Searching Sentiments in Blogs

Hongcheng Mi and I-Heng Mei
Opinmind, Inc
3370 Victor Court
Santa Clara, CA 95054
charles.mi@opinmind.com, heng.mei@opinmind.com

Abstract

Sentiment mining is a computational approach used to identify expressions made about topics within a span of text. The blogosphere is a particularly useful corpus for sentiment mining because bloggers express a wide variety of opinions and sentiments in their online journals. Previous works on sentiment identification and extraction have been primarily focused on using machine-learning methods to extract sentiment patterns. Annotating text corpuses, however, is a time-consuming process. In this paper, we present a streamlined approach to extract sentiments from untagged text. We use heuristic models to quickly identify sentiment expressions and target subjects. This is an enabling approach to the rapid identification and extraction of expressions about topics.

1 Introduction

There are an estimated fifty millions blogs today. Furthermore, it is estimated that thirty to forty thousand new blogs are created each day (Sifry, 2005). Blogs are online journals that allow people to share and publish their thoughts and daily experiences on the web. A typical blog consists of a simple profile about the author and a series of entries in which the author writes about any topic (s)he chooses. Similar to traditional written diaries, blogs tend to contain a wide array of expressions about topics (i.e. opinions and sentiments). These expressions are often about products, places and people, though there is no limitation to the topic of the expression. Extracting opinions from blogs can be particularly useful for individuals seeking to understand consumer sentiment about a particular topic of interest. Because the opinions expressed in blogs are unsolicited and voluntary, they also tend to be more genuine and unbiased than opinions solicited by surveys or focus groups.

The size and exponential growth rate of the Blogosphere require that sentiment extraction occur at a sufficiently high speed to ensure that such sentiments can be extracted soon after blog entries are published. Unlike traditional keyword indexing, however, sentiment mining involves a sufficiently complex semantic and syntactic analysis of the text. The process of sentiment mining includes identifying whether the span of text contains sentiments, associating expressions to the correct topic, and determining the strength of such expressions. Once expression-topic sets have been identified, they are stored and indexed so they may be readily accessed in response to user search queries.

In this paper, we present two stages of our sentiment mining system and the architecture of our sentiment index. In Section 2, we review some related work. In sections 3 and 4, we discuss our sentiment mining approach and sentiment index architecture. Section 5 concludes with a discussion of ongoing work and future challenges.

The infrastructure described in this paper is currently implemented on production servers at www.opinmind.com and is serving more than twenty-five million opinions from more than three million bloggers.

2 Related Work

Hatzivassiloglou and McKeown (1997) first produced a list of seed words to determine whether a sentence expresses positive or negative sentiment. Turney (2002) suggested an approach to extend this list by using value phrases composed of six syntactic patterns. Similar to these two approaches, Yi and Niblack (2005) also used word clues to determine the polarity of the sentence. Riloff and Wiebe (2003) proposed a method that bootstraps more subjectivity patterns using a Naïve Bayes classifier to distinguish between subjective and objective sentences. Pang and Lee (2004) used Support Vector Machines to classify subjective sentences by incorporating the degree of proximity between sentences as a feature. Traditional machine learning approaches often suffer from the time consuming manual annotation of corpus training and bias towards various domains. The approach described within this paper also operates at the sentence level. Unlike (Riloff and Wiebe, 2002) and (Pang and Lee, 2004), however, we use various syntactic patterns as indicators for sentiment features thereby eliminating the need for a corpus of training documents.

3 Sentiment Mining

The approach described here consists of two unique stages – (1) the parsing of raw text and (2) the extraction of sentiment features. In the first stage, raw text is broken into sentences by heuristically identifying sentence punctuation points. Some punctuation instances are ignored (i.e. periods used in abbreviations). Within each sentence, each word is tagged with its associated part of speech through the use of an electronic lexicon. For words that have more than one POS, the tagger disambiguates based on the POS of the surrounding words (Riloff and

Philips, 2004). Once tagged, the sentence is then broken into clauses using a heuristic approach based on word groupings and various clause indicators. The presence of relative pronouns, for example, indicates the location of independent and dependent clauses. Each clause has a subject and verb, and optionally, other constructs such as prepositional phrases and objects. The parsing process returns a sentence tree composed of clauses and words tagged with roles. Roles roughly correspond to grammatical constructs such as subject and direct object.

In the second stage, the sentence tree is analyzed for polarity features. We reference an extensive lexicon of polarity terms similar to those produced by (Hatzivassiloglou and McKeown, 1997). We have enhanced the lexicon by adding *strength* and *behavior* factors to each polarity term. Strength factors are used to differentiate degrees of expressed sentiments. “Love”, for example, has a stronger positive connotation than the word “prefer”. Behavior factors are used to indicate the contextual use of each polarity term so that the associated topics can be correctly identified. For example, the word “loves” is a positive polarity term used as verb. As such, the associated topic is likely to appear after the term “loves” as the object of the sentence (as opposed to the subject of the sentence). For each sentence tree, polarity terms and their associated topics are identified. The polarity term is referred to as the *sentiment feature* and the related topic is known as the *target feature*.

4 Sentiment Indexing

Sentiment and target features are electronically stored in a sentiment index. A sentiment index stores more information about a search term than traditional keyword indexes. Additional information including the polarity of the sentiment, the original sentence, the blog URL and date/time of the entry are stored along with the sentiment and target features in the main index. The main index is segmented and replicated across multiple physical systems to optimize performance and provide redundancy. The run-time system is a distributed network of low-cost, commodity computers.

5 Ongoing Work

Precision and recall are inherent tradeoffs when working with any text mining system. In our current production system, we have optimized precision at the expense of recall. Currently, our research focus is on improving our accuracy and recall metrics and to allow for the improvement of recall with little effect on precision.

The detection of sarcasm is also a focus on ongoing research. Sarcastic statements are often mis-categorized as it is difficult to identify a consistent set of features to identify sarcasm. One proposed method is to develop a more holistic approach which uses previous sentiments expressed by the blogger to determine the probability of a sarcastic sentiment.

Finally, disambiguation is also a focus of ongoing work. For example, “apple” could refer to a piece of fruit or it could refer to Apple, Inc. This is less of a problem for traditional sentiment mining systems where the topics are predefined, but in a general system for searching opinions on any topic, these disambiguation problems become more evident.

6 Summary

Work at Opinmind to date has focused on optimizing the processing speed of sentiment extraction and building a scalable and efficient sentiment index. We aim to continually improve our accuracy and recall in order to fulfill our goal of collecting and presenting the opinions of bloggers all over the world.

References

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the ACL, 2004*.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105-112.

Ellen Rilloff and W. Philips. 2004. An Introduction to the Sundance and AutoSlog Systems. In *University of Utah School of Computing Technical Report #UUCS-04-015.*

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinions questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 129-136.

David Sifry. 2005. State of The Blogosphere, March 2005, Part I: Growth of Blogs. In *Sifry’s Alerts.*
http://www.sifry.com/alerts/archives/000298.html

Peter Turney. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Pennsylvania.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL,* pages

174-181, Madrid, Spain, July. Association for Computational Linguistics.

Jeonghee Yi, Wayne Niblack. 2005. "Sentiment Mining in WebFountain," *icde*, pp. 1073-1083, 21st International Conference on Data Engineering (ICDE'05), 2005.