RMIT at TREC 2010 Blog Track: Faceted Blog Distillation Task

Zhixin Zhou
School of CS & IT
RMIT University
VIC 3001, Australia
zhixin.zhou@rmit.edu.au

Xiuzhen Zhang
School of CS & IT
RMIT University
VIC 3001, Australia
xiuzhen.zhang@rmit.edu.au

Phil Vines
School of CS & IT
RMIT University
VIC 3001, Australia
phil.vines@rmit.edu.au

Abstract *This paper reports RMIT's participation in the TREC Blog Track 2010. For the baseline task, we adopted the BM25 model implemented in the Zettair search engine to establish a retrieval system of blog posts based on topic relevance. We then experimented with a number of different approaches to aggregate the post similarity scores to retrieve the most relevant blogs. Similarly, for the faceted distillation task we built a system at the post level first. After that, scores are aggregated for blogs to re-rank the most relevant blogs for the facet inclinations. A SVM classifier has been trained on Blog 06 collection to produce the opinion scores for each post. The cross entropy is used to evaluate posts for the in-depth versus shallow facet. For the personal versus official facet, we assumed blogs which are opinionated are also personal.*

Keywords post score aggregation, faceted distillation

1 INTRODUCTION

With the emergence of Web 2.0 technologies, user generated data has become one of the major sources of information in the web. Unlike content generated by mass media, this type of data is usually not subject to censorship, nor is it conforming to strict formats or standard writing style. More often than not, the authors of these articles are motivated solely by the desire to express opinions or basic instincts to release information, which makes the articles an ideal source of information to draw a true picture of the social response to various matters. The TREC Blog Track allows researchers to conduct research on user generated data through a set of tasks. Two collections have been used in the Blog Track [10, 6, 9, 7], namely Blog06 and Blog08. While the Blog06 collection had been used mainly for post-level distillation, new tasks on distillation at blog level had been proposed with the introduction of Blog08 [9]. These tasks mimic the user behavior of finding and subscribing to blogs using feeds, with facets enabled to perform advanced searching. Blog Track 2010 has three tasks. The baseline task is about finding blogs which are relevant to a user query. Faceted blog distillation task is extended beyond the baseline task to find blogs which shows recurring interest on given topics. Three facets have been considered since Blog Track 09, namely, opinionated versus factual, personal versus official and in-depth versus shallow. Top stories identification task is defined to investigate the blogospheres response to news stories as they develop, and find the most relevant blogs for headlines of each query day.

In this year RMIT university participates in blog track and submits runs for both baseline task and faceted distillation. We did not participate in the top stories identification task due to time limitations. In the remainder of this paper we will explain our approaches in the baseline and faceted distillation tasks. First we shall introduce the methods we used in both tasks in section two, after which we will describe our systems implemented in section three. Following that we shall discuss the evaluation methods and produce a conclusion in section four, and conclude with some discussions in section five.

2 METHODS

2.1 Overview

Most existing document retrieval systems are designed to retrieve documents. When a query is given, the similarity between a document is often estimated with a score that could be calculated with a certain query model. However, a blog consists of a number of posts, each of which is a single document. As such, document-based retrieval systems are usually not directly applicable to the blog distillation task. Let us consider two blogs F_A and F_B, where F_A contains m relevant post, and F_B contains n posts. With an existing system we can evaluate how likely each post p_{A_i} or p_{B_i} is related to the query, but we will not be able to tell which one of F_A and F_B is more relevant. On the other hand, for a user who intends to subscribe to a blog, the judgment is made at the blog level, not on any single post. To address this need, we experimented a number of ways to aggregate the relevance score of all posts belonging to each blog. Our experiment shows that the probability-based approach renders the best MAP.

Users are also interested in search tasks refined by facets inclinations. For a given topic X, some users might want to find blogs which contains opinionated content, whereas others might be looking for those which provide an in-depth coverage of it. In these cases, the topic relevance and facet inclinations must both be evaluated. Two strategies were considered, namely the bottom up strategy and the two-step strategy. The former combines the relevance score and the facet score at the post level, and then aggregates the combined post scores for the blogs to produce a final ranking. The latter involves two steps, which first retrieves the blogs by topic relevance, and then re-rank the blogs by their aggregated facet sores. However, as TREC organizers have provided the participating teams with standard baselines for the faceted distillation task, it is easier to evaluate the effectiveness of our systems with the two-step process. Thus, we adopted the latter this year. The bottom-up strategy is subject to further study.

