A Study of Faceted Blog Distillation
-- PRIS at TREC 2009 Blog Track

Si Li, Huiji Gao, Hao Sun, Fei Chen, Oupeng Feng, Sanyuan Gao,
Hao Zhang, Xinsheng Li, Caili Tan, Weiran Xu, Guang Chen, Jun Guo
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications
Beijing, P.R. China, 100876
ls198cf@gmail.com

Abstract:

This paper describes BUPT (pris) participation in faceted blog distillation task at Blog Track 2009. The system adopts a two-stage strategy in faceted blog distillation task. In the first stage, the system carries out a basic topic relevance retrieval to get the top k blogs for each query. In the second stage, different models are designed to judge the facets and ranking.

1 Introduction

The Blog track [1] had two tasks in the TREC 2009 and we participate in the faceted blog distillation task. This task involves locating blogs that contain relevant information about a given target topic, and judging the facet of each relevance blog.

The PRIS system is submitted by Pattern Recognition and Intelligent System Lab at Beijing University of Posts and Telecommunications.

This system adopts a two-stage strategy in faceted blog distillation. The goal of the blog distillation task, which is the baseline of the faceted blog distillation, is to find relevant blogs for specific topics. And the faceted blog distillation is used to explore the facets of the topic-relevant blogs. In order to achieve the in-depth blog distillation, a two-stage strategy is employed. In the first stage, the PA (Posts Average) algorithm is involved into the blog distillation. In PA algorithm, the relevance degree of the blog to the topic is decided by its posts' similarity scores. Besides, a Learning Query Expansion (LQE) algorithm is designed to improve the precision of topic-relevance retrieval. In the second stage, different facets are automatically identified by different models. There are three facet groups, which are opinionated vs. factual (o.f), personal vs official (p.o) and in-depth vs. shallow (i.s). For opinionated vs. factual, the o.f model combining Maximum Entropy [2] based classifiers is used to implement opinion polarity judging and ranking. The named entity recognition [3] is employed in p.o model. The boundary between in-depth and shallow is defined according to the document length in our i.s model. We use a criterion related to two factors. One factor is documents lengths. Because the lengths of in-depth documents are usually long, and the long documents are usually in-depth after we exclude those with a lot of spam information or only a small part of relevant content. However, it is difficult to clean these spam, at the same time, if we give more consideration on the length, these spam will take negative influence on in-depth analysis. The other factor is topic-relevance information which is important but usually neglected. We propose the *L-Qtf* (Length-Query term frequency) coefficient with considering these factors. In this coefficient, the relationship between the lengths and average

length is involved to reduce the spam's influence. Meanwhile, the query term frequency is employed to present the relevant information. The in-depth analysis model with the *L-Qtf* coefficient is used in the retrieved blogs form the first step. Finally, for each facet, the retrieved top 100 blogs are submitted. The framework of the overall processing is shown in Figure 1.

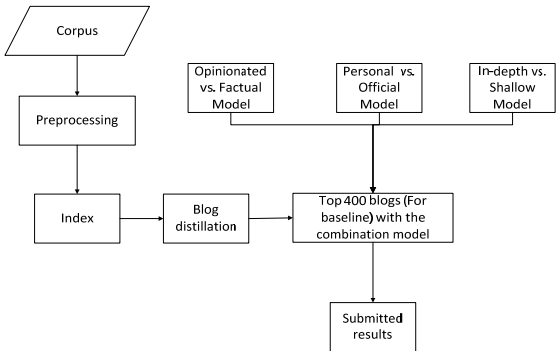

Figure 1. System Framework

In Section 2 and Section 3, we introduce the blog distillation algorithm and facets models respectively. In Section 4, the evaluation of the faceted blog distillation system is presented. Finally in Section 5, conclusions and comments on the future work are given.

2 Blog distillation

Our system contains two stages. In the first stage, the blog distillation, which is concerned with the search of blogs rather than blog posts and it is the baseline for the in-depth analysis, is performed to find the topic-relevant blogs. A blog's relevance degree and in-depth degree are evaluated by checking its containing posts. Traditional blog distillation approaches view the whole blog collection as a complete document [4], which use Eq. (1), called Baseline A, to sum up each post's similarity score for the given topic query Q. The score is defined as the ranking function of the topic-relevant blogs.

$$Score(blog_X, Q) = \sum_{i=1}^{n} Score(post_i, Q) \qquad (1)$$

where n is the number of the posts in blog x, and the similarity scores of posts are supplied by Indri [5].

