FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track

Giambattista Amati[1], Edgardo Ambrosi[2], Marco Bianchi[2], Carlo Gaibisso[2], and Giorgio Gambosi[3]

[1] Fondazione Ugo Bordoni, Rome, Italy, gba@fub.it
[2] IASI "Antonio Ruberti" - CNR, Rome, Italy, *lastname*@iasi.cnr.it
[3] Matematics Department of University "Tor Vergata", Rome, Italy, gambosi@mat.uniroma2.it

Abstract We present a fully automatic and weighted dictionary to be used in topical opinion retrieval. We also define a simple topical opinion retrieval function that is free from parameters, so that the retrieval does not need any learning or tuning phase.

1 Introduction

A blog is basically a diary made up of items (posts) that are posted on a regular basis and, typically, displayed to visitors in reverse chronological order. Posts comprises text, hypertext, images, and links to other web pages, video, audio and other files. The TREC Blog track [6] was first introduced in TREC 2006 and ran again in 2007, with the claimed purpose of to explore information seeking behavior in the blogosphere.

A test collection, called Blog06, was created for the TREC Blog track [3]. The blog collection was crawled over a period of 11 weeks (December 2005 - February 2006). The total size of the collection amounts to 148 GB with three main different components: feeds (38.6 GB), permalinks (88.8GB)[1], and homepages (20.8 GB). The collection contains spam as well as possibly non-blogs and non-English pages. For our experimentation we considered only the permalink component, consisting of 3.2 million of Web pages, each one containing a post and its related comments.

Blog track 2006 only ran the opinion retrieval task. This task has been ran again in 2007, with a polarity subtask. Additionally, a new task has been added, the Blog Distillation (Feed Search) task. We have only taken part in the opinion retrieval task. This task focuses on a specific aspect of blogs: the opinionated nature of posts, that is it requires locating those blog posts that express an opinion about a given target. It can be summarized as "What do people think about

[1] Permalinks are URL links of blogging entries that exist even after they pass from the front page into the blog archives.

<target> ?". The target can be a "traditional" named entity (a name of a person, location, or organization), but also a concept (such as a type of technology), a product name, or an event.

Blog track 2007 adopted the same assessment procedure defined in 2006. The retrieval unit is document from the permalink component of the Blog06 test collection. The content of a blog post is defined as the content of the post itself and the contents of all comments to the post: if the relevant content is in a comment, then the permalink is declared to be relevant. The objective is to run again the opinion retrieval task with 50 new topics, that NIST selected from query logs provided by commercial blog search engines.

The following scale has been used for the assessment:

- *-1*, as not judged. The content of the post was not examined due to offensive URL or header (spam).
- *0*, as not relevant. The post and its comments were examined, and does not contain any information about the target, or refers to it only in passing.
- *1*, as relevant. The post or its comments contain information about the target, but do not express an opinion towards it.

Furthermore, if the post or its comments are not only relevant, but also contain an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), then the document is alternatively labeled as follows:

- *2*, as relevant, and containing negative opinions. The post contains an explicit expression of opinion or sentiment about the target, showing some personal attitude of the writer(s), and the opinion expressed is explicitly negative about, or against, the target.
- *3*, as relevant, and containing mixed positive and negative opinions. Same as *2*, but contains both positive and negative opinions.
- *4*, as relevant, and containing positive opinions. Same as *2*, but the opinion expressed is explicitly positive about, or supporting, the target.

Evaluation metrics are precision/recall based, such as the Mean Average Precision (MAP), but we focused our attention on Precision at 10 documents (P@10), because it is often used for Web search evaluation.

