Automatic Generation of Personalized Annotation Tags for Twitter Users

Wei Wu, Bin Zhang, Mari Ostendorf
Electrical Engineering
University of Washington, Seattle, WA
{weiwu, binz, ostendor}@uw.edu

Abstract

This paper introduces a system designed for automatically generating personalized annotation tags to label Twitter user's interests and concerns. We applied TFIDF ranking and TextRank to extract keywords from Twitter messages to tag the user. The user tagging precision we obtained is comparable to the precision of keyword extraction from web pages for content-targeted advertising.

1 Introduction

Twitter is a communication platform which combines SMS, instant messages and social networks. It enables users to share information with their friends or the public by updating their Twitter messages. A large majority of the Twitter users are individual subscribers, who use Twitter to share information on "what am I doing" or "what's happening right now". Most of them update their Twitter messages very frequently, in which case the Twitter messages compose a detailed log of the user's everyday life. These Twitter messages contain rich information about an individual user, including what s/he is interested in and concerned about. Identifying an individual user's interests and concerns can help potential commercial applications. For instance, this information can be employed to produce "following" suggestions, either a person who shares similar interests (for expanding their social network) or a company providing products or services the user is interested in (for personalized advertisement).

In this work, we focus on automatically generating personalized annotation tags to label Twitter user's interests and concerns. We formulate this problem as a keyword extraction task, by selecting words from each individual user's Twitter messages as his/her tags. Due to the lack of human generated annotations, we employ an unsupervised strategy. Specifically, we apply TFIDF ranking and TextRank (Mihalcea and Tarau, 2004) keyword extraction on Twitter messages after a series of text preprocessing steps. Experiments on randomly selected users showed good results with TextRank, but high variability among users.

2 Related Work

Research work related to Twitter message analysis includes a user sentiment study (Jansen et al., 2009) and information retrieval indexing. To our knowledge, no previously published research has yet addressed problems on tagging user's personal interests from Twitter messages via keyword extraction, though several studies have looked at keyword extraction using other genres.

For supervised keyword extraction, (Turney, 2000; Turney, 2003; Hulth, 2003; Yih et al., 2006; Liu et al., 2008) employed TFIDF or its variants with Part-of-Speech (POS), capitalization, phrase and sentence length, etc., as features to train keyword extraction models, and discriminative training is usually adopted. Yih et al. (2006) use logistic regression to extract keywords from web pages for content-targeted advertising, which has the most similar application to our work. However, due to the lack of human annotation on Twitter messages, we have to adopt an unsupervised strategy.

For unsupervised keyword extraction, TFIDF ranking is a popular method, and its effectiveness has been shown in (Hulth, 2003; Yih et al., 2006). TextRank and its variants (Mihalcea and Tarau, 2004; Wan et al., 2007; Liu et al., 2009) are graph-based text ranking models, which are derived from Google's PageRank algorithm (Brin and Page, 1998). It outperforms TFIDF ranking on traditional keyword extraction tasks. However, previous work on both TFIDF ranking and TextRank has been done mainly on academic papers, spoken documents or

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 689–692,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

web pages, whose language style is more formal (or, less "conversational") than that of Twitter messages. Twitter messages contain large amounts of "noise" like emoticons, internet slang words, abbreviations, and misspelled words. In addition, Twitter messages are a casual log of a user's everyday life, which often lacks of a coherent topic sequence compared to academic papers and most spoken documents. Hence, it remains to see whether TFIDF ranking and TextRank are effective for identifying user's interests from Twitter messages.

3 System Architecture

Figure 1 shows the framework of our system for tagging Twitter user's interests. A preprocessing pipeline is designed to deal with various types of "noise" in Twitter messages and produce candidate words for user tags. Then the TFIDF ranking or TextRank algorithm is applied to select user tags from the candidate words.

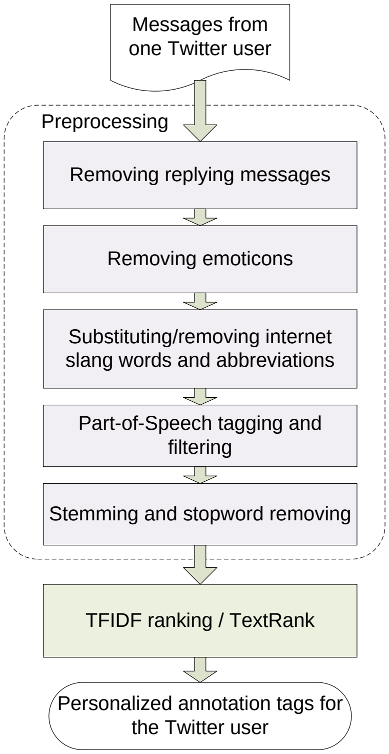
Messages from one Twitter user
Preprocessing
Removing replying messages
Removing emoticons
Substituting/removing internet slang words and abbreviations
Part-of-Speech tagging and filtering
Stemming and stopword removing
TFIDF ranking / TextRank
Personalized annotation tags for the Twitter user

