Using very simple statistics for review search: An exploration

Bo Pang
Yahoo! Research
bopang@yahoo-inc.com

Lillian Lee
Computer Science Department, Cornell University
llee@cs.cornell.edu

Abstract

We report on work in progress on using very simple statistics in an unsupervised fashion to re-rank search engine results when review-oriented queries are issued; the goal is to bring opinionated or subjective results to the top of the results list. We find that our proposed technique performs comparably to methods that rely on sophisticated pre-encoded linguistic knowledge, and that both substantially improve the initial results produced by the Yahoo! search engine.

1 Introduction

One important information need shared by many people is to find out about opinions and perspectives on a particular topic (Mishne and de Rijke, 2006; Pang and Lee, 2008). In fact, locating relevant subjective texts was a core task in the 2006 and 2007 TREC Blog tracks (Ounis et al., 2006; Ounis et al., 2008). Most participants considered a two-phase re-ranking approach, where first topic-based relevancy search was employed, and then some sort of filtering for subjectivity was applied; these filters were based on trained classifiers or subjectivity lexicons.

We propose an alternative approach to review search, one that is unsupervised and that does not rely on pre-existing dictionaries. Rather, it in essence simply re-ranks the top k topic-based search results by placing those that have the *least idiosyncratic* term distributions, with respect to the statistics of the top k results, at the head of the list. The fact that it is the least, not the most, rare terms with respect to the search results that are most indicative of subjectivity may at first seem rather counterintuitive; indeed, previous work has found rare terms to be important subjectivity cues (Wiebe et al., 2004). However, reviews within a given set of search results may tend to resemble each other because they tend to all discuss salient attributes of the topic in question.

2 Algorithm

Define a *search set* as the top n webpages returned in response to a review- or opinion-oriented query by a high-quality initial search engine, in our case, the top 20 returned by Yahoo!. As a question of both pragmatic and scientific value, we consider how much information can be gleaned simply from the items in the search set itself; in particular, we ask whether the subjective texts in the search set can be ranked above the objective ones solely from examination of the patterns of term occurrences across the search-set documents.

The idea we pursue is based in part on the assumption that the initial search engine is of relatively high quality, so that many of the search-set documents probably are, in fact, subjective. Therefore, re-ordering the top-ranked documents by how much they resemble the other search-set documents in aggregate may be a good way to identify the reviews. Indeed, perhaps the reviews will be similar to one another because they all tend to discuss salient features of the topic in question.

Suppose we have defined a *search-set rarity function* $\text{Rarity}_{ss}(t)$ (see Section 2.1 below) that varies inversely with the number of documents in the search set that contain the term t. Then, we define the *idiosyncrasy* score of a document d as the average search-set rarity of the most common terms it contains:

$$I(d,k) = \frac{1}{k} \sum_{t \in k\text{-commonest-terms}(d)} \text{Rarity}_{ss}(t)\,, \tag{1}$$

where $k\text{-commonest-terms}(d)$ is the k commonest terms in the search set that also occur in d. For example, when we set k to be the size of the vocabulary of d, the idiosyncrasy score is the average search-set rarity of all the terms d contains. Then, to instantiate the similarity intuition outlined above, we simply rank by decreasing idiosyncrasy.

The reason we look at just the top most common terms is that the rarer terms might be noise. For example, terms that occur in just a few of the search-set documents might represent page- or site-specific information that is irrelevant to the query; but the presence of such terms does not necessarily indicate that the document in question is objective.

One potential problem with the approach outlined above is the presence of stopwords, since all documents, subjective or objective, can be expected to contain many of them. Therefore, stopword removal is indicated as an important preprocessing step. As it turns out, the commonly-used InQuery stopword list (Allan et al., 2000) contains terms like "less" that, while uninformative for topic-based information retrieval, may be important for subjectivity detection. Therefore, we used a 102-item list[1] based solely on frequencies in the British National Corpus.

2.1 Defining search-set rarity

There are various ways to define a search-set rarity function on terms. Inspired by the efficacy of the inverse document frequency (IDF) in information retrieval, we consider several definitions for $\text{Rarity}_{ss}(t)$. Let $n_{ss}(t)$ be the number of documents in the *search set* (not the entire corpus) that contain the term t. Due to space constraints, we only report results for:

$$\text{Rarity}_{ss}(t) \stackrel{def}{=} \frac{1}{n_{ss}(t)},$$

which is linearly increasing in $1/n_{ss}(t)$, (as befits a measure of "idiosyncrasy"). The other definitions we considered were logarithmic or polynomial in $1/n_{ss}(t)$, and performed similarly to the linear function.

