Characterizing the Subjectivity of Topics

Marc-Allen Cartright, Elif Aktolga, and Jeffrey Dalton
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
Amherst, Massachusetts
{irmarc, elif, jdalton}@cs.umass.edu

ABSTRACT

A document or web page in isolation may appear completely reasonable, but may represent a biased perspective on the topic being discussed. Given the topic of a document, we propose new metrics *provocativeness* and *balance* that suggest when the topic could be controversial. We explore the use of these metrics to characterize the subjectivity of the topics in the TREC Blog Track.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Selection Process

General Terms: Algorithms, Measurement, Theory

Keywords: Opinion Classification, Sentiment Analysis, Subjectivity/Polarity Detection

1. INTRODUCTION & PREVIOUS WORK

We begin in the context of a person browsing the web. Our goal is to inform the user about the web document they are reading. In particular, it would be useful to know if the document is about a highly polarized topic. For example, topics like "flag burning" and "NAFTA" have a high degree of subjective documents and the reader should proceed carefully.

Recently, there has been growing interest in opinion analysis and sentiment classification, especially in the context of blogs and social media [2]. The Text Retrieval Conference Blog track explored the retrieval of opinionated documents in the blogosphere, including the subtask of determining sentiment polarity [1]. Beyond polarity, there were attempts to analyze the subjective strength of individual words and phrases [3], but only little research for promoting comprehension of controversial topics discussed online.

In this poster we introduce two new metrics, *provocativeness* and *balance*, which we use to characterize the subjectivity of the topics in the TREC Blog Track. This contributes a new level of analysis to the opinion finding task.

2. METRICS

We propose the following metrics for tasks with arbitrary units of retrieval. In this work, a *unit* is a blog post, but it could be defined at other levels.

Definition 1. The **provocativeness** (PROV) of a topic measures the degree of subjectivity of the topic, which de-

scribes the quantity of subjective versus objective content on the topic. Topics with a high provocativeness should caution a reader to seek multiple perspectives on the topic.

We define the PROV of a topic to be the average subjectivity of all units relevant to the topic:

$$\mathbf{PROV} = \frac{\sum_{r \in R} \text{subjectivity}(r)}{|R|} \quad (1)$$

where R is the set of units relevant to the topic. The subjectivity of one unit is a real value in the range 0 to 1. A unit with a subjectivity of 0 is completely objective on the topic, while a unit with a score of 1 is completely subjective. By bounding this score, we limit its amount of influence towards the provocativeness of the topic. This is done to counteract the possibility that units may be of different size.

Definition 2. The **balance** (BAL) of a topic is the degree to which opinion on the topic differs. Let P and N be the set of positive and negative units on the topic. We define the balance as:

$$\mathbf{BAL} = \frac{|P| - |N|}{|P \cup N|} \quad (2)$$

BAL describes the amount of imbalance between the negative and positive opinions of a topic. Negative values indicate that the analyzed unit set contains more negative than positive content. Likewise for positive values. Note that by construction, the balance is bounded between -1.0 and 1.0. A value of 0 indicates that the topic is evenly balanced between units containing positive and negative sentiment. A reader that is aware of the calculated balance of a topic will be able to discern if a particular document is more or less likely to reflect the majority opinion on a topic, if one exists.

3. ANALYSIS ON BLOG TRACK

We use PROV and BAL to characterize the subjectivity of the topics used in the TREC Blog Track for the Opinion Finding task. We use the relevance judgments from the TREC 2008 Blog Track to calculate provocativeness and balance for topics 851 to 950 and 1001 to 1050. Each topic comes with a list of documents judged to be relevant to that topic, as well as whether the assessor deemed the document positive, negative, mixed (containing both positive and negative content), or neutral (no opinion). As defined by the data provided, a document acts as the unit of retrieval. A brief statistical summary of the topics is given in Table 1.

For this analysis, we define the subjectivity score of a document to be a binary value: 1 if the document expresses

Table 1: Basic judgment statistics per topic

Class:	rel.	opin.	pos.	neg.	mixed
min.	12	4	0	0	0
max.	950	826	392	533	455
avg.	292	182	70	56	57

an opinion on the topic (i.e. is non-neutral) and 0 if it does not. We note that this provides a very coarse measure of subjectivity and ignores the degree of opinionatedness.