Figure 1: Framework of the personalized annotation tag generation system for Twitter users

3.1 Preprocessing

In addition to messages describing "What am I doing" or "what's happening right now", Twitter users also write replying messages to comment on other users' messages. This kind of message generally contains more information about the users they reply to than about themselves, and therefore they are removed in the preprocessing pipeline.

Emoticons frequently appear in Twitter messages. Although some of them help express user's sentiment on certain topics, they are not directly helpful for keyword analysis and may interfere with POS tagging in the preprocessing pipeline. Therefore, we designed a set of regular expressions to detect and remove them.

Internet slang words and abbreviations are widely used in Twitter messages. Most of them are out-of-vocabulary words in the POS tagging model used in the next step, and thus will deteriorate the POS tagging accuracy. Hence, we build a lookup table based on the list of abbreviations in the NoSlang online dictionary,[1] which we divide by hand into three sets for different processing. The first set includes 422 content words and phrases, such as "bff" (best friend forever) and "fone" (phone), with valid candidate words for user tags. The second set includes 67 abbreviations of function words that usually form grammatical parts in a sentence, such as "im" (i'm), "abt" (about). Simply removing them will affect the POS tagging. Thus, the abbreviations in both these sets are replaced with the corresponding complete words or phrases. The third set includes 4576 phrase abbreviations that are usually separable parts of a sentence that do not directly indicate discussion topics, such as "lol" (laugh out loud), "clm" (cool like me), which are removed in this step.

We apply the Stanford POS tagger (Toutanova and Manning, 2000) on Twitter messages, and only select nouns and adjectives as valid candidates for user tags. At the end of the preprocessing pipeline, the candidate words are processed with the rule-based Porter stemmer[2] and stopwords are filtered using a publicly available list.[3]

[1] www.noslang.com/dictionary

[2] tartarus.org/ martin/PorterStemmer/

[3] armandbrahaj.blog.al/2009/04/14/list-of-english-stop-words/

3.2 User Tag Extraction

3.2.1 TFIDF ranking

In the TFIDF ranking algorithm, messages from user u are put together as one document. The TFIDF value of word i from this user's messages is computed as

$$tfidf_{i,u} = \frac{n_{i,u}}{\sum_j n_{j,u}} \log(\frac{U}{U_i})$$

where $n_{i,u}$ is the count of word i in user u's messages, U_i is the number of users whose messages contain word i, and U is the total number of users in the Twitter corpus. For each user, words with top *N* TFIDF values are selected as his/her tags.

3.2.2 TextRank

According to the TextRank algorithm (Mihalcea and Tarau, 2004), each candidate word is represented by a vertex in the graph; edges are added between two candidate words according to their co-occurrence. In the context of user tag extraction, we build a TextRank graph with undirected edges for each Twitter user. One edge is added between two candidate words if they co-exist within at least one message; the edge weight is set to be the total count of within-message co-occurrence of the two words throughout all messages of this user.

Starting with an arbitrarily assigned value (e.g. 1.0), the rank value $R(V_i)$ of the candidate word at vertex V_i is updated iteratively according to the following equation,

$$R(V_i) = (1-d) + d \sum_{V_j \in E(V_i)} \frac{w_{ji}}{\sum_{V_k \in E(V_j)} w_{jk}} R(V_j)$$

where w_{ji} is the weight of the edge that links V_j and V_i, $E(V_i)$ is the set of vertices which V_i is connected to, and d is a damping factor that is usually set to 0.85 (Brin and Page, 1998). The rank update iteration continues until convergence. The candidate words are then sorted according to their rank values. Words with top-*N* rank values are selected as tags for this user.

4 Experiment

4.1 Experimental Setup

We employed the Twitter API to download Twitter messages. A unigram English language model was used to filter out non-English users. We obtained messages from 11,376 Twitter users, each of them had 180 to 200 messages. The word IDF for TFIDF ranking was computed over these users.

