Twibo: Comparing Very Large Communities via Massive Social Media Datasets

Tian Xia[1], Miao Chen[2], Xiaozhong Liu[2]
[1]School of Information Resource Management, Renmin University of China, Beijing, 100872, China
[2]Scholl of Informatics and Computing, Indiana University Bloomington, Bloomington, IN 40475

Abstract
Online social media are becoming the standard infrastructure for social communication and dissemination of information. As social media platforms not only passively provide infrastructure but also actively perform algorithmic curation for their profit and user experience, an important concern, often called "filter bubble" arises: people are trapped in their own personalized bubble–being exposed only to the opinions that conform their beliefs and political positions, thus potentially creating social polarization and information "islands". Although the adoption of social media is an international phenomenon, language difference and policy barrier also create information islands. The goal of this paper is to develop methods/system to cross-link concepts and communities in different social media, and leverage them to study the extent and impact of filter bubbles. To accomplish this goal, the main objectives in this paper are to develop text/graph mining methods to connect concepts and entities in Twitter and Weibo through Wikipedia knowledge base; and to compare two social media in the dimension of topics and networks to quantify the significance of language bubble.

Keywords: Social media; computational social science; community comparison; Twitter; Weibo

Contact: xiat@ruc.edu.cn

1 Background and Significance

The proliferation of social media is bringing about significant changes in how people perceive and make sense of their world (Pak & Paroubek, 2010; Shuai, Liu, Xia, Wu, & Guo, 2014). Millions of individuals communicate with each other through a variety of social media platforms, sharing pertinent information about the world as well as the most minute details of their social lives, thereby collectively shaping each others' culture and worldview. Studies of cultural sense-making are increasingly focused on how social media affect how we communicate, when, and with whom, and how their influence is modulated by a range of intercultural, political, and social factors. Fortunately, social media platforms at the same time provide a unique opportunity to study human communication and sense-making *in vivo* by virtue of the scale, quantity, and detail of data that they generate about the social and informational environments of millions of individuals. The development of social media mining algorithms and methods has over the past decade made significant contributions to social science as well as computer science (Baucom, Sanjari, Liu, & Chen, 2013; Jansen, Zhang, Sobel, & Chowdury, 2009; Morozov, 2012; Turkle, 2012).

In spite of social media being an inherently borderless and international online phenomenon, it is still marked by strong geographical and cultural divisions induced by linguistic, social, and policy barriers. Twitter and Facebook, for example, are strictly forbidden in mainland China due to political reasons [1], meaning that 21.97% [2] of internet users in the world are excluded from participating in these platforms. This policy has resulted in the fact that the world's second largest microblogging system—Sina Weibo[3], with more than 249 million active users in 2014 [4]—is serving mostly Chinese users since the default system language is Chinese. As a result, the world's social media platforms are to a large degree segregated according to linguistic, social, cultural, geo-graphical, and political barriers, in spite of their promise to transcend such distinctions and in spite of the many areas in which topics, users, interests, and preferences do actually overlap.

In this paper, by leveraging the gap between different social media systems, we propose and develop a new system, Twibo, to enable social comparison between Twitter and Weibo. Behind the system, we employed

[1] Wikipedia block list:http://en.wikipedia.org/wiki/Lis_of_websites_blocked_in_China
[2] Statistics of China Internet Users (2014) from internet live stats site
[3] Weibo: World second largest microblogging system. Default language is Chinese.
[4] CNNIC: http://www.cnnic.cn/hlwfzyj/hlwxzbg/201502/P020150203551802054676.pdf

novel method to interconnect very large Twitter and Weibo datasets from text and graph viewpoints. By using novel text and graph mining algorithms, we extract information and knowledge from a variety of very large social media datasets, i.e., Twitter and Weibo, and we propose novel factors to efficiently compare massive users from different communities, a.k.a. **Computational Social Comparison**. For instance, by using very large Twitter and Weibo datasets, with the proposed method and system, social scientists can collect the evidence to answer the questions like *"What are the similarly/differences between the responses of the Chinese and US community when relating to socio-economic and political news events, such as 'North Korea develops nuclear weapons', 'Hillary Clinton participates present election' or 'US legalizes same sex marriage'?"* with a very low cost.

