Limits of Opinion-Finding Baseline Systems

Craig Macdonald, Ben He, Iadh Ounis
Department of Computing Science
University of Glasgow, Scotland, UK
{craigm,ben,ounis}@dcs.gla.ac.uk

Ian Soboroff
NIST
Gaithersburg, MD, USA
ian.soboroff@nist.gov

ABSTRACT

In opinion-finding, the retrieval system is tasked with retrieving not just relevant documents, but which also express an opinion towards the query target entity. Most opinion-finding systems are based on a two-stage approach, where initially the system aims to retrieve relevant documents, which are then re-ranked according to the extent to which they are detected to be of an opinionated nature. In this work, we investigate how the underlying 'baseline' retrieval system performance affects the overall opinion-finding performance. We apply two effective opinion-finding techniques to all the baseline runs submitted to the TREC 2007 Blog track, and draw new insights and conclusions.

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

General Terms: Performance, Experimentation

Keywords: Opinion finding, baselines

1. INTRODUCTION

Blog posts often contains opinions by bloggers about various entities, and these can be mined and searched to understand public opinion about an entity. In 2006, TREC initiated the Blog track, and in particular an opinion-finding task [3, 4]. Most of the participating groups at TREC approached the problem in two stages. Firstly, a standard IR system was deployed to retrieve as many relevant documents as possible for a given query. Secondly, the system then re-ranked the retrieved documents based on the extent to which they are detected to be of an opinionated nature.

In TREC 2007, there was a requirement for all participating groups in the opinion-finding task to submit at least one baseline run where all opinion-finding features were disabled. We aim to determine if the baseline retrieval approach has an effect on the overall opinion-finding performance. Our methodology is as follows: We use two different but effective opinion-finding techniques, and apply these to all submitted baseline runs from TREC 2007. In this way, we are able to fix one experimental parameters of the TREC setting - i.e. the opinion finding approach used.

2. OPINION-FINDING TECHNIQUES

We apply two different approaches for detecting opinionated documents, which were both shown to be effective in TREC 2007 [1]. The first approach ('DICT') consists in automatically building a weighted dictionary from a training dataset where the distribution of terms in relevant and opinionated documents is compared to their distribution in relevant but not necessarily opinionated documents. The resulting weight of each term in the dictionary estimates its opinionated discriminability. The weighted dictionary is then submitted as a query to generate a score predicting how opinionated each document of the collection is. The second technique ('OF') is based on OpinionFinder, a freely available natural language processing toolkit, which identifies subjective sentences in text [5]. For a given document, the OpinionFinder tool is adapted to produce an opinion score for each document, based on the identified opinionated sentences. Using both techniques, we integrate the opinion score using the function proposed in [2]. In particular, this combination has a parameter k - we use the values in [2].

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

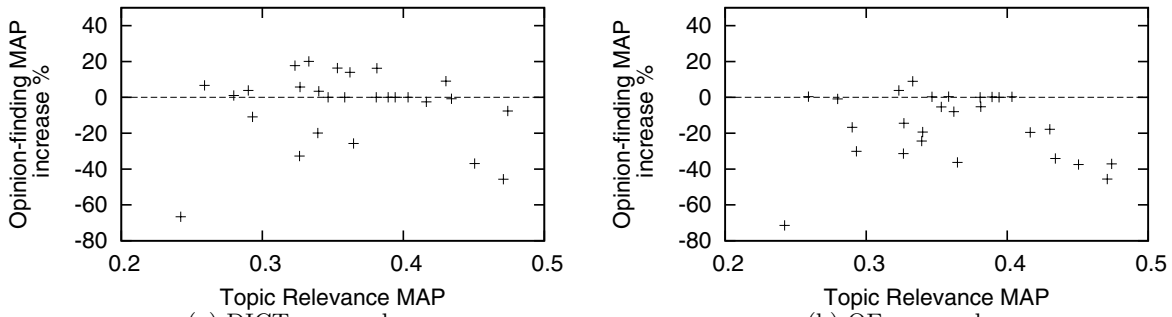
3. EXPERIMENTS AND ANALYSIS

We apply both opinion finding techniques, on 26 baseline runs submitted to the TREC 2007 task[1]. In the following, we compare the topic relevance Mean Average Precision (TMAP), which is calculated using any relevant documents retrieved, with opinion-finding (OMAP), which is calculated using only opinionated relevant documents. A successful application of an opinion-finding will result in an increase in OMAP. Firstly, we measure the correlation between runs ranked by TMAP and OMAP, using Spearman's ρ and Kendall's τ. We observe $\rho = 0.565, \tau = 0.442$ for the DICT approach, and $\rho = 0.410, \tau = 0.293$ for OF. These fairly marked correlations indicate that a strong baseline retrieval system is useful for achieving a high OMAP.

