THUIR at TREC2008: Blog Track[1]

Tong Zhu, Min Zhang, Yiqun Liu, Shaoping Ma

State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China

Tsinghua National Laboratory for Information Science and Technology, Tsinghua University

Zhutong000@gmail.com

Abstract. This is the second time we participate in TREC Blog Track. There are three main tasks in the track, relevant finding task, opinion finding task and polarity task. In this year, we use multi-field relevance ranking in relevant finding task; and in opinion finding task, we focused on the combination of relevance score and opinionate score use a unified generation model; in polarity task, we develop two new methods to find out positive and negative blogs.

1 Introduction

This is the second year that the IR groups of Tsinghua University participated in TREC Blog Track. Different from the previous track, TREC introduced a new task, the polarity finding task. So, we focus on 3 main tasks this year. The opinion retrieval task involves locating blog posts that express an opinion about a given target. The target can be a "traditional" named entity -- a name of a person, location, or organization -- but also a concept (such as a type of technology), a product name, or an event. The topic of the post does not necessarily have to be the target, but an opinion about the target must be present in the post or one of the comments to the post. The polarity task is to locate blog posts that express an idea either positive or negative about a target.

For relevant task, a multi-field relevance ranking based on probabilistic retrieval model has been used. Both feed content and permalink content are used. Two kinds of information fusion have been experimented. One is the result combination on both parts. Another is to combine the two corpus in the weighting phase with improved algorithms. Experimental results on training set showed that both methods are proved to be effective and the second way seemed to be more stable.

For opinion finding tasks, the combination of relevance score and opinionate score use a unified generation model is emphasized. The final score of one document is a quadratic combination of sentiment score given by an opinion generation model and the relevance score given by document generation model. HowNet has been used as the sentimental lexicon.

For polarity task, several algorithms on using sentiment words co-occurrence frequency are implemented. The selection of the sentiment dictionaries and the effectiveness of co-occurrence window size are studied. The approach of using polarity words as query terms on first-step relevance results is also performed.

2 Relevant Finding Task

The task of relevant finding is defined to retrieve those blogs which are relevant to the given query. To do this task, we need to make some pretreatment to the original corpus.

2.1 Pre-Processing

The main purpose of processing permalinks components is to remove noisy data from corpus. We found blog posts are written in various languages, and some blog posts are spam. So cleaning the corpus is very necessary for further work. We processed the permalinks components in the following two aspects.

One is to remove blogs written in languages other than English. We did this by examining the letter of the blog posts. By checking the content of set, we removed more than forty thousand blogs from permalinks

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141).

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
THUIR at TREC2008: Blog Track	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Tsinghua University,State Key Laboratory of Intelligent Technology and Systems,,Beijing 100084, China,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**5**	

components. we also removed strings which contained more than 20 letters. This value is fixed through experience. The other aspect is that we removed some obviously useless tags from the corpus. These tags include <script> <link> <dochdr> <style> et al.

For feed extract, a feed parser is designed to extract information from feed items, in this year, we use this part of data, only for multi-field search.

2.2 Retrieval Process

Two kinds of information fusion have been experimented. One is the result combination on both parts. Another is to combine the two corpus in the weighting phase with improved algorithms. Experimental results on our training set showed that both methods are proved to be effective and the second way seemed to be more stable. Hence we used the second way (named as multi-field search) to merge permalink and feed information.

2.3 Submitted Relevant Finding results

This task requests each group to submit 2 runs which regard as baseline in Blog track. In this task, we just use BM2500 formula and some expanded features such as word pair. The TMiner search engine, from IR group of Tsinghua University, is used as our text retrieval system. Two runs are submitted. One is retrieved in only one field (permalinks field), while the other is multi-field search (permalinks field & feeds field). Table 1 shows the map of these two runs. The results are similar and no obvious improvement can be made.

