The Media Feature Analysis of Microblog Topics

Xing Chen[1], Lin Li[2], and Shili Xiong[3]

School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
{rebecca_lymx,cathylilin}@whut.edu.cn
Communication University of China
slxiong@cuc.edu.cn

Abstract. As microblogging grows in popularity, many research articles are exploring and studying the micro-blogs, especially the English micro-blogging, i.e., twitter. However, Chinese micro-blog service starts rather late and a few research is about its data and characteristics. In this paper, we give out the media feature analysis of microblog topics. Firstly we present our observations tweets and users from Sina by crawling 14 topics and their 74,662 tweets and give out the topic evolution in a certain interval. Then considering the microblogs under a topic exist a lot of redundant information, so in order to reduce the trainning datasets for studying on microblogging , we respectively select diffierent data source as our datasets and give out the evaluation method. We have also studied the microblog semantic extraction based on the topic model of LDA (Latent Dirichlet Allocation). we conclude that the active period of most micro-blog topics is about a month and the out-dated topics will be replaced by the upcoming and related new topics and find that the tweets that appeared in the peak time or the tweets from authenticated users can reflect the whole tweets situation of a topic. However, due to the microblog text is so short that LDA for semantic extraction is not ideal.

Keywords: Microblogging, Media Feature, LDA.

1 Introduction

Microblogging is a new form of communication in which users describe their current status in short posts distributed by instant messages, mobile phones, email or the Web. One of the popular microblogging platforms is Twitter [1]. This model was co-founded by Evan Williams,etc and started in March 2007. The use of Twitter of Obama in his presidential campaign has become an important turning point for the development of the Twitter history. When Michael Jackson died, within one hour,the message has risen up to 65000 on Twitter, which was a peak of development for Twitter. However, Chinese Mirco-blogging service has originated in the English twitter model. The arrival of Sina Micro-blog service marks that micro-blogging has formally crowed into Chinese people’s sights. Updates or posts are made by succinctly describing one’s current status through a short message that is limited to

B. Hong et al. (Eds.): DASFAA Workshops 2013, LNCS 7827, pp. 193–206, 2013.

140 characters. Topics range from daily life to current events, news stories, and other interests. Instant messaging (IM) tools including QQ and MSN have features that allow users to share their current status with friends on their buddy lists. Microblogging tools facilitate easily sharing status messages either publicly or within a social network.

By May 2012, Sina Micro-blog users has exceeded 300 million. Nowadays microblogging as a widely used medium platform, its diverse features meet the people's information, interpersonal information and other aspects of the new requirements. Users usually revolves around a certain hot topic to spread their own opinions. So studying the microblog media feature is becoming more and more important.

To study the the characteristics of Chinese tweets and users and its power as a new medium of information sharing, We have crawled 14 topics and its 74,662 tweets from Sina.We analyze the topic evolution and compare the methods that getting data from different sources, then study the microblog semantic extraction based on the topic model of LDA.

To the best of our knowledge this work is the first study on the Chinese mirco-blog topic sphere. This paper is organized as follows. We first discuss some related works in Section 2. Section 3 describes our topic evolution and compare how to select our studying data source.we study the Selection and Preprocessing of Dataset in Section 4. Last ,we give out Experimental Results and Analysis in Section 5,and conclude our work in Section 6.

2 Related Work

The rising popularity of online social networking services has spurred research into their characteristics and recent work has forayed into characteristics beyond crawled data [2], [3]. Recently, there are a number of articles about exploring and studying the micro-blogging, especially the English microblogging, i.e., twitter. For example, (Krishnamurthy, Gill, & Arlitt, 2008) summarize general features of the Twitter social network such as topological and geographical properties, patterns of growth, and user behaviors. Others such as (Java, et al., 2007), argue from a network perspective that user activities on Twitter can be thought of as information seeking, information sharing, or as a social activity. Newman et al. [4] have made the first quantitative study on the entire Twitter sphere and information diffusion on it. They studied the topological characteristics of Twitter and its power as a new medium of information sharing and have found a non-powerlaw follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. In 2007, Java et al. [5] conduct preliminary analysis of Twitter. They find user clusters based on user intention to topics by clique percolation methods. Krishnamurthy et al. [6] also analyze the user characteristics by the relationships between the number of followers and that of followings. In 2010, Huberman et al. [7] reports that the number of friends is actually smaller than the number of followers or followings. Jansen [8] conducts preliminary analysis of word-of-mouth branding in Twitter.

