Topical differences between Chinese language Twitter and Sina Weibo

Qian Zhang
Northeastern University
Boston, MA, USA
qi.zhang@neu.edu

Bruno Gonçalves
Center for Data Science
New York University
New York, NY, USA
bgoncalves@gmail.com

ABSTRACT

Sina Weibo, China's most popular microblogging platform, is considered to be a proxy of Chinese social life. In this study, we contrast the discussions occurring on Sina Weibo and on Chinese language Twitter in order to observe two different strands of Chinese culture: people within China who use Sina Weibo with its government imposed restrictions and those outside that are free to speak completely anonymously. We first propose a simple ad-hoc algorithm to identify topics of Tweets and Weibos. Different from previous works on micro-message topic detection, our algorithm considers topics of the same contents but with different #tags. Our algorithm can also detect topics for Tweets and Weibos without any #tags. Using a large corpus of Weibo and Chinese language tweets, covering the entire year of 2012, we obtain a list of topics using clustered #tags and compare them on two platforms. Surprisingly, we find that there are no common entries among the Top 100 most popular topics. Only 9.2% of tweets correspond to the Top 1000 topics of Weibo, and conversely only 4.4% of weibos were found to discuss the most popular Twitter topics. Our results reveal significant differences in social attention on the two platforms, with most popular topics on Weibo relating to entertainment while most tweets corresponded to cultural or political contents that is practically non existent in Weibo.

Keywords

Topic detection on short-message; Social Attention; Twitter; Weibo; Social media; Online Behavior

INTRODUCTION

China is known for its rich internal Internet ecosystem where Chinese alternatives to most foreign Internet services flourish. This is due not only to cultural differences that prevent foreign websites from gaining a large market share, but also due to stringent government controls that sometimes prevent foreign Internet companies from selling their services or that outright block access to them.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890562

Sina Weibo, as China's most popular microblogging platform, is perhaps the most visible face of China's own internal version of the Internet. It origins date back to 2009, but wasn't until 2011 that it rose to prominence. Since July 2009, Twitter has been blocked in China [7], leaving national services such as Sina Weibo as the only alternative. It is currently used by over $500M$ users and, similarly to its foreign counterpart Twitter [1] that is widely considered to be a proxy for its users social life and interests [2, 3], it has recently started to draw the attention of researchers everywhere [4, 5, 6, 7]. Different from Twitter, in March 2012, Weibo started requiring its users to associate their profile with their true identity [4].

The previous works on topic detection on microblogs are usually designed for pre-selected specific topics [8, 9] or only for short-messages with #tags [10]. However, the majority of Tweets are not # tagged [8], and there is few work focusing on automatic topic detection for microblogs. We first propose a simple ad-hoc algorithm to identify topics on microblogs without pre-selection, and cluster those microblogs without #tag into the detected topics. More importantly, without the assumption that a #tag represents a unique topic, our algorithm merges posts of the same contents but with different #tags.

Past studies on Chinese microblogging platforms [5, 7] mainly focused on censorship and analyzed deleting practices on microblogs containing censored key words. Others compared the user behaviors, texture features of posts and temporal dynamics of re-posting [1] and an artificially selected categorical events [11] on Sina Weibo and Twitter. There is little research on comparing the collective attention of Chinese microbloggers in a large scale. Here, we take a first step in this direction by proposing an algorithm to model and compare Sina Weibo and Twitter. We have observed two different versions of Chinese culture: people inside China (94.8% of geotagged Weibos are within China) and those outside (93.7% of geotagged Tweets are located outside China). Due to the complexity of its language, the number of people outside China learning Chinese as a second language is still very small. For instance, in 2009 10 times more students in US colleges studied Spanish than Chinese [12], indicating that people who Tweet in Chinese outside of China are likely either Chinese expats or from Chinese heritage. While some analyses have been performed on geographically distributed populations speaking the same language [13], this combination of technically equivalent services serving populations with a similar cultural background that are isolated from each other is unique. It provides us

with the perfect opportunity to study the cultural differences in the virtual world of Chinese. Our comparison results reveal significant differences in social attention distribution across both platforms, with the most popular topics on Sina Weibo relating to entertainment while the most topics in Twitter corresponded to cultural or political contents.