More specifically, we investigate the following research questions in this paper:

- [**RQ1**] Developing methods to cross-link concepts and topics that are discussed in different social media sites and different languages through Wikipedia, the most prominent multi-lingual knowledge-base. This effort will require sophisticated text and heterogeneous graph mining algorithms, natural language processing tools, and knowledge representation methods to link users and information that are not physically connected nor cannot be identified with existing information identification methods.
- [**RQ2**] By using the theoretical and empirical framework developed in RQ1, we will quantify the similarities and differences of the topical attention and responses in multiple social media bubbles/platforms, to study their evolving dynamics in terms of topics, sentiment, and information diffusion. We will use the outcomes of RQ1 to conduct effective statistical and systemic comparison of social media environments with their counterparts, across a large sample of observed social media bubbles. The proposed method will help social scientists to answer their questions via big social data comparison.

2 Previous Studies in Social Media Mining and Comparison

When Pariser defined the concept of filter bubble (Pariser, 2011) as "the personal ecosystem of information", it intended to describe a phenomenon that is more likely caused by algorithms, e.g., Google personalized search and Facebook's personalized news stream or Twitter limited user's information access to those "local topics" and narrow her outlook. Prior studies indeed have shown that users may get less exposure to conflicting viewpoints and are isolated intellectually in their own informational bubble (Weisberg, 2011).

This problem of filter bubble and social media bubbles raises a critical questions regarding the public access to the information and polarization: how strong the bubbles are and how can we let people to access information outside their bubble? It also puts technical challenges: how can we map concepts and topics across social media, particularly when they are in different languages? Prior studies, e.g., (Shuai et al., 2014), showed that social media data enable social scientists to answer very challenging questions with a low lost. However, processing very large social media datasets while designing sophisticated mining algorithms can be challenging.

In the prior studies, studies about integrating and comparing multiple social media sites data are quite sparse. Take Twitter and Weibo comparison as an example, not until recently, some researchers investigated the basic statistics of Weibo and Twitter corpora, i.e., basic sentiment comparison (Gao, Abel, Houben, & Yu, 2012), hashtag distribution comparison (Gao et al., 2012; Li et al., 2012), and user gender comparison (Guan et al., 2014). All of these studies utilized small datasets. For instance, (Guan et al., 2014) explored the statistics of Weibo users by collecting messages of 32 Weibo users. Meanwhile, no prior study investigated topical or categorical Twitter and Weibo comparison during hot events, which is important; the nature of Weibo (or Twitter) users' responses to e.g. Political news can be very different from that of Science or Entertainment (Shuai et al., 2014).

3 Twibo: Compare Twitter and Weibo

As aforementioned, language, network and policy restrictions can keep users and information isolated in social network filter bubbles. However, users from different social media systems may be interested in similar topics. For instance, the populations of Weibo and Twitter users may both be interested in *"Obama's asian policy"* or *"Gay marriage legalization"*, but the topic's popularity, dynamics, and sentiment information may vary significantly. The context of topics associated with this particular topic may differ as well. Understanding and comparing how different group of users, from different filter bubbles, interact with similar topics therefore remains an interesting but challenging research question. In this paper, we propose a new system **Twibo**. The goal is straightforward and clear: help social and information scientists to effectively compare China and US communities by interconnecting very large Twitter and Weibo datasets. As shown in figure 1, Twibo has two parts: frontend for target users and backend for system designers.

