HIT_LTRC at TREC 2010 Blog Track: Faceted Blog Distillation

Jinfeng Yang, Xishuang Dong, Yi Guan, Chengzhen Huang, Sheng Wang
School of Computer Science and Technology,
Harbin Institute of Technology, 150001, Harbin, China
{yangjinfeng2010, dongxishuang}@gmail.com,
guanyi@hit.edu.cn, huangcheng_zhen@126.com,
maxwell_blueing@yahoo.com.cn

Abstract

This paper describes our participation in the faceted blog distillation task at Blog Track 2010. In our approach, indri toolkit is applied for basic topic relevance retrieval. Then the Maximum Entropy (ME) model is adopted to judge the relevance of each blog to specified facet. Feed faceted relevance is calculated by integrating the average relevance of all blogs within a feed and the average relevance of the most relevant N blogs. Two implementations are applied to calculate feed faceted relevance. Experimental results on Blogs08 dataset show the effectiveness of our approach.

1 Introduction

Blog track explores information seeking behavior in the blogosphere [1]. In TREC 2010, the Blog track has two tasks: Faceted Blog Distillation Task and Top Stories Identification Task. Faceted blog distillation is a more refined version of the blog distillation task, addressing the quality aspect of the retrieved blogs [2]. Our group from Language Technology Research Center of Harbin Institute of Technology (HIT_LTRC) participated in the first task.

We considered our implementation as a two-step procedure. The first step was to retrieve blogs that were relevant to a topic or query. This was an Information Retrieval (IR) process, and the Indri[1] toolkit was used. The second step was to assign particular facets of interest to the retrieved blogs. For the second step, a Maximum Entropy (ME) model toolkit was applied for text classification on specified facet and blogs extracted according to the Qrel (query-relevance files) of blog track 2009 were used as training documents.

The rest of this paper is structured as follows: Section 2 describes the preprocessing of the data, Section 3 and Section 4 introduce the methods and runs of baseline blog distillation and faceted blog distillation, in section 5, the evaluation results of our runs are given, and section 6 concludes our work.

2 Preprocessing

The Blog track 2010 used Blogs08 dataset[2]. The Blogs08 dataset contains permalinks, feeds and blog homepages, and only permalink pages were used in our participation for the faceted blog distillation task. We applied Indri to build index, and specify some fields as metadata, such as title, feedno, docno, for the index so that we could search and combine blogs to feeds flexibly. When building in-

[1] http://www.lemurproject.org/indri.php
[2] http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html

dex, the DiffPost algorithm [3] was applied to filter the non-relevant contents. Porter stemmer [4] and 415 stop words were also used.

3 Baseline Blog Distillation

3.1 Query construction

In the baseline stage, our queries were constructed by two ways: automatically and manually.

3.1.1 Automatic Construction

Automatic query construction follows three steps below:

1) Extract the content of *Query* field of a topic with stop words removed and construct indri-style query as the first sub-query;
2) Considering the importance of blog title, construct title context query according the result of step 1) as the second sub-query;
3) Combine the two sub-queries in different weights. According our experiments, the weight of the first sub-query is given 0.7, and the second is given 0.3.

An example of an automatic query is shown below:

<query>
<number>1102 </number>
<text>#weight(0.3 #combine(Wal-Mart.(title) communities.(title)) 0.7 #combine(Wal-Mart communities))</text>
</query>

3.1.2 Manual Construction

Manual query construction follows five steps below:

1) Extract the content of *query* field of a topic with stop words removed and construct the first sub-query;
2) Recognize the keywords in *desc* field by manual work, combine these words by "or" and "combine" operator, then construct the second sub-query;
3) Recognize the keywords in *narr* field, combine them by "or" and "combine" operator, then construct the third sub-query;
4) The content of *query* field of a topic can be regarded as the most concise version of the topic. So, construct a compound query by combining the three sub-queries with different weights, 0.6 for the 1st, 0.2 for the 2nd and 3rd;
5) Considering to the importance of blog title, construct title context query, then combine the title context query and the compound query with weights 0.2 and 0.8 according experiments.

