On Performance of Topical Opinion Retrieval

Giambattista Amati
Fondazione Ugo Bordoni
Rome, Italy
gba@fub.it

Giuseppe Amodeo
University of L'Aquila
L'Aquila, Italy
gamodeo@fub.it

Valerio Capozio
University "Tor Vergata"
Rome, Italy
capozio@mat.uniroma2.it

Carlo Gaibisso
IASI - CNR
Rome, Italy
carlo.gaibisso@iasi.cnr.it

Giorgio Gambosi
University "Tor Vergata"
Rome, Italy
gambosi@mat.uniroma2.it

ABSTRACT

We investigate the effectiveness of both the standard evaluation measures and the opinion component for topical opinion retrieval. We analyze how relevance is affected by opinions by perturbing relevance ranking by the outcomes of opinion-only classifiers built by Monte Carlo sampling. Topical opinion rankings are obtained by either re-ranking or filtering the documents of a first-pass retrieval of topic relevance. The proposed approach establishes the correlation between the accuracy and the precision of the classifier and the performance of the topical opinion retrieval. Among other results, it is possible to assess the effectiveness of the opinion component by comparing the effectiveness of the relevance baseline with the topical opinion ranking.

Categories and Subject Descriptors: H.3.3 Information Search and Retrieval: Performance evaluation (efficiency and effectiveness)

General Terms: Theory, Experimentation

Keywords: Sentiment Analysis, Opinion Retrieval, Classification

1. INTRODUCTION

Opinion mining aims to classify sentences or documents by polarity of opinions. The application of opinion mining to IR (named Topical Opinion Retrieval) deals with ranking documents according to both topic relevance and opinion content. Topical Opinion Retrieval goes back to the novelty track of TREC 2003 [11] and the Blog tracks of TREC [7, 4, 8].However, there is not yet a comprehensive study of the interaction and the correlation between relevance and sentiment assessments. For example, the best runs based on the best official topic relevance baseline (baseline4) in the blog track of TREC 2008 (short topics 1001-1050) [8] achieve the MAP_R value equal to 0.4724, that drops to the $\text{MAP}_{O|R}$ of opinion equal to 0.4189, and to MAP equal to 0.1566 and 0.1329 for the polarity tasks (positive and negative opinionated rankings respectively). Performance degradation is intuitively expected because any variable which is additional to relevance, for example opinion, deteriorates system performance.

There is no way to separate and evaluate the effectiveness of the opinion detection component, or to determine whether and to which extent the relevance and opinion detection components are influenced by each other. It seems evident that an evaluation methodology or at least some benchmarks are needed to assess how much effective the opinion component is. At the moment, the only way to assess $\text{MAP}_{O|R}$ after opinionated re-ranking is to compare the increment of $\text{MAP}_{O|R}$ with respect to the relevance baseline, that is to assume the relevance baseline as a random ranking of opinionated documents about a given topic. Continuing with the same example, the $\text{MAP}_{O|R}$ of the relevance baseline is 0.3822, so that the actual $\text{MAP}_{O|R}$ increment is 9.6% after opinion re-ranking. Changing baselines and thus initial MAP_R, one would have different increment rates even if the same polarity or opinion mining and re-ranking techniques were used. It is also a matter of fact that opinion $\text{MAP}_{O|R}$ seems to be highly dependent on the initial relevance MAP_R of the first-pass retrieval [7, 4, 8]. To exemplify: how effective is the performance value of opinion $\text{MAP}_{O|R}$ 0.4189 when we start from an initial relevance MAP_R of 0.4724? What would be the $\text{MAP}_{O|R}$ and $\text{P@10}_{O|R}$ by filtering documents as in a binary classification approach, and what would be the accuracy of such opinion classifier?

In conclusion, can absolute values of MAP be used to compare different tasks, such as topical opinion and the ad hoc retrieval, and to compare and assess the state of the art of different techniques on opinion finding? At this aim, we introduce a completely novel methodological framework to:
- provide best achievable topical opinion $\text{MAP}_{O|R}$, for a given relevance document ranking MAP_R;
- predict the performance of topical opinion retrieval given the performance of topic retrieval and opinion classification;
- reciprocally, assess the opinion detection component accuracy from the overall topical opinion retrieval performance;
- study the robustness of standard evaluation measures (MAP, P@10 etc,) for opinion retrieval;
- study best re-ranking or filtering strategies on top of opinion classifiers independently from the adopted ad hoc relevance model.

This paper focuses on a few of these issues.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

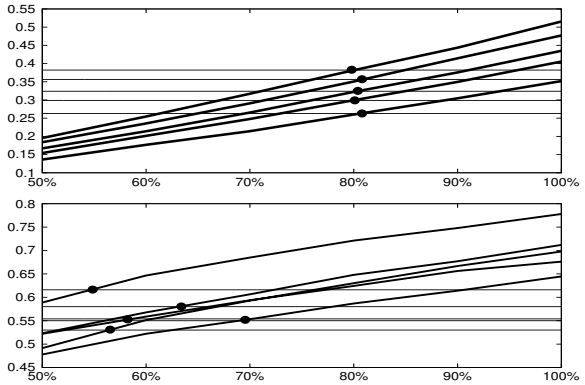

Figure 1: $\text{MAP}_{O|R}$ and $\text{P@10}_{O|R}$ by classifier accuracy. Opinionated document filtering with Monte Carlo sampling from the five TREC official baselines.

