A Microblog Unbalanced Data Evolution Analysis Method Based on LDA Model

Shaorong Feng
School of Informatics, Xiamen University, Xiamen 361005, China
shaorong@xmu.edu.cn

Abstract. As a social network platform, Microblog has a large amount of information that can be mined. Microblog differs from static texts, which is time-sensitive, and the popularity of themes is changing with time. In addition, the popularity of Microblog varies from theme to theme, and the number of Microblogs varies greatly. Therefore, it has obvious unbalance characteristics. The traditional LDA model can not mine the theme intensity evolution process, and the theme mixing phenomenon will appear in the theme of unbalanced data mining. In view of the above problems, this paper proposes a method to analyze the theme evolution law of unbalanced Microblog data by introducing time variable. The experiment shows that the problem of theme words mixing is effectively solved after the balance treatment. The introduction of time variables can mine Microblog popularity and theme intensity evolution process.

1. Introduction
Microblog is an Internet social platform for real-time sharing of short information. As one of the important network media, Microblog has the advantages of simple operation, strong communication, wide range of use, and rapid development of themes. From national policies to entertainment news, users express their opinions and discussions while producing hot themes. The huge amount of information has an important impact on people's lives. Therefore, mining the information in Microblog has important commercial value. In addition, it can monitor the hot spots of network public opinion in real time, strengthen the information carding and filter the bad information.

Theme model is a statistical model used to find abstract themes in a series of documents. The theme model generally refers to LDA theme model [1-2]. LDA theme model is often used to mine potential themes in large-scale corpus. In recent years, the research on LDA theme model has been deepened, such as continuous time dynamic theme model [3] Most theme mining for Microblogs is based on twitter data [4-5]. There are also many studies on Microblog themes in China. Shi Qingwei and others proposed a word pair theme evolution model [6], It is assumed that the theme follows a Beta distribution over time, He combined BTM(bitterm topic model) model to fit the dynamic evolution process of Microblog themes. The DC-DTM model[7] based on Microblog text is proposed by Liu Bingyu et al. He comprehensively considers the Microblog forwarding relationship and the author theme distribution to assist the theme mining of Microblog

In text mining, it is better to use theme model to mine text with balanced data, but when it is applied to unbalanced data, it will lead to the phenomenon of theme words mixing, and words that should not belong to the same theme are divided into the same theme. The processing algorithms of unbalanced data are divided into oversampling and undersampling. In the oversampling algorithm, He et al. proposed adaptive composite sampling ADASYN[8]. Barua et al. proposed the MWMOTE algorithm [9] to assign different weights to samples. In the undersampling algorithm, The nearest
neighbor rule (ENN) [10], neighborhood cleaning (NCL) [11] and so on are used to sample by the distance between samples. After spectral clustering for most samples, Spectral clustering down sampling algorithm [12] selects information points according to the cluster size and the distance from the minority samples. These methods effectively solve the imbalance problem.

In view of the problems existing in the theme modeling of unbalanced data, as well as the strong timeliness of Microblog, this paper proposes a theme evolution analysis method of unbalanced Microblog data based on LDA model to balance the Microblog data with great disparity in quantity. The time variable is introduced to model, and the evolution process of theme intensity of Microblog is found. The results show that after balancing the data, the mixed phenomenon of Microblog theme words is effectively solved, and the introduction of time variable can truly reflect the evolution trend of theme intensity.

2. Unbalanced Data Processing Based on Theme Model

2.1. LDA Theme Model

LDA theme model is a potential Dirichlet assignment model proposed by Blei et al. The structure of LDA theme model is shown in Figure 1.

\[
\begin{align*}
\theta_m & \sim \text{Dir}(\alpha), \quad m \in [1, M] \\
\varphi_k & \sim \text{Dir}(\beta), \quad k \in [1, K]
\end{align*}
\]

Figure 1. LDA model

Let \( M \) be the number of documents, \( K \) be the number of themes, \( w_{mn} \) be the \( n \)th word of document \( m \), \( z_{mn} \) be the theme of the \( n \)th word in document \( m \), \( \alpha \) be the prior parameter of theme distribution and \( \beta \) be the prior parameter of word distribution.

The theme distribution of a document is extracted from the Dirichlet distribution with parameter \( \alpha \), \( \theta_m \sim \text{Dir}(\alpha) \), \( m \in [1, M] \). The word distribution of theme is extracted from the Dirichlet distribution with parameter \( \beta \), \( \varphi_k \sim \text{Dir}(\beta) \), \( k \in [1, K] \).