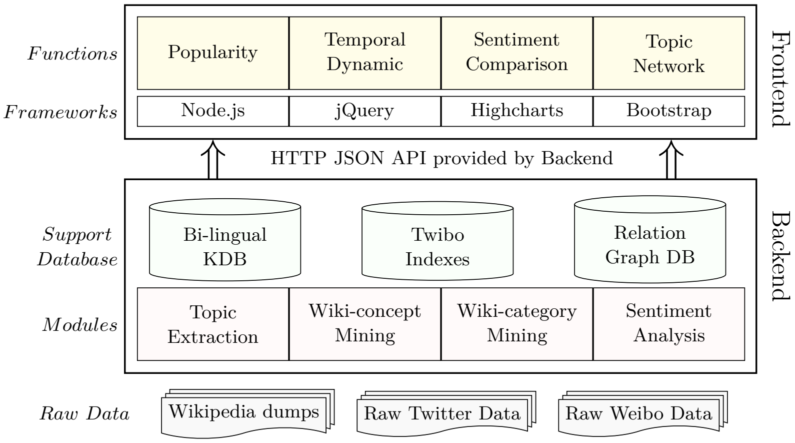


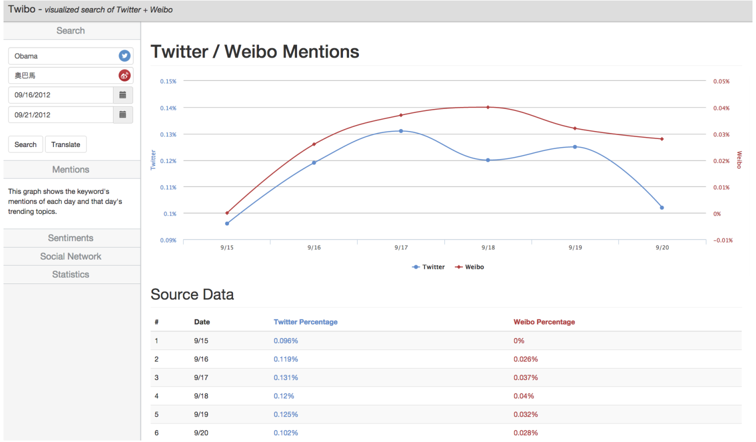
Figure 1: Components and Workflow of Twibo.

3.1 Frontend Interface

In the system frontend (prototype), users can compare chronological Twitter and Weibo datasets via the following indicators:

1. Popularity

   The popularity of a concept or query, E_p, represents the degree of collective attention it receives, which can be characterized by the sum of daily or hourly probabilities that this concept or query is mentioned on Twitter or Weibo during the time period under consideration. We can further estimate the concept categorical popularity for a given target category, C, by averaging the values of $P(E_p)$ for all $E_p \in C$. A system screenshot is depicted in Figure 2.



This graph shows the keyword's mentions of each day and that day's trending topics.

Source Data

#	Date	Twitter Percentage	Weibo Percentage
1	9/15	0.096%	0%
2	9/16	0.119%	0.026%
3	9/17	0.131%	0.037%
4	9/18	0.12%	0.04%
5	9/19	0.125%	0.032%
6	9/20	0.102%	0.028%

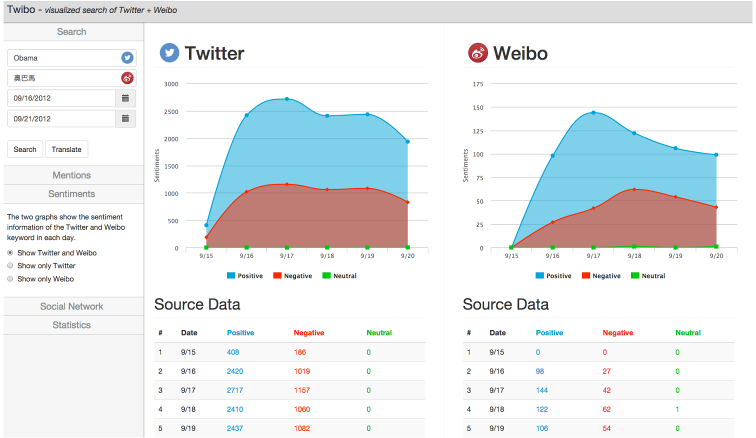
Figure 2: Popularity and Temporal Dynamic Comparison (*Query: Obama*).

2. Temporal Dynamic

By using user-topic dynamic interesting probability distributions, and Twitter and Weibo bridge index, we can compare topic temporal dynamic. For instance, we can compare the topic peak, when the discussion of topic reaches the maximum degree of interest on Twitter and Weibo, and how the spiky discussion date is temporally related to the topic. Meanwhile, Twibo can depict the community interest probability change over time on the target topic (see Figure 2).

3. Sentiment Comparison

Based on the sentiment index for each Twitter or Weibo message in the indexation (positive or negative probability), we can aggregate and compare user or community sentiment for a given topic or event. Meanwhile, we can also compare the sentiment change on Twitter and Weibo over time given the target topic or event. Sentiment comparison results can be particularly useful for social scientists since they provide an orthogonal signal with respect to how a particular topic is emotionally evaluated by the community under consideration. See Figure 3.



Obama

奥巴馬

09/16/2012

09/21/2012

The two graphs show the sentiment information of the Twitter and Weibo keyword in each day.

● Show Twitter and Weibo
○ Show only Twitter
○ Show only Weibo

Twitter — Source Data

#	Date	Positive	Negative	Neutral
1	9/15	408	186	0
2	9/16	2420	1019	0
3	9/17	2717	1157	0
4	9/18	2410	1060	0
5	9/19	2437	1082	0

