation results of our runs with different measure depths.

Runs	Data	Num of clusters
ISCAS-S-C-1	SogouQ, Baidu, Google, Baidu encyclopedia	10
ISCAS-S-C-2	SogouQ,	10
ISCAS-S-C-3	SogouQ, Baidu, Google, Baidu encyclopedia	5
ISCAS-S-C-4	SogouQ	5

Table 1: description of our runs

Runs	I-rec	D-nDCG	D#-nDCG
ISCAS-S-C-1	0.5022*#	0.6336*#	**0.5679***#
ISCAS-S-C-2	0.3019	0.4491	0.3755
ISCAS-S-C-3	0.491*#	0.6386*#	0.5648*#
ISCAS-S-C-4	0.3062	0.481	0.3936*

Table 2(a): evaluation results with top 10 subtopics

Runs	I-rec	D-nDCG	D#-nDCG
ISCAS-S-C-1	0.6406*#	0.6387*#	0.6397*#
ISCAS-S-C-2	0.3922	0.4434	0.4178
ISCAS-S-C-3	0.6478*#	0.637*#	**0.6424***#
ISCAS-S-C-4	0.4053	0.4626	0.434

Table 2(b): evaluation results with top 20 subtopics

Runs	I-rec	D-nDCG	D#-nDCG
ISCAS-S-C-1	0.6861*#	0.5783*#	**0.6322***#
ISCAS-S-C-2	0.432	0.4059	0.4189
ISCAS-S-C-3	0.6884*#	0.5419*#	0.6152*#
ISCAS-S-C-4	0.4394	0.4066	0.423

Table 2(c): evaluation results with top 30 subtopics

Table 2: evaluation results with different measure depths. * and # indicate significant improvements over ISCAS-S-C-2 and ISCAS-S-C-4 respectively ($p<0.01$).

4.2 Analysis

From the evaluation results illustrated above, we can analyze our approach and draw some conclusions.

Firstly, the number of clusters may not affect the results consistently. Comparing "ISCAS-S-C-1" with "ISCAS-S-C-3" at different measure depths respectively, we find that 5 clusters may improve the results at depths of 20, while 10 clusters are better at depths of 10 and 30. But things are different when we compare "ISCAS-S-C-2" with "ISCAS-S-C-4". We can conclude that 5 clusters are better than 10 clusters at all of these three depths. Even so, the improvement brought by this variable is not distinct, so it is difficult for us to determine which one is more suitable for clustering the subtopics.

Another conclusion is that the web data may improve the results dramatically. For some queries of this task, we cannot find any subtopics in Sogou query log, which is the key that "ISCAS-S-C-2" and "ISCAS-S-C-4" are worse than "ISCAS-S-C-1" and "ISCAS-S-C-3". Table 3 shows the count of topics which have no subtopic in each run at different depths.

	Top 10	Top 20	Top 30
ISCAS-S-C-1	2	1	1
ISCAS-S-C-2	20	18	17
ISCAS-S-C-3	1	1	1
ISCAS-S-C-4	18	18	17

Table 3: count of topics which have no subtopic

In our approach, we judge whether a query S is a subtopic of another query Q by calculating their similarity. This may introduce a problem that some query S may be similar to query Q, but it is not the subtopic of query Q, for example, query "汶川地震原因" and "汶川地震校舍倒塌原因" are lexically similar, but they are of different topics, the former is to find the reason of earthquake, and the later is to find the reason of collapse of the school building. To solve this problem, we would consider some more information, such as user clicks recorded by the query log, or some semantic resources.

5. FUTURE WORK

As discussed above, there still are a lot work to do to mine and rank subtopics. We will focus mainly on two aspects. The first one is to discriminate those queries which are lexically similar to the topic but with different intents. And the second is to organize these subtopics into a hierarchy structure according to their semantic relationships.

6. ACKNOWLEDGEMENTS

This research is supported by the National Natural Science Foundation Program of China under grants 60736044, 90920010, and the National Science and Technology Major Project (No. 2010ZX01037-001-002).

7. REFERENCES

[1] M. A. Hearst and J. O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of SIGIR1996.*

[2] R. Jones, B. Rey, O. Madani, and W. Greiner. 2006. Generating query substitutions. In *Proceedings of WWW '06*.

[3] H. Lang, D. Metzler, B. Wang, J-T. Li. 2010. Improved Latent Concept Expansion Using Hierarchical Markov Random Fields. In *Proceedings of SIGIR2010*.

[4] D. Metzler and W. B. Croft. 2007. Latent Concept Expansion Using Markov Random Fields. In *Proceedings of SIGIR2007*.

[5] P. Pirolli, P. K. Schank, M. A. Hearst, and C. Diehl. 1996. Scatter/gather browsing communicates the topic structure of a very large text collection. In CHI, pages 213-220, 1996.

[6] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, N. Orii. 2011. Overview of the NTCIR-9 INTENT Task. In *Proceedings of NTCIR-9*.

[7] E. Uluhan and B. Badur. 2008. Developmetn of a Framework for Sub-topic Discovery from the Web. In *Proceedings of PICMET2008*.

[8] X. Wang and C. Zhai. 2007. Learn from Web Search Logs to Organize Search Results. In *Proceedings of SIGIR 2007*.

[9] X. Wang and C. Zhai. 2008. Mining term association patterns from search logs for effective query reformulation. In *Proceedings of CIKM '08*.

[10] H.J. Zeng, Q.C. He, Z. Chen, W.Y. Ma, and J. Ma. 2004. Learning to cluster web search results. In *Proceedings of SIGIR2004*.