2.2 Baseline Task

Existing information retrieval systems are capable of ranking blog posts by their relevance to a user generated query, where the relevance is often estimated in the form of a similarity score. In this paper, we use Zettair search engine to for indexing, and adopt the Okapi BM25 model implemented by Zettair to calculate the similarity scores for each blog post. We assume that the similarity score produced by Zettair is a reasonably good indicator of the topic relevance of a post to the query.

As blogs are collections of blog posts, there arises the need of aggregating the scores for blogs. We propose two approaches for aggregation,

1. The Two-step Baseline A successful aggregation method would effectively simulate how human beings, when given the information on post similarities to the query, make judgements on blog relevance. Intuitively, a blog with at least one relevant post should be judged as relevant; and we assume that a blog with a larger amount of relevant content is considered more relevant. Following these intuitions, we proposed a two-step approach. With this approach, blogs are retrieved through two stages. First, the highest post similarity score is recorded for each blog, by which all the blogs in the collection are ranked.

$$S_F = s_{p_{top}} \tag{1}$$

where $s_{p_{top}}$ is the highest score among all posts in the feed.

The top 100 blogs in this ranking are kept in the pool. Note that the ranking is only meant to retrieve a pool of candidate blogs. While it is intuitively valid to claim that blogs with a highly relevant post are more likely to be relevant to the query, it would be insufficient to assume that a higher top post score implies stronger topic relevance at the blog level. As such, we use the sum of similarity scores in each blog to re-rank the blogs in the pool.

$$S_F = \sum_{i=1}^{n} s_i \tag{2}$$

where s_i is the similarity score of the i^{th} post.

The size of the pool is a factor that could be tuned in order to achieve optimal performance. An advantage of this approach is that it provides a flexible framework. As there is often an inverse relationship between recall and precision, this approach employs a high-recall approach in the first step followed by a high-precision approach in the second to optimize the retrieval performance at the cost of time and computation power.

2. The Probablistic Baseline Another approach we propose is also based on the assumption that "A blog with at least one relevant post is relevant", but holistic. We estimate the likeilihood of a blogs relevance to a given query from the degree of relevance of its post entries. According to our assumption, a blog is considered irrelevant only if all posts in the blog are irrelevant. To calcluate the probability of the event that a blog be relevant to the query, we first transformed the post similarity scores in a feed F into probabilistic values,

$$p_i = \frac{s_i - s_{Q_{lower}}}{s_{Q_{upper}} - s_{Q_{lower}}} \tag{3}$$

where p_i is the probability of the i^{th} post being relevant to the query, Q_{upper} is the highest similarity score of all posts relevant to the query Q, Q_{lower} is the lowest similarity score of all posts relevant to the query Q, and s_i is the similarity score of the i^{th} post. We performed the transformation on a per-topic basis, as we expected different distributions of post similarity scores for each topic. Based on our assumption, the probability of the blog being irrelevant can then be calculated as,

$$\bar{P}_F = \prod_{i=1}^{n} (1 - p_i) \tag{4}$$

As a blog can be either relevant or irrelevant, the probability of a blog being relevant is thus,

$$P_F = 1 - \prod_{i=1}^{n} (1 - p_i) \tag{5}$$

Intuitively, this approach would not work well with blogs with a large count of irrelevant posts, as $\prod_{i=0}^{n}(1 - p_i)$ shrinks dramatically even if p_i is sufficiently small. We circumvent this problem by applying a threshold on p_i, so that only the relevant posts are considered. Here, the probability of the blog being irrelevant is no longer the probability of all its posts being irrelevant. Instead, it is calculated as the probability of all relevant posts in this feed being irrelevant, where the relevant posts are selected by the threshold applied on the similarity score of the posts. Effectively this is setting the probability of low-score posts being relevant to zero. And since we believe the similariy score is a reasonable indicator of the query relevance, the low-scored posts can be assumed to be irrelevant.

2.3 Faceted Distillation

We considered the faceted distillation task as a re-ranking task. For each pair of the facets, we re-ranked the top 100 feeds retrieved in the baseline task for the positive inclinations (*opinionated, in-depth, personal*), and then reverse the ranking for the negative inclinations (*factual, shallow, official*).

2.3.1 Opinionated versus Factual

Support vector machine has been proven to be an efficient classifier in text mining [1]. In the faceted distillation task, we use the support vector machine to evaluate the extent to which a blog post is opinionated. We used an opinionated lexicon consisting of 389 words, which is a subset complied from the MPQA subjective lexicon [11]. Each post was represented by a vector, the components of which were binary values that indicated whether a word existed in the post or not. The classifier was trained on the Blog06 text collection first, and then applied to the posts in the Blog08 text collection to estimate the probability of each post being relevant to the query.