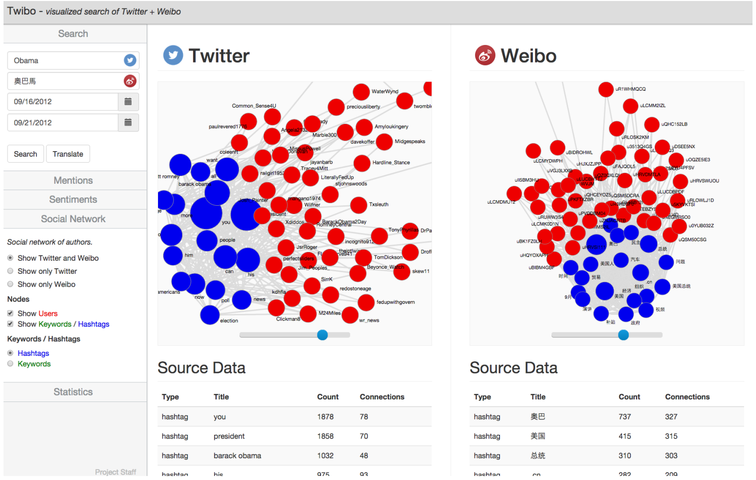
Weibo — Source Data

#	Date	Positive	Negative	Neutral
1	9/15	0	0	0
2	9/16	98	27	0
3	9/17	144	42	0
4	9/18	122	62	1
5	9/19	106	54	0

Figure 3: Sentiment Comparison (*Query: Obama*).

4. Associated Topic Network

For the target query topic, i.e., represented by an event or concept, researchers may be interested in a number of associated topics. Given the same topic, users from different social media may be interested in different associated topics for different time periods. For this study, we will compare and visualize different associated topics (network) given the target query. See Figure 4. Currently, each topic is represented by a hashtag.



Obama

奥巴馬

09/16/2012

09/21/2012

Social network of authors.

● Show Twitter and Weibo
○ Show only Twitter
○ Show only Weibo

Nodes

☑ Show Users
☑ Show Keywords / Hashtags

Keywords / Hashtags

● Hashtags
○ Keywords

Twitter — Source Data

Type	Title	Count	Connections
hashtag	you	1878	78
hashtag	president	1858	70
hashtag	barack obama	1032	48
hashtag	his	975	93

Weibo — Source Data

Type	Title	Count	Connections
hashtag	奥巴	737	327
hashtag	美国	415	315
hashtag	总统	310	303
hashtag	.cn	282	209

Figure 4: Topic Network Comparison (*Query: Obama*).

3.2 Backend Algorithm

As shown in figure 1, all text/graph mining and data processing algorithms are implemented at backend. The backend API enable frontend search and visualization functions via HTTP and JSON format data transmission. In order to achieve the goals of this system, we will interconnect and bridge Twitter and Weibo by employing a global or domain-specific multilingual knowledge-base. Over the past decade, Wikipedia has become an increasingly important store of the world knowledge. It provides unique features that can potentially integrate different kinds of social media data for cross-language information discovery. Therefore, we utilize Wikipedia as a global knowledge-base because of the following reasons.

Multi-lingual Wikipedia provides concept definitions in multiple languages. For instance, in Wikipedia 2014 May dumps, we find 380,000 important concepts (those that have at least 3 incoming links) defined in both English and Chinese, which cover essential universal knowledge base. Take the Weibo and Twitter instance, while the English article (content) can be projected into Twitter topic space, the Chinese counterpart for the same concept can be used to bridge the Weibo topics by using sophisticated text mining algorithms.

Concept hierarchies All concepts in Wikipedia are interlinked via Wikipedia hierarchical categories and incoming/outgoing links among Wikipedia articles. For instance, the concepts *“NBA”* and *“LeBron James”* are connected via the path *“[Wikipedia Concept: NBA]* $\xrightarrow{b}$ *[Wikipedia Category: Basketball]* $\xleftarrow{b}$ *[Wikipedia Concept: LeBron James]”* and path *“[Wikipedia Concept: LeBron James]* $\xrightarrow{l}$ *[Wikipedia Concept: Basketball]”* ($\xrightarrow{b}$ represents *“belong to”* relation, and $\xrightarrow{l}$ represents *“link to”* relation). In other words, all concepts in Wikipedia are inter-connected through heterogeneous links and cross-language equivalents. From this viewpoint, all the topics, in multiple filter bubbles, are also interlinked via Wikipedia serving as a bridge, which provides important potential for random walks on the heterogeneous graph.

Spoken language recognition Most social media textual data are generated in spoken language, and Wikipedia provides spoken-language-like Redirected Links in different languages. For instance, the concept *“Patient Protection and Affordable Care Act”* can be redirected from *“obamacare”* and article “Barack_Obama” can be redirected from *“‘o’bama”* or *“obamma”* (spell mistake), which can be helpful for social media text mining.