DATA DESCRIPTION

We use the dataset of Sina Weibo from Open Weiboscope Data Access [5, 6]. The dataset contains 226.8 million Weibo posts (Weibos for short) collected over the full course of 2012. The Twitter dataset used in this study was extracted from the raw Gardenhose feed, an unbiased sample of 10% of the entire Twitter dataset that provides a statistically significant real time view of all Twitter account activity [14]. To identify Tweets and Weibo in Chinese language, we perform language detection using the "Chromium Compact Language Detector" [15, 16]. This way, we collected 12.3 Million Tweets and 216.8 Million Weibo in both simplified and traditional Chinese language covering the entire year of 2012. The Sina Weibo dataset also include microblogs which are not accessible to the public, either censored or self deleted. Following [5], we consider weibos deleted by the censorship (with message "permission denied" from API). In total, we considered 74,132 deleted weibos for our study.

CLUSTERING MICROBLOGS INTO TOPICS

Topic modeling on micro-messages is still challenging due to its inherent sparseness [10] and noise [9]. In this study, we take microblogs with #tags as potential known topics. There are $2.06M$ tweets and $18.9M$ weibos with #tags. We build a vocabulary vector space on each platform with words of high TF-IDF score, and cluster similar #tags into a specific topic. For instance, for the top 100 #tags on Sina Weibo and Twitter, we merge them into 83 and 58 topics respectively. We assign the rest of microblogs without #tags to topics that are closest to them in the vocabulary vector space. To reduce statistical fluctuations we restrict our study to the Top 1000 topics in each platform. In total, 20.8% weibos are classified into popular topics on Weibo and 34.5% of tweets discuss popular topics on Twitter.

Preprocessing: We first filter microblogs by removing the words representing short URLs and mentioning other users ("@username"). Filtered microblogs in traditional Chinese are then converted to simplified Chinese with the python-jianfan library [17]. Chinese word segmentation is performed using Jieba [18] and part-of-speech tagging (POS) is performed following [9]. This way each microblog is represented as a set of words tagged as noun, name, location, organization, time, place word, position word or verb.

Vector representation: We merge all microblogs with the same #tag h_i as a document D_{h_i}, and calculate its TF-IDF (term frequency–inverse document frequency). For each D_{h_i}, we exclude the words with length less than 2 since a single character word in Chinese can be noisy and under-representative, and choose the first 10 words $t^j_{h_i}, j = 0 \dots 9$ with highest frequency and their TF-IDF weights $w^j_{h_i}$, and its vocabulary vector can be written as $V_{h_i} = (0:0, ..., 0:0, t^0_{h_i}:w^0_{h_i}, t^1_{h_i}:w^1_{h_i}, ..., t^9_{h_i}:w^9_{h_i}, 0:0, ..., 0:0)$, and $|V_{h_i}| = 10N$ where N is the number of #tags we select.

Clustering: Since several similar #tags likely refer to the same topic, we further cluster #tags into topics using hierarchical clustering. In Figure 1, we show the dendrogram for the Top 100 #tags on Sina Weibo platform, calculated using cosine distances in the embedding vector space $d_{h_i,h_j} = V_{h_i} \cdot V_{h_j} / \|V_{h_i}\| \|V_{h_j}\|$. Interestingly, most clades are simplicifolious, indicating that distribution of words for each #tag is substantially different from the distribution in others. We observe similar dendrogram for Top 100 #tags on Twitter (figure not shown). Thus, we apply a modified divisive clustering method (Algorithm 1), where we iteratively divide the largest cluster into a small cluster and a large one, until the size of the small cluster is 0.

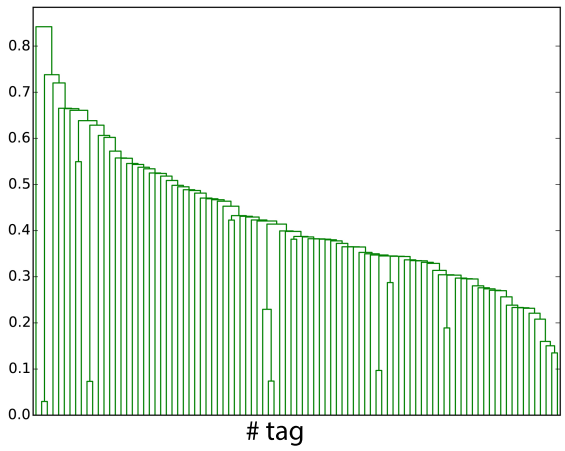

Figure 1: Cluster dendrogram for Top 100 #tags on Sina Weibo contains many simplicifolious clades, indicating most #tags are distant from each other.