With the rise of Sina microblogging, more and more researchers works to study the new Chinese micro-blog media. Liu et al. [9] combine a translation-based method with a frequency-based method for keyword extraction. They extract keywords for microblog users from the largest microblogging website in China, Sina Weibo. Our work is also based on the background of chinese Micro-blog topics from Sina. We describe our topic evolution and compare how to select our studying data source , then study the microblog semantic extraction based on the topic model of LDA .

3 Dataset Description

In this section, we first introduce how to extract microblog data by crawling. Then we discuss and analysis the distribution of the tweets and users for each topic in the time interval of 20 days.

3.1 Extracting Microblog Data by Crawling

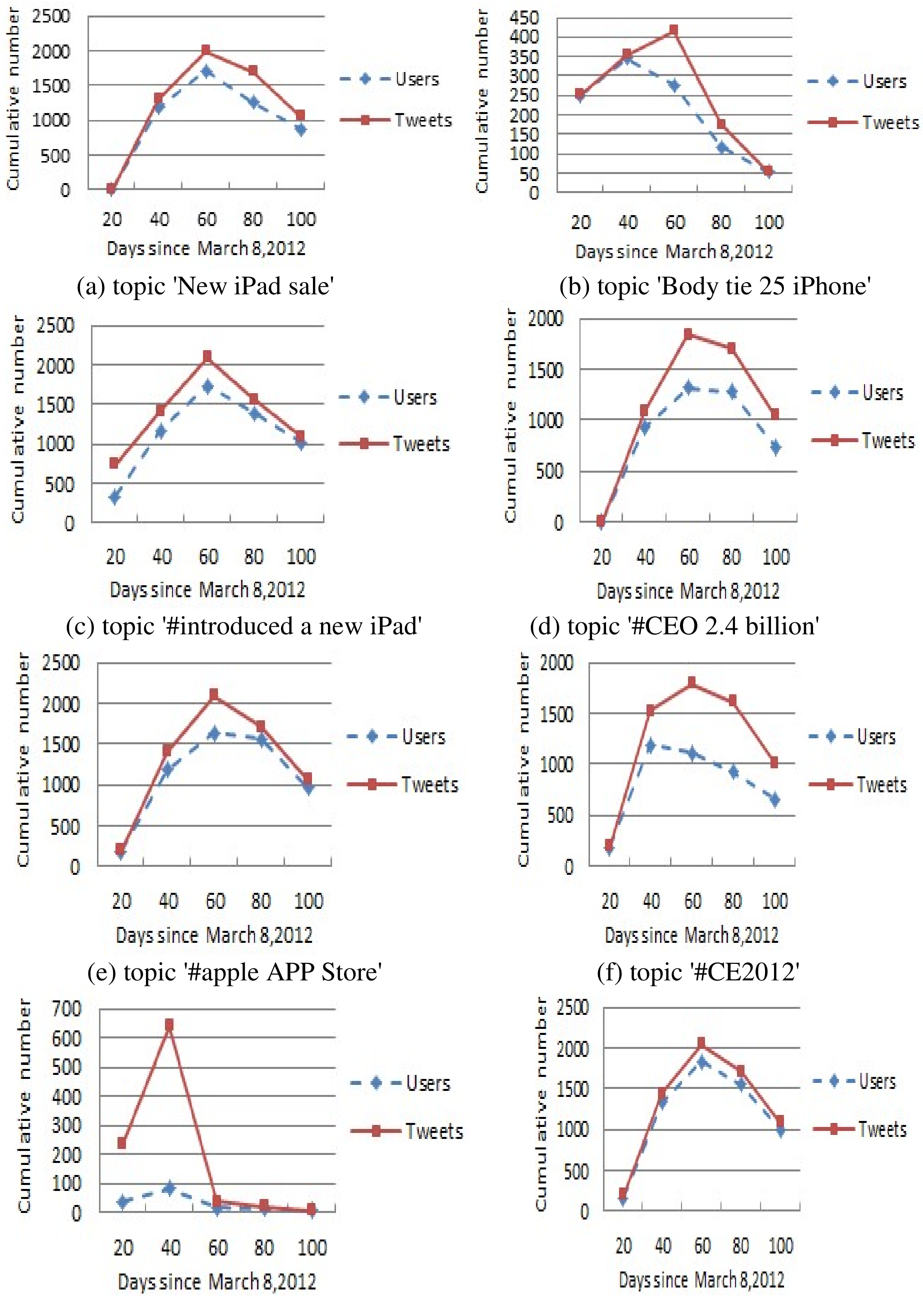
For benefiting from our professional point of view of computer science technology, we select the technology/IT Internet as our data source. The chosen 14 topics all are the most popular topics at the crawling time as our experimental data. We extract the tweets of these 14 topics. Sina microblog only gives 10 pages space capacity to display the tweets of each micro-topic and each page only exists 20 tweets. From the end of March 2012, we start collecting micro-topic data. In order to ensure nonduplication of data, nearly every two days, we download the html webpages of each Micro-topic, and then save them in the local disk folder as .txt file format. By the end of June 2012, we obtain almost 3750 web pages. But how to extract our needed information from these html files? Here we use HtmlParser[1]. HtmlParser is an open-source project used to parse the HTML document. It is small, fast, simple and has a powerful function.

