Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Recommending Related Microblogs: A Comparison Between Topic and WordNet Based Approaches

Xing Chen, Lin Li, Huifan Xiao
School of Computer Science & Technology
Wuhan University of Technology, China
{xingchen, cathylilin, huifan_xiao}@whut.edu.cn

Guandong Xu
Centre for Applied Informatics
Victoria University, Australia
guandong.xu@vu.edu.au

Zhenglu Yang, Masaru Kitsuregawa
Institute of Industrial Science
The University of Tokyo, Japan
{yangzl, kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract

Computing similarity between short microblogs is an important step in microblog recommendation. In this paper, we investigate a topic based approach and a WordNet based approach to estimate similarity scores between microblogs and recommend top related ones to users. Empirical study is conducted to compare their recommendation effectiveness using two evaluation measures. The results show that the WordNet based approach has relatively higher precision than that of the topic based approach using 548 tweets as dataset. In addition, the Kendall tau distance between two lists recommended by WordNet and topic approaches is calculated. Its average of all the 548 pair lists tells us the two approaches have the relative high disaccord in the ranking of related tweets.

Introduction

Similarity computation between texts, such as various kinds of documents and queries is a long-history research direction in Information Retrieval. Recently, microblogs as a new social media have attracted researchers' interests (Kwak et al. 2010). One of microblog recommendation methods is to suggest micro-blogs related to what a user has issued or trending topics. Therefore, computing similarity between microblogs is an essential step in making recommendations.

Traditional term-based similarity computing measure perform poorly on such tasks because of data sparseness, the lack of context, and low term occurrence in both two microblogs. If two texts do not have any terms in common, then they receive a very low similarity score, regardless of how topically related they actually are. This is well-known as the vocabulary mismatch problem. This problem is only exacerbated if we attempt to use traditional measures to compute the similarity of two short segments of text (Metzler, Dumais, and Meek 2007). According to conventional measures, the more overlaps of words two texts have, the higher similarity score they will receive, however it sometimes is unreasonable and inaccurate. For example, apple pie and apple cellar phone share one word apple yet have low semantic relation.

To overcome the above difficulty, in this paper, we investigate a topic based approach and a WordNet based approach to estimate similarity scores between microblogs and recommend top related ones to users. The WordNet based approach utilizes dictionary-based algorithms to capture the semantic similarity between two texts based on the WordNet taxonomy dictionary. Latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003), a topic model for text or other discrete data, allows us to analyze of corpus, and extract the topics that combined to form its documents. Empirical study is conducted to compare their recommendation effectiveness in terms of precision measure.

Copyright © 2012, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

WordNet based Approach

WordNet is a lexical database which is available online and provides a large repository of English lexical items. The whole dictionary can be treated as a large graph with each node being a synset and the edges representing the semantic relations. Here, we treat each microblog as a sentence. Similarity computation between microblogs is converted to sentence similarity calculation. Steps for computing semantic similarity between two sentences are as follows:

1. First each sentence is partitioned into a list of words and we remove the stop words. Stop words are frequently occurring, insignificant words that appear in a database record, article or a web page, etc. As a result, each sentence is turned to be a list of tokens.
2. Second the task is part-of-speech disambiguation (or tagging) to identify the correct part of speech (POS - like noun, verb, pronoun, adverb...) of each word in the sentence.
3. Stemming words. We use the Porter stemming algorithm. Porter stemming is a process of removing the common morphological and inflexional endings of words.
4. Third we find the most appropriate sense for every word in a sentence (Word Sense Disambiguation)
5. Finally, the similarity between the sentences is computed based on the similarity of the pairs of words. And we capture semantic similarity between two word senses based on the path length similarity, in which we treat taxonomy as an undirected graph and measure the distance between them in WordNet.

Topic based Approach

We chose LDA to do topic analysis on microblog texts. The intuition behind LDA is to discover this latent topic structure

via estimating the probability distribution of the original co-occurrence activities. Here, we represent each microblog as a topic vector. Similarity computation between microblogs is equivalent to the dot product of two normalized topic vectors. The following is steps we take.