For the Blog track of TREC 2007, we create a dictionary of opinionated words from the Blog 2006 relevance data. We have two information theoretic measures to accomplish the automatic construction of an opinionated vocabulary. The first measure is based on a DFR (Divergence From Randomness) model and defines the weight of opinionated terms within documents. The second measure is based on the maximization of the entropy in the set $Opin =$

$\{\mathbf{d} : \mathbf{d}$ is relevant and opinionated$\}$ of all relevant and opinionated documents, and it is used to filter the words into a sequence of classes $\mathbf{V}_k \supset \mathbf{V}_{k+1}$ with $1 \leq k \leq |Opin|$. Our aim is to define the optimal k such that $\mathbf{V}_k$ is both as small as possible (for a real-time implementation purpose) and as effective as possible for opinion detection.

We take three assumptions for an automatic construction of an opinionated dictionary:

- Content-bearing words *maximize* the probability $\text{Prob}(\text{posterior}|\text{prior})$ of observing the posterior probability of occurrence in the subset $Opin$ of opinionated and relevant documents given the prior probability of occurrence in all relevant documents. In other words, content-bearing words occur with similar relative frequencies in both opinionated relevant and strictly relevant sets of documents.
- Opinion-bearing words instead *minimize* the probability $\text{Prob}(\text{posterior}|\text{prior})$ of observing the posterior probability of occurrence in the subset $Opin$ of opinionated and relevant documents given the prior probability of occurrence in all relevant documents. The weight of an opinion-bearing word is provided by the $-\log$ of such a probability (DFR model), and thus opinion-bearing words maximize the divergence $-\log_2 \text{Prob}(\text{posterior}|\text{prior})$.
- Best opinion-bearing words also *maximize* the entropy in the subset $Opin$ of opinionated and relevant documents. In other words opinion-bearing words occur more randomly than content-bearing words in the subset $Opin$ of opinionated and relevant documents. An approximating way to maximize entropy is to consider terms with highest divergence and that belong to a large number k of opinionated documents.

Our experimentation in the BLOG TREC 2007 consists of three phases:

1. data pre-processing and topic relevance retrieval;
2. semi-automatic construction of a sentimental dictionary with weighted terms;
3. topic opinionated relevance retrieval.

These phases are detailed described in Sections 2, 3 and 4, respectively. Finally, in Section 5 we report and discuss on the experimentation activity and results.

2 Data pre-processing and topic relevance retrieval

The data pre-processing phase is aimed to remove not English documents from the collection. This goal is achieved by Lingpipe [1], a suite of Java libraries for the linguistic analysis of human language. LingPipe's text classifiers learn by example. For each language being classified, a sample of text is used as training data. LingPipe learns the distribution of characters per language using character language models. Character language models provide state-of-the-art accuracy for text classification. Character-level models are particularly well-suited to language ID because they do not require tokenized input; tokenizers are often language-specific. A text classifier has been trained using the Leipzig Corpora Collection [2] with the aim to recognize up to 15 other than English. The Leipzig Corpora Collection is a freely available resource for corpora and corpus statistics covering more than 20 languages at the time being. The corpora are identical in format and similar in size and content. They contain randomly selected sentences in the language of the corpus and are available in sizes of 100,000 sentences, 300,000 sentences, 1 million sentences etc.. By means of Lingpipe classifier we removed 541.725 permalinks (16.86% of all documents) but only 74 were relevant (0.61% of all relevant documents). On the other hand, we have eliminated 8.09% of false positive documents (documents retrieved but labeled 0). This proves that language classification is very useful in both topical and opinion finding retrieval.

For the topic relevance retrieval we adopted Terrier [5]. Due to the large dimension of BLOG track collection we have developed a distributed version of Terrier, and we run our experiments on a cluster of 13 machines (1 broker + 12 nodes). From a software architecture perspective, we adopted a document partitioning strategy, and we solved the ”results merging” problem providing global statistics to each server query. Thanks to the distributed version of Terrier we can index the entire collection in less than 45 minutes: this is very useful to tune the indexing Terrier parameters and to quickly test the effectiveness of pre-processing activities. The index that we used for BLOG TREC 2007 was generated indexing all permalink files except for those marked as not English document by the LingPipe classifier and using the weak Porter stemmer. As retrieval models we use the parametric model PL2, with its parameter c set to 9, and the parameter-free model DPH.

