# Understanding the Top Grass Roots in Sina-Weibo

Ze Huang[1], Bo Yuan[1], and Xuelei Hu[2]

[1] Intelligent Computing Lab, Division of Informatics,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, P.R. China
workthy@hotmail.com, yuanb@sz.tsinghua.edu.cn
[2] School of Computer Science and Technology,
Nanjing University of Science and Technology, Nanjing 210094, P.R. China
xlhu@njust.edu.cn

**Abstract.** Microblogging is now popular among everyday web users in China who have a common name called *grass roots* in Sina-Weibo, a major microblogging service similar to Twitter. In this paper, we investigate the properties of messages published by this group of users and classify the messages into various topic categories using text classification methods based on the Bag of Words (BOW) model. We find that, using Naïve Bayes, it is possible to achieve high accuracy in recognizing the topic of a message but the popularity of a message cannot be reliably predicated based on its contents. These findings are also further explored with visualization techniques.

**Keywords:** Microblogging, Text Classification, Bag of Words, Visualization

## 1 Introduction

The popularity of social media is expected to be growing continuously world-wide. According to the recent report by Nielsen on American internet users, social networks and blogs account for 23% of time spent online, compared to 9.8% for online games and 7.6% for email [1]. In China, Sina-Weibo is one of the most popular microblogging services. According to the quarterly reports of SINA Corporation, Sina-Weibo had more than 100 million registered users in March 2011, and this number doubled five months later with nearly 90 million messages published each day. Similar to Twitter, it allows users to post messages with a character limit, with optional links to other sources of information. There are also some differences between Sina-Weibo and Twitter due to local conventions.

Existing studies on applying text classification methods to English microblogging sites have been conducted in several aspects, such as sentiment analysis [2], topic detection [3], information filtering [4, 5] and performance comparison using different classifiers and feature selection methods [6, 7]. There are also some studies on understanding the information diffusion in Twitter and the structure of Twitter [8-12]. In Chinese context, an interesting case study was conducted in [13] to investigate how

Chinese users used microblogging services in response to the 2010 YuShu Earthquake in China. The influence of Part-Of-Speech (POS) features on Chinese webpage classification was analyzed in [14]. A recent study based on 43,000 volunteer ratings on tweets shows that contents on information sharing, self-promotion and questions to followers were often valued highly [15].

Ordinary users in Sina-Weibo have a common name, grass roots, to be distinguished from famous users such as celebrities whose real identities are manually verified. Our study focused on analyzing the messages published by grass roots as they are representatives of the vast majority of microblog users. With a close observation of the message contents, we found that most of the non-private messages, especially those relatively long messages, may be mapped to a few topics. In this paper, messages were classified into five categories: Living Tips (LT), Design & Originality (DO), Fashion (F), Entertainment (E), Quotation & Sayings (QS). Certainly, for microblogging services where all messages have user specified tags, it is preferable to use the tags as class labels.

As a popular representation model used in text classification, the Bag of Words (BOW) model usually cannot achieve satisfactory performance on short text classification as the texts do not provide sufficient word occurrences, making the feature space quite sparse. To address this issue, one solution is to inflate the text by integrating meta-information and word-occurrence information from other sources, such as Wikipedia or search results returned by web search engines [16, 17].

However, we found that the messages published by top grass roots in Sina-Weibo had special properties different from ordinary short texts: they seemed to be well structured and contain good quality information for indicating a topic. By contrast, we found that the popularity of a message cannot be reliably predicated based on its contents and we concluded that there are some distinct characteristics of grass roots' messages in Sina-Weibo, which may reflect the special social phenomena behind the sociocultural system in China.

## 2 Data Preparation

Similar to other microblogging services, Sina-Weibo displays a list of the most influential grass roots based on their numbers of followers. This ranked list shows the top 300 grass roots and is updated on a daily basis since the number of followers may change continuously. We collected the messages of the top 300 grass roots and some randomly selected followers using Sina-Weibo API. The API had an access limit and only returned up to 2,000 historical messages before the date that the API was invoked. As a result, the messages of some grass roots were collected completely while the others were not.

A closer look into the collected dataset revealed some special features in the messages of top grass roots:

- *Username*. Their usernames often directly implied a certain topic, such as "*The digest of cold joke*", "*Classic Quotations*" and "*Beauty and Health*". However, their messages were not always consistent with the topics revealed by the

usernames. For example, many top grass roots published messages related to quotations.

- *Content.* They rarely published private messages, or original contents. Instead, they often shared information about fashion, constellation, jokes and classic quotations. In many occasions, they even used software to automatically post messages. The contents of most of these messages were from the Web and had no connection with the social events at the time of publishing.
- *Hashtags and Personal Description*. Top grass roots widely used hashtags and personal signatures in their profiles for further explanation of the topics on which their microblogs were focused.

