The Report on Subtopic Mining and Document Ranking of NTCIR-9 Intent Task

Wei-Lun Xiao,
CSIE, Chaoyang University of Technology
No. 168, Jifong E. Rd., Wufong, Taichung, Taiwan, R.O.C
s9927632@cyut.edu.tw

Shih-Hung Wu*,
CSIE, Chaoyang University of Technology
No. 168, Jifong E. Rd., Wufong, Taichung, Taiwan, R.O.C
shwu @cyut.edu.tw (*Contact author)

Liang-Pu Chen, and Tsun Ku
IDEAS, Institute for Information Industry
8F., No. 133, Sec. 4, Minsheng E. Rd.,Taipei, Taiwan, R.O.C
{cujing, eit}@iii.org.tw

ABSTRACT

In this paper we report our approach and result as a participant of the NTCIR-9 Intent task. INTENT task is a new NTCIR task which consists of two subtasks: (1) Subtopic Mining subtask: given a query, a system lists all possible subtopics that might cover users' different intents. Our approach is mining the query log to find subtopics candidates and rank them according to the frequencies of each candidate. (2) Document Ranking subtask: given a query, a system returns diversified document URLs that might cover users' diversified intents. Since the document set is larger than the capacity of PC. Our approach is to construct a distributed framework that can search a partial document set by one PC at a time and merge the partial search results to get the final ranking list.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Intent, Subtopic Mining, Document Ranking

General Terms

Experimentation

Keywords

Intent, Subtopic Mining, Document Ranking.

Team Name

III_CYUT_NTHU

Subtasks/Languages

Chinese Subtopic Mining, Chinese Document Ranking

External Resources Used

No

1. INTRODUCTION

The search result of a typical search engine is usually very large and users cannot browse all the results; therefore, the ranking of the search result is very important. Since different users might use the same query term to search for different objectives, it is very important that the search engine can diversify the search result to satisfy the diversified information need. If the search engine does not diversify the search result, the users might need to browse many pages or re-issue a new query to find the information.

The INTENT task in NTCIR-9 is dealing the problem via two subtasks. First, the subtopic Mining subtask, i.e., given a query, a system lists all possible subtopics that might cover users' different intents. Second, the document Ranking subtask, i.e., given a query, a system returns diversified document URLs that might cover users' diversified intents.

We will describe the related works in the second section. Then we will describe the data set in the third section, our approach in the fourth section, and show the result in the fifth section. Finally, we will give our conclusions and future work.

2. Related works

In the previous work, Bordogn et al. suggested an approach that can generate disambiguated queries by combining clustering and personalized ranking score. The system forms several clusters from the initial search result and extract terms form the titles and snippets of documents in each cluster [1]. Song et al. proposed a way on mining subtopics by log-based method and collection-based method [2]. The first step of log-based method is to observe each clicked document and find the relation to the intents. The second step is to calculate the number of clicks for each intent:

$$\mathrm{count}(\mathrm{i},\mathrm{q}) = \sum_{\mathrm{rel(d.i)=1, d}\in C} \mathrm{click(d)}$$

where C is the clicked document set, click(d) is the number of times a document d is clicked. The third step is to estimate the probability of an intent i for a given query q:

$$\mathrm{P(i|q)} = \frac{\mathrm{count(i,q)} + 1}{\sum_{\mathrm{i}\in I+}(\mathrm{count(i,q)} + 1)}$$

On the other hand, collection-based method is to calculate the p(i|q) via observing document. The first step is creating a new sub-query that can serve as a better query for the disambiguation of each intent. The second step is to count the term frequency N_i from the documents of the initial query result. Finally, the probability is estimated by:

$$\mathrm{P(i|q)} = \frac{\mathrm{N_i} + 1}{\sum_{\mathrm{i}\in I+}(\mathrm{N_i} + 1)}$$

3. Data set of NTCIR-9 Intent task

In this section, we describe the data set and the preprocessing of the data set.