3 Automatic construction of a sentimental dictionary with weighted terms

The automatic construction/weighting of a sentimental dictionary is a particularly ambitious objective: to automatically identify sentiment-bearing terms, and to automatically assign them an opinion weight.

To achieve these goals we learned from the set of relevant documents (labeled 1, 2, 3 or 4 from TREC 2006 relevance data) and the set of the opinionated ones (labeled 2, 3 or 4). Our hypothesis is that opinion-bearing words distribute more randomly in the set of opinionated documents than semanticbearing terms, but less randomly than the non-informative terms. Starting from this idea, we used a Divergence From Randomness (DFR) query expansion model to filter several levels of candidate terms, and we selected the level that maximized the number of "opinion-bearing" words.

More precisely, we first compute the asymmetric Kullback-Leibler divergence (KL) in the opinionated set with respect to the set of relevant documents obtaining a set of candidate opinion-bearing terms. Then, we compute different layers of opinionated terms by document frequency, as a way to maximize the opinion entropy for all such candidate terms. A generic k level contains all index terms occurring in at least k relevant and opinionated documents. Therefore the higher the number of documents containing a term, the higher is the probability that the term is opinionated. Furthermore, the larger k is chosen, the less the terms that are selected.

Our goal was to find the optimal level k that maximize the number of "opinion-bearing" words. To evaluate a generic k level, we executed the following steps:

1. we reduce the noise caused by the presence of many proper names, numerical terms, dates, and typos words by excluding all the index terms that contain numerical substrings or that are not included in the WordNet [4] dictionary.
2. we use a *sentimental dictionary* $Sent\mathbf{V}$ semi-manually built by third parts [7,8] to assess the automatic dictionary by precision and recall with respect to the sentimental dictionary $Sent\mathbf{V}$.
3. we study the effectiveness of combining both dictionaries.

All runs presented in Section 5 have been performed using all (not filtered) index terms occurring at the *100*-th level. We refer to this set of terms as *learned dictionary*. Note that the learned dictionary also contains terms that does not occur in the sentimental dictionary.

Each term in the learned dictionary has a *learned sentimental weight* coinciding with the weight returned by the query expansion model.

Table 1 shows an excerpt from the used learned dictionary of all terms classified as strong subjective in the sentimental dictionary $Sent\mathbf{V}$, and reports some statistics.

4 Topic opinionated relevance retrieval

In this section we show how we compute the final topic-opinionated relevance documents list. Let us assume to have an *opinionated vocabulary*, $Opin\mathbf{V}$ such that each term $\mathbf{t} \in Opin\mathbf{V}$ has a weight $w_{\mathbf{t}}$. Opinionated and relevant document ranking is obtained in four steps:

Table 1. The excerpt is from the set of index terms extracted at level $k = 100$ with their opinionated scores. The size of the learned dictionary amounts to 4222 index terms. We reduced the noise caused by the presence of proper names, numerical terms, dates, and meaningless words removing 3505 index terms (45.36% of the index terms occurring at level $k = 100$). The number of "opinion-bearing" terms occurring both in the learned and in the sentimental dictionary $Sent\mathbf{V}$ is 1529 (36% of the learned dictionary terms), with 945 terms classified as *strong subjective* and 584 terms classified *weak subjective* by the sentimental dictionary (61% and 39% of the terms in the intersection of the learned and the sentimental dictionaries, respectively).