2.3.2 In-depth versus Shallow

Gerani et.al [3] employed the Cross Entropy (CE) between a blog post and the whole collection to measure the extent to which a post is in-depth. We adopted the same approach to produce the facet inclination score for the in-depth facet. The score of a post d being in-depth is calculated as,

$$s_i = CE(p(.|d), p(.|c)) = \sum_{t \in d} p(t|d) log(\frac{1}{p(t|c)}) \quad (6)$$

where $p(t|d)$ is the probability of a term t existing in the post d, $p(t|c)$ is the probability of a term t existing in the collection c.

Here, $p(t|c)$ was calculated by,

$$p(t|c) = \frac{tf(t,d)}{\sum_{t' \in d} tf(t',d)} \quad (7)$$

and $p(t|c)$ was calculated by,

$$p(t|c) = \frac{df(t,d)}{|c|} \quad (8)$$

2.3.3 Personal versus Official

Due to time limitations, we assumed that opinionated blogs were also likely to be personal blogs, whereas factual blogs tended to be official blogs. Thus, we applied the same approach used in the opinionated versus factual facet to this pair of facets.

Table 1: Baseline Recall

Run label	Num of docs returned
rmitprob	213
rmit2step	169
stdbaseline1	267
stdbaseline2	196
stdbaseline3	148

3 IMPLEMENTATION

The Blog08 collection contains three types of documents, namely, permalinks, feeds, and homepages. We used Zettair[1] to index the permalink section of the whole collection.

With the Okapi BM25 [2] query system implemented in Zettair, we retrieved the top 100,000 relevant posts for each query. We evaluated the post score aggregation methods on the topics from Blog Track 2009 to select the baseline for submission. Graph 1 shows our results against the three standard baselines,

It can be seen from the graph that both of our approaches performed worse than the three standard baselines. However, as is shown in Table 3, with regard to the number of documents returned, textitrmitprob is better than two out of three baselines, and textitrmit2step is better than one. This implies that our approaches work reasonably well in finding the documents, but fail in ranking them. With the two-step approach, a possible cause of failure is that the sum of the similarity scores of all posts is not a good indicator of the blog's relevance to the topic, as blogs with numerous low-score posts could be ranked higher than blogs with a small number of high-score posts, despite of the fact that the latter is more relevant.

Similarly, the probablistic approach tends to favor the blogs with a large number of posts, regardless of their similarity score. This is because $\prod_{i=1}^{n}(1 - p_i)$ decreases dramatically as n grows, even if p_i were very small. With the probablistic approach, we have applied a threshold to include only the relevant posts in the pool. We defined relevant posts as those posts which have a possibility $p_i > p_T$, where p_T is the threshold. Figure 2 shows an independant experiment with the topics used in TREC 2009. It is clear in the graph that the threshold has a positive effect on the MAP. However, the recall was compromised when the threshold grew higher. This is mainly due to the fact that many relevant blogs don't have any post with a probability above the threshold. Since the top 100 feeds would be re-ranked in the faceted distillation task, we submitted the run with $p_T = 0.382$, as it had a reasonably high recall.

Figure 3 shows the performance of our approaches compared against the best, medium and worst runs from TREC 2010. Our approaches were consistently worse