But this algorithm isn't sufficiently reasonable. Let's look at a failure example using Eq. (1) by considering the following two feeds. Feed X contains 10 posts and all are relevant to a given query, feed Y contains 100 posts including the 10 of X. According to (1), feed X should have the same similarity score as feed Y for the given query if the other 90 posts of feed Y are totally irrelevant. However, considering the obviously different relevancies (100% vs 10%), we should judge that feed X is much more relevant to the given query than feed Y. To address this problem, we take the average value of the sum of posts' similarity scores as a new ranking function as Eq. (2).

$$Score(blog_X, Q) = \frac{\sum_{i=1}^{n} Score(post_i, Q)}{n} \qquad (2)$$

We call Eq. (2) Posts Average (PA) as well as Baseline B.

In our system, query expansion is added automatically to improve the retrieval accuracy. From the aspect of topic understanding, the Learning Query Expansion (LQE) model based on

semi-machine learning method is designed.

The topic given by the Blog track is as shown in [1]. The topic is composed of 5 parts: number, title which is the original query, description, facet and narrative. With considering the Indri query language, expansion words and their weighting, we introduce two kinds of features into the LQE model. These two kinds of features are extracted from the sentences in the topic description and topic narrative which supply training data and testing data. One kind is syntactic feature which contains part of speech (POS) tags and syntax analysis tags. The other kind is distance feature which is a novel feature in the query expansion. Firstly, we define an ordered center word list which contains the words regarded as the center word of a sentence, such as “not”, “relevant”, “find”, and so on. Each sentence has one center word at most. The words “not” and “no” have the highest priority to be the center word in one sentence, while the “find” has the lowest. Secondly, in a sentence, we calculate the distance from each word to the center word and this distance value is deemed as the distance feature for the word.

We trained LQE model based on CRFs with the manual Blog track 2006 queries which were expanded based on the human common sense and comprehension. After the classifier was trained, it was applied to the whole Blog track 2009’s queries for query expansion which contains both expansion words and their weightings with Indri query language. One of the final query examples is as the following:

```
<query>
    <number>1101</number>
    <text>genealogical sources #5(genealogical sources ).(title) #weight(1.0
    #combine(genealogical sources) 0.8 Death 0.8 Security 0.8 Social 0.8 genealogical 0.8
    registration 0.8 sources )
    </text>
</query>
```

3 Facets Models

3.1 Opinionated vs. Factual Model

This model contains two stages. In the first stage, the sentiment analysis model [2] is applied to the posters in the baseline feeds to generate opinion judgments. The analysis model we used is the same as the model we used in the Blog Track 2008 [6]. In the second stage, the facet of the feed is judged based on the poster sentiment analysis results. For the factual facet, two rules are defined:

1. If objective posters occupy more than 50% in a feed, this feed is deemed as factual.
2. If the number of posters in a feed is more than 150, the feed is regarded as factual.

For the opinionated, we calculate the feed by this function:

$$S_o = \frac{|N_{pos} - N_{neg}|}{N_{pos} + N_{neg}} \quad (3)$$

N_{pos} is the number of the positive posters, while N_{neg} is the negative posters number. The lager the opinionated score S_o is, the higher level of opinionatedness has.

3.2 Personal vs. Official Model

We find that the difference between personal blogs and official blogs is the frequency of the same organization named entity. First, organization entity type is identified by Stanford Named

Entity Recognizer [7]. Because there are a lot of advertisements in the blog web pages, and the advertisements contain too many organization named entities, it is necessary to detect the spam organization entities. We regard one organization entity as spam entity if it appears in the same position of each poster in the same feed. An organization entity list without spam can be got orderly for each poster. Then according to the position in the list, the weight of each entity is given as the following.

$$w = \begin{cases} 1.0 & 0 < p \le \frac{L}{3} \\ 0.8 & \frac{L}{3} < p \le \frac{2L}{3} \\ 0.2 & \frac{2L}{3} < p << L \end{cases} \quad (4)$$

P is the position of the entity in the list. L represents the length of the list. W is the weight given to the entity. Then, in a feed, each entity's weight WF is calculated using the Eq. (5).

$$WF_j = \sum_{i=1}^{n} w_{i,j} \quad (5)$$

In this function, n is the number of the relevance posters in the feed j. w_i can be got from the Eq. (4). Finally, the score of the feed is measured using the Eq. (6),

$$S_p = \frac{WF_{max}}{n} + \frac{lg\, n}{lg\, n_{max}} \times \frac{1}{2} \quad (6)$$

n is defined the same as the Eq. (5). n_{max} is the maximum posters number in the whole relevance feeds to a topic. WF_{max} is the maximum entity value in the feed j. We think larger S_p corresponds to higher possibility that feed might be an official one.