An excerpt from the 945 strong subjective terms of the learned dictionary					
abide	0,0023	inaccurate	0,0064	⋮	⋮
abject	0,0031	inane	0,0009	wish	0.0060
absolute	0,0029	inappropriate	0,0028	wonder	0.0068
absurd	0,0076	incapable	0,0072	wonderful	0.0025
abusive	0,0047	incessant	0,0052	woo	0,0024
abyss	0,0008	inclin	0,0043	worri	0,0020
acclaim	0,0008	incoherent	0,0010	worse	0,0044
accuse	0,0012	incompetent	0,0012	worst	0,0041
activist	0,0023	incomprehensible	0,0018	worth	0,0018
actual	0,0069	inconvenient	0,0026	worthless	0,0097
admir	0,0024	incredible	0,0048	worthwhile	0,0016
admirable	0,0030	indefensible	0,0011	wound	0,0046
admire	0,0011	indicative	0,0017	wrath	0,0021
admit	0,0063	indifferent	0,0029	yeah	0,0070
⋮	⋮	indispensable	0,0033	yearn	0,0049

1. A content-only scored document list is obtained from Section 2. It means that each document has already associated a content score: we use the DFR models PL2 and DPH as retrieval functions to provide the content score of the documents

$$content\_score(\mathbf{d}||\mathbf{q}) = score_{DFR}(\mathbf{d}||\mathbf{q}).$$

We obtain a *content rank* for all documents:

$$content\_rank(\mathbf{d}||\mathbf{q}).$$

2. We submit all terms of $Opin\mathbf{V}$ as query-terms with their weights $w_{\mathbf{t}}$ and assign a query-independent score to the set of retrieved documents by means of a DFR model:

$$opinion\_score(\mathbf{d}||Opin\mathbf{V}) = score_{DFR}(\mathbf{d}||Opin\mathbf{V}).$$

The opinionated score with respect to a topic $\mathbf{q}$ is defined as follows:

$$opinion\_score(\mathbf{d}||\mathbf{q}) = \frac{opinion\_score(\mathbf{d}||Opin\mathbf{V})}{content\_rank(\mathbf{d}||\mathbf{q})}.$$

We thus obtain a *opinion rank* for all documents:

$$opinion\_rank(\mathbf{d}||\mathbf{q}).$$

3. We further boost document ranking with the dual function of $opinion\_score(\mathbf{d}||\mathbf{q})$:

$$content\_score^{+}(\mathbf{d}||\mathbf{q}) = \frac{content\_score(\mathbf{d}||\mathbf{q})}{opinion\_rank(\mathbf{d}||\mathbf{q})}.$$

4. Finally, we re-rank the documents by $content\_score^{+}(\mathbf{d}||\mathbf{q})$.

5 Results and discussion

Results of submitted runs are shown in Table 2.

- The run labeled as *FIUbPL2* represents the baseline. It is computed adopting the PL2 model (C=9) with all opinion-finding features turned off.
- The second run (*FIUdPL2*) is a second baseline based on *FIUbPL2*. It rearranges the topic-based rank with respect to the sentimental query, by using the algorithm described in Section 4 with DPH as model for the first opinion score. In this run we do not use a learned dictionary $\mathbf{V}_k$, for some level k, but only the strong subjective terms of the sentimental dictionary $Sent\mathbf{V}_{|Strong}$ (the use of the whole sentimental dictionary performs less). Each term in $Opin\mathbf{V} = Sent\mathbf{V}_{|Strong}$ has a weight $w_{\mathbf{t}}$ equal to 1. This run is useful to compare with the runs obtained by using a learned dictionary $\mathbf{V}_k$. It also provides us useful indications of the effectiveness of the re-ranking score function $content\_score^{+}(\mathbf{d}||\mathbf{q})$.
- The run labeled *FIUlPL2*, performed best in terms of MAP. It has been designed as described in Section 4, where the initial ranking is given by the baseline *FIUbPL2* and DPH as model for the opinion score. In this run we use both learned dictionary $\mathbf{V}_k$, for $k = 100$, and $Sent\mathbf{V}$ restricted to only the strong subjective terms. Each term in $Opin\mathbf{V} = Sent\mathbf{V}_{|Strong}$ has a weight equal to $1 + \alpha w_{\mathbf{t}}$, where α is set to the value ~ 1000 that normalizes the greatest opinionated term weight $w_{\mathbf{t}}$ in $\mathbf{V}_{100}$ to 1, and $w_{\mathbf{t}}$ is added when the term is also in $\mathbf{V}_{100}$.
- The run labeled as *FIUlDPH* differs from *FIUlPL2* because we use the parameter free model DPH instead of PL2 in the step 1 of the algorithm described in Section 4.