In addition to the text contents, messages obtained through Sina-Weibo API contained other types of information. For example, URL links to websites and videos were widely used in grass roots' messages. In this paper, the focus was on text classification and all non-text information was removed.

The main steps of data preprocessing for our experiments are as follows:

- Removed messages that contained "*@username*". Each user had a unique username, and "*@username*" was linked to the user's microblog. However, messages with multiple occurrences of "*@username*" often had only a few greeting or commentary words with little association with the message topic.
- Messages with string length less than 60 were removed to filter out those less meaningful messages and reduce the burden of manually labeling the messages.
- Messages with English words were discarded. Although there were some English messages in the dataset, we wanted to focus on Chinese text classification and also avoid the curse of dimensionality, as the feature space would expand a lot when considering English words.
- Removed URL links starting with "http:// ".

After this preprocessing procedure, the final dataset contained totally 40,636 messages (Table 1).

**Table 1.** The categorization of message topics

| Category | Description | Count |
|---|---|---|
| Living Tips (LT) | Knowledge of daily life such as cooking and health | 8989 |
| Design & Originality (DO) | Novel design, new science & technology inventions | 5665 |
| Entertainment (E) | News and comments of movies | 3884 |
| Fashion (F) | Latest fashion trends and dressing advices | 6029 |
| Quotations & Sayings (QS) | Classic quotations, excerpts from literary works | 16069 |

## 3 Classification and Prediction

Since in Chinese language there are no fixed separators between words, Chinese lexical analysis is required to segment the string of words into meaningful units, each of which is considered as a feature and then built into a vector space. We adopted a widely used Chinese lexical analysis system ICTCLAS in our experiment, since ICTCLAS supports word segmentation, Part-Of-Speech (POS) tagging and unknown word entities recognition and has achieved satisfactory segmentation accuracy compared to other Chinese lexical analysis systems.

The experiments were conducted using the Multinomial Naïve Bayes classifier implemented in WEKA 3.6, using 10-fold cross validation. In Table 2, the first column shows the different groups of features adopted in classification (*/n*: nouns, /v: verbs and /a: adjectives). It can be seen that when only using nouns and verbs as the features, the size of the feature space reduced from 49,441 to 35,576 with little loss in accuracy. We also tested the effect of feature selection techniques such as information gain (IG) and $\chi^2$-test (CHI). When only using nouns as the original features, both techniques achieved good accuracies close to 93.8% with around 15,000 features and the accuracy was already above 93% with only 4,000 features.

**Table 2.** Classification accuracy with different POS elements

| **Features** | **#Features** | **Accuracy (%)** | | | | | |
|---|---|---|---|---|---|---|---|
| | | ***DO*** | ***E*** | ***LT*** | ***F*** | ***QS*** | ***Overall*** |
| All | 49441 | 91.2 | 92.7 | 96.5 | 97.8 | 95.6 | **95.2** |
| /n | 24236 | 88.4 | 89.9 | 96.2 | 95.3 | 94.6 | **93.7** |
| /n/v | 35576 | 90.2 | 91.6 | 97.2 | 97.0 | 95.1 | **94.8** |
| /n/a | 26885 | 89.2 | 89.4 | 96.2 | 95.7 | 95.1 | **94.1** |

Another interesting question is on the correlation between the contents of messages and their popularity, which can be measured by two figures: the number that a message was forwarded (*#relay*) and the number that a message was commented (*#cmt*). We focused on the top 50 grass roots as of 15/06/2011 in Sina-Weibo and collected 169,775 historical messages published in the original authorship. All URL links in messages were removed and messages with string length more than 80 were selected. There were 52,508 messages in the dataset, which were uniformly assigned to 3 levels or classes (low popularity, medium popularity and high popularity) according to their *#relay* and *#cmt* values.

Similarly, these messages were classified using Naïve Bayes and the performance was evaluated using 10-fold cross validation. According to the results in Table 3, the accuracies for both measures as well as for all of the levels were quite moderate (random guess: 33.3%). This evidence may indicate that using the pure text contents of messages to predicate their potential popularity is not reliable. We also observed that messages published in the early stage of the top 50 grass roots received little

attention, regardless of their topics and their quality. By contrast, after these grass roots became influential (getting into the top list), their messages tended to have some good chance of being forwarded and commented.