3.1.1 Chinese document Collections

The Chinese document collections in the NTCIR-9 Intent task was provided by Tsinghua-Sohu Joint Laboratory. The contents are the original webpages collected from the Internet. The number of the pages is more than 130 million and the size of the pages is more than 5TB. The document format is shown in the figure 1, and an example is shown in Figure 2.

```
<doc>
<docno>Page ID</docno>
<url>Page URL</url>

original page contents

</doc>
```

Figure 1: Chinese document format

```
<DOC>
<DOCNO>01486710a6e73c36-66e5d9a362314a50</DOCNO>
<URL>http://zhengguo156.blog.163.com/blog/static/12956274200611893150933/</URL>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
 <meta http-equiv="Content-type" content="text/html; charset=GBK">
   <link href="http://st.blog.163.com/style/common/error/error.css" type="text/css"
rel="stylesheet"/>
   <link href="http://st.blog.163.com/style/common/error/color.css" type="text/css"
rel="stylesheet"/>
<title>操作失败</title></head>
<body>
<div id="pagebody">
                                        <div class="logo"
onclick="window.location.href='http://blog.163.com/login.html';return false;"></div>
                                                     <div class="login">
                                                                <a target="_blank"
href="http://blog.163.com/">博客首页</a><span> | </span>
                                                                <a target="_blank"
href="http://help.163.com/special/007525FT/blog.html">帮助</a>
                                                     </div>
                                                     <div style="clear:both;"></div>
                                                     <div class="errorinfo">博主设置
了该日志的访问权限，你暂时不能查看。</div>
                                                     <div style="text-
align:center;color:#000000;font-size:14px;">
                                                     你还没有登录网易博客，请先
<a style="color:#3366cc;margin-right:20px;" href="http://blog.163.com/login.html">登录
</a>
 <input type="button" value="返 回" onclick="history.back();"/></div>
</div>
</body>
</html>
</DOC>
```

Figure 2: Chinese document example

3.1.2 Chinese query log

The Chinese query log in the NTCIR-9 Intent task was also provided by Tsinghua-Sohu Joint Laboratory. The contents are the click log collected from the Internet user during June 2008. The number of clicks is 0.3 million. The log format is: time \t user ID \t [query] page rank \t click rank \t the clicked URL. Query log examples are shown in Figure 3.

```
00:00:00    11515839301781111     [2008年运程]   4 3     2008.gif123.com/
00:00:00    01532039495118448     [不锈钢]       1 1     www.51bxg.com/
00:00:01    0014362172758659586   [明星合成]     64 21   link.44box.com/
00:00:01    01532039495118448     [不锈钢]       1 2     www.51bxg.com/
00:00:01    7812322229275207      [link:jejie.cn] 1 3    jejie.cn/
00:00:02    1421205460982763      [健美]         8 2     www.mandf.cn/
00:00:02    09766368846776196     [英语培训]     1 1     www.glvchina.com/
```

Figure 3: Chinese query log examples

3.2 Preprocessing

The preprocessing flowchart of our system is shown in Fig 4. Since the document set consist of a lot of HTML tags of the original pages such as <html>, <head>, …, etc. which are not necessary to our task. The first step of our preprocessing is to filter out HTML tags. Then our system segments the documents by word segmentation toolkit. Finally, our system constructs the index for search engine.

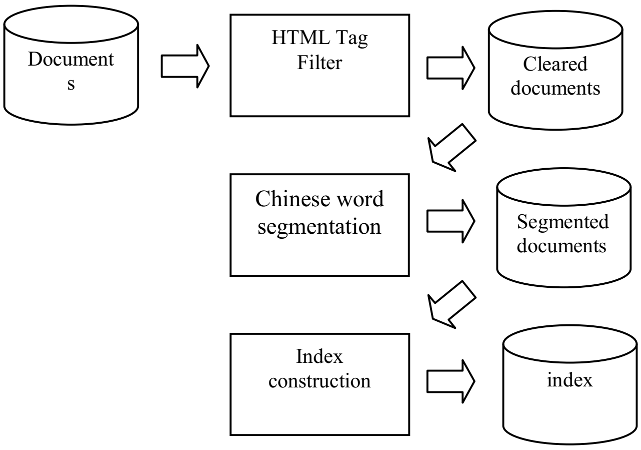

Figure 4: Preprocessing flowchart

3.3 Chinese Word Segmentation toolkit

The word segmentation toolkit in used is the ICTCLAS word segmentation system, which is provided by the Institute of Computing Technology Chinese Academy of Sciences. The toolkit functions includes word segmentation, POS tagging, NE recognition, new word identification, and customized dictionary [7].

3.4 Index toolkit

The index and search engine in used is the Lucene system, which is an open source full text search engine provided by Apache software foundation. Lucene is written in JAVA and can be called by JAVA program easily to build various applications [8].