For balance, we used the judgments from the polarity sub-task of the opinion retrieval task. P and N are the sets of documents that were judged to contain positive and negative content on the topic, respectively. Using this definition, mixed documents are included in both the positive and negative sets. The inclusion of the mixed documents in our calculation is not incongruous with our above definitions, as the mixed documents cancel themselves out in the numerator of the expression. In the limit where all of the opinionated documents are mixed, the balance is 0, indicating a balance of positive and negative content. This result exactly corresponds with what we would expect if all opinions were considered split between the two sides.

We show a scatter plot of these two metrics for the blog topics in Figure 1 to illustrate the score distributions. For purposes of analysis, we omitted 13 topics which had fewer than fifty relevant documents due to insufficient sample size.
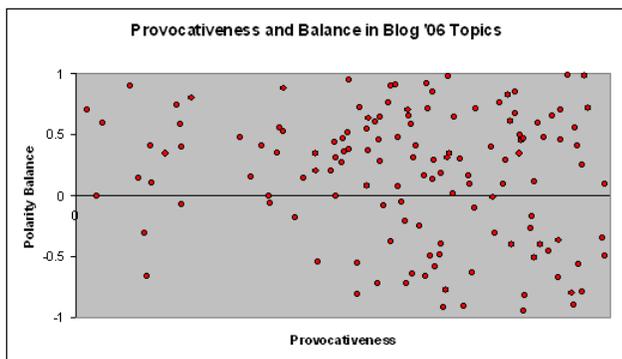

Figure 1: The provocativeness and balance for the 150 TREC blog opinion topics.

Figure 1 illustrates that the topics used in the blog track tend to be provocative. This is expected, as the topics were selected for the opinion retrieval task. We also observe that topics tend to express a higher degree of positive opinion than negative.

Table 2 shows the most provocative topics in the blog track according to our measure. The most subjective topics include the NAFTA treaty and the WTO, but also movie titles, and some people and brands. Finding a positive document on "Barry Bonds" in isolation would seem reasonable, but the data indicates that the document is in the minority and does not actually reflect popular opinion.

In contrast, Table 3 shows the topics with the lowest calculated proportion of provocativeness. While the types of entities may be the similar, posts on these topics tend to be news reports and do not express explicit opinions.

Table 2: Blog Opinion Topics with High PROV

Topic	Title	PROV	BAL
895	Oprah	0.990	0.059
1027	NAFTA	0.990	-0.454
887	World Trade Organization	0.985	-0.269
1032	I Walk the Line	0.958	0.682
925	mashup camp	0.952	0.983
867	cheney hunting	0.947	-0.768
869	muhammad cartoon	0.946	0.218
1043	A Million Little Pieces	0.941	-0.490
1049	YouTube	0.937	0.360
1040	TomTom	0.934	0.553
1008	UN Commission on Human Rights	0.932	-0.864
870	"barry bonds"	0.929	-0.726

Table 3: Blog Opinion Topics with Low PROV

Topic	Title	PROV	BAL
899	cholesterol	0.022	0.714
950	Hitachi Data Systems	0.040	0.000
898	Business Intelligence Resources	0.051	0.600
901	jstor	0.102	0.909
891	intel	0.116	0.125
897	ariel sharon	0.129	-0.238
868	"joint strike fighter"	0.134	-0.533
879	"hybrid car"	0.139	0.378
862	blackberry	0.143	0.080

4. CONCLUSIONS

In this work, we propose new metrics for characterizing subjective topics via their relevant content. We calculate provocativeness and balance using the Blog '06 corpus using the relevance judgments from the 2008 blog opinion finding task. We would like to apply these measures to several other corpora where we expect different characterizations, such as Wikipedia or a collection of newswire documents.

We also hope to evaluate on data that includes the degree of opinionatedness. Our future work includes measuring the strength of an opinion to distinguish between topics with trivial disagreement and those that are very emphatic. In addition, we would like to extend the presented measures to accommodate associations which are not bipolar (e.g. positions on religion or politics). We believe that these measures will readily evolve to provide a spectrum for automatic subjectivity analysis.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2008 blog track. In *Seventeenth Text REtrieval Conference (TREC 2008)*. NIST, February 2009.

[2] B. Pang and L. Lee. *Opinion Mining and Sentiment Analysis*. Now Publishers Inc, July 2008.

[3] T. Wilson. Just how mad are you? finding strong and weak opinion clauses. In *Proceedings of AAAI*, pages 761–769, 2004.