3.2 Topic Evolution

In order to discuss the distribution of users and tweets in a period of time. In Figure 1, we plot a graph that every 20 days the numbers of users and tweets in a topic.

First, from the overall curve, we can see that at the beginning the number of the tweets and the users is increasing gradually as the time goes on. It means that when a topic is just emerging, it will attract a lot of people. Then, the number of tweets and users reached the maixmum which means the peak period is arrived. After the peak period, the number of users and tweets begin to reduce slowly until the topic die, which means no users are involved in this topic.

[1] http://htmlparser.sourceforge.net/



(a) topic 'New iPad sale'

(b) topic 'Body tie 25 iPhone'

(c) topic '#introduced a new iPad'

(d) topic '#CEO 2.4 billion'

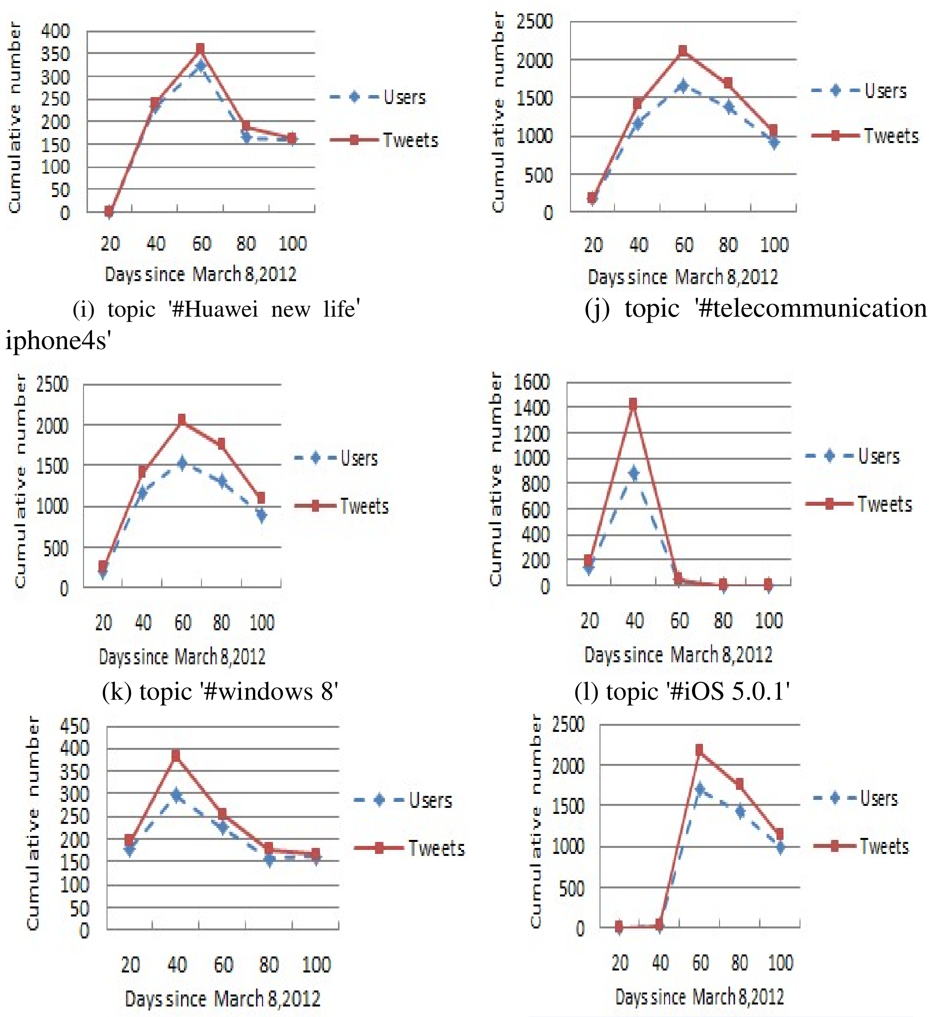
(e) topic '#apple APP Store'