An example of query is shown below:

<query>
<number>1101</number>
<text>#weight(0.2 #combine(genealogical.(title) sources.(title)) 0.8 #weight(0.6 #combine(genealogical sources) 0.2 #combine(genealogical research) 0.2 #or(#combine(genealogical information) #combine(Social Security Death Register) #combine(Mormon record keeping) #combine(Cemetery records))))</text>
</query>

3.2 Retrieval Method

In this step, the relevant feeds were retrieved by queries without facet. Since a feed could be regarded as a collection of blogs [5], relevant score of a feed could be calculated according to the relevant scores of blogs within the feed. We retrieved 10,000 blogs for a query, and call the result as topic-relevant blogs list. Then we calculated feeds relevant score based on the hypotheses of Global Evidence Model (GEM) and Local Evidence Model (LEM) [5], and then ranked the feeds list.

$$Score_{feed} = \alpha Score_{GEM} + (1-\alpha) Score_{LEM} \quad (1)$$

$$Score_{GEM} = \frac{\sum Score_{rel}}{|Feed|} \quad (2)$$

$$Score_{LEM} = \frac{\sum Score_{rel(topn)}}{N} \quad (3)$$

In Eq. (2), $Score_{rel}$ is the relevant score of a blog

belong to the feed, and |*Feed*| is the count of blogs in the feed. In Eq. (3), $Score_{rel(topn)}$ is the top *N* relevant score in the feed, and α =0.1. During our experiments, we found that the mean average precision (MAP) of experiments' results will be higher while N is given 1 or 2. These events are also consistent with the hypotheses of LEM. Also, Statistics for |*Feed*| shows that the |*Feed*| of half of feeds in the dataset is less than 50. So the value of N is determined by the following conditions:

$$N = \begin{cases} 1 & |Feed| < 10 \\ 2 & 10 =< |Feed| < 50 \\ 3 & |Feed| >= 50 \end{cases} \quad (4)$$

3.3 Baseline Runs

In the baseline blog distillation task, we submitted two runs as follows:

1) **hitQuerybl** denotes automatic run, with the query field of a topic.
2) **hitTDNbl** denotes manual run, with the query, desc, narri field of a topic.

4 Faceted Blog Distillation

The basic idea of this stage was adopting the Maximum Entropy model to predict a blog's inclination on a specified facet, and then calculating the faceted relevance in feed level. Two methods were applied to implement the idea: One method was based on the topic-relevant blogs list, and the other was based on the baseline result.

By analyzing documents extracted according to Qrel of blog track 2009, many obvious features could be found:

1) Opinionated, personal, shallow blogs generally had more subjective words than factual, official, in-depth blogs;
2) Personal, shallow blogs might have shorter length than official, in-depth blogs;
3) The first person words, 'I', 'We' and 'Our', were more likely to appear in personal, shallow blogs than in official, in-depth blogs.

According to manual analysis, subjective words were chosen as major features, also including document length and the count of first person words in a document. SentiWordNet [6] is a lexical resource for opinion mining, which assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity. In this part, SentiWordNet3.0[3] was employed to obtain a set of subjective words. A word in SentiWordNet3.0 was regarded as subjective word if its average subjective value (a word may have different subjective values in distinct synonym sets) was above 0.5.

To calculate the real length of a blog, regular expressions were applied to identify the time when the blog was submitted or the time of the first comment.

For opinionated facet, subjective words and their statistics were considered as major features. Besides subjective words, the first person words 'I', 'We', 'Our' and the length of blog were also taken as feature for in-depth facet and personal facet.

Based on these features, blogs' faceted inclination could be predicted, and blogs' faceted relevance can be calculated.

$$Score_{Facet1} = \lambda Score_{Rel} + (1-\lambda) Score_{Rel} * P \quad (5)$$

$$Score_{Facet2} = \lambda Score_{Rel} + (1-\lambda) Score_{Rel} * (1-P) \quad (6)$$

In Eq. (5) and Eq. (6), $Score_{Rel}$ is the topic relevant score; *P* is the probability of a document determined opinionated, shallow or personal by ME model, and 1-*P* is the probability of a document determined factual, in-depth or official; $Score_{Facet1}$ is the first faceted relevant score of a blog, and $Score_{Facet2}$ is the second faceted relevant score of a blog.

4.1 Training documents Extraction

Training documents for ME model were ex-

[3] http://sentiwordnet.isti.cnr.it/

tracted according to the Qrel of blog track 2009. Qrel could be regarded as manual faceted tags in feed level. However, not all blogs in the feeds are highly relevant to specified facet. A trial method to extract training data was taking advantage of the features of blogs in different facet shown above.

The training data in our system was extracted as following steps:

1) Filter the blogs with length less than 100 words;
2) Calculate the percentage of subjective words (subjective words divided by total words) in each blog;
3) Sort the blogs by the percentage of subjective words in descending order when facet value are opinionated, personal, shallow, and in ascending order in other cases(factual, in-depth, official).
4) Choose the top N blogs as training data, where N = |Feed|*10%, then refine the value of N to 1 if N < 1, and to10 if N > 10;

4.2 Features Selection

There were four kinds of features selected in our method.