- The run labeled as *FIUlP22* differs from *FIUlPL2* because we use PL2 in the step 2 of the algorithm described in Section 4.
- Finally, *FIUDDPH* differs from *FIUlDPH* mainly because we conside both the title and the description topics fields in the step 1 of the algorithm described in Section 4. Furthermore, the content-only score is obtained applying the query expansion technique (we used a parameter free model of query expansion with 3 top ranked documents and 20 expansion terms).

All our official runs were evaluated by `trec_eval` as they were baselines, because we updated the final ranks but not the final topical-opinion scores. Using the Terrier evaluation tool, which instead evaluates runs by ranks and not by scores, our best title-only run is FIUIDPH for precision at 10 (topic 0.7160, opinion 0.5360 with +13.7% and +18% respectively over the baseline FIUbPL2), FIUIPL2 for MAP (topic 0.3910, opinion 0.3210 with +9.3% and +17.7%) and for R precision (topic 0.4264, opinion 0.3679 with +6.6% and +14.8%). Our actual results are reported in Table 2. It is worth to note that, with respect to the results presented in the “Overview of the TREC2007 Blog Track” paper [9], the FIUIPL2 is the best run in terms of improvements over the baseline, and is the fourth run among all the automatic title-only runs submitted to the opinion track.

Table 2. Summary of the FIU Blog track automatic runs.

		Relevance		Opinion	
Name	Topic field used	MAP	P@10	MAP	P@10
Median run of the Blog track 2007		0.3340	0.6480	0.2416	0.3031
FIUbPL2 (baseline)	title	0.3621	0.6300	0.2727	0.4540
FIUdPL2	title	0.3831	0.6760	0.3045	0.5180
FIUlP22	title	0.3930	0.6760	0.3177	0.5240
FIUlDPH	title	0.3861	0.7160	0.3188	0.5360
FIUlPL2	title	0.3956	0.6860	0.3210	0.5260
FIUDDPH	title, description	0.3977	0.7340	0.3315	0.5420

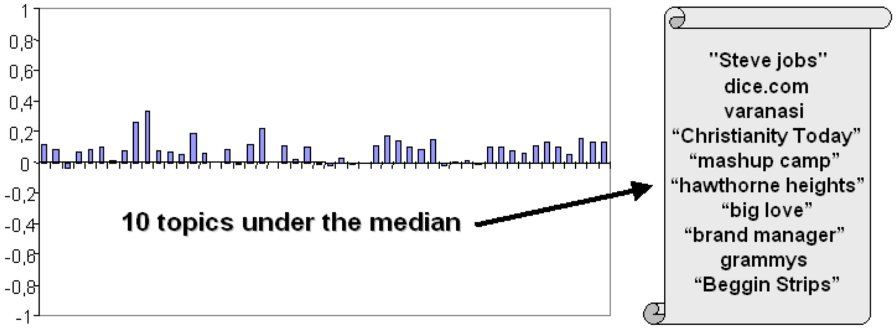
Considering the difference from median average precision per topic of the FIUlPL2 run (see Figure 1), we notice that only 10 topics are under the median and that we could improve the performance of most of them adopting some proximity strategy.