(f) topic '#CE2012'

(g) topic '#Tablet PC'

(h) topic '#Kodak bankruptcy protection'

Fig. 1. The Evolution Curves of users and tweets in different topics



(i) topic '#Huawei new life' (j) topic '#telecommunication
iphone4s'

(k) topic '#windows 8' (l) topic '#iOS 5.0.1'

(m) topic '#HTC' (n) topic '#Facebook'

Fig. 1. *(Continued)*

4 Selection and Preprocessing of Dataset

In this section, we first get data source from diffierent aspects.We collected a sample of almost three months tweets between March 25th and June 17th from Sina Microblogging platform, which contains 22,724 tweets from authenticated users and 63,354 tweets from total users, and we also collected tweets that from diffierent period for each of topics. It is worth mentioning that the so-called authenticated users mainly includes the users of Sina microbog VIP, Sina approved Sina agencies and authenticated Sina individuals. Then we give out our approach for text processing and evaluation method. Finally we present semantic extraction for all tweets under 14 topics based on LDA.

4.1 Text Processing and Evalution Method

For each of topics and each of tweets, we conduct the preprocessing of removing stopwords and chinese participle preprocessing. To complete text processing, two NLP tools (i.e.,MyT xtSegT ag andMyZiCiFreq) is used[2]. The next step is to remove all stopwords. We remove the words in our tweets for each topic that appeared in a stopwords list. After making stopwords processing for authenticated tweets, the whole tweets and the tweets that from diffierent period of each topic, we make Chinese participle preprocessing for the filtered tweets by a set of word segmentation and POS tagging tool which is named MyT xtSegT ag. The big advantage of this software is that it can identify proper and newly appeared nouns and minimize the word granularity, such as the new word of Mirco-Letters, weixin(a mobile phone chat software) which is a new application launched by Tencent company in 2011. If the option of starting proper nouns is not selected, then after making participle processing, the word of weixin will be divided into two words that ”Micro” and ”letters”. So using this software can improve the precision and accuracy of participle processing results. Last, we save these produced words in a .txt formatted file, making a preparation for word frequency statistics and analysis. We adopt a word frequency statistic tool named MyZiCiFreq. This software can not only make character frequency statistics but also make word frequency statistics.

Then word frequency statistics are done and top 10 representative nouns or verb-nouns are extracted. In most query suggestion articles, people usually adopt query logs to determine possible query suggestions. But nowadays, with the rise of social network media, many people begin to take microblog as their study background instead of traditional query logs. For a given query, the precision of a query suggestion method is defined as the fraction of suggestions generated that are meaningful. Note that since an exhaustive set of all possible suggestions for a given query is not available, recall cannot be computed. Also, for the query suggestion task, precision is a much more important metric than recall as the number of suggestions that can be offered is limited by the screen space. Here we take our selected top 10 representative nouns or verb-nouns as suggestion list. Hence in order to measure the quality of query suggestion, we decide to select the precision value as our evaluation method which is defined as

$$\text{Precision@N}=\frac{\#\text{related words in a suggestion list}}{\text{N}} \tag{1}$$

In Equation 1, we take the extracted top 10 representative nouns or verbnouns as the words that can represent a certain topic. We manually judge whether these words can be considered to accurately reflect the topic. The precision value of each topic is computed.