1) Subjective words in body text;
2) Statistics of subjective words in body text:

①Count of subjective words.

② Count of sentences containing subjective words.

③Percentage of subjective words (count of subjective words are divided by total words count).

④Percentage of subjective sentences (count of subjective sentences containing subjective words are divided by total sentences count).

⑤Sum of values of subjective words.

⑥Sum of values of subjective words are divided by total words count.

⑦Sum of values of subjective words are divided by subjective words count.

⑧Sum of values of subjective words are divided by total sentences count.

⑨Sum of values of subjective words are divided by subjective sentences count.

3) Count of subjective words in title.
4) Statistics of subjective words in title:

①Count of subjective words in title.

Content (submitted by blog owner) and comments (submitted by guests) of a document may have different contribution to the document's faceted inclination. So the features of document's content and of document's comments are extracted respectively. Finally, the entire document's feature, the content's feature, and the comments' feature are combined.

4.3 Faceted Distillation on Topic Relevant Blogs List

Topic-relevant blogs list contained 10000 relevant blogs per topic. In this method, a blog in topic-relevant blogs list was predicted by a ME model corresponding to its topic facet, with the predicted result P ($P_{opinionated}$, $P_{shallow}$ and $P_{personal}$). Then the faceted relevant score of blog can be calculated applying Eq. (5) and Eq. (6).

The faceted relevant score of feeds can be calculated by applying $Score_{Facet1}$ and $Score_{Facet2}$ into Eq. (1).

4.4 Faceted Distillation on Baseline

Baseline system retrieved results without facet consideration, and faceted inclination of each blog in baseline feeds could be predicted by ME model. According to LEM-GEM hypotheses, we could calculate the faceted inclination of baseline feeds.

$$P_{feed-facet1} = \alpha \frac{\sum P_{doc-facet1}}{|Feed|} + (1-\alpha)\frac{\sum P_{doc-facet1(topn)}}{N} \quad (7)$$

$$P_{feed-facet2} = \alpha \frac{\sum P_{doc-facet2}}{|Feed|}$$

$$+(1-\alpha)\frac{\sum P_{doc-facet2(topn)}}{N} \quad (8)$$

In Eq. (7) and Eq. (8), α = 0.1, $P_{feed\text{-}facet1}$ and $P_{feed\text{-}facet2}$ is the first and second faceted inclination probability of a feed, $P_{doc\text{-}facet1}$ and $P_{doc\text{-}facet2}$ is the first and second faceted inclination probability of a blog, and $P_{doc\text{-}facet1} = 1 - P_{doc\text{-}facet2}$.

The faceted relevant score of a feed could be calculated with Eq. (9) and Eq. (10).

$$Score_{facet1} = \lambda Score_{Rel} + (1-\lambda) Score_{Rel} * P_{feed-facet1} \quad (9)$$

$$Score_{facet2} = \lambda Score_{Rel} + (1-\lambda) Score_{Rel} * P_{feed-facet2} \quad (10)$$

In Eq.9 and Eq.10, λ = 0.8, $Score_{Rel}$ is the relevant score of a feed in baseline, $Score_{Facet1}$ and $Score_{Facet2}$ is the first and second faceted relevant score of a feed.

4.4 Faceted Runs

In the faceted blog distillation task, we submitted seven runs as follows:

1) **hitQFeedR** used topic-relevant blogs list retrieved according automatic query.
2) **hitTDNfeedR** used topic-relevant blogs list retrieved according manual query.
3) **hitQFeedbl** uses hitQuerybl baseline.
4) **hitTDNfeedbl** used hitTDNbl baseline.
5) **hitFeeds1** used stdbaseline1 baseline.
6) **hitFeeds2** used stdbaseline2 baseline.
7) **hitFeeds3** used stdbaseline3 baseline.

