RGU at the TREC Blog Track

Malcolm Clark, Ulises Cerviño Beresi, Stuart Watt, David Harper
The School of Computing
The Robert Gordon University
Aberdeen, Scotland, United Kingdom
{mc,ucb,sw,djh}@comp.rgu.ac.uk

November 8, 2006

Abstract

Blogs are highly rich in opinion making their automatic processing appealing to marketing companies, the media, costumer centres, etc. TREC ran a Blog track in 2006 with two tasks: opinion retrieval and an open task. This document reports the experiments conducted at The Robert Gordon University (RGU) where we used Statistical Language Models combined with shallow parsing techniques for the opinion retrieval problem.

Keywords: TREC, Blog, Opinion Retrieval, Information Retrieval

1 Introduction

The simplicity of Internet publishing has resulted in an increase of the information available online. Individuals have taken this opportunity to post their thoughts in a particular form of publishing known as blogging on Blogs (short for Web log). It has been claimed that Blogs are the latest form of self expression and it has been noted that they are rapidly changing the face of the Web. The most attractive aspect of Blogs, and the key to their success, is that they are mainly authored by independent individuals who have, as their sole purpose, the desire to make their opinions known to the world[3]. Consequently, Blogs are highly rich in opinion and often up to date with current affairs. This factor makes them quite useful to industries, such as marketing, or the media, where more direct feedback is an invaluable resource[4]. TREC ran a Blog track this year (2006) aimed at discovering the particularities involved in opinion search where participants were instructed to retrieve documents containing *opinions* pertaining a set of topics. The collection used in the 2006 TREC Blog Track consisted of over a million posts collected over 77 days; this offers a test-bed for Blog analysis research. In this study, we intend to use this collection to investigate the extraction of subjective, positive and negative opinions. The rest of this report is structured as follows: section 2 defines the problem and the framework, section 3 introduces the corpus, how it was indexed and our proposed runs. Sections 4 and 5 finalise by discussing the results and our conclusions.

2 Definition of the problem

The Blog track defined two problems to be tackled this year: opinion retrieval and an open task. The opinion retrieval task objective was to retrieve documents containing opinions about certain given topics. This suggests that content based retrieval would not be enough since there is an extra factor of difficulty which is to be able to identify the nature of the content, i.e. whether it is opinionated or not. An opinion is something that radically differs from a fact and that it is easy for a human being to identify but not for a computer, therefore special mechanisms or heuristics had to be implemented to enable our systems to be able to do so. This is due to the fact that some concepts are *"difficult"* to understand by computers. During the Reliable Information Access (RIA) Workshop, several categories of *difficult* topics and their properties were

proposed[1]. In particular, systems did not perform well when the systems needed extra information to correctly interpret the information need contained in a topic. An example of these topics is Topic 413 *What are new methods of producing steel?* The concept of a *new method* would be very difficult for a system to infer. The track's open task gave freedom to participants to propose a problem, do a preliminary analysis of it and present the results. The aim of the task is to find a suitable problem to be run as an official task in next year's Blog track.

3 Experiments

3.1 Indexing the corpus

The corpus was compiled by the University of Glasgow by harvesting blog sites over a 77 day period of time. Each document was constructed of information such as:

- **Permalinks:** A permalink is a URL(web address) that points to a specific blog posting despite the fact that the entry has passed from the front page into the relevant blogging archive.
- **Feeds (URL and number):** The URL provides the original feed web address and relevant identity number.
- **Blog (URL and number):** this information indicates the original web address of the blog page and relevant id number.
- **Webpage:** The actual webpage which contains all the blog data such as the original post by the blog author and any comments made by blog visitors.

We indexed the webpages (approximately 3 million documents) keeping for each document the following statistics:

- **Term Frequencies:** this is the traditional *tf* statistic.
- **Number of subjective terms:** the number of subjective terms taken from a pre-mined list of terms. This list was mined from a corpus of reviews (movie reviews, restaurant reviews, politicians reviews, etc.). These are the terms that usually precede an *opinionated* term, i.e. *I, you, we, think, believe, guess* etc.
- **Number of positive opinionated terms:** the number of positive *opinionated* terms taken from the same pre-mined list of terms from the previous bulletpoint. These are the terms that are likely to express a positive opinion, i.e. *like, love, brilliant, rocks* etc.
- **Number of negative opinionated terms:** the number of negative *opinionated* terms taken from the same pre-mined list of terms from the previous bulletpoint. These are the terms that are likely to express a negative opinion, i.e. *hate, sucks, lame, boring* etc.

We also calculated global statistics in the form of *document frequencies* (standard df).