Table1. The official results of relevant finding runs

Run No.	Run Tag	**MAP**
1	THUrelTwp , baseline	0.3909
2	THUrelTwpmf, multi-filed search	0.3988

3 Opinion Finding Task

The task of opinion finding is defined to retrieve those blogs which are opinionated to given query. Relevant finding task is the foundation of this task. In this task, each participating group was required to submit a compulsory automatic run, using only the title field of the topics, with all opinion finding features of the retrieval system turned off[1]. Four runs are required.

3.1 Opinion Finding retrieval process

In opinion finding task, users want to find the documents that is both relevant and with subjective opinions. Thus to the retrieval system, it is to find the document with the high probability of $p(d|q,s)$. For simplicity, when we discuss the lexicon-based sentiment analysis, the latent variable s is assumed to be a pre-constructed bag-of-word sentiment thesaurus, and all sentiment words s_i are uniformly distributed. Then the prior probability that the document d contains relevant opinions to query q is given by [2]:

$$p(d \mid q,s) = \sum_i p(d \mid q,s_i)\, p(s_i,s) \;\propto\; \frac{1}{|S|}\sum_i p(s_i \mid d,q)\, p(q \mid d)\, p(d) \qquad (1)$$

where $|s|$ is the number of words in sentiment thesaurus s.

Define $p(d \mid q,s) = I_{op}(d,q,s) I_{rel}(d,q), \quad where \quad I_{op}(d,q,s) \equiv \frac{1}{|S|}\sum_i p(s_i \mid d,q), \quad I_{rel}(d,q) \equiv p(q \mid d)\, p(d)$ (2)

where $I_{op}(d,q,s)$ is the opinion generation probability, and the $I_{rel}(d,q)$ is the document relevance probability. For relevance score, we use BM 25 ranking function [4]:

$$ScoreI_{rel}(d,q) = \sum_{w \in q \cap d} (\ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1 + 1) \times c(w,d)}{k_1(1-b) + b\frac{|d|}{avdl} + c(w,d)} \times \frac{(k_3 + 1) \times c(w,q)}{k_3 + c(w,q)}) \quad (3)$$

For opinion score, totally two heuristic ranking functions have been used in our experiments:

1. $ScoreI_{op}(d,q,s) = \sum_{S_i \in d} (1-\lambda) \frac{co(s_i, q \mid W)}{c(q,d) \cdot |W|} + \lambda$ (for details, see [2]) (4)

where $co(s_i,q|W)$ is the frequency of sentiment word s_i which is co-occurred with query q within window W, $c(q,d)$ is the query term frequency in the document.

2. Use sentiment words as query terms, searching on top returned documents and get the opinion score. i.e.

$$ScoreI_{op}(d,q,s) = \sum_{w \in s \cap d} (\ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \times \frac{(k_1 + 1) \times c(w,d)}{k_1(1-b) + b\frac{|d|}{avdl} + c(w,d)} \times \frac{(k_3 + 1) \times c(w,s)}{k_3 + c(w,s)}) \quad (5)$$

where s means the sentiment words in the dictionary.

3.2 Parameter settings

In our experiments, λ is set to 1.35.

And the sentiment dictionary we used is HowNet[5]. it is a knowledge database of Chinese, and some of the words in the dictionary have properties of positive or negative. We use the English translation of those sentiment words provided by HowNet. Finally there are 4621 English sentiment words selected.

For co-occurrence window size design, best performance is got when window size is full text according to training result on Blog07 data. A possible explanation is that the majority of authors in a blog article on only one thing to express their notions, so generally the topic diversity is much smaller than ordinary web pages. In all the experiment, we fixed the window size to full text.

3.3 Submitted Opinion Finding results

In this year opinion finding task, we submitted 4 runs which are listed in table 2. The results show the effectiveness of our runs.