2.2 Comparison algorithms

OpinionFinder is a state-of-the-art publicly available software package for sentiment analysis that can be applied to determining sentence-level subjectivity (Riloff and Wiebe, 2003; Wiebe and Riloff, 2005). It employs a number of pre-processing steps, including sentence splitting, part-of-speech tagging, stemming, and shallow parsing. Shallow parsing is needed to identify the extraction patterns that the sentence classifiers incorporate.

We used OpinionFinder's sentence-level output[2] to perform document-level subjectivity re-ranking as follows. The result of running OpinionFinder's sentence classifier is that each valid sentence[3] is annotated with one of three labels: "subj", "obj", or "unknown". First, discard the sentences labeled "unknown". Then, rank the documents by decreasing percentage of subjective sentences among those sentences that are left. In the case of ties, we use the ranking produced by the initial search engine.

We also considered a more lightweight way to incorporate linguistic knowledge: score each document according the percentage of adjectives within the set of tokens it contains. The motivation is previous work suggesting that the presence of adjectives is a strong indicator of the subjectivity of the enclosing sentence (Hatzivassiloglou and Wiebe, 2000; Wiebe et al., 2004).

[1] www.eecs.umich.edu/∼qstout/586/bncfreq.html

[2] There are actually two versions. We used the accuracy-optimized version, as it outperformed the precision-optimized version.

[3] OpinionFinder will only process documents in which all strings identified as sentences by the system contain fewer than 1000 words. For the 31 documents in our dataset that failed this criterion, we set their score to 0.

	p@1	p@2	p@3	p@4	p@5	p@10	p@S	MAP
Search-engine baseline	.536	.543	.541	.554	.554	.528	.538	.612
OpinionFinder (accuracy version)	.754	.717	.729	.725	**.733**	**.675**	**.690**	**.768**
% of adjectives (type-based)	.710	.703	.696	.681	.678	.625	.633	.715
$idiosyncrasy(linear)$, $k = 50$	**.797**	**.783**	.739	.717	.696	.613	.640	.729
$idiosyncrasy(linear)$, $k = 100$	.754	**.783**	**.768**	<u>.739</u>	<u>.716</u>	<u>.630</u>	<u>.665</u>	<u>.743</u>
$idiosyncrasy(linear)$, $k = 200$	<u>.768</u>	<u>.761</u>	.744	**.746**	<u>.716</u>	.623	.653	.731
$idiosyncrasy(linear)$, $k = 300$	.754	<u>.761</u>	<u>.749</u>	.736	.704	.614	.641	.724

Table 1: Average search-set subjective-document precision results. "S": number of subjective documents. Bold and underlining: best and second-best performance per column, respectively.

3 Evaluation

Our focus is on the quality of the documents placed at the very top ranks, since users often look only at the first page or first half of the first page of results (Joachims et al., 2005). Hence, we report the precision of the top 1-5 and 10 documents, as well as precision at the number of subjective documents and mean average precision (MAP) for the subjective documents. All performance numbers are averages over the 69 search sets in our data, described next.

Data Here, we sketch the data acquisition and labeling process. In order to get real user queries targeted at reviews, we began with a randomly selected set of queries containing the word "review" or "reviews"[4] from the the query log available at `http://www.sigkdd.org/kdd2005/kddcup/KDDCUPData.zip`. We created a search set for each query by taking the top 20 webpages returned by the Yahoo! search engine and applying some postprocessing. Over a dozen volunteer annotators then labeled the documents as to whether they were subjective or objective according to a set of detailed instructions. The end result was over 1300 hand-labeled documents distributed across 69 search sets, varying widely with respect to query topic. Our dataset download site is `http://www.cs.cornell.edu/home/llee/data/search-subj.html`.

For almost every annotator, at least two of his or her search sets were labeled by another person as well, so that we could measure pair-wise agreement with respect to multiple queries. On average, there was agreement on 88.2% of the documents per search set, with the average Kappa coefficient (κ) being an acceptable 0.73, reflecting in part the difficulty of the judgment.[5] The lowest κ occurs on a search set with a 75% agreement rate.

Results A natural and key baseline is the ranking provided by the Yahoo! search engine, which is a high-quality, industrial-strength system. We consider this to be a crucial point of comparison. The results are shown in the top line in Table 1.

OpinionFinder clearly outperforms the initial search engine by a substantial margin, indicating that there are ample textual cues that can help achieve better subjectivity re-ranking.

The adjective-percentage baseline is also far superior to that of the search-engine baseline at all ranks, but does not quite match OpinionFinder. (Note that to achieve these results, we first discarded all terms contained in three or fewer of the search-set documents, since including such terms decreased performance.) Still, it is interesting to see that it appears that a good proportion of the improvements provided by OpinionFinder can be achieved using just adjective counts alone.