3.2 Retrieval model

We decided to work with the Statistical Language Modelling (LM) for Information Retrieval (IR) model[5]. This model proposes that the score of a document is proportional to the probability of generating the query from the document model. Therefore the core of this approach is to have the best estimate possible of the model of the document.

$$score(q,d) \propto P(q|M_d)P(d) \tag{1}$$

where M_d is the model estimated for document d. We chose to estimate $P(d|M_d) = \lambda P_{mle}(q|M_d) + (1-\lambda)P_{Coll}(q)$ where

$$P_{mle}(q|M_d) = \prod_{w \in q} P_{mle}(w|M_d) = \prod_{w \in q} \frac{tf(w,d)}{|d|} \tag{2}$$

with $tf(w)$ being the term frequency of term w in document d and $|d|$ is the document length. We approximate $P_{Coll}(w) = \frac{df(w)}{|D|}$ being $df(w)$ the document frequency for term w and $|D|$ the size of the collection. The standard LM model disregards the prior probabilities $P(d)$ for scoring purposes making, effectively, all documents to be a-priori equally relevant. In our runs we decided to approximate the prior based on some belief on the opinion contained in the document.

3.2.1 Using classes of terms

In order to explore some of the subjective, positive and negative content of the blog corpus we conducted manual experiments with random samples of blog pages to establish the nature of it. In this experiment we asked three judges to manually classify a random sample of two thousand documents into two classes; whether they contained an opinion or not. We then applied shallow parsing techniques on the documents to extract the features we were interested in, such as the number of subjective terms, the number of opinionated terms, etc. Processes like this have been successfully applied to classify documents according to the sentiment expressed in them[6, 2]. Documents were, therefore, represented using only these features. A binary classifier was trained using a 10-fold procedure producing very good results. We found that the classifier was able to filter the documents based on the sentiment expressed in them irrespectively of the topic discussed in them, encouraging us to embed this information in our experiments which are detailed below.

3.2.2 Our baseline run

Our baseline run (run 1) used the standard language modelling model extending the query with opinionated terms from the previously mined list.

3.2.3 Run 2 - Subjective

For run 2 we took advantage of the statistics gathered at indexing time and we estimated the prior probability of a document, $P(d)$, to be proportional to the number of subjective terms in it. Mathematically speaking, this can be written as

$$P(d) \propto \frac{number\ of\ subjective\ terms}{document\ length} \tag{3}$$

The rationale behind this is that if a document contains a high proportion of subjective terms, it is more likely to contain an opinion on a certain topic than one that has a lower proportion.

3.2.4 Run 3 - Positive vs. Negative term counts

Run 3 was designed to identify whether an opinion had a positive characteristic or a negative one. We, therefore, approximate the document prior to be proportional to the proportion of positive opinionated terms versus the negative opinionated terms:

$$P(d) \propto \frac{pos(d)}{neg(d)} \tag{4}$$

where $pos(d)$ is the number of positive opinionated terms and $neg(d)$ is the number of negative opinionated terms. There is an issue with this approach: a document may not contain negative opinionated terms (or no opinionated terms at all). Moreover, if a document doesn't contain positive opinionated terms, the prior will be zero making the document's score zero as well. Therefore we modified equation 4 and we approximate $P(d)$ as

$$P(d) \propto \frac{pos(d) + neg(d) + 1}{document\ length} \tag{5}$$

which we interpret as the probability of the document containing *any* kind of opinion. To assess whether the opinion is positive or negative we simply evaluate $pos(d) < neg(d)$.

4 Results

Official TREC results were absolutely discouraging and suspicious, suggesting us that our systems had major bugs. This turned out to be the case, therefore we encourage the reader not to pay attention to them. We are currently re-running our experiments where we expect our systems to produce good results which will be available on the final paper.

5 Conclusions and future work

We saw that shallow parsing techniques can improve document filtering, in particular when the classes are not dependent on topical words but rather on a higher, more abstract, concept such as an *opinion*. We believe that incorporating these features into a formal model such as Language Modelling would prove to be highly beneficial not only for opinion retrieval but also for other types of retrieval tasks. In this paper we presented three simple heuristics but we would also like to point out that there is a lot of room for improvement. On

this matter, we plan to conduct more experiments where we will take into account opinionated term positions regarding content-carrying terms (query terms) to assess if the opinion expressed in the document is actually about the topic being searched for.

References

[1] C. Buckley. Why current ir engines fail. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 584–585, New York, NY, USA, 2004. ACM Press.

[2] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624, New York, NY, USA, 2005. ACM Press.

[3] S. C. Herring, L.A. Scheidt, S. Bonus, and E. Wright. Bridging the gap: A genre analysis of weblogs. *hicss*, 04:40101b, 2004.

[4] J. Hill. The Voice of the Blog: The attitudes and experiences of small business bloggers using blogs as a marketing and communications tool.

[5] J.M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Research and Development in Information Retrieval*, pages 275–281, 1998.

[6] C. Whitelaw and N. Garg. Using appraisal taxonomies for sentiment analysis. *Proc. Second Midwest Computational Linguistic Colloquium*, 2005.