Algorithm 1: Merging #tags into topics

Input: The hierarchical cluster L
Output: The set of topics S_T
initialization;
– $S_T = \emptyset$;
while $L \neq \emptyset$ **do**
  Partition L into a larger cluster B_L and a small cluster B_S;
  $L \leftarrow B_L$;
  $S_T \leftarrow S_T \bigcup \{B_S\}$;

After merging # tags into topics, each topic t_i in vocabulary vector space now is defined as $V_{t_i} = (0:0, ..., 0:0, t^0_{t_i}:w^0_{t_i}, t^1_{t_i}:w^1_{t_i}, ..., t^9_{t_i}:w^9_{t_i}, 0:0, ..., 0:0)$, and $|V_{t_i}| = 10T, T \leq N$. $t^j_{t_i}, j = 0 \dots 9$ is now the first 10 words with highest frequency in a topic t_i and their TF-IDF weights $w^j_{t_i}$. The centroids of the final clusters are taken to represent topics in the vector space of each platform. To classify the remaining microblogs, we measure the cosine distance between the centroid of a topic t_j, and each microblog m, $d_{t_j,m}$. If $d_{t_j,m}$ is less than a threshold d_t, we consider the microblog m is discussing the topic t_j, shown in Algorithm 2. To determine the threshold d_t, we measure the distribution of distance between a centroid of a topic and microblogs inside that topic. About 65% of tweets and 76% of weibos have distances less than 0.9 to their topical centroids. Meanwhile, on average

only 9.2% of microblogs outside a topical centroid have distances less than 0.9. Therefore, we set $d_t = 0.9$.

Algorithm 2: Clustering microblogs into topics

Input: The set of topics S_T and a set of microblogs M
foreach *topic* t_j *in* S_T **do**
  computing its centroid C_{t_j} ;
  label t_j with its # tags clustered;
foreach *microblog* m *without # tag in* M **do**
  set $V_m = (t^j_{t_0} : w^j_{t_0}, ...t^j_{t_i} : w^j_{t_i}, ..., t^j_{t_{T-1}} : w^j_{t_{T-1}})$;
  let $T_m = None$ be the topic for m;
  let $D^{min}_m = \infty$ be the minimum distance from m to any centroid of clustered topic ;
  foreach *topic* t_j *in* S_T **do**
    $d_{t_j,m} = V_m \cdot C_{t_j} / \|V_m\| \|C_{t_j}\|$;
    if $d_{t_j,m} < d_t$ **then**
      $D^{min}_m \leftarrow d_{t_j,m}$;
      $T_m \leftarrow t_j$;
  if T_m *is not None* **then**
    Assign the microblog m into the topic T_m ;

RESULTS AND DISCUSSION

Our analysis aims to compare topical spaces in Chinese language on different microblogging systems. With identified centroids in the vocabulary vector space defined in the last section, we first calculate the distance between the centroids of the Top 1000 topics on the two platforms. We define d_{ij} as the cosine distance in the vocabulary vector space between the centroid of topic i on Twitter and topic j on Sina Weibo. Figure 2-A shows the cumulative distribution function of distance d_{ij} for $10^3 \times 10^3$ pairs of topical centroids. Surprisingly, only 8.9% pairs of topics have distance less than 0.9. In Figure 2-B, we show the distance between Top 100 topics on Weibo and Top 100 topics on Twitter. Surprisingly, the distance between 91% of pairs of Top 100 topics on the two platforms is larger than 0.9, indicating that microbloggers in each platform have significantly different conversation topics and interests.

In Table 1, we provide the Top 10 topics in Chinese language on Sina Weibo and Twitter to illustrate the differences. On Sina Weibo, $\sim$ 1.88% of entire datasets can be classified into top 10 topics ($\sim$ 9.01% for the Top 100); while on Twitter $\sim$ 13.43% over all tweets are categorized into top 10 topics ($\sim$ 20.8% for Top 100). The microblogs on Sina Weibo focus on entertainment (singers, actors and games) and advertising. In contrast, on Twitter, there is no commercial advertisement appearing, and the most of others are all corresponding to political contents.