4. Method

4.1 Substring matching approach to Subtopics mining

In the subtopics mining subtask, we use the query log as the only information resource. Our method consists of two steps:

First, find the subtopic candidates by exact substring matching. For example, suppose the query is "红酒 (red wine)" the candidates will be “红酒酒具”, “红酒酒架”, “红酒面膜”, and etc.. The query term is a substring of each candidate. On the other hand, “粉紅氣泡酒” will not be our candidate, since the characters of the query are not adjacent in this term.

Second, find the frequency of each candidate in the query log by exact matching. That is, the frequency of “红酒酒具” will be the number of times that the exactly query appeared in the query log and will not be mixed with the frequency of a long query, such as “红酒酒具商”. And output our candidate lists ranking according to the frequency.

4.2 A distributed approach to the Document Ranking subtask

Since the document set is larger than the capacity of an ordinary PC, we designed a distributed system flowchart to overcome the problem. In the Document Ranking subtask, we divide the document set into k subsets arbitrarily. Our system builds k separated indices. A query is sent to k separated search engine and finally a merger merges the results from k sub-results. The flowchart is shown in Figure 5. There will be k separated index files i_1 to i_k, and the system will search the k index files and get k separated Top N search results r_1 to r_k. The system then merges the results according to the rank in each search result. The rank 1 result in set r_1 will be the top 1 in the merged rank list, followed by the rank 1 result in set r_2, and the rank 1 result in set r_k will be the top k. The rank 2 result in set r_1 will be ranked k+1th in the merged result.

5. Experimental results

5.1 Output format

In the NTCIR-9 Intent task, we submit one run for the Subtopic Mining subtask and one run for Document Ranking subtask. The filenames are III&CYUT&NTHU-S-C-1 for Subtopic Mining subtask and III&CYUT&NTHU-D-C-1 for Document Ranking subtask. The format is as following:

5.1.1 Subtopic Mining subtask

First line give a simple description:

<SYSDESC>this is a dummy description.</SYSDESC>

The following lines are formatted as shown in Figure 6:

[TopicID];0;[Subtopic String];[Rank];[Score];[RunTag]