[1] http://www.seg.rmit.edu.au/zettair/

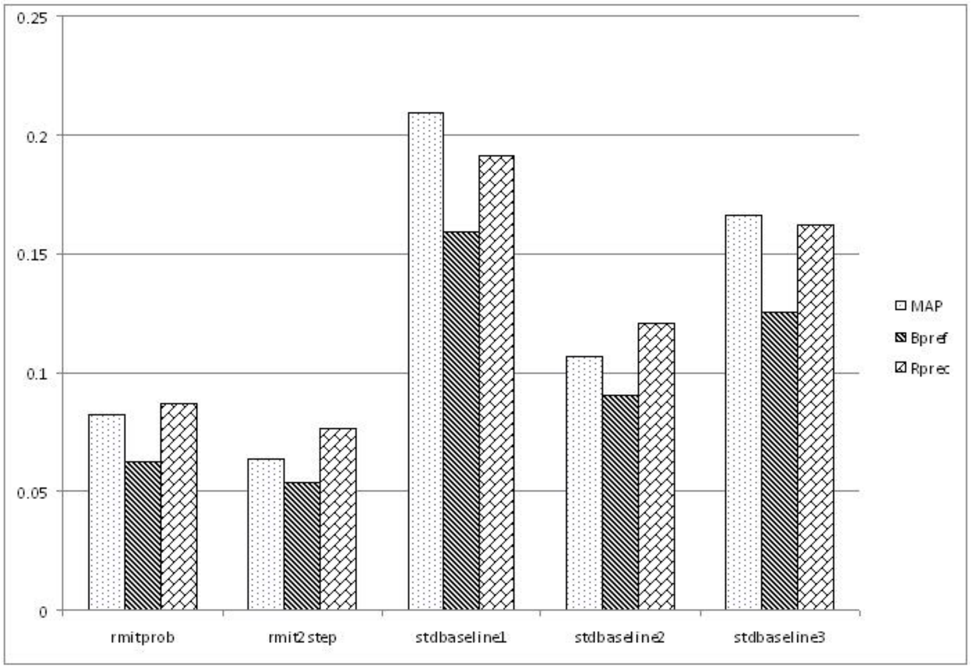

Figure 1: The Baseline Task

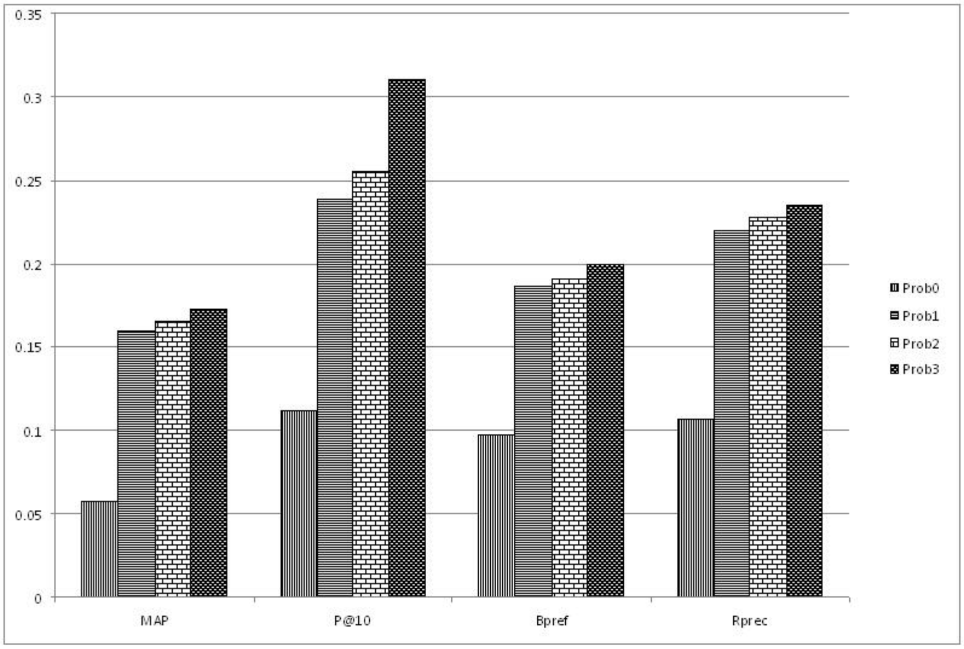

Figure 2: Probablistic approach with thresholds

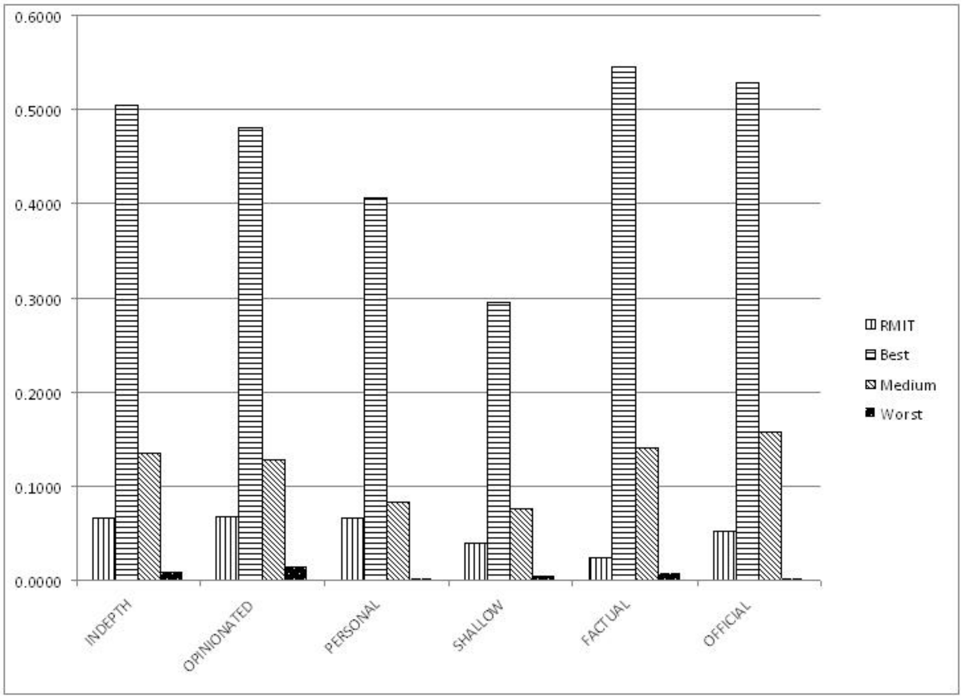

Figure 3: The Faceted Distillation Task

than average in all facets. As the performance of the faceted distillation system is closely related to the performance of the baseline, the failure of our system is largely due to the weak baseline we used. While most other teams used *standard baseline 1* as the basis for faceted distillation, we used our own baseline *rmitprob*, whose MAP was only 0.0887, whereas that of the *standard baseline 1* was 0.2174.

4 DISCUSSIONS

This is our team's first attempt at the Blog Track, and our approaches suffered from several flaws. For the baseline task, we retrieved 100,000 posts for each topic, and used all of them for blog retrieval. We assumed that the similarity score produced by the search engine is a reasonable indicator of the post relevance to the query. The assumption is more likely to be valid if the size of our pool is smaller, as the precisions of the search engines are generally higher with the high-ranking posts, but declines sharply when the pool gets larger. Therefore, a pool of less than 10, 000 posts would have been a better choice. Another way to effectively exclude the low-ranking posts is to use a threshold on the post similarity score. Our later work [13] has shown that a higher threshold, which selects a smaller pool of documents with relatively higher query similarity scores, leads to better results in the baseline task.

Moreover, a query expansion module would have been helpful. Most other teams [12, 8, 5, 4] have used search engines with a query expansion module in TREC Blog Track 2009 and earlier, which lead to better baseline performance.

In aggregating post similarity scores, we could have applied a flexible threshold for each blog instead of using a fixed threshold for all posts in the collection. Other methods to effectively select a better pool of documents are subject to further study.

