Fitting Score Distribution for Blog Opinion Retrieval

Ben He
Department of Computing Science
University of Glasgow
Glasgow, The United Kingdom
ben@dcs.gla.ac.uk

Jie Peng
Department of Computing Science
University of Glasgow
Glasgow, The United Kingdom
pj@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, The United Kingdom
ounis@dcs.gla.ac.uk

ABSTRACT

Current blog opinion retrieval approaches cannot be applied if the topic relevance and opinion score distributions by rank are dissimilar. This problem severely limits the feasibility of these approaches. We propose to tackle this problem by fitting the distribution of opinion scores, which replaces the original topic relevance score distribution with the simulated one. Our proposed score distribution fitting method markedly enhances the feasibility of a state-of-the-art dictionary-based opinion retrieval approach. Evaluation on a standard TREC blog test collection shows significant improvements over high quality topic relevance baselines.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Performance, Experimentation

Keywords: Blog search, Opinion finding, Distribution fitting

1. INTRODUCTION

Blog opinion retrieval requires finding not only relevant, but also opinionated blog posts for a given topic. Success in blog opinion retrieval is measured by the improvement brought by an opinion retrieval technique over a topic relevance baseline [3].

Most current effective approaches to blog post opinion finding are dictionary-based, using a vocabulary of opinionated words for mining subjectivity in the blogosphere [3]. An example is the statistical dictionary-based approach proposed in [2], which is currently one of the most effective approaches. For example, this approach is able to improve the best baseline run in the TREC 2007 Blog opinion finding task [2]. This approach derives an opinion dictionary from the TREC Blogs06 test collection, and assigns an opinion score to each document in the collection by considering highly opinionated terms in the dictionary as a query.

The opinion scores are combined with the topic relevance scores given by the baseline, using methods derived from standard data fusion techniques, such as linear combination, or logarithmic combination. Documents are then re-ranked by the combined scores. An underlying assumption of the score combination methods is that the opinion scores and the topic relevance scores follow similar distributions. This assumption holds when the topic relevance scores and opinion scores are given by the same system. However, such an assumption may not hold if the baseline is produced by an external system, which may apply a weighting model that is different from the local one. Therefore, such an opinion retrieval approach has a limited feasibility and can only be applicable to specific topic relevance retrieval systems.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-483-6/09/07.

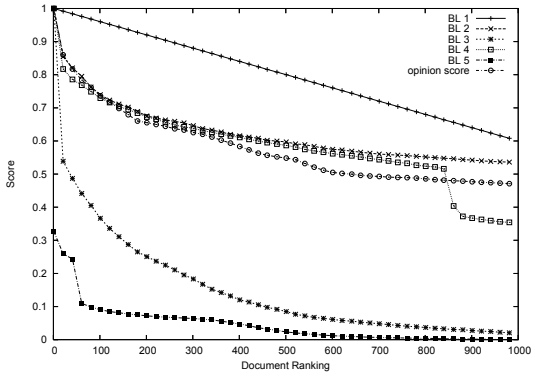

Figure 1: Score distributions of the top-1000 retrieved documents of the standard baselines (BL) 1-5 for TREC topic 851, and the opinion score distribution produced by the dictionary-based approach in [2]. All scores are normalised by the absolute value of the maximum score. For baseline 5, which has negative scores, all normalised scores are added by 2 for presentation purpose.

2. RESEARCH PROBLEM

An illustration of our research problem is the latest Blog opinion finding task in TREC 2008, where 5 high quality standard baselines were provided by the track organisers. Participants are required to apply their opinion finding techniques over these standard baselines. In this case, the score distribution of the baselines is a key factor that affects the effectiveness of the dictionary-based approaches. As shown in Figure 1, only baselines 2 & 4 have a similar score distribution with the opinion scores produced by the approach proposed in [2]. The score distributions of the other three baselines are highly different from that of the opinion scores. In this case, the dictionary-based approach in [2] is not applicable to the other three baselines. In other words, the dictionary-based approaches assume that the opinion scores and topic relevance scores are given by the same weighting model, which are likely to have similar score distributions. However, this approach is not applicable if the two scores are produced by statistically different weighting models, hence the need for simulating scores.

A possible solution for this problem is to apply a rank-based combination, ignoring the salient implication on rele-

vance and subjectivity of the actual scores. However, rank-based methods do not perform very well as shown by the TREC experiments [3]. In this paper, we propose to overcome this problem by fitting the opinion score distribution. By doing so, the baseline scores are replaced by the simulated ones obtained from the fitting function. As an illustration, we apply the score distribution fitting method on the dictionary-based approach in [2]. We evaluate the proposed method on the standard TREC 2008 Blog track test data. Using our proposed method, we show how the dictionary-based approach becomes applicable for baselines with different score distributions. In particular, the evaluation results show significant improvements given by the dictionary-based approach over 4 out of 5 standard baselines.