Table2. The official results of opinion finding runs

Run Tag	Description	Relevance Baseline Run in task1	MAP
THUopnTwpGen	Use co-occurrence MLE of senti-word and query term (Eq. 4). Weight on permalinks.	Run 1	0.3155
THUopnTwpRRM	Re-search by using sentiment words as query (Eq 5). Weight on permalinks.	Run 1	0.3169
THUopnTmfRQ	Use co-occurrence MLE of senti-word and query term (Eq. 4). Weight on permalinks and feeds.	Run 2	0.3120
THUopnTmfRmf	Re-search by using sentiment words as query (Eq 5). Weight on permalinks and feeds.	Run 2	0.3283

4 Polarity Task

The task of opinion finding is defined to retrieve those blogs which are opinionated whether positive or negative.

The polarity should be identified. This task was introduced in TREC 2008 as a natural extension of the opinion finding task, and it required 2 runs from each group.[3]

4.1 Polarity Task retrieval process

The retrieval process likes the process of opinion finding task. We also computed two scores of each blog posts. One is positive score, the other is negative score. The calculation is the same as we did in opinion finding task, i.e. the co-occurrence MLE based opinion score, and the re-search based score. Different with the previous task, the sentiment dictionary is divided into two polarity ones: positive dictionary and negative dictionary.

For combination of relevance score and polarity score, three algorithms are implemented in this task. Aussume *pos* is positive score, *neg* is negative score, *A* is the const threshold. Then the three algorithms are:

Alg. 1

If ((pos > A) or (pos > 0) and (neg = 0))
  Then it is positive;
If ((neg > A) or (neg > 0) and (pos = 0))
  Then it is negative;
For Other conditions
  It is mixed, and neither in the positive set nor in the negative set.

Alg. 2

If ((pos –neg > A + α) or (pos > 0) and (neg = 0))
  Then it is positive;
If ((pos - neg < A – α) or (neg > 0) and (pos = 0))
  Then it is negative;
For Other conditions
  It is mixed, and neither in the positive set nor in the negative set.

Alg. 3

If ((pos – neg > A) or (pos > 0) and (neg = 0))
  Then it is positive;
If ((pos / neg < 1/A) or (neg > 0) and (pos = 0))
  Then it is negative;
For Other conditions
  It is mixed, and neither in the positive set nor in the negative set.

Comparative experiments have been made on the training set we constructed on Blog 06 & 07 topics with the three algorithms. Table 3 shows the differences between these algorithms on the training sets.

Table3. The results of 3 polarity algorithms

	Blog 06 Racc	**Blog 07 Racc**
Alg. 1	0.107	0.1537
Alg. 2	0.1092	0.2041
Alg. 3	***0.1141***	***0.2066***

In the table, all the polarity scores are got by co-occurrence MLE approach. And the choice of window size was the same as opinion finding task. It's found that the Alg. 3 gets best result.

4.2 Submitted Polarity Task results

In this year polarity task, we submitted 2 runs which are listed in table 4.

Table4. The official results of polarity runs

Run Tag	Description	Relevance Baseline Run in task1	Polarity	MAP
THUpolTwpRD	Use co-occurrence MLE of polarity words and query terms (Eq.4). Combination Alg. 3	Run 1	positive	0.1149
			negative	0.0807
THUpolTmfPNR	Re-search by using polarity words as query (Eq 5). Combination Alg. 3	Run 2	positive	0.1399
			negative	0.1055

5 Discussion and Future Work

In relevant finding task, we will use more blog-specific features in blog data.

In opinion finding and polarity finding task, we will make further analysis on different algorithms, and a classify of query can be taken into account.

References

1. Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, Ian Soboroff(2006). Overview of the TREC-2006 Blog Track. TREC Conference.
2. Min Zhang, Xingyao Ye, A generative model to unify topic relevance and lexicon-based sentiment for opinion retrieval, The 31st Annual International ACM SIGIR Conference (SIGIR 2008), 20-24 July 2008, Singapore, pp411-419
3. Craig Macdonald, Iadh Ounis,Ian Soboroff(2007). Overview of the TREC2007 Blog Track. TREC Conference.
4. Robertson, S. E., and Walker, S., Okapi/Keenbow at TREC-8. In TREC-8.
5. Dong, Z. HowNet. http://www.HowNet.org