3.3 In-depth vs. Shallow Model

In the second stage, we use the in-depth analysis model. The facet of a blog is judged based on all the posts in it. What kind of posts is considered as in-depth? In common sense, an in-depth post expresses author's opinion on the given topic in detail with a long length in ideal situation. For minimizing the impact of spam contents, the length with average length is considered as a feature of the in-depth degree. But only using the length feature isn't sufficient, to confirm the relevance degree, considering the query term frequency in the post is also necessary. We combine the posts' length and the query term frequency in the following *L-Qtf* coefficient:

$$L\text{-}Qtf = \sum_{t \in Q \cap D} \frac{1 + \ln(1 + \ln(tf))}{(1 - s) + s\frac{dl}{avdl}} \times qtf \quad (7)$$

where *tf* and *qtf* represent the query term frequency in the post and in the query respectively. dl is the post length and $avdl$ is the average-length of the whole relevant posts for the topic. s is a parameter which is set as 0.2 in our experiments. *L-Qtf* coefficient is a kind of pivoted weighting coefficient [8].

Based on the whole posts of the topic-relevant blogs given by the blog distillation, the posts are ranking according to the in-depth coefficient. In this ranking list, the top 45% of topic-relevant posts are considered as the in-depth, while the last 45% posts are the shallow. The in-depth degree (*ID*) of each blog is calculated according to the relationship between the in-depth posts and

shallow posts as Eq. (4), where $in(post_i, Q)$ is 1 if post i is in-depth, and 0 otherwise. Similarly, $sh\ (post_i, Q)$ is 1 if post i is shallow. The larger the *ID* is, the deeper the feed is. Otherwise, the shallower the feed is.

$$S_i = ID(blog_x, Q) = \frac{\sum_{i=1}^{n} in(post_i, Q) - \sum_{i=1}^{n} sh(post_i, Q)}{n} \quad (8)$$

3.4 Combination model

In the task, the feed should be judged not only the topic relevance but also the facets. By considering these two points, the combination model is adopted.

$$S_j = \mu \times Score(blog_x, Q) + (1-\mu) \times S(feed_j) \quad (9)$$

S_j is the final confidence value of the feed j. *Facet* means the facet's value from different model. $S(feed_j)$ is got from function 1. μ is a weighting parameter distributing in the interval [0, 1]. μ is a parameter balancing the scores of facet level and similarity.

4 Submission and Evaluation Results

We submitted 2 runs. The difference between the 2 runs is query. The first run prisb's query is without query expansion, the words in the title field are only used. The second run pris's query is expanded by LQE. The evaluation results of the 2 submitted runs are listed in the following Table 1. From these data, it proves that the LQE is effective.

We have done some experiments after this track. In this section, we present empirical evaluation results to assess the effectiveness of our technique for the in-depth blog distillation. In particular, we conducted experiments on the permalink HTML pages of Blog08 [1] Collection to show that our algorithm is effective. We selected Indri as our information retrieval platform and preprocessed the data collection. A post is very similar to a web page which contains many HTML tags and scripts, so we parsed the HTML pages and reserved the texts in the same way dealing with a web page. In addition, we applied some rules for abbreviations, for example "I'm" was processed to "I am". And we stemmed the texts by Indri.

4.1 Blog distillation

To evaluate the performance of Baseline A, Baseline B and the LQE model on blog distillation, we made an experiment with 39 queries given by Blog track 2009, among them 18 queries were used for in-depth blog distillation.

The evaluation results are illustrated in Table 2. We employed four performance metrics: mean average precision (MAP), P@10, binary preference (bPref) and rPrec. It is obvious that Baseline B outperforms Baseline A on MAP, bPref and rPrec. From the results, we can see that PA algorithm is an effective algorithm for blog distillation, and that the results with query expansion outperform the baseline results in four performance metrics. We believe that the LQE model is an effective model for query expansion and information retrieval if the query is given as in Blog track task.

4.2 In-depth analysis model

Experiments with in-depth analysis coefficients were done on the retrieved blogs by Baseline B with LQE. The in-depth blogs were ranked according to their ID values. The top 100 blogs with positive ID values were evaluated.

