Twititude: Message Clustering and Opinion Mining on Twitter

Keling Chen and Huasha Zhao

March 12, 2012

1 Introduction

The growing availability of public opinion on the web makes a fine-grained analysis of "what people think" possible. Twitter (twitter.com), as a particularly popular micro-blog service that enables users' to express their opinion in a real time manner, has attracted a great body of research works recently [6, 7].

Twitter-based content analysis tasks can be divided into four catergories: sentiment analysis, summarization, event extraction and understanding twitter as a social network graph [twitternlp]. The first two are most related to our work. Previous research has exploited the use of both unsupervised and supervised methods for topic categorization and tweets sentiment classification. For example, Frederking et al [7] proposed a clustering technique to classify tweets into topics such as News, Sports and etc. Pak and Paroubek [6] developed an Naive Bayes algorithm with discretionary feature selection to analyze the emotion embedded in each individual tweet. Though topic classification and sentiment polarization analysis provide some useful information on the public behavior and mood, they fail to answer questions like "why people are happy?" and "which aspect people like the subject of interest?".

We propose Twititude to address the above problem: Instead of offering general sentiment and content analysis, our system will provide a fine-grained query based opinion analysis, which will automatically extracts different aspects people are talking about a subject, and summarizes people's attitude and comments towards each aspect. A potential application would be summarizing product feedbacks from Twitter. For example, Apple may want to know users' experience on its iPhone product: whether the screen is bright enough or whether it cost too much. It might seem easy to cluster topics for a specific product like iPhone, however, our system should be able to cluster themes for arbitrary queries.

One work that is most close to our system is TweetMotif proposed by O'Connor et al [5]. They applied word-frequency based heuristics to cluster topics, but they did neither report any metrics on system performance nor compare their approach with existing clustering techniques. Further more, feature selection and model tuning has not been discussed in their paper either.

2 Contribution

Our work is related to traditional topic clustering and multi-document summarization. Document clustering has been a well developed research area, and algorithms including Latent Semantic Analysis (LSA), Non-negative Matrix Factorization (NMF), Latent Dirichlet Allocation (LDA) and Gamma-Poisson (GaP) [4, 2, 1] are considered to be adequate for clustering tasks. However, the nature of Twitter documents set two new challenges to our task: first, words used in tweets are often ill-formed which introduces additional noisy to the documents; second, the length of the message is relatively short, and this may reduce the robustness of existing clustering algorithms.

The contribution our Twititude system has three folds. Firstly, we will try to tackle with noisy input by first translating out-of-vocabulary words into normal words. Furthermore, a thorough comparison of applying different clustering models to short text data will be studied. Finally, efforts will also be spent on model tuning and parameter selection for unsupervised topic learning models.

3 Method

3.1 Tweets Normalization

The noisy channel model [8] is the most widely accepted framework for text normalization tasks. Given the ill-formed word T, the standard form of this word S is found by

$$S = \arg\max_t Pr(S|T) = \arg\max_t Pr(T|S)Pr(S) \tag{1}$$

where $Pr(S|T)$ model the noisy channel. In practice, the channel model can be either trained with sophisticated statistical learn method or determined by linguist or speech experts. Recent years sees a surge in research effort on noisy channel modeling to tackle with ill-formed text in both SMS messages and social network sites. A variety of noisy channel has been proposed [8, 3], including Grapheme Channel, Phoneme Channel, Acronym Channel and etc.

Coming up with a state-of-the-art microtext normalization system is out of the scope of this paper, however, we would like to propose a simple and light-weighted tweets normalization technique to facilitate our clustering and summarization task.

3.2 Feature Extraction

Bag-of-words, part-of-speech, punctuations and emoticons are all potentially considered as features of our system. However, function word elimination and word stemming may also help boost clustering performance.

3.3 Clustering and Summarization

We carefully study the similarity and difference between models such as NMF, LDA and GaP, and compare their performance on short text clustering.

3.4 Performance Evaluation

Two major performance measure for clustering are topic coverage and distinctiveness; mathematically, they can be characterized by quantities such as entropy and mutual information. Summarization performance can be measured by the amount of information represented in summaries. If time permits, it is also interesting to collect subjective evaluation from peers using our system for performance comparison.

References

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] J. Canny. Gap: a factor model for discrete data. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 122–129. ACM, 2004.

[3] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics*, 2011.

[4] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.

[5] B. OConnor, M. Krieger, and D. Ahn. Tweetmotif: Exploratory search and topic summarization for twitter. *Proceedings of ICWSM*, pages 2–3, 2010.

[6] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010, 2010.

[7] K.D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical clustering of tweets. 2011.

[8] Z. Xue, D. Yin, and B.D. Davison. Normalizing microtext. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.