We adopted an evaluation measure similar to the one proposed in (Yih et al., 2006) for identifying advertising keywords on web pages, which emphasizes precision. We randomly selected 156 Twitter users to evaluate the top-*N* precision of TFIDF ranking and TextRank. After we obtained the top-*N* outputs from the system, three human evaluators were asked to judge whether the output tags from the two systems (unidentified) reflected the corresponding Twitter user's interests or concerns according to the full set of his/her messages.[4] We adopted a conservative standard in the evaluation: when a person's name is extracted as a user tag, which is frequent among Twitter users, we judge it as a correct tag only when it is a name of a famous person (pop star, football player, etc). The percentage of the correct tags among the top-*N* selected tags corresponds to the top-*N* precision of the system.

Precision (%)	TFIDF	TextRank
top-1	59.6	67.3
top-3	61.5	66.0
top-5	61.2	63.0
top-10	59.0	58.3

Table 1: Tagging precision on all users in the test set

4.2 Experimental Results

Table 1 gives the top-*N* precision for TFIDF and TextRank for different values of *N*, showing that TextRank leads to higher precision for small *N*. Although Twitter messages are much "noisier" than regular web pages, the top-*N* precision we obtained for Twitter user tagging is comparable to the web page advertising keyword extraction result reported in (Yih et al., 2006).

Figure 2 shows an example of the candidate word ranking result of a Twitter user by TextRank (the font size is set to be proportional to each word's TextRank value). By examining the Twitter messages, we found that this user is an information tech-

[4] The pairwise Kappa value for inter-evaluator agreement ranged from 0.77-0.83, showing good agreement.

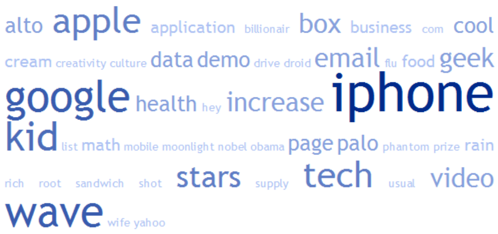

Figure 2: Example of a Twitter user's word ranks (the font size is proportional to each word's TextRank value)

Precision (%)				
top-N	σ >0.6	σ ≤0.6	H>5.4	H≤5.4
top-1	71.6	60.7	78.5	50.8
top-3	71.9	56.8	74.2	54.0
top-5	69.3	53.1	69.2	53.7
top-10	65.1	47.7	63.8	50.2

Table 2: TextRank tagging precision on users with different Top-10 TextRank value standard deviation (σ) and user message text entropy (H).

nology “geek”, who is very interested in writing *Apple's iPhone applications*, and also a user of *Google Wave*. In this work, we use only isolated words as user tags, however, “google”, “wave”, and “palo”, “alto” extracted in this example indicate that phrase level tagging can bring us more information about the user, which is typical of many users.

Although most Twitter users express their interests to some extent in their messages, there are some users whose message content is not rich enough to extract reliable information. We investigated two measures for identifying such users: standard deviation of the top-10 TextRank values and the user's message text entropy. Table 2 shows a comparison of tagging precision where the users are divided into two groups with a threshold on each of the two measures. It is shown that users with larger TextRank value standard deviation or message text entropy tend to have higher tagging precision, and the message text entropy has better correlation with the top-10 tagging precision than TextRank value standard deviation (0.33 v.s. 0.20 absolute).

5 Summary

In this paper, we designed a system to automatically extract keywords from Twitter messages to tag user interests and concerns. We evaluated two tagging algorithms, finding that TextRank outperformed TFIDF ranking, but both gave a tagging precision that was comparable to that reported for web page advertizing keyword extraction. We noticed substantial variation in performance across users, with low entropy indicative of users with fewer keywords, and a need for extracting key phrases (in addition to words). Other follow-on work might consider temporal characteristics of messages in terms of the amount of data needed for reliable tags vs. their time-varying nature, as well as sentiment associated with the identified tags.

References

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

A. Hulth. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proc. EMNLP*, pages 216–223.

B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

F. Liu, F. Liu, and Y. Liu. 2008. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *Proc. IEEE SLT*, pages 181–184.

F. Liu, D. Pennell, F. Liu, and Y. Liu. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proc. HLT/NAACL*, pages 620–628.

R. Mihalcea and P. Tarau. 2004. Textrank: Bringing order into texts. In *Proc. EMNLP*.

K. Toutanova and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proc. EMNLP*, pages 63–77.

P. D. Turney. 2000. Learning algorithms for keyphrase. *Information Retrieval*, 2(4):303–336.

P. D. Turney. 2003. Coherent keyphrase extraction via web mining. In *Proc. IJCAI*, pages 434–439.

X. Wan, J. Yang, and J. Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proc. ACL*, pages 552–559.

W.-T. Yih, J. Goodman, and V. R. Carvalho. 2006. Finding advertising keywords on web pages. In *Proc. 15th International Conference on World Wide Web*, pages 213–222.