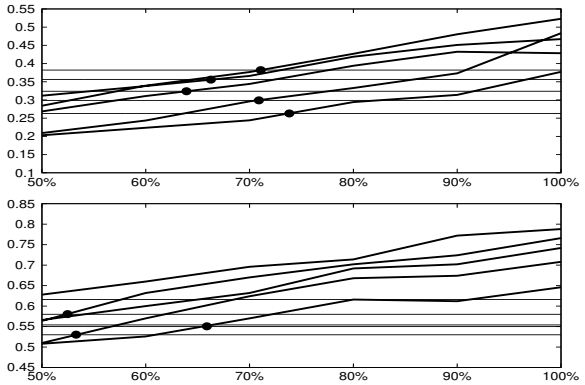

Figure 2: $\text{MAP}_{O|R}$ and $\text{P@10}_{O|R}$ by classifier accuracy. Relevant document re-ranking of the five TREC official baselines with lexicon-based scores and Monte Carlo sampling.

2. METHODOLOGY

In the first step all documents that are relevant to the test queries $\mathcal{R} = \cup_{q \in Q} \mathcal{R}(q)$ are pooled. From the subset $\mathcal{O} = \cup_q \mathcal{O}(q) \in \mathcal{R}$ of opinionated and relevant documents one obtains the conditional probability of occurrence of opinions with respect to relevance, $P(\mathcal{O}|\mathcal{R})$. Assuming that each document has an unknown topic as a hidden variable, $\mathcal{O}$ is a sample of opinionated documents of the whole collection, and the prior for opinionated but not relevant documents, i.e. $P(\mathcal{O}|\overline{\mathcal{R}})$, is provided by $P(\mathcal{O}|\mathcal{R})$ (i.e. it is *de facto* postulated the independence between relevance and opinion content). The second step consists in constructing a Monte Carlo sampling of opinionated documents from test data, and in obtaining thus opinion-only classifiers with different accuracy k. The $\text{MAP}_{O|R}$ is averaged on rankings built with all classifiers with the same accuracy k. The relevance rankings are modified according to:

- *filtering*, that is not opinionated documents are removed from the relevance baseline;
- *re-ranking*, that is opinionated documents receive a "reward" in their relevance ranking.

Topical opinion retrieval by Monte Carlo sampling is easily conducted with the filtering approach. Re-ranking approach requires the combination of two different scores, related to content and to opinion (e.g. [2, 9]). Opinion scores can be easily obtained with a lexicon-based approach [5, 10, 6, 3, 1]. The official relevance baselines and the topical opinion scores are provided by the blog TREC, while the opinion scores are not available, so that we use the lexicon-based scores and the re-ranking methodology that can be found in [1]. This lexicon-based approach has a good performance and achieves the $\text{MAP}_{O|R}$ of 0.4006 with respect the same baseline (baseline4 of the blog TREC 2008). Monte Carlo sampling consists in assigning the lexicon-based score to opinionated documents and a null score to non opinionated ones, assuming an accuracy $0 \leq k \leq 1$ of the classifier.

3. CONCLUSIONS

To improve $MAP_{O|R}$ with respect to its baseline requires a high accuracy of the opinion-only classifiers (at least around 80% with the filtering approach and 70% with the re-ranking technique of [1] as shown in Figures 1 and 2). However, smaller accuracy even close to that of the relevance baseline improves P@10 more easily than for MAP. The best value of $MAP_{O|R}$ with the filtering approach achieves an empirical value around the MAP_R of the relevance baseline. There is almost a linear correlation between the $MAP_{O|R}$ and the accuracy k of the opinion-only classifier, both in filtering and re-ranking approaches. Finally, interpolating values of Figures 1 and 2 one can show that the best run of TREC 2008 based on the re-ranking approach must have an opinion-only classifier accuracy greater than or equal to 78%, while using the lexicon-based classifier of [1] an accuracy of 74%.

Within this evaluation framework we can now compare performances of different opinion mining techniques, investigate how relevance and opinion are influenced by each other, and assess effectiveness of re-ranking strategies.

4. REFERENCES

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In *Proc. 30th ECIR*, vol. 4956, LNCS, pages 89–100, 2008.

[2] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *Proc. ACL-EMNLP Conf.*, pp. 345–354, 2006.

[3] X. Huang and W. B. Croft. A unified relevance model for opinion retrieval. In *Proc. 18th ACM-CIKM*, pp. 947–956, 2009.

[4] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 Blog Track. In *Proc. 16th TREC*, 2007.

[5] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *Proc. 15th TREC*, 2006.

[6] S. H. Nam, S. H. Na, Y. Lee, and J. H. Lee. Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. In *Proc. 31st BCS-ECIR*, vol. 5478, *LNCS*, pp. 791–795, 2009.

[7] I. Ounis, M. de Rijke, C. Macdonald, G. A. Mishne, and I. Soboroff. Overview of the TREC-2006 Blog Track. In *Proc. 15th TREC*, 2006.

[8] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC-2008 Blog Track. In *Proc. 17th TREC*, 2008.

[9] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[10] J. Skomorowski and O. Vechtomova. *Proc. 29th BCS-ECIR*, vol. 4425, *LNCS*, pp. 405–417, 2007.

[11] I. Soboroff and D. Harman. Overview of the TREC-2003 Novelty Track. In *Proc. 12th TREC*, pp. 38–53, 2003.