We have classified weibos and tweets into the topical space in their own vocabulary vector space. For an unclassified weibo or tweet, we calculate its distance to centroids on both platforms, and assign it to the closest topic. Interestingly, we find there are only $\sim$ 9.2% of tweets correspond to the Top 1000 topics on Sina Weibo platform, and only $\sim$ 4.4% of weibos were discussing the most popular topics on Twitter. Chinese microbloggers speaking the same languages on two platforms share a few social attentions.

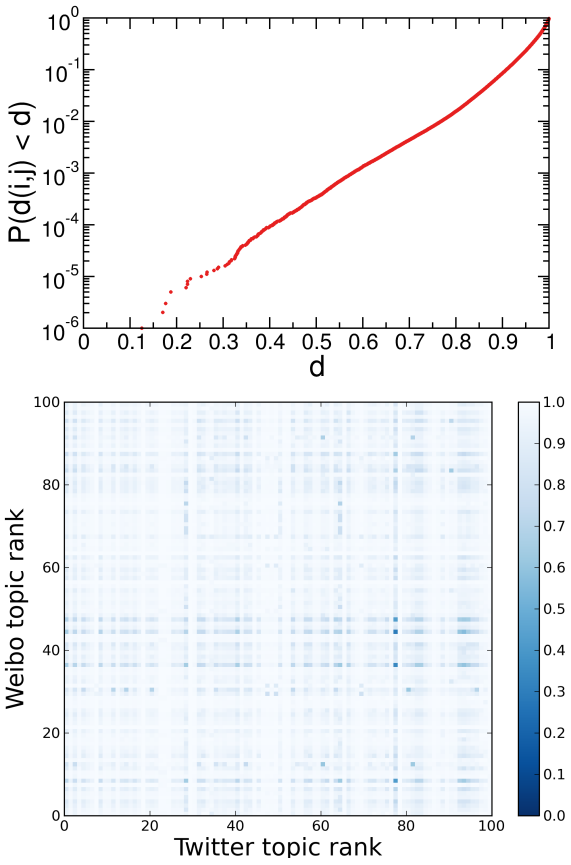

Figure 2: Distance between popular topics on Twitter and Sina Weibo. (A) The cumulative distribution function of the distance d_{ij} as the cosine distance between the centroid of topic i in Top 1,000 topics on Twitter and topic j in Top 1,000 topics on Sina Weibo. (B) The distance matrix between the centroid of topic i in Top 100 topics on Twitter and topic j in Top 100 topics on Sina Weibo. The distance 0 refers two exactly similar topics, and 1 indicates that two topics are completely different.

We further investigate deleted weibos that were likely censored [5] by checking if they belong to topics which appear on Twitter. In total, 1,558 deleted weibos can be classified into the Top 100 topics on Twitter. We re-rank the topics in accordance with the frequency of deleted weibos. The Kendall rank correlation coefficient between the top 100 topics for all tweets and for deleted weibos is $\tau = 0.31$, with p-value 4×10^{-6}. In Table 2, we list Top 10 topics for deleted weibos on Twitter's vector space. Compared with popular topics on Sina Weibo, the deleted weibos are significanlty more likely to discuss political issues.

CONCLUSION

The social attention of online users from the same cultural backgrounds but living in different countries might be different due to the changes of social environments. In this study, we take the first steps toward understanding such differences. Sina Weibo is used almost exclusively within China while most Chinese language use of Twitter occurs almost

Rank	Sina Weibo	in English	%	Twitter	in English	%
1	三国来了	an online game	0.51	陈光诚	Chen Guangcheng (a Chinese civil rights activist)	3.56
2	林峰	Raymond Lam (a singer from Hongkong)	0.39	乌坎	Wukan protest	2.56
3	晚安/早安	good morning/night	0.38	Freetibet	Free Tibet	1.62
4	微博客户端	Sina Weibo app	0.36	李旺阳	Li Wangyang (a Chinese dissident labor rights activist)	0.09
5	搞笑	joke	0.07	温云	@wenyunch	0.97
6	美图秀秀	Meitu (an iOS/Android app to edit pictures)	0.04	抗暴	Tibetan Uprising Day	0.88
7	有奖转发	re-posting to win a prize	0.04	达赖喇嘛	Dalai Lama	0.68
8	WeicoLomo	An iOS/Android app for Weibo to record video	0.04	钓鱼岛	Uotsuri Jima Diaoyu Dao / Diaoyutai	0.66
9	韩庚	Han Geng (a Chinese singer and actor)	0.03	ipadgame	iPad game	0.61
10	新版微博	new version of Sina Weibo	0.02	武士朝代	an Andorid game	0.48

