MicroScholar: Mining Scholarly Information from Chinese Microblogs

Yang Yu and Xiaojun Wan
Institute of Computer Science and Technology, Peking University, Beijing 100871, China
The MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China
{yu.yang, wanxiaojun}@pku.edu.cn

Abstract

For many researchers, one of the biggest issues is the lack of an efficient method to obtain latest academic progresses in related research fields. We notice that many researchers tend to share their research progresses or recommend scholarly information they have known on their microblogs. In order to exploit microblogging to benefit scientific research, we build a system called MicroScholar to automatically collecting and mining scholarly information from Chinese microblogs. In this paper, we briefly introduce the system framework and focus on the component of scholarly microblog categorization. Several kinds of features have been used in the component and experimental results demonstrate their usefulness.

Introduction

In recent years, microblogging has become a popular method for many researchers to publish their research progresses and thoughts or recommend new scholarly information they have known. For example, many prestigious researchers and professors post scholarly messages on Sina Weibo (the most popular Chinese microblogging system) in their daily life. These scholarly messages may introduce or recommend latest research papers, upcoming academic conferences and events, useful research tools and datasets, etc. We feel that it would be very meaningful for helping researchers to automatically obtain such scholarly information from microblogs because scholarly information are scattered across the vast amount of microblogs. We aim to build a novel system called MicroScholar to automatically collecting and mining scholarly information from Sina Weibo, which is deemed to be a beneficial tool for scientific research activities.

As shown in Figure 1, our proposed MicroScholar system consists of three parts: a real-time microblog crawling component which crawls recent microblogs from a predefined list of users (including researchers, professors or organizations related to scientific research, currently limited in the computer science field), a scholarly microblog categorization component that recognizes scholarly microblogs and categorizes them into four types (Conference/Event, Paper/Report, Tool/Resource and Other/Irrelevant), and a scholarly microblog ranking component to rank academic microblogs in each type according to each microblog's publishing time or popularity measured by the number of retweets and comments. Finally, the ranked results in each type are shown to researchers. With our system, a researcher can easily find the important scholarly information he has interest in and he will be informed with new scholarly information. Due to page limit, we focus on introducing the techniques and experiments for the component of scholarly microblog categorization.

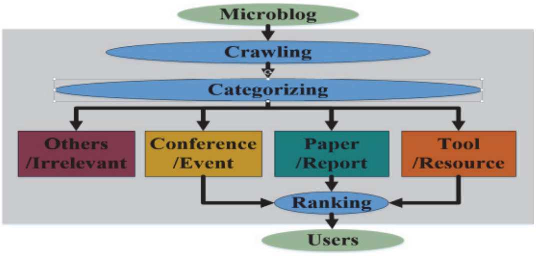

Figure 1. Framework of MicroScholar

Scholarly Microblog Categorization

We treat the scholarly microblog recognition problem as a multi-class classification task and aim to categorize each microblog text into one of the following four types:

- ***Conference/Event***: Includes microblogs introducing a conference or an academic event being held or will take place soon.
- ***Paper/Report***: Includes microblogs introducing or sharing a research paper or report.
- ***Tool/Resource***: Includes microblogs recommending an academic tool or some academic resources, such as an

academic application, or a dataset which may be useful in some particular fields of research.

- ***Other/Irrelevant***: Includes all microblogs that are irrelevant to academic research, or related to academic research but do not belong to any of the above three important categories. The microblogs in this type are much less important and they will be filtered out.

We utilize the popular SVM classifier for the categorization task and apply the SMO algorithm in WEKA toolbox[1] for implementation. The following different kinds of features are extracted and used by the classifier:

- **Textual features (T)**: Microblog texts are segmented with an in-house Chinese word segmenter and then binary unigram features are extracted.
- **Domain features (D)**: Two academic dictionaries are constructed from the DBLP database: one is composed of all words in article titles (words with low TF-IDF value are discarded), and the other is composed of all abbreviates of conference and journal names in the computer science field. Based on the above dictionaries, two features are extracted by counting the dictionary words in the microblog text.
- **LDA features (LDA)**: Inspired by the work introduced in (Phan et al., 2008), to deal with the short microblog texts, we further apply Latent Dirichlet Allocation (Blei et al., 2003) with Gibbs Sampling to discover hidden topics from an unlabeled large microblog text corpus. This microblog corpus consists of 113,925 general microblogs crawled from Sina Weibo. We perform topic analysis on that unlabeled corpus, then use the topic model trained on the general corpus to do topic inference on our labeled dataset. After doing that, we get a topic distribution for each labeled text over hidden topics ($\vartheta_1, \vartheta_2, .., \vartheta_M$, where M is the number of topics) and then use the inferred topic distribution as new feature vector consisting of M new features. We apply GibbsLDA++[2] for the LDA implementation, and the number of topics is set to 300.

Experiments

In order to evaluate the classification performance, we crawl several thousand microblog texts and manually annotate them into four types described above, then construct a balanced evaluation dataset of 2,142 microblog texts (592: 491: 514: 545 for the four categories) by sampling from the whole annotation corpus.

We develop a keyword matching method as a simple baseline. We construct three keyword list with no intersection, each for recognizing one of the first three type. Microblog texts with no words matched in any of the three list will be treated as Other/Irrelevant.

	F-Measure	ROC Area
Keyword Matching	0.519	0.473
SVM(T)	0.789	0.806
SVM(T+D)	0.798	0.819
SVM(T+LDA)	0.831	0.857
SVM(T+D+LDA)	**0.838**	**0.865**

Table 1. Experiment results

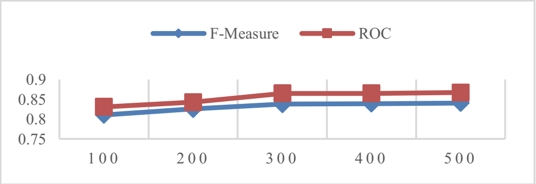

Figure 2. Performance w.r.t. number of topics

In order to test the effectiveness of the different feature types, we perform SVM classification with different feature combinations (T, T+D, T+LDA, T+D+LDA). We perform all SVM experiments in 10-fold cross validation.

The evaluation results are shown in Table 1. As we can see, all the SVM approaches outperform the keyword matching baseline. The SVM approach with all features performs best, and both SVM(T+D) and SVM(T+LDA) outperform SVM(T), which demonstrates the usefulness of every kind of features, especially the LDA features. Figure 2 plots the performance values of SVM(T+D+LDA) with respect to different number of topics ranging from 100 to 500. We can see the performance improvement slows down a lot when the number of topics reaches 300.

Conclusion

In this paper, we introduce a novel system - MicroScholar for automatic scholarly information harvesting. We develop different kinds of features for scholarly microblog categorization. In future work, we will polish the microblog ranking component and learn to identify the significant scholarly microblogs.

Acknowledgments

This work was supported by 863 Program of China (2015AA015403) and NSFC (61170166, 61331011). Xiaojun Wan is the corresponding author.

References

Blei, D., Ng, A. & Jordan, M. 2003. Latent Dirichlet Allocation. *JMLR*, 3:993–1022.

Phan, X. H., Nguyen, L. M., & Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of WWW*.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://gibbslda.sourceforge.net/