Moreover, Table 3 summarizes some new automatic runs performed as variations of the FIUlPL2 run. Comparing Tables 2 and 3 we can draw the following conclusions:

Table 3. Summary of new automatic runs. Variations of the FIUIPL2 run

			Relevance		Opinion	
Opinion Query	Name	Topic field used	MAP	P@10	MAP	P@10
Median run of the Blog track 2007			0.3340	0.6480	0.2416	0.3031
$\mathbf{V}_{100}$	weights of $\mathbf{V}_{100}$	title	0.3943	0.6800	0.3088	0.5120
$Sent\mathbf{V}$	weights of $\mathbf{V}_{100}$	title	0.3945	0.6880	0.3145	0.5260
$Sent\mathbf{V}_{\vert Strong}$ (FUIlPL2)	weights of $\mathbf{V}_{100}$	title	0.3956	0.6860	0.3210	0.5260
$Sent\mathbf{V}_{\vert Strong} \cap \mathbf{V}_{100}$	weights= 1	title	0.3864	0.6760	0.3089	0.5220
$Sent\mathbf{V}_{\vert Strong} \cap \mathbf{V}_{100}$	weights of $\mathbf{V}_{100}$	title	0.3981	0.6960	0.3218	0.5420
$Sent\mathbf{V} \cap \mathbf{V}_{100}$	weights=1	title	0.3880	0.6820	0.3074	0.5180
$Sent\mathbf{V} \cap \mathbf{V}_{100}$	weights of $\mathbf{V}_{100}$	title	0.3948	0.6920	0.3166	0.5320

Figure 1. Difference from median average precision per topic of the FIUlPL2 run. The 10 topics under the median are also listed.



1. Semi-manually built dictionaries can be successfully used in opinion finding retrieval. However, the fully automatic dictionary performs better than the semi-manual one (compare the submitted run *FIUdPL2* to the new run of Table 3 labeled by $\mathbf{V}_{100}$).
2. Different combinations of semi-manually built dictionaries with our fully automatic dictionary improves topical opinion retrieval.
3. Automatic weighting of opinionated terms by DFR improves topical opinion retrieval.
4. It is not the exhaustively of the semi-manual dictionary but its quality that improves performance. Using the intersection $Sent\mathbf{V}_{|Strong} \cap \mathbf{V}_{100}$ of the semi-manual dictionary containing all terms labeled as strong subjective with the automatic one, and weighting the opinionated terms as in the run *FIUlPL2*, we obtain the best combination strategy.
5. The parameter free model DPH provides the best precision at 10.

6 Conclusions

We have shown a very simple opinion retrieval model free from parameters to re-rank the set of retrieved documents. The opinion retrieval function requires the computation of an absolute or query-independent opinion rank of the retrieved documents. To obtain this sentimental score we have learned a dictionary of opinion-bearing words from blog track data of TREC 2006. The technique to extract and weight terms is that used in query expansion. We have then submitted such a sentimental dictionary as a query. However, dictionaries may contain many thousand of words, and we have thus further filtered the words obtaining smaller and smaller sentimental dictionaries. It is possible to achieve as good performance as we have achieved with the official runs with a very restricted number of opinion-bearing terms, but this issue is not fully presented here and will be studied in the next participation to the blog track.

References

1. *Alias-i. Lingpipe named entity tagger. In http://www.alias-i.com/lingpipe/*.
2. C. Biemann, G. Heyer, U. Quasthoff, and M. Richter. The Leipzig corpora collection - monolingual corpora of standard size. In *Proceedings of Corpus Linguistic 2007*, Birmingham, UK, 2007.
3. C. Macdonald and I. Ounis. The trec blogs06 collection : Creating and analyzing a blog test collection. Dcs technical report, Department of Computing Science of the University of Glasgow, 2006.
4. George A. Miller. Wordnet: a lexical database for English. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 483–483, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
5. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.
6. I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *In Proceedings of the Text REtrieval Conference (TREC 2006)*. National Institute of Standards and Technology, 2006.
7. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT-EMNLP*, 2005.
8. Theresa Wilson, Janyce Wiebe, and Rebecca Hwa. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99, 2006.
9. I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC2007 Blog Track In *In Proceedings of the Text REtrieval Conference (TREC 2007)*. National Institute of Standards and Technology, 2007.