1. We need to pre-process the original data into required data format. During the process of analyzing our dataset(548 tweets), the term index and term-document matrix would be created, which provide great convenience for the transformation of all the original 548 tweets into the data format that LDA-c [1] implementation required.
2. We employ the variational EM algorithm (Blei, Ng, and Jordan 2003) to find the variational parameters that maximize the total likelihood of the corpus with respect to model parameters. Meanwhile, the calculated estimation parameters can be used to infer topic distribution of a new document by performing the variational inference.
3. After topic inference, each tweet can be represented by a topic vector. The dot product of two normalized topic vectors is the similarity score of corresponding tweets.

Experiments

Dataset and Evaluation Methodology

We demonstrate the working of the two approaches on the dataset extracted from (Han and Baldwin 2011). It contains 548 English messages sampled from Twitter API (from August to October, 2010) and contains 1184 normalized tokens. All ill-formed words had been detected, and recommended candidates are generated based on morphophonemic similarity. Both word similarity and context are then exploited to select the most probably correct candidate for the word.

Popular precision and Kendall tau distance are used as evaluation metrics. Given the recommended list for each tweet in dataset, we manually judge how many of these relatively high related tweets are really related. The Kendall tau distance is a metric of comparing the disagreement between two lists by counting the number of pairwise disagreements between two lists. The larger the distance, the more dissimilar the two lists are.

The Kendall tau distance between two lists τ_1 and τ_2 is:

$$K(\tau_1, \tau_2) = \sum_{i,j \in P} \bar{K}_{i,j}(\tau_1, \tau_2), \quad (1)$$

where P is the set of unordered pairs of distinct elements in τ_1 and τ_2. $\bar{K}_{i,j}(\tau_1, \tau_2)$ will be equal to 0 if the two lists are identical and 1 if one list is the reverse of the other. Often Kendall tau distance is normalized by dividing by $n(n-1)/2$ so a value of 1 indicates a maximum disagreement. Here it is used to measure the agreement of the two recommendation lists produced by our WordNet and topic based approaches.

Experimental Results

By varying the number of latent topics (K) in LDA, our results say that the highest precision score is obtained when K=15 as shown in Table 1.

Table 1: Precision scores at different number of topics(K)

	K=10	**K=15**	**K=20**
Top5	0.8246	0.8246	0.8231
Top10	0.8261	0.8261	0.8231

Table 2: Precision scores of the two approaches at top 5 and 10 recommendations

	Topic-based	**WordNet-based**
Top5	0.8246	0.8306
Top10	0.8261	0.8366

The comparison between the two approaches is given in Table 2. We can see the results from Table 2 that the WordNet-based approach is higher than the topic based one.

The average Kendall tau distance between 548 WordNet lists and corresponding LDA lists is 0.570537528674173, which indicates the relative high disaccord in the ranking of lists by two approaches. In other words, the recommendation made by the WordNet based approach is different from that made by the topic based approach. We reckon the rational for this observation is that these two approaches tackle the recommendation from different aspects. The essence of topic based method lies on the assumption that there exists an unseen structure of "topics" or "themes" in the text corpus, which governs the co-occurrence observations, while WordNet-based method is more concerned with the semantics of words.

Conclusions

From our experiment analysis, the WordNet-based method outperforms the topic based method in finding related short microblogs. We think that topic models are applicable for long and rich training texts, but not effective for short and sparse text. WordNet shows stable and acceptable performance in both long and short texts. Combining the results of WorNet and topic approaches might be an interesting topic for our future work. More about our research can be found on the Web [2].

Acknowledgments

This research was undertaken as part of Project 61003130 funded by National Natural Science Foundation of China.

References

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

Han, B., and Baldwin, T. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *ACL*, 368–378.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. B. 2010. What is twitter, a social network or a news media? In *WWW*, 591–600.

Metzler, D.; Dumais, S. T.; and Meek, C. 2007. Similarity measures for short segments of text. In *ECIR*, 16–27.

[1] http://www.cs.princeton.edu/ blei/lda-c/

[2] http://e-research.csm.vu.edu.au/files/xu/