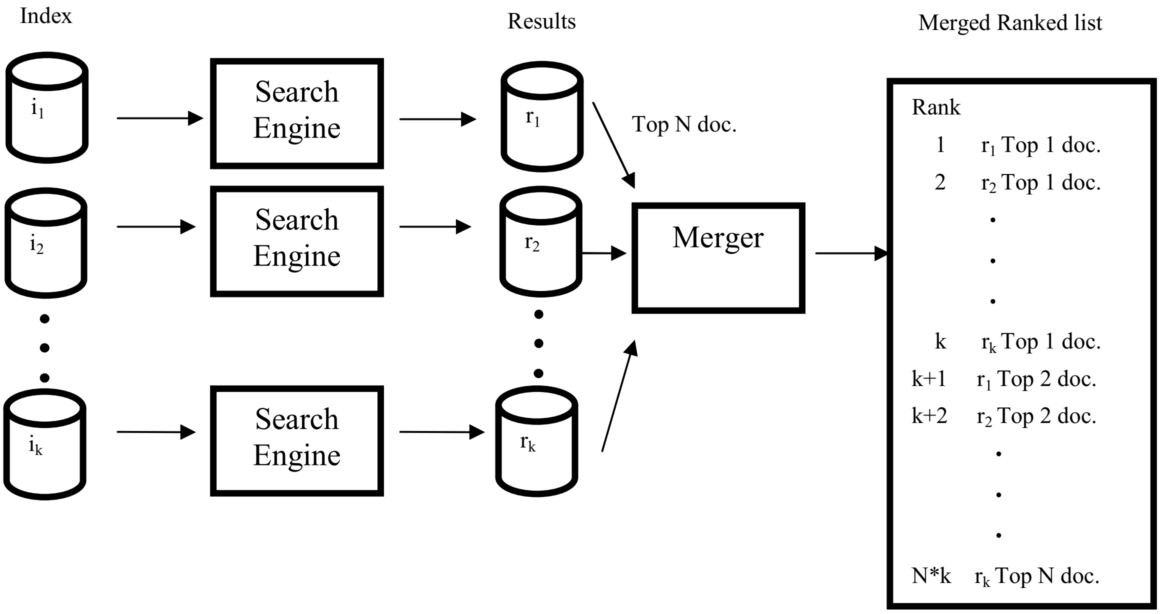

Figure 5: Document Ranking System flowchart

```
0007;0;巧克力冰;1;1;Run1
0007;0;巧克力城;2;1;Run1
0007;0;diy 巧克力;3;1;Run1
0007;0;巧克力球;4;1;Run1
0007;0;LG 巧克力;5;1;Run1
0007;0;黑巧克力;6;1;Run1
0007;0;toffifee 巧克力;7;1;Run1
0007;0;KINDER 巧克力;8;1;Run1
0007;0;LG+巧克力;9;1;Run1
0007;0;巧克力酱;10;1;Run1
0007;0;巧克力++MP3;11;1;Run1
0007;0;lindt 巧克力;12;1;Run1
0007;0;dove 巧克力;13;1;Run1
0007;0;hershey 巧克力;14;1;Run1
0007;0;HuHu 巧克力;15;1;Run1
0007;0;巧克力卷;16;1;Run1
0007;0;巧克力店;17;1;Run1
0007;0;巧克力 DIY;18;1;Run1
0007;0;MILKA+巧克力;19;1;Run1
0007;0;巧克力 ice;20;1;Run1
```

Figure 6: Subtopic Mining subtask output format example

5.1.2 Document Ranking subtask

First line give a simple description:

<SYSDESC>this is a dummy description.</SYSDESC>

The following lines are formatted as shown in Figure 7:

[TopicID] 0 [DocumentID] [Rank] [Score] [RunTag]

```
0001 0 01e462c75528364c-ba34aec0a46c4f90 1 1 Run1
0001 0 0249e173806b3f6a-4b74aec0a46c4f90 2 1 Run1
0001 0 45a330f31bde3f5f-4dc8547d6df9ecf0 3 1 Run1
0001 0 06ee3ad1f9ac9643-efd7276b50c7c930 4 1 Run1
0001 0 20ececbf5fe769e9-dd40d12d983f5ee0 5 1 Run1
0001 0 22cbbb09896cf772-80429576244c9da0 6 1 Run1
0001 0 25590e8aa47bd129-9c2de80ad0c786d0 7 1 Run1
0001 0 465cbc51a0886c64-d495d9a362314a50 8 1 Run1
0001 0 400fc2cef5673b70-5d5e53cd09768520 9 1 Run1
0001 0 421a76f6236b3f6a-4b74aec0a46c4f90 10 1 Run1
0001 0 01a1d748c8f189f5-123ea3bb382bb360 11 1 Run1
0001 0 02f045281a3212d0-2ecdaa0c03222580 12 1 Run1
0001 0 44f87c79f7759c9c-92c5c177ab6cc660 13 1 Run1
0001 0 07d460d2ff707f8b-bb672760f9ef5240 14 1 Run1
0001 0 208e82162a241260-9488547d6df9ecf0 15 1 Run1
```

Figure 7: Document Ranking subtask output example

The Chinese topics in NTCIR-9 Intent task are 100 queries selected from Sogou query log, June 2008. Examples are shown in Figure 8.

0001	日俄战争
0002	象棋秘籍
0003	雅虎中国
0004	减肥粥
0005	红酒
0006	千秋
0007	巧克力
0008	糖尿病症状
0009	金素妍
0010	杭州西湖
0011	越南旅游
0012	永乐
0013	多晶硅
0014	羽毛球规则
0015	莫扎特
0016	血型
0017	十二生肖来历
0018	蟑螂
0019	原油
0020	微型车

Figure 8: Chinese topics examples

5.2 Evaluation Metrics

In the official evaluation of two subtasks, the primary index is the D#-nDCG. *D#-nDCG* is a linear combination of *intent recall*(or "I-rec" , which measures diversity) and *D-nDCG* (which measures overall relevance across intents)。The advantages of *D#-nDCG* over other diversity metrics such as *α-nDCG* [3] and *Intent-Aware metrics*[4] are discussed elsewhere [5]. The result is computed according to the default setting of the NTCIREVAL [4]. *D#-nDCG* is a simple average of I-rec and D-nDCG. The gain values for the per-intent graded relevance were set linearly: 1, 2, 3 and 4 for *L*1, *L*2, *L*3 and *L*4, respectively. (the per-topic relevance assessments for Subtopic Mining only have *L*0 and *L*1.) The *measurement depths* (i.e. number of top ranked items to be evaluated) are of *l* = 10, 20 and 30 for both Subtopic Mining and Document Ranking.