We now turn to subjectivity re-ranking based on term-distribution (idiosyncrasy) information. For

[4] Subsequent manual filtering discarded some non-opinion-oriented queries, such as "alternative medicine review volume5 numer1 pages 28 38 2000".

[5] One source of disagreement that stems from the specifics of our design is that we instructed annotators to mark "sales pitch" documents as non-reviews, on the premise that although such texts are subjective, they are not valuable to a user searching for unbiased reviews. (Note that this policy presumably makes the dataset more challenging for automated algorithms.) There are several cases where only one annotator identified this type of bias, which is not surprising since the authors of sales pitches may actively try to fool readers into believing the text to be unbiased.

consistency with the adjective-based method just described, we first discarded all terms contained in three or fewer of the search-set documents.

As shown in Table 1, the idiosyncrasy-based algorithm posts results that are overall strongly superior to those of the initial, high-quality search engine algorithm and also generally better than the adjective-percentage algorithm. Note that these phenomena hold for a range of values of k. The overall performance is also on par with OpinionFinder; for instance, according to the paired t-test, the only statistically significant performance difference (.05 level) between the accuracy-emphasizing version of OpinionFinder and the idiosyncrasy-based algorithm for $k = 100$ is for precision at 10. In some sense, this is a striking result: just looking at within-search-set frequencies yields performance comparable to that of a method that utilizes rich linguistic knowledge and external resources regarding subjectivity indicators.

Another interesting observation is that term-distribution information seems to be more effective for achieving high precision at the very top ranks (precision at 1, 2, 3, and 4), whereas in contrast, relatively deep NLP seems to be more effective at achieving high precision at the "lower" top ranks, as demonstrated by the results for precision at 5, 10, and the number of subjective documents, and for MAP. These results suggest that a combination of the two methods could produce even greater improvements.

4 Concluding remarks

We considered the task of document-level subjectivity re-ranking of search sets, a task modeling a scenario in which a search engine is queried to find reviews. We found that our proposed term-distributional, idiosyncrasy-based algorithm yielded the best precision for the very top ranks, whereas the more linguistically-oriented, knowledge-rich approach exemplified by OpinionFinder gave the best results for precision at lower ranks. It therefore seems that both types of information can be very valuable for the subjectivity re-ranking task, since they have somewhat complementary performance behaviors and both outperform the initial search engine and an adjective-based approach.

Our motivation that within a search set, reviews tend to resemble one another rather than differ is reminiscent of intuitions underlying the use of *pseudo relevance feedback (PF)* in IR (Ruthven and Lalmas, 2003, Section 3.5). Future work includes comparison against PF methods and investigation of ways to select the value of k.

Acknowledgments We thank Eli Barzilay, Rich Caruana, Thorsten Joachims, Jon Kleinberg, Ravi Kumar, and the reviewers for their very useful help. We are also very grateful to our annotators, Mohit Bansal, Eric Breck, Yejin Choi, Matt Connelly, Tom Finley, Effi Georgala, Asif-ul Haque, Kersing Huang, Evie Kleinberg, Art Munson, Ben Pu, Ari Rabkin, Benyah Shaparenko, Ves Stoyanov, and Yisong Yue. This paper is based upon work supported in part by the NSF under grant no. IIS-0329064, a Yahoo! Research Alliance gift, Google Anita Borg Memorial Scholarship funds, a Cornell Provost's Award for Distinguished Research, and an Alfred P. Sloan Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of any sponsoring institutions, the U.S. government, or any other entity.

References

Allan, James, Margaret E. Connell, W. Bruce Croft, Fang-Fang Feng, David Fisher, and Xiaoyan Li. 2000. INQUERY and TREC-9. In *Proceedings of TREC*, pages 551–562. NIST Special Publication 500-249.

Hatzivassiloglou, Vasileios and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*.

Joachims, Thorsten, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of SIGIR*, pages 154–161.

Mishne, Gilad and Maarten de Rijke. 2006. A study of blog search. In *Proceedings of ECIR*.

Ounis, Iadh, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. 2006. Overview of the TREC-2006 Blog Track. In *Proceedings of TREC*.

Ounis, Iadh, Craig Macdonald, and Ian Soboroff. 2008. On the TREC Blog Track. In *Proceedings of ICWSM*.

Pang, Bo and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval series. Now publishers.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of EMNLP*.

Ruthven, Ian and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145.

Wiebe, Janyce M. and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing*, number 3406 in LNCS, pages 486–497.

Wiebe, Janyce M., Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308, September.