Improving Microblog Retrieval from Exterior Corpus by Automatically Constructing a Microblogging Corpus

Wenting Tu, David Cheung, and Nikos Mamoulis
Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong
E-mail: {wttu, dcheung, nikos}@cs.hku.hk

Abstract

A large-scale training corpus consisting of microblogs belonging to a desired category is important for high-accuracy microblog retrieval. Obtaining such a large-scale microblgging corpus manually is very time and labor-consuming. Therefore, some models for the automatic retrieval of microblogs from an exterior corpus have been proposed. However, these approaches may fail in considering microblog-specific features. To alleviate this issue, we propose a methodology that constructs a simulated microblogging corpus rather than directly building a model from the exterior corpus. The performance of our model is better since the microblog-special knowledge of the microblogging corpus is used in the end by the retrieval model. Experimental results on real-world microblogs demonstrate the superiority of our technique compared to the previous approaches.

1 Introduction

The lack of a good microblogging corpus for microblog retrieval can be alleviated by using an *exterior corpus* instead, which contains non-microblog documents that are specific to the desired category. For example, we could use as a corpus a set of news articles posted by newspapers (which can be easily crawled) instead of news microblogs to train a classifier for retrieving news microblogs. Still, the exterior corpus may not be effective because of the differences in the language patterns between non-microblogs and microblogs. Transfer learning theory has recently been used for improving microblog retrieval from an exterior corpus. For example, *OCHI* (Feng et al. 2011) improves microblog retrieval using an exterior corpus by designing a domain-adapted similarity metric between the exterior corpus and microblog documents. Still, traditional transfer learning relies only on transferable or shared knowledge from the exterior corpus to the microblogging documents; i.e., many microblog-special language features (e.g., special notations, such as #hashtags# and @someone, transpositions, etc.) cannot be utilized. Thus, transfer learning that only relies on common or transferred knowledge could be inadequate. Our work attempts to improve microblog retrieval, by automatically constructing a corpus consisting of real microblogs, with the use of the exterior corpus. We exploit a known characteristic of current microblogging platforms: for a given category, typically there exists a group of users whose microblogs belong exclusively to this category. Twitter, one of the world's largest microblogging platform, contains examples supporting this observation: The user *@BBCWorld* always post microblogs related to news events (news microblogs); The user *@Stock Predict* from Twitter always post their opinions to stock movements (investor-opinion microblogs); The user *@8joke* from Twitter always post joke microblogs; The user *@Discount LA* from Twitter always post advertisement microblogs. Similar examples also can be found in other microblogging platforms (e.g., SinaWeibo) and on other categories (e.g., product comments). We propose a framework to take use the exterior corpus and previous approaches to identify these users rather than directly build the microblog retrieval model. Then, we use their posts to construct a microblogging corpus; this corpus is finally used to build a target-special model for the microblog retrieval task. Our experiments on real-world microblogs and an exterior corpus show that the microblog retrieval model of our method can obtain better retrieval performance than the ones built by previous approaches embedded in our framework.

2 Our methodology

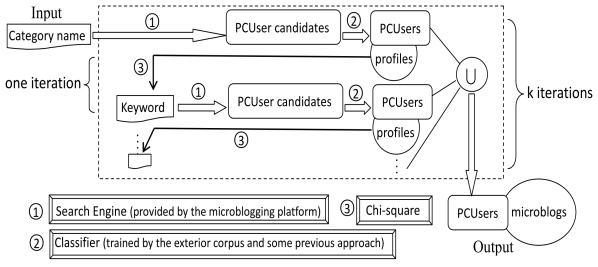

Figure 1: Automatically constructing a microblogging corpus by selecting a set of PCUsers.

Our objective is to use an exterior corpus (e.g. an corpus of news articles) $\mathcal{D}^e$ to build a model for retrieving microblogs that belong to a special category ℓ (e.g., ℓ=news). The key step is to identify a group of users, called PCUsers, who post microblogs that belong to category ℓ. Our microblogging corpus is then simulated by their posts and be

used to train the retrieval model.

As Figure 1 shows, we find the PCUsers by a boosting approach with k iterations. In the i-th iteration ($i = 1, \cdots, k$), three steps are performed. ①: given a keyword w_i related to the category ℓ, we obtain a set $\mathcal{C}_i^{PC}$ of candidate PCUsers by selecting from the set of users those who have w_i in their profiles (formed by names, descriptions, and tags). ②: we use any previous work (any common classifier or transfer-learning model) to find whether a post by c^{PC} belongs to ℓ. The *professionalism* of a candidate $c^{PC} \in \mathcal{C}_i^{PC}$ is proportional to the percentage of his/her microblogs that belong to ℓ. The users in $\mathcal{C}_i^{PC}$ whose professionalism surpasses a threshold (e.g., we use 0.6 in our experiments) are selected as the set of PCUsers $\mathcal{U}_i^{PC}$ in this iteration. ③: After a set of PCUsers are selected, we generate the keywords used in the next iteration from their profiles. We calculate the dependencies between the profiles of $\mathcal{U}_i^{PC}$ and each word w in the profiles of users in $\mathcal{C}_i^{PC}$ by Chi-Square $\mathcal{X}^2$ statistic:

$$\mathcal{X}^2(w, \mathcal{U}_i^{PC}) = \frac{C \times (C_1 C_4 - C_2 C_3)^2}{(C1 + C3) \times (C2 + C4) \times (C1 + C2) \times (C3 + C4)}, \quad (1)$$

where C_1 is the number of times w occurs in the profiles of all users in $\mathcal{U}_i^{PC}$, C_2 is the number of times w occurs in the profiles of other users, C_3 is the number of users in $\mathcal{U}_i^{PC}$ whose profiles do not contain w, C_4 is the number of other users whose profiles do not contain w, and C is the total number of word occurrences. Then, the word w with highest $\mathcal{X}^2(w, \mathcal{U}_i^{PC})$ is selected as the keyword in the next iteration. In the first iteration, the category name ℓ is used as the keyword.

After selecting PCUsers k times, the final PCUser set $\mathcal{U}^{PC}$ is defined as $\bigcup_{i=1}^{k} \mathcal{U}_i^{PC}(w_i)$. For performing classification, we combine the microblogs posted by PCUsers to train a model.

3 Experiments

To evaluate the effectiveness of our framework, we conducted experiments on the tasks of retrieving investor-opinion microblogs and news microblogs from SinaWeibo.

Data Preparation and Experiment Setup

Data Preparation Caiku (www.caiku.com) posts investment opinions provided by invited financial experts. After crawling the Caiku website, we obtained a exterior corpus containing 17,745 investment opinions. For retrieving news microblogs, we crawled 70,000 news titles for the period 2008/01/01-2012/12/31 from a news website (www.news.sina.com.cn) as the exterior corpus. To evaluate the retrieval performance, after manual labeling, the test data contain 10,000 real investor-opinion microblogs, 10,000 news microblogs, 10,000 non-investor-opinion microblogs, and 10,000 non-news microblogs.

Our Methods and Competitors We use the methodology of Section 2 to select PCUsers. We use Multinomial-Naive-Bayes (*MNB*) theory (Rennie et al. 2003) and *OCHI* as the approaches for constructing the classification module for calculating professionalism score. Then, we compare the retrieval performance of our method to that of using *MNB* and *OCHI* only.

Table 1: Performance (F-score) of *OCHI*, *MNB*, and our method on retrieving investor-opinion&news microblogs.

OCHI	Our method (Use of *OCHI* to select PCUsers)		
/	k=1	k=2	k=3
0.72&0.87	0.76&0.89	0.79&0.92	0.83&0.95
MNB	Our method (Use of *MNB* to select PCUsers)		
/	k=1	k=2	k=3
0.57&0.85	0.65&0.83	0.71&0.90	0.75&0.93

Experimental Results and Analysis

Table 1 compares the retrieval performance (F-score) on the test microblogs. We can see that our method improves the retrieval performance (F-score) except for case when $k = 1$ and *MNB* is used for selecting the PCUsers. When $k = 1$, the selected PCUsers may not be enough to provide a good microblogging corpus.

4 Related Work

Most of the previous work on automatically constructing a corpus for text classification (Read 2005; Davidov et al. 2010) focus on sentiment classification, using emotion icons to distinguish between positive and negative sentiments. However, these methods can hardly be used for classification tasks on other categories. There is some recent work on the use of microblogs posted by special classes of users (Sankaranarayanan et al. 2009) or containing some linguistic patterns defined by domain experts (Read 2000). These methods cannot achieve fully-automatic corpus extraction, as opposed to our approach, which only requires to know the category.

5 Conclusion

In this paper, we proposed a framework which greatly improves the effectiveness of classification and transfer learning in the retrieval of microblogs that belong to a specific category, given an exterior corpus. We plan to conduct a more detailed study with a wide range of microblog categories in our future work.

References

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*

Feng, X.; Shen, Y.; Liu, C.; Liang, W.; and Zhang, S. 2011. Chinese short text classification based on domain knowledge. In *IJNLP*.

Read, J. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *ACL*.

Rennie, J.D.; Shih, L.; Teevan, J.; and Karger, D.R.. 2003. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*.

Sankaranarayanan, J.; Samet, H.; Teitler, B.E.; Lieberman, M.D.; and Sperling, J. 2009. Twitterstand: news in tweets. In *GIS* .

Yangarber, R.; Grishman, R.; Tapanainen, P.; and Huttunen, S. 2000. Automatic acquisition of domain knowledge for information extraction. In *COLING* .