References

[1] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137142, 1998.

[2] K. Sparck Jones, S. Walker and S. E Robertson. A probabilistic model of information retrieval: development and comparative experiments:: Part 2. *Information Processing & Management*, Volume 36, Number 6, pages 809840, 2000.

[3] M. Keikha, M. Carman, R. Gwadera, S. Gerani, I. Markov, G. Inches, A. A Alidin, F. Crestani and LUGANO UNIV (SWITZERLAND). University of lugano at TREC 2009 blog track. 2009.

[4] Y. Lee, S. H Na, J. Kim, S. H Nam, H. Jng, J. H Lee, POHANG UNIV OF SCIENCE and TECHNOLOGY (SOUTH KOREA). Kle at trec 2008 blog track: Blog post and feed retrieval. 2008.

[5] S. Li, H. Gao, J. Gao, H. Sun, F. Chen, O. Feng, S. Gao, H. Zhang, X. Li, C. Tan et al. A study of faceted blog DistillationPRIS at TREC 2009 blog track. 2009.

[6] C. Macdonald, I. Ounis and I. Soboroff. Overview of the TREC 2007 blog track. In *Proceedings of TREC 2007*, 2007.

[7] C. Macdonald, I. Ounis and I. Soboroff. Overview of the TREC 2009 blog track. *Proceedings of TREC 2009*, 2010.

[8] R. McCreadie, C. MacDonald, I. Ounis, J. Peng, R. L Santos and GLASGOW UNIV (UNITED KINGDOM). University of glasgow at TREC 2009: experiments with terrier. 2009.

[9] I. Ounis, C. Macdonald, I. Soboroff and GLASGOW UNIV (UNITED KINGDOM). Overview of the trec-2008 blog track. 2008.

[10] I. Ounis, M. De Rijke, C. Macdonald, G. Mishne and I. Soboroff. Overview of the TREC-2006 blog track. In *Proceedings of TREC*, Volume 6, 2006.

[11] T. Wilson, J. Wiebe and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, page 347354, 2005.

[12] X. Xu, Y. Liu, H. Xu, X. Yu, L. Song, F. Guan, Z. Peng, X. Cheng and CHINESE ACADEMY OF SCIENCES BEIJING INST OF COMPUTING TECHNOLOGY. ICTNET at blog track TREC 2009. 2009.

[13] Z. Zhou, X. Zhang and P. Vines. Seeing the forest from trees: Blog retrieval by aggregating post similarity scores. *ADCS 2010*, pages 12–19, 2010.