Next, Fig. 1 contains scatterplots showing how the increase in OMAP resulting from the application of the opinion-finding technique varies with the TMAP of the run. We observe that the DICT approach improves many runs by an equal margin of 20%, while some other runs are not improved. Indeed for 5 of 26 runs, applying the DICT approach results on a decrease in OMAP. For the OF approach, similar observations can be made, although there is more variance, and more runs are not improved. As both of opinion-finding approaches are integrated with the document retrieval score using a function that takes a parameter k, we hypothesise that this parameter, which was trained for one retrieval system, is not applicable for all systems.

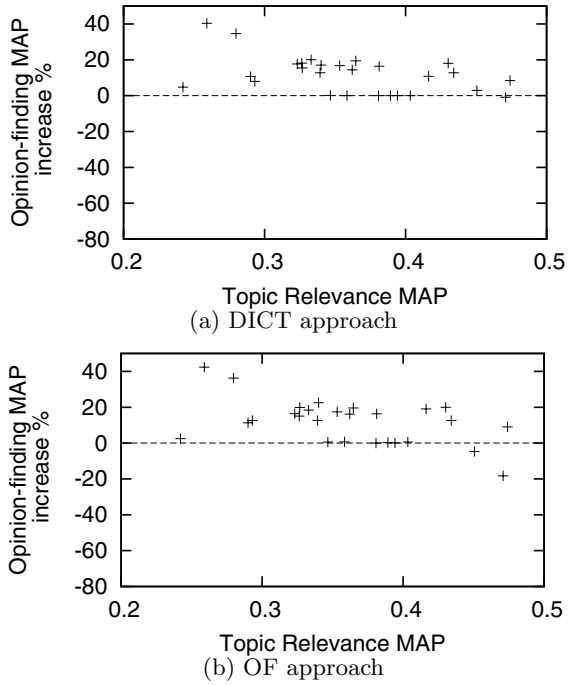
In the scatterplots in Fig. 2, each point represents the OMAP increase % achieved by applying an opinion-finding approach to a retrieval system, where the k has been opti-

[1]A baseline run with TMAP 0.001 is omitted.



(a) DICT approach (b) OF approach

Figure 1: Scatterplots showing the overall correlation between Topic Relevance MAP, and an improvement in Opinion-Finding MAP when the opinion-finding technique is applied to each run.



(a) DICT approach (b) OF approach

Figure 2: As Fig. 1 (k has been optimally set).

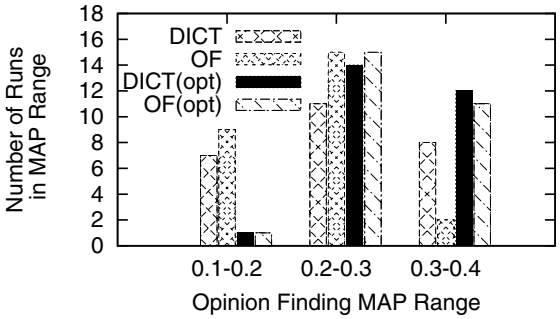


Figure 3: Histogram of absolute achieved Opinion Finding MAP by each approach. 'opt' denotes an approach where k parameter is optimally set.

mally set for each run by a wide-range value sweep. We observe that almost all runs can be improved by applying an opinion-finding technique. Of interest is that two runs, with high TMAP, have improved opinion-finding performance over the default k setting, yet still did not achieve the OMAP of their corresponding baseline runs. Conversely, the run with highest TMAP among others has improved OMAP over its baseline run when k is optimally set.

Fig. 3 contains a histogram of the absolute OMAP achieved by all runs when each approach is applied, including with optimal k values. We observe that higher absolute OMAP can be achieved by optimal approaches. As would be expected given the high correlations observed earlier, the run with highest TMAP achieves a 9% improvement in OMAP when k is optimally set, giving it the highest OMAP observed.

4. CONCLUSIONS

In this paper, we found that while there was a strong correlation between TMAP and OMAP, the application of either opinion-finding approach could fail to result in an increase in OMAP for some systems. In contrast, when the combination function was optimally tuned for each input system, there was more chances of improved retrieval performance. However, because some strong baseline retrieval systems could not be improved by applying effective approaches, we believe there is still further research to be done to fully understand the opinion-finding task.

For TREC 2008 Blog track, we (the organisers) have proposed a shared baseline task, where any proposed opinion-finding techniques has be to applied on the shared baseline. Based on the results in this paper, we conclude that it is important that participants are provided with the baseline runs on appropriate training data, so that they can train their re-ranking function. In the future, we will investigate if the opinion finding approach can be combined using a rank combination technique, instead of relying on the retrieval systems having compatible score distributions.

5. REFERENCES

[1] D. Hannah, C. Macdonald, B. He, J. Peng, and I. Ounis. University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier. In *Proceedings of TREC 2007*.

[2] B. He, C. Macdonald, and I. Ounis. Ranking Opinionated Blog Posts using OpinionFinder. In *Proceedings of SIGIR 2008*.

[3] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *Proceedings of TREC 2007*.

[4] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC 2006 Blog Track. In *Proceedings of TREC 2006*.

[5] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. OpinionFinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demos*, 2005.