Table 1. Faceted blog distillation results

	MAP	R-pre	b-pref	P@10
prisb.none	0.2756	0.3206	0.2767	0.3821
pris.none	0.2821	0.3420	0.2852	0.3949
prisb.first	0.1026	0.1249	0.1098	0.0923
pris.first	0.1243	0.1669	0.1324	0.1026
prisb.second	0.0626	0.0715	0.0511	0.1000
pris.second	0.0616	0.0792	0.0547	0.0846

Table 2. Blog distillation results

	MAP	P@10	bPref	rPrec
Baseline A	0.2145	0.3231	0.2372	0.2753
Baseline B	0.2756	0.2767	0.3206	0.3821
% improvement over Baseline A	+6.11	-4.64	+8.34	+10.68
B with LQE	0.2821	0.2852	0.3420	0.3949
% improvement over Baseline B	+2.36	+3.07	+6.67	+3.35

To find which factors are more efficient for in-depth analysis, we used four kinds of in-depth coefficients for comparison: (1) *Length* (baseline, coefficient A), (2)$Length \times$*L-Qtf* (coefficient B), (3)$\sqrt{Length} \times$*L-Qtf* (coefficient C), and (4) *L-Qtf* (coefficient D).

Table 3 shows MAP of each ID's result, comparing with the Length coefficient. From Table 3, we can see that the results of coefficient B and coefficient C significantly outperform coefficient A. The improvements by coefficients B and C over A are found to be statistically significant for MAP with large margin, but the results from C and D are not significantly different. The *L-Qtf* (coefficient D) achieves the best MAP, because it considers the average lengths of the posts that counteract parts of the impact produced by spam information. At the same time, query term frequency in *L-Qtf*, added as a factor for in-depth analysis, considers the relevance to the topic. From the results of the four coefficients, it can be concluded that the length isn't the most important factor for in-depth analysis, because there are a lot of spam information which are difficult to be detected in the blog posts. The comparative results from the coefficients B and D indicate that the lengths can't be seen as a single factor in the in-depth analysis. When the length without average length is considered as a single factor, spam is involved.

4.3 In-depth blog distillation system

We conducted experiments to examine the effect of our in-depth blog distillation system, and show the results in Table 4. We see that the combination model outperforms the model only using the in-depth analysis. While the combination model with the coefficient B was the best run in Trec 2009 in-depth blog distillation [5], from Table 4 the improvement over the Trec 2009 best run for the combination model with the coefficient D is large, showing more than 6% increase of MAP. This improvement also proves that it is not necessary to give high weighting on the length. With considering the performance of the combination model, we think that the topic-relevance information is a kind of much more important factor for in-depth analysis.

Table 3. Different in-depth coefficients for *ID*

	A	B	C	D
%MAP	0.78	2.38	4.17	4.45
% MAP improvement over Baseline	0	+1.6	+3.39	+3.67

Table 4. In-depth blog distillation results

	L-Qtf	Fused with coefficient B	Fused with *L-Qtf*
MAP	0.0445	0.1955	**0.2614**
P@10	0.0500	0.1167	**0.2278**
bPref	0.0475	0.2050	**0.2461**
rPrec	0.0380	0.2290	**0.2620**

5 Conclusion

In this paper, we present a system for the faceted blog distillation, propose a novel *L-Qtf* coefficient for the in-depth analysis and make a discussion on what kinds of factors may influence the in-depth analysis. The experiments prove that the topic-relevance information is an important factor for in-depth analysis, while the length factor should be considered but not too much.

Our system still has some weak points. The results of opinionated blog distillation and personal blog distillation are not good. In the future research, more factors influencing the faceted analysis should be explored and key words representing the faceted meanings should also be taken into consideration.

References

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

[2] B. Chen, H. He, J. Guo. Constructing maximum entropy language models for movie review subjective analysis. Journal of Computer Science and Technology, 23(2), pp. 231-239, 2008.

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

[4] J. Callan. Distributed information retrieval. In W. B. Croft, Editor, Advances in Information Retrieval. Kluwer Academic Publishers, Norwell, pp.127-150, 2000.

[5] D. Metzler, T. Strohman, H. Turtle, and W. Croft. Indri at TREC 2004: Terabyte track. In Proc. of the 2004 Text Retrieval Conf, 2004.

[6] H. He, B. Chen, L. Du, S. Li, et al.. PRIS in TREC 2008 Blog Track, In proceeding of TREC 2008, 2009.

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

[8] A. Singhal, C. Buckley, M. Mitra. Pivoted document length normalization. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval. pp21-29,1996.