[2] They are recommended by the website of http://www.chinalanguage.gov.cn/index.htm

Table 1. Explanation for the abbreviations in Table 2

	Specific meaning
Authenticated tweets	The tweets that from authenticated users
Total tweets	The tweets that from all of users involved in a topic
Sum1	The sum of tweets that from these period 3.08--3.23,4.10-4.25,4.26-5.11
Authenticated/Total	The tweets that from authenticated users/The tweets that from all of users involved in a topic
Atop10 precision	The precision value of top10 terms that produced by the tweets that from authenticated users
Ttop10 precision	The precision value of top10 terms that produced by the tweets that from all of users involved in a topic
Top10/Top5	The precision value of top10 terms and The precision value of top5 terms

Table 2. The precision values of each topics

Topics	New ipad sale	Iphone news	Ipad show	Apple ceo salary	App Store	CES2012	HTC
Authenticated tweets	2079	471	2420	1831	3005	1866	487
Total tweets	6874	1419	8313	6324	8175	7453	1385
Authenticated/Total	0.3024	0.3319	0.2911	0.2895	0.3676	0.2504	0.3516
ATop10 precision	0.4	0.6	0.4	0.5	0.4	0.4	0.5
TTop10 precision	0.5	0.6	0.4	0.5	0.4	0.3	0.5
ATop5 precision	0.3	0.4	0.3	0.2	0.3	0.3	0.4
TTop5 precision	0.3	0.3	0.3	0.4	0.4	0.2	0.5

Topics	Tablet pc	Kodak bankrupt	Huawei for new life	Iphone 4s sale	Windows 8	iOS jailbreak	Facebook
Authenticated tweets	443	2472	340	2491	2454	604	1761
Total tweets	1556	9084	1448	8085	7125	1932	5639
Authenticated/Total	0.2847	0.2721	0.2348	0.3081	0.3444	0.3126	0.3123
ATop10	0.4	0.4	0.4	0.3	0.5	0.6	0.3

precision							
TTop10 precision	0.4	0.4	0.4	0.3	0.4	0.6	0.3
ATop5 precision	0.4	0.4	0.4	0.3	0.4	0.6	0.3
TTop5 precision	0.4	0.2	0.2	0.3	0.3	0.5	0.2

Topics Top10/Top5	New ipad sale	Iphone news	Ipad show	Apple ceo salary	App Store	CES2012	HTC
3.08--3.23		0.5/0.3	0.3/0.3		0.1/0.1	0.2/0	0.4/0.2
3.24--4.09	0.4/0.3	0.4/0.2	0.1/0.1		0.4/0.3	0.5/0.4	0.6/0.4
4.10--4.25	0.4/0.3	0.6/0.4	0.2/0.2	0.4/0.3	0.4/0.4	0.5/0.3	0.5/0.4
4.26--5.11	0.4/0.3	0.5/0.3	0.1/0.1	0.4/0.4	0.4/0.4	0.5/0.3	0.5/0.4
5.12--5.27	0.3/0.2	0.3/0.3	0.1/0.1	0.5/0.3	0.4/0.4	0.5/0.3	0.4/0.4
5.28--6.13	0.3/0.3	0.3/0.3	0.1/0.1	0.3/0.2	0.4/0.3	0.4/0.3	0.5/0.5
6.14--6.29	0.4/0.3	0.3/0.3	0.1/0.1	0.3/0.3	0.4/0.3	0.4/0.2	0.6/0.4
TTop10 precision/TTop5 precision	0.4/0.3	0.6/0.3	0.4/0.3	0.5/0.4	0.4/0.4	0.3/0.2	0.5/0.5

Topics Top10/Top5	Tablet pc	Kodak bankrupt	Huawei for new life	Iphone 4s sale	Windows 8	iOS jailbreak	Face book
3.08--3.23	0.4/0.4						0
3.24--4.09	0.4/0.4	0.4/0.3		0.3/0.3	0.4/0.3	0.6/0.5	0
4.10--4.25	0.3/0.3	0.3/0.2	0	0.3/0.3	0.5/0.4	0.6/0.5	0.3/0.2
4.26--5.11	0.4/0.2	0.3/0.2	0	0.3/0.3	0.5/0.4	0.4/0.2	0.3/0.2