Table 1: Relevant results of the baseline runs

runid	MAP	P@10	bPref	rPrec
hitQuerybl	0.2493	0.3152	0.2384	0.2814
hitTDNbl	0.2822	0.3674	0.2674	0.3106

Table 2: Faceted results of the baseline runs

runid	All	Opinionated	Factual	Official	Personal	In-depth	Shallow
hitQuerybl	0.1804	0.1624	0.2584	0.2196	0.1467	0.1902	0.1051
hitTDNbl	0.1815	0.1751	0.265	0.2171	0.1607	0.143	0.1279
stdbaseline1	0.2061	0.2128	0.3678	0.255	0.1418	0.1661	0.0929
stdbaseline2	0.1498	0.1179	0.2338	0.2079	0.0785	0.1467	0.1137
stdbaseline3	0.1221	0.0927	0.1323	0.221	0.0982	0.0993	0.089

Table 3: Results of the faceted blog distillation

runid	Baseline	All	Opinionated	Factual	Official	Personal	In-depth	Shallow
hitQFeedbl	hitQuerybl	0.1769	0.1605	0.2607	0.2093	0.1461	0.177	0.1075
hitQFeedR	N/A	0.1725	0.1564	0.257	0.1988	0.149	0.1713	0.1027
hitTDNFeedbl	hitTDNbl	0.1811	0.1782	0.2682	0.2057	0.1678	0.1464	0.1204
hitTDNFeedR	N/A	0.1738	0.1716	0.2714	0.1743	0.1701	0.1307	0.1246
hitFeeds1	stdbaseline1	0.2071	0.2142	0.368	0.2578	0.1424	0.1663	0.0937

hitFeeds2	Stdbaseline2	0.1498	0.1176	0.2339	0.2081	0.0786	0.1469	0.1136
hitFeeds3	Stdbaseline3	0.1234	0.0928	0.1323	0.2197	0.0983	0.1059	0.0913

Table 4: Improvement results compared by the baseline runs

runid	Mean Facet	Baseline Mean Facet	Improvement
hitQFeedbl	0.1769	0.1804	-1.94%
hitTDNFeedbl	0.1811	0.1815	-0.22%
hitFeeds1	0.2071	0.2061	0.49%
hitFeeds2	0.1498	0.1498	0.02%
hitFeeds3	0.1234	0.1221	1.06%

5 Results

The primary measure for evaluating submitted runs is the mean average precision (MAP). With our methodology, we achieved the following results: Table 1 shows the query relevant results of the baseline runs. Table 2 and table 3 show the faceted results of the baseline runs and the faceted runs. In table 2, the three runs, stdbaseline1, stdbaseline2, and stdbaseline3 denote the three provided standard baselines. In table 3, the result of the run based on topic-relevant blogs list (that means no baseline) was inferior to the results of the run based on our baseline. Comparing table 3 with table 2, our faceted results can make improvements in some facets. According to table 4, the mean facet MAP values of faceted runs are compared with the mean MAP values of the baseline runs. Three runs have consistently improved upon the faceted performances of the three provided standard baselines, but the other two runs failed to improve the faceted performance of our own baselines.

6 Conclusions

This is the first time that our group participates in TREC Blog track. In this paper, we present our method for the faceted blog distillation. Our approach may be simple but straightforward. In the baseline stage, the indri was used as retrieval platform, and the title context query was built for the consideration to the importance of blog title. Experiment results proved that the title context query help to retrieve more relevant blogs. In the faceted stage, our Maximum Entropy model toolkit was applied to predict documents inclination, and subjective words from SentiWordNet were used as main features. The LEM-GEM assumption was implemented in two ways to calculate feeds' facet relevance. Results show that the second, based on baseline, is more effective.

Although our method performed consistent improvements upon the three provided standard baselines, the improvements were not notable. The feature selection and the training documents extraction may cause this limitation. The features, mainly included subjective words, were shared by three facets. This seemed to be arbitrary. Training documents were extracted by the Qrel of blog track 2009. The extracted documents may not be accurately tagged according to the facet. Future work will mainly focus on these two aspects.

7 Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 60975077 and Grant 60736044.

References

[1] I. Ounis, C. Macdonald, and I. Soboroff. On the TREC Blog Track. Proceedings of AAAI, 2008

[2] C. Macdonald, L. Ounis, and L. Soboroff. Overview of the TREC-2009 Blog Track. In proceeding of TREC 2009. 2010.

[3] Nam, S. H., Na, S. H., Lee, Y., Lee, J. H.: Diffpost: Filtering non-relevant content based on con-

tent difference between two consecutive blog posts. In: ECIR. Volume 5478 of Lecture Notes in Computer Science, Springer (2009)791–795
[4] Peter Willett, (2006) "The Porter stemming algorithm: then and now", Program: electronic library and information systems, Vol. 40 Iss: 3, pp.219 - 223
[5] Lee, Y., Na, S. H., Kim, J., Nam, S. H., young Jung, H., Lee, J. H.: Kle at trec 2008 blog track: Blog post and feed retrieval. In: Proceedings of TREC2008. (2008)
[6] Baccianella Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2200–2204. European Language Resources Association.