Table 1: Top 10 Topics on Sina Weibo and Chinese language Twitter in 2012**.**

rank	topic on Twitter	in English
1	fb	Facebook
2	乌坎	Wukan protest
3	np	Now Playing
4	hitbag	-
5	AutoShare	-
6	GFW	Great Firewall
7	HK71	Hong Kong 1 July march
8	bot	-
9	A片	Adult movie
10	JapanLife	-

Table 2: Top 10 topics in deleted weibos on Twitter's vector space .

exclusively outside Chinese borders. By comparing the most popular topics in these two platforms we can, for the first time, observe how the interests of two populations, with similar cultural backgrounds, differ. Surprisingly, we find that there is very little overlap between the two attention profiles. Weibo users speak mostly about popular culture and games while Twitter users focus mostly on political issues.

The reasons behind this divergence are hard to discern but can likely be attributed to one of two factors: lack of interest for political topics within China or a high degree of self-censorship that prevents Chinese from discussing politics in public. A small indication towards this second hypothesis is the list of topics seen in deleted Weibos (see Table 2) that have higher political content. It is worth to remark that our algorithm of detecting topics still depends on # tags, and some of such # tags may not necessarily be a social topic but likely represent some commercial web/mobile applications. Manual annotations could be included in the future work to improve the topic detection results. Another key datapoint we are missing to fully clarify this question is the usage of foreign VPN services to reach Twitter. An analysis of this interesting factor will be the subject of future study. The proposed methodology in this paper can be easily applied to any other languages across different online conversation platforms if data are available.

Acknowledgments

BG thanks the Moore and Sloan Foundations for support as part of the Moore-Sloan Data Science Environment at NYU.

REFERENCES

[1] Q. Gao, F. Abel, G. Houben, and Y. Yu. A comparative study of users' microblogging behavior on Sina Weibo and Twitter. In *User modeling, adaptation, and personalization*, pp 88–101, 2012.

[2] O. Phelan, K. McCarthy, and B. Smyth. Using Twitter to recommend real-time topical news. In *RecSys'09*, page 385, 2009.

[3] P.T. Metaxas, E. Mustafaraj, and D. Gayo-Avello. How (not) to predict elections. In *IEEE SocialCom*, page 165, 2011.

[4] A. Rauchfleisch and Mike S. Schäfer. Multiple public spheres of Weibo: a typology of forms and potentials of online public spheres in China. *Information, Communication & Society*, 18:139, 2014.

[5] K. Fu, C. Chan, and M. Chau. Assessing censorship on microblogs in China: discriminatory keyword analysis and the real-name registration policy. *Internet Computing*, 17(3):42–50, 2013.

[6] K. Fu and M. Chau. Reality check for the Chinese microblog space: a random sampling approach. *PLoS One*, 8(3):e58356, 2013.

[7] D. Bamman, B. O'Connor, and N. Smith. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3), 2012.

[8] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *WI-IAT*, 2010.

[9] G. Li, K. Meng, and J. Xie. An improved topic detection method for chinese microblog based on incremental clustering. *Journal of Software*, 8(9):2313–2320, 2013.

[10] O. Tsur, A. Littman, and A. Rappoport. Efficient clustering of short messages into general domains. In *ICWSM*, 2013.

[11] X. Shuai, Xi. Liu, T. Xia, Y. Wu, and C. Guo. Comparing the pulses of categorical hot events in twitter and weibo. In *HT*, 2014.

[12] N. Furman, D. Goldberg, and N. Lusin. Enrollments in Languages Other Than Engish in United States Institutions of Higher Education, Fall 2009. Technical report, Modern Language Association, 2010.

[13] B. Gonçalves and D. Sánchez. Crowdsourcing dialect characteriation through twitter. *PLoS One*, 9:E112074, 2014.

[14] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *WWW*, pages 249–252. ACM, 2011.

[15] M.M. Candless. Chromium Compact Language Detector. http://code.google.com/p/chromium-compact-language-detector, 2012.

[16] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS One*, 8(4):e61981, 2013.

[17] Jianfan. https://code.google.com/p/python-jianfan, 2013.

[18] Jieba: Chinese word segmentation module. https://github.com/fxsjy/jieba, 2014.