5.12--5.27	0.4/0.3	0.3/0.1		0.3/0.3	0.5/0.4		0.3/0.2
5.28--6.13	0.4/0.3	0.3/0.2	0	0.3/0.2	0.5/0.3		0.3/0.2
6.14--6.29	0.3/0.2	0.3/0.1	0	0.3/0.2	0.5/0.3		0.3/0.2
TTop10 precision/TTop5 precision	0.4/0.4	0.4/0.3	0.4/0.2	0.3/0.3	0.4/0.3	0.6/0.5	0.3/0.2

Topics	New ipad sale	Iphone news	Ipad show	Apple ceo salary	App Store	CES2012	HTC
3.24-4.09	482	234	841	0	644	690	239
4.10-4.25	1502	352	1546	1513	1514	1493	303
4.26-5.11	1364	268	1504	1482	1517	1381	183
Sum1	3290	854	3891	2995	3675	3564	725
Total tweets	6874	1419	8313	6324	8175	7453	1385
Sum1/Total tweets	0.4786	0.6018	0.4681	0.4735	0.4495	0.4782	0.5235

Topics	Tablet pc	Kodak bankrupt	Huawei for new life	Iphone 4s sale	Windows 8	iOS jailbreak	Facebook
3.24-4.09	670	645	0	636	780	780	0
4.10-4.25	82	1609	397	1564	1529	852	609
4.26-5.11	17	1424	199	1485	1467	18	1567
Sum1	769	3678	596	3685	3776	1650	2176
Total tweets	1556	9084	1448	8085	7125	1932	5639
Sum1/Total tweets	0.4942	0.4049	0.4116	0.4558	0.5299	0.8540	0.3858

Table 3. The average precision values of all the 14 topics

	Average
Authenticated/Tota l	0.3038
Sum1/Total tweets	0.4665
Atop10 precision	0.4357
Ttop10 precision	0.4286
Atop5 precision	0.3286
Ttop5 precision	0.3214

4.2 Semantic Extraction Based on LDA

LDA is a hierarchical probabilistic model of documents. Here we utilize a C implementation of variational EM for latent Di richlet allocation (LDA), a topic model for text or other discrete data. It allows us to analyze of corpus, and extract the topics that combined to form its documents. The following is steps we take when doing experiment with LDA-c code:

1. Transform original data into required data format.

During the process of analyzing our dataset(14 topics,74,662 tweets), the term index and term-document matrix would be created, which provide great convenience for the transformation of all the original tweets of 14 topics into the data format that LDA-c implementation required. The final data is a file where each line is of the form:

```
[M] [term_1]:[count] [term_2]:[count] ...  [term_N]:[count]
```

where [M] is the number of unique terms in one of 14 topics, and the [count] associated with each term is how many times that term appeared in this topics. Note that [term_1] is an integer which indexes the term; it is not a string.

2. Topic estimation and inference

After compiling the code, we could estimate the model by executing:

```
lda est [alpha] [k] [settings] [data] [random/seeded/*] [directory].
```

To perform inference on a different set of data execute:

```
lda inf [settings] [model] [data] [name]
```

Variational inference is performed on the data using the model produced by estimation. After inference, [name].gamma will be created, it's the variational Dirichlet parameters for each topic, with which the top N words from each topic in a .beta file being printed.

3. Printing topics

Here we can use the Python script topics.py to print out the top N words from each topic in a .beta file. The usage is as following:

```
python topics.py <beta file> <vocab file> <n words>
```

In our experiment, usually we take the value of alpha as 1. Parameter k indicates that program according to these 14 topics extract and statistic the k hottest topics. Then through taking the diffierent k value,we can see the distribution of 14 topics under the k hottest topics. Finally we select the topic distribution that probability greater than 1,and print out top 5 and top 10 words from each topic for 14 topics. The experimental results were shown in Table4.