3. SCORE DISTRIBUTION FITTING

The basic idea of our score distribution fitting method is to map the score distribution of a given opinion retrieval method to a given topic relevance baseline, so that the opinion and topic relevance scores follow similar distributions.

We have experimented with several possible functions for fitting the distribution of the opinion scores of the retrieved documents. An exponential function was empirically chosen, given as follows:

$$f(rank) = a \cdot e^{b \cdot rank} + c \cdot e^{d \cdot rank} \quad (1)$$

where $f(rank)$ is the opinion score with a given rank; a, b, c and d are the parameters, which are obtained from the fitting procedure on a per-query basis.

In the fitting procedure, we maximise both R-square and adjusted R-square, as well as minimising both the residual of the Sum of Square due to Error (SEE) and Root Mean Squared Error (RMSE). The R-square statistic measures the success of the regression in predicting the values of the dependent variable within the sample. It can be interpreted as the fraction of the variance of the dependent variable explained by the independent variables. One problem with using R-square as a measure of goodness of fitting is that it never decreases in that it adds more regressors. The adjusted R-square, on the other hand, penalises R-square for the addition of regressors, which do not contribute to the explanatory power of the model.

Given a topic relevance score, for each query, the score of each retrieved document in the baseline is given by the above exponential function $f(rank)$ with the parameter values obtained in the fitting procedure.

4. EVALUATION

On the standard Blogs06 TREC collection, we use the 50 title-only topics from the latest TREC 2008 Blog track opinion finding task. For indexing and retrieval, we use the Terrier IR platform[1], and apply standard stopword removal and the Porter's stemming algorithm for English. We index the permalink component of Blogs06, which is the standard retrieval unit in the TREC Blog opinion finding task. The evaluation of our proposed approach is done over the 5 standard baselines in this task.

We apply the two score combination methods proposed in [2], namely the linear combination and the logarithmic combination. The parameter settings are taken from those recommended by He et al. in [2]. By fitting the opinion score distribution produced by their approach, we aim to show that i) this dictionary-based approach is now applicable to some of the 5 standard baselines as shown in Figure 1, and ii) this approach is able to improve the opinion retrieval performance over the 5 standard baselines using the simulated scores.

[1] `http://terrier.org`

Table 1: The MAP values provided by the 5 baselines (BL), and by the dictionary-based approach using the linear or logarithmic (Log.) combinations. An asteroid indicates a statistically significant improvement over the baseline according to the Wilcoxon Matched-pairs Signed-rank test. Mean is computed over the MAP values obtained for the 5 baselines.

	BL 1	BL 2	BL 3	BL 4	BL 5	Mean
BL	0.2335	0.2666	0.3074	0.3403	0.3211	0.2938
Linear	0.2724*	0.2723*	0.3181*	0.3466	0.3345*	0.3088*
Log.	0.2772*	0.2731*	0.3185*	0.3449	0.3343*	0.3096*

The evaluation results are summarised in Table 1. We find that using the simulated scores, the dictionary-based approach statistically outperforms 4 out of the 5 standard baselines in TREC 2008. Note that the dictionary-based approach is not applicable at all for baselines 1, 3 & 5 without the application of our score distribution fitting method. Therefore, our approach indeed enhances the feasibility of the dictionary-based approach for opinion retrieval, which is now applicable to any given baseline.

For baselines 2 & 4, the use of the original topic relevance scores and the simulated scores leads to similar retrieval performances[2]. We find no statistically significant difference between them. This is expected since we have shown in Figure 1 that the topic relevance scores in baselines 2 & 4 and the opinion scores follow similar distributions.

5. CONCLUDING REMARKS

Current opinion retrieval approaches cannot be applied to baselines that have a different topic relevance score distribution from the opinion score distribution. We have addressed this problem by fitting the distribution of opinion scores. The main contributions of this paper are twofold. First, our proposed score distribution fitting method has markedly enhanced the feasibility of a state-of-the-art dictionary-based opinion retrieval approach. There is no longer a limit on the score distribution in the topic relevance baselines for the application of such an approach. Second, our evaluation has shown statistically significant improvements over baselines, which were not applicable for the dictionary-based approach without the use of the simulated scores. Our proposed method can be applied to tasks that involve the combination of different sources of evidence. For example, we can combine the expertise evidence from external resources, as well as the internal resource, to boost retrieval performance in expert search.

6. REFERENCES

[1] G. Amati. *Probabilistic models for information retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, 2003.

[2] B. He, C. Macdonald, J. He, and I. Ounis. An Effective Statistical Approach to Blog Post Opinion Retrieval. In *Proceedings of ACM CIKM 2008*.

[3] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 Blog Track. In *Proceedings of TREC 2008*.

[2] We do not report the related results for brevity.