**Table 3.** Results of popularity prediction

| **Measures** | **Prediction Accuracy** | | |
|---|---|---|---|
| *#relay* | 58.0% (Low) | 41.5% (Medium) | 56.0% (High) |
| *#cmt* | 59.5% (Low) | 52.4% (Medium) | 60.2% (High) |

## 4 Visualization

In order to provide deeper insights into the experiment results on topic classification and popularity prediction, we conducted some interesting text visualization with the help of an open-source software package called *Gephi*, which has previously been used in similar work [18]. Each message in the dataset was shown as a node in the graph. Edges were added only if two messages (nodes) were similar enough, which was measured by the cosine distance in our experiment (only nouns, adjectives and verbs were used as features). When the cosine similarity was above a threshold, an edge was drawn between the two nodes.

In Fig. 1 and Fig. 2, the size of a node was determined by its degree (the number of other nodes connected to it), and gray levels were used to distinguish messages in various categories. The Fruchterman-Reingold algorithm was used to create the layouts to bring together nodes with strong ties. Since a large number of nodes and edges would make the graph difficult to read, *Gephi* provides a filtering method to hide certain nodes and edges. For example, we can make a node or an edge invisible if the node's degree or the edge's weight is less than a threshold.

To create Fig. 1, 2,000 messages were randomly selected from each category. For the sake of clarity, messages belonging to QS, LT and E are shown in Fig. 1 (left) with threshold 0.6 while other messages are shown in Fig. 1 (right) with threshold 0.45. Nodes with degrees less than 4 were kept invisible (their edges were also kept invisible). Note that even a seemingly isolated node in Fig. 1 actually had at least 4 invisible edges linked to other invisible nodes.

The most important observation from Fig. 1 is that nodes within the same sub-graph (a set of connected nodes) often had the same gray level, which means that messages were generally similar to those in the same class and different from messages in other classes. We believe that this inherent similarity can largely explain the good classification performance observed in Section 3.

Fig. 2 was created in a similar manner, showing 391 nodes and 6640 edges generated from the original 52,508 messages in the dataset for popularity prediction (Section 3). There were three types of nodes in terms of *#relay*: white for low popularity, grey for medium popularity and black for high popularity. It is clear that each sub-graph typically contained messages with mixed popularity levels (messages

similar to each other often had different levels of popularity). This phenomenon is in distinct contrast to Fig. 1 and it would be very challenging to accurately predict the popularity of messages.

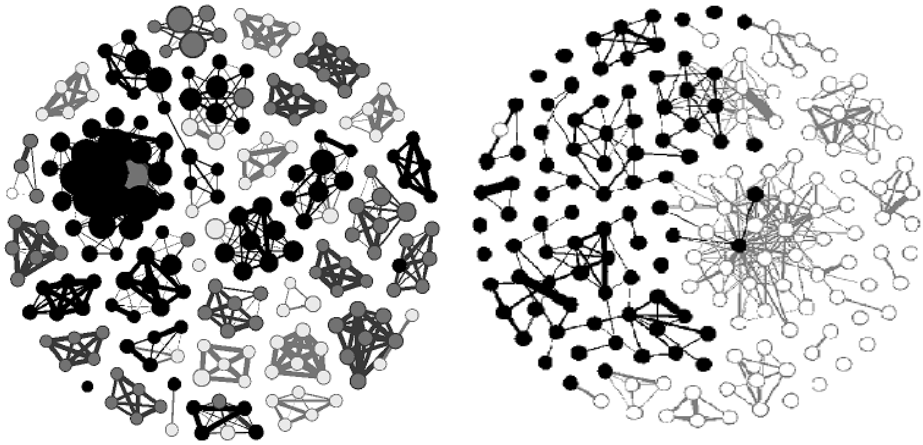

**Fig. 1.** Visualization of messages for topic classification: QS, LT & E (left) and DO & F (right). It shows that messages similar to each other often belong to the same topic category.

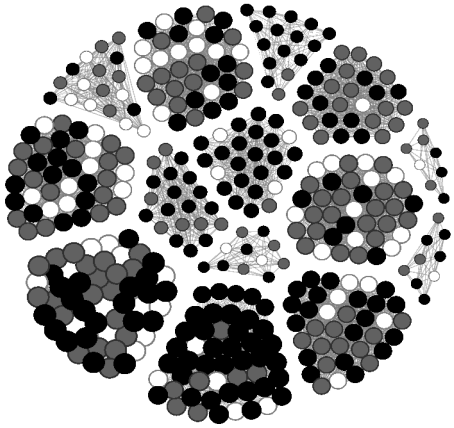

**Fig. 2.** Visualization of messages for popularity prediction: white (Low), gray (Medium) and black (High). It shows that similar messages often have different levels of popularity.

To better understand the forwarding relationship among users, Fig. 3 shows a directed graph describing the relationship among the top 100 grass roots as of 15/06/2011 and the authors from whom they have forwarded messages before this date. The size of a node was determined by the total times that his/her messages were forwarded by any of the top 100 grass roots. The white nodes are the top 100 grass roots and the grey ones are the non-top grass roots whose messages had been forwarded for more than 300 times and the black ones are other ordinary grass roots.