Table 4. ,when k=5, probability>1,top 5 and top 10 words from each topic for 14 topics

K=5	probability> 1	Top5	Top10
CE2012	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
Facebook	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company, Software, Release,China
IOS jailbreak	0 1 2 4	Phone, Product, America, Telecom, China, Phone	Phone, Product, America, HK, Kodak, Telecom, Company, Software, China, Release
Huawei for new life	1 2 3 4	Telecom, China, Phone	Telecom, Phone, Company, Software, Kodak, Release, China
HTC	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
Windows 8	0 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Software, China, Release
Kodak	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
Telecom	0 1 2 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Company, Software, China, Release

Table 4. *(Continued)*

Apple Store	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
New IPAD	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
New IPAD Sale	0 1 2 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company,Software, Release, China
Apple ceo salary	0 1 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company, Release,China
Iphon e news	0 1 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Company, Release,China
Talet PC	0 3 4	Phone, Product, America, Telecom, China	Phone, Product, America, HK, Kodak, Telecom, Phone, Release, China

5 Experimental Results and Analysis

Here, in Table 1, we have made a specific explanation for all abbreviations appeared in Table 2 which show the results of each microblog topic. The average results of all the 14 topics are listed in Table 3. From the result that presented in Table 2, we can clearly see that the differences between Atop10 precision and Ttop10 precision are not particularly obvious in terms of precision. For further observation, we decide to make an average for the precision values of 14 topics. To our surprise, the precision value of top 10 words that produced by the tweets that from authenticated users is higher than that produced by the tweets from all of users involved in a topic on average. At the beginning, we think that the number of the total tweets for one topic actually not only contains the tweets from authenticated users, but also contain others tweets from common users. In comparison, it has the larger suggestion context source and the richer content information. Thus, it should output higher precision scores. However, the results run adversely to what we might intuitively expect the average precision value of top 10 words that produced by the tweets that authenticated users, i.e., slightly higher.

However, we also do the job of computing the top10 and top5 percision value of the tweets that from different periods. By comparing with the value of both TTop10 precision and TTop5 precision, we find that the highest persion value are most concentrated in the 3.24 to 5.11,according to the curves of topic evolution ,they are appeared around the peak time.

So what does this show? It illustrates that we need not to select all tweets of a topic as our suggestion context source. Considering the final result, the tweets that from authenticated users and the tweets that around the peak time could be on behalf of the entire tweets under a topic. During the preprocessing, we observed that under the background of computer configuration with a 32-bit operating system, dual-core CPU and 3.00GB memory, it takes about 4 hours for all users' tweets of a certain topic, but for processing authenticated users' tweets, it just takes about 30 minutes. Taking tweets that from authenticated users as our suggestion context source saves not only the processing time, but also the storage space. How much storage space does it save at all? From experimental data, we can see that average Authenticated/Total is about 0.3038 and Sum1/Total tweets is 0.4665. In other words, the authenticated context accounts for around 1/3 in total tweets and almost save 2/3 storage space. The tweets that appeared around the peak time accounts for around 1/2 in total tweets and almost save 1/2 storage space.

In Table 4,we can see that the top5 and top 10 words do not reflect some topic. Due to the microblog text is too short, making the analysis of the hiden topic analysis for the tweets does not seem ideal.

6 Conclusion

In this paper, firstly we introduced how to extract microblog data by crawling. Then we discussed and analysised the distribution of the tweets and users for each topic in the time interval of 20 days. Then we gave out our selection of datasets and the approach for text processing and evaluation method. Finally we presented semantic extraction for all tweets under 14 topics based on LDA. Considering the final results, we can see that the tweets that from authenticated users and the tweets that around the peak time could be on behalf of the entire tweets under a topic. Due to the microblog text is too short, making the analysis of the hiden topic analysis for the tweets does not seem ideal. In the further, We will combine the media characteristic of mircoblog to da some jobs of social network, such as query suggestion and so on.

References

[1] Pontin, J.: From many tweets, one loud voice on the internet. The New York Times (April 22, 2007)

[2] Benevenut, F., Rodrigues, T., Cha, M., Almeida, V.: Characterizing user behavior in online social networks. In: Proc. of ACM SIGCOMM Internet Measurement Conference. ACM (2009)

[3] Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P., Zhao, B.Y.: User interactions in social networks and their implications. In: Proc. of the 4th ACM European Conference on Computer Systems. ACM (2009)

[4] Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. Phys. Rev. E 68(3), 036122 (2003)

[5] Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. ACM (2007)

[6] Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about twitter. In: Proc. of the 1st Workshop on Online Social Networks. ACM (2008)

[7] Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2009)

[8] Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Microblogging as online word of mouth branding. In: Proc. of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems, pp. 3859–3864 (2009)