In this proposed work three methods are employed to link Wikipedia concepts to different topics and users across different languages and various social media: **Entity Recognition (NR)**; **Explicit Semantic Analysis (ESA)** (Gabrilovich & Markovitch, 2007); and **Explicit Semantic Path Mining (ESPM)** (Xia, Chen, & Liu, 2014). The former method relies on an exact match method by Wikipedia title, and the latter two methods are semantic match techniques that leverage Wikipedia article content, article category and concept links.

In this step, each topic k is associated with its text context, and, by using various methods, each topic can be indexed by a number of Wikipedia articles $P(A_1|k), P(A_2|k), ..., P(A_n|k)$ or Wikipedia categories $P(C_1|k), P(C_2|k), ..., P(C_m|k)$, where $P(A_i|k)$ and $P(C_j|k)$ are the probability of the Wikipedia article A_i or category C_j given the topic k. **Entity Recognition** utilizes mainly Wikipedia article title and their redirected page titles, and greedy match and language model is used calculate $P(A_i|k)$. For this method, a Wikipedia article title (or redirected titles) should explicitly exist in the topic text context. Meanwhile, we will use Wikipedia disambiguation and Wikipedia's link structure to enhance performance of our entity recognition performance. This method has been proved useful in Wikifier system (Ratinov, Roth, Downey, & Anderson, 2011).

3.3 Data for Preliminary System Testing

We collect one month Twitter and Weibo posts from Sep 15, 2012 to Sep 21, 2012 for test purpose. All Weibo posts are fetched through Weibo open API, about 3 million each day. For Twitter, we have data use agreement, and we sample 40 million tweets each day of that week. Therefore, more than 300 million posts are processed in this prototype system.

English and Chinese Wikipedia snapshot dumps of March 4, 2014 are also used in Twibo. These two dump files can be downloaded via Wikipedia website(http://dumps.wikimedia.org/). That English dump contains 10 million articles while Chinese dump contains 1 million. After language alignment process, we get 400 thousand articles which have both Chinese and English content.

4 Conclusion

In this paper, we propose Twibo system to assist social and information scientists to compare very large social media datasets, which mirrors the comparison of China and US. While different comparison indicators are utilized, social and information scientists can easily issue any query to address their hypothesis. At the backend, we use complex and efficient models to process very large Twitter and Weibo datasets. Most existing studies focus on a single service and a single language, mainly because of the lack of methods to cross-link concepts and online communities between social media. This project will offers new methods to address this challenge and will open up a new avenue of research on cross-cultural social media studies. The algorithms and datasets created in this project will not only kindle new algorithms from computer science but also provide social scientists with data and toolsets to ask novel sociological and cultural questions. This research will contribute to the ability of underrepresented groups to fully participate in the global cultural and political conversation that is now increasingly taking place online and through social media. Our results may mitigate the digital divide that results from social and linguistic disparities.

References

Baucom, E., Sanjari, A., Liu, X., & Chen, M. (2013). Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In Proceedings of the 2013 international workshop on mining unstructured big data using natural language processing (pp. 61–68).

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Ijcai (Vol. 7, pp. 1606–1611).

Gao, Q., Abel, F., Houben, G.-J., & Yu, Y. (2012). A comparative study of users' microblogging behavior on sina weibo and twitter. In User modeling, adaptation, and personalization (pp. 88–101). Springer.

Guan, W., Gao, H., Yang, M., Li, Y., Ma, H., Qian, W., ... Yang, X. (2014). Analyzing user behavior of the micro-blogging website sina weibo during hot social events. Physica A: Statistical Mechanics and its Applications, 395, 340–351.

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11), 2169–2188.

Li, D., Zhang, J., Sun, G. G.-z., Tang, J., Ding, Y., & Luo, Z. (2012). What is the nature of chinese microblogging: Unveiling the unique features of tencent weibo. arXiv preprint arXiv:1211.2197.

Morozov, E. (2012). The net delusion: The dark side of internet freedom. PublicAffairs Store.

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In The international conference on language resources and evaluation.

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin UK.

Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1 (pp. 1375–1384).

Shuai, X., Liu, X., Xia, T., Wu, Y., & Guo, C. (2014). Comparing the pulses of categorical hot events in twitter and weibo. In Acm hypertext.

Turkle, S. (2012). Alone together: Why we expect more from technology and less from each other. Basic Books.

Weisberg, J. (2011). Bubble trouble: Is web personalization turning us into solipsistic twits. Slate. com, 10–06.

Xia, T., Chen, M., & Liu, X. (2014). Explicit semantic path mining via wikipedia knowledge tree. In The annual meeting of the association for information science and technology.