5.3 Chinese Subtopic Mining Results

Tables 1-3 show the mean intent recall, D-nDCG and

D#-nDCG values for *l* = 10*,* 20*,* 30.

5.4 Chinese Document Ranking Results

In the formal run, due to the time limitation, our system indexed only 1.1 TB data from Chinese document collections. Tables 4-6 show the mean intent recall, D-nDCG and D#-nDCG values for *l* = 10*,* 20*,* 30. In the additional run, our system indexed all the Chinese document collections. Tables 7-9 show the mean intent recall, D-nDCG and D#-nDCG values for *l* = 10, 20, 30.

6. Conclusions and Future Work

This paper described our system in the NTCIR-9 intent task. In the Formal Run, submit one run for Subtopic Mining subtask and one run for Document Ranking subtask. The result is shown in Table 1-6. In the additional run, we indexed all the Chinese document collections for Document Ranking subtask. However, the performance decreases a little. We believed that our method tend to rank documents in the same subtopic together into a big block in ranking result. This is not good for diversity.

This paper reported our implementation on the subtopic mining task and document ranking task. There are still many ideas have not been realized. We have not connected the link between the two tasks. In the future, we might mine subtopics from the document clustering results. Also, in the Document Ranking subtask, currently we divide the document set into k subsets arbitrarily; it might be k clusters of the clustering result or search results of k subtopics.

7. ACKNOWLEDGMENTS

This study was conducted under the ‘Intelligent Web-Enabled Service Research and Development Project’ of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

8. REFERENCES

[1] Gloria Bordogna, Alessandro Campi, Giuseppe Psaila, and Stefania Ronchi. Disambiguated query suggestions and personalized content-similarity and novelty ranking of clustered results to optimize web searches. Information Processing & Management, In Press, Corrected Proof, Available online 19 April 2011..

[2] Ruihua Song, Dongjie Qi, Hua Liu, Tetsuya Sakai, Jian-Yun Nie, Hsiao-Wuen Hon, Yong Yu. Constructing a Test Collection with Multi-Intent Queries. The Third International Workshop on Evaluating Information Access (EVIA), June 15, 2010.

[3] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A comparative analysis of cascade measures for novelty and diversity. In Proceedings of ACM WSDM 2011, 2011.

[4] R. Agrawal, S. Gollapudi, A. Halverson, and S. Leong. Diversifying search results. In Proceedings of ACM WSDM 2009, pages 5–14, 2009.

[5] T. Sakai and R. Song. Evaluating Diversified Search Results Using Per-Intent Graded Relevance. In Proceedings of ACM SIGIR 2011, pages 1043–1052, 2011.

[6] T. Sakai. NTCIREVAL: A Generic Toolkit for Information Access Evaluation. In Proceedings of the Forum on Information Technology 2011, to appear, 2011.

[7] http://baike.baidu.com/view/1215398.htm “ICTCLAS”.

[8] http://zh.wikipedia.org/wiki/Lucene “Lucene”

Table 1: formal run Chinese Subtopic Mining for *l* = 10.

I-rec@10	D-nDCG@10	D#-nDCG@10
0.3085	0.4099	0.3592

Table 2: formal run Chinese Subtopic Mining for *l* = 20.

I-rec@20	D-nDCG@20	D#-nDCG@20
0.3890	0.3946	0.3918

Table 3: formal run Chinese Subtopic Mining for *l* =30.

I-rec@30	D-nDCG@30	D#-nDCG@30
0.3890	0.3042	0.3466

Table 4: formal run Chinese Document Ranking for *l* = 10.

I-rec@10	D-nDCG@10	D#-nDCG@10
0.4630	0.2040	0.3335

Table 5: formal run Chinese Document Ranking for *l* = 20

I-rec@20	D-nDCG@20	D#-nDCG@20
0.5658	0.2179	0.3919

Table 6: formal run Chinese Document Ranking for *l* =30

I-rec@30	D-nDCG@30	D#-nDCG@30
0.5821	0.1869	0.3845

Table 7: Additional run Chinese Document Ranking for *l* = 10.

I-rec@10	D-nDCG@10	D#-nDCG@10
0.4430	0.1784	0.3107

Table 8: Additional run Chinese Document Ranking for *l* = 20

I-rec@20	D-nDCG@20	D#-nDCG@20
0.5070	0.1659	0.3365

Table 9: Additional run Chinese Document Ranking for *l* =30

I-rec@30	D-nDCG@30	D#-nDCG@30
0.5613	0.1586	0.3599