We tracked for 25 days the top 300 users with the most followers (Top-300) between August and September 2011 in Sina-Weibo. Interestingly, we found that, out of these 43 grey nodes, 16 entered into Top-300. So, it seems important for an ordinary grass root to become popular if his/her messages can be forwarded frequently by the top ones. We believe that this is the reason that many ordinary users actively contribute good quality messages to top grass roots.

**Fig. 3.** The forwarding relationship among users

## 5 Conclusion

The motivation of our work was to gain some appreciation of the characteristics of grass roots in Sina-Weibo, which represent the vast majority of microblog users. Experimental results showed that: (a) standard text classification methods performed well on topic classification with overall accuracy more than 95%; (b) using part of the POS features (e.g., nouns and verbs) can effectively decrease the feature dimension with little sacrifice of accuracy; (c) the pure text contents cannot provide sufficient information to accurately predicate a message's popularity.

These results suggest that the topics of non-private messages from top grass roots can be identified effectively with simple classification methods (e.g., Naïve Bayes). The reason behind this interesting phenomenon is, in our opinion, that the contents of such messages are often carefully organized to be concise (e.g., the frequent use of keywords) and of good quality compared to private messages to attract followers and improve their popularity. A possible incentive for publishing this type of messages may be the potential advertising value of these microblog accounts.

**Acknowledgements.** This work was supported by the National Natural Science Foundation of China (No. 60905030).

## References


1. The Nielsen Company: State of the Media: The Social Media Report – Q3 2011. http://www.nielsen.com/us/en/insights/reports-downloads.html
2. Bermingham, A., Smeaton, A.: Classifying Sentiment in Microblogs: Is Brevity an Advantage? In: 19th ACM Conference on Information and Knowledge Management, pp. 1833−1836 (2010)
3. Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., Sperling, J.: TwitterStand: News in Tweets. In: 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, pp. 42−51 (2009)
4. Laboreiro, G., Sarmento, L, Teixeira, J., Oliveira, E.: Tokenizing Micro-Blogging Messages Using a Text Classification Approach. In: Fourth Workshop on Analytics for Noisy Unstructured Text Data, pp. 81−88 (2010)
5. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short Text Classification in Twitter to Improve Information Filtering. In: 33rd ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841−842 ( 2010)
6. Ramage, D., Dumais, S., Liebling, D.: Characterizing Microblogs with Topic Models. In: Fourth International Conference on Weblogs and Social Media, pp. 130−137 (2010)
7. Rosa, K., Ellen, J.: Text Classification Methodologies Applied to Micro-Text in Military Chat. In: 2009 International Conference on Machine Learning and Applications, pp. 710−714 (2009)
8. Wu, S., Hofman, J., Mason, W., Watts, D.: Who Says What to Whom on Twitter. In: 20th International Conference on World Wide Web, pp. 705−714 (2011)
9. Naaman, M., Boase, J., Lai, C.: Is It Really About Me? Message Content in Social Awareness Streams. In: 2010 ACM Conference on Computer Supported Cooperative Work, pp. 189−192 (2010)
10. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: Understanding Microblogging Usage and Communities. In: 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis, pp. 56−65 (2007)
11. Krishnamurthy, B., Gill, P., Arlitt, M.: A Few Chirps about Twitter. In: First Workshop on Online Social Networks, pp. 19−24 (2008)
12. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? In: 19th International Conference on World Wide Web, pp. 591−600 (2010)
13. Qu, Y., Huang, C., Zhang, P., Zhang, J.: Microblogging after a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In: 2011 ACM Conference on Computer Supported Cooperative Work, pp. 25−34 (2011)
14. Huang, W., Xu, L., Duan, J., Lu, Y.: Chinese Web-Page Classification Study. In: IEEE International Conference on Control and Automation, pp. 1553−1558 (2007)
15. Andre, P., Bernstein, M., Luther, K.: Who Gives a Tweet? Evaluating Microblog Content Value. In: 2012 ACM Conference on Computer Supported Cooperative Work, pp. 471−474 (2012)
16. Schonhofen, P.: Identifying Document Topics Using the Wikipedia Category Network. In: 2006 International Conference on Web Intelligence, pp. 456−462 (2006)
17. Broder, A., Fontoura M., Gabrilovich E., Joshi A., Josifovski V., Zhang T.: Robust Classification of Rare Queries Using Web Knowledge. In: 30th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 231−238 (2007)
18. Stray, J.: A Full-text Visualization of the Iraq War Logs. http://jonathanstray.com/a-full-text-visualization-of-the-iraq-war-logs