LDA parameters are estimated by approximate reasoning algorithm. Firstly, LDA uses variational expectation maximization algorithm, and later finds that Gibbs sampling[13] is easier to solve. Gibbs sampling is a special case of Monte Carlo Markov (MCMC). Gibbs samples the themes of all words until convergence, and then obtains the theme distribution and the word distribution of Microblog. The calculation of probability of theme \( k \) under document \( m \) and probability of word \( v \) under theme \( k \) are shown in formula (1).

\[
\begin{align*}
\theta_{m,k} &= \frac{n_{m,k} + \alpha}{\sum_{i=1}^{K} (n_{m,i} + \alpha)}, \\
\varphi_{k,v} &= \frac{n_{k,v} + \beta}{\sum_{i=1}^{V} (n_{k,i} + \beta)}
\end{align*}
\]

(1)

2.2. Unbalanced Microblog Data Balancing Method Based on Theme Model

Microblog data is real and effective data, and the synthesized data is easy to distort the results. Therefore, combined with the idea of distance similarity between the same class samples in undersampling algorithm of two kinds of unbalanced data, this paper proposes a balanced method of
unbalanced micro blog data based on theme model. The specific steps are as follows

Step 1 calculates the theme distribution and word distribution of unbalanced Microblog text, and observes the distribution of subject words

The mixed distribution of subject words mainly occurs in the case of large sample gap. The more hot Microblogs are, the more Microblogs there are. When the gap is large enough, there will be a mixture of the most popular and the least popular

Step 2 calculates the probability $\theta_{m,1}$ of the theme corresponding to the most popular Microblog, the probability $\theta_{m,2}$ of the theme mixed with the subject words, and difference value $\Delta \theta$ to calculate the sample retention degree $E$. The calculation of sample retention degree is shown in formula (2).

$$E = \sqrt{\Delta \theta^2} - \sum_{i=3}^{K} \theta_{m,i}$$

(2)

In step 3, set the threshold $\varepsilon, \mu$, delete the Microblogs whose $\theta_{m,1}, \theta_{m,2}$ are greater than the threshold $\varepsilon$ and the retention $E$ is less than the threshold $\mu$. The threshold $\varepsilon$ calculation is shown in formula (3).

$$\varepsilon = \frac{1}{K}$$

(3)

When the theme words are mixed, it will lead to errors in the theme assignment of hot Microblog samples. The closer the probability of belonging to the above two themes, the smaller the difference of theme probability. The length of Microblog is different and the feature is sparse. When the theme is not clear, the probability value of each theme is similar, while the probability of the above two themes is greater than the threshold $\varepsilon$, The larger the probability value of other themes, the smaller the sample retention.

Step 4: After deleting the sample, the theme modeling is carried out again. Repeat steps 1 and 2 until the JS(Jensen Shannon) distance between the distribution of subject words tends to be stable.

3. Research on the Evolution of Microblog Theme Intensity

Traditional LDA model is suitable for dealing with static long text, and can not judge the intensity and evolution trend of themes. Microblog text is short and has strong timeliness. Therefore, it is of great significance to measure the popularity of Microblog themes and mine the evolution law of theme strength.

In the research on the evolution of Microblog theme strength, the corresponding time stamp of each Microblog is introduced into the model, and the theme intensity value in different time periods is calculated, and the evolution law of theme intensity is mined out, and the development trend of theme is grasped.

Definition 1 time interval refers to the division of time period and represents the span from one time point to another.

Definition 2 Microblog timestamp refers to the generation time of Microblog, and $t_m$ represents the time stamp of Microblog $m$. the Microblog with different time intervals is divided by using the timestamp.

$L$ is the total number of time intervals. Theme intensity reflects the heat of themes in different time intervals, so it is reflected in different time intervals.

After the unbalanced processing, part of the noise Microblog is removed to improve the accuracy of the model and make the Microblog theme be correctly divided, which plays an important role in accurately calculating the theme intensity. The calculation steps of Microblog theme intensity based on LDA are as follows:

Step 1 calculates the theme distribution $\theta_m$ of each Microblog

Step 2 extracts the Microblog time stamp and divides the Microblog in different time intervals according to the time stamp.

Step 3 calculate the theme intensity of different themes at different time intervals.

The theme intensity of different time intervals is equal to the sum of probability distribution of
theme $k$ in unit time slice divided by the total number of Microblogs in all time intervals. The number of Microblogs in different time intervals is different. Divided by the total number of Microblogs can reflect the different popularity of themes in different time slices. The intensity of theme $k$ in $l$ time interval is shown in formula (4).

$$Q_{k,l} = \frac{\sum_{m=1}^{\infty} \theta_{m,k,t_m}}{\sum_{j=1}^{L} n_{k,j}}$$

(4)

$\theta_{m,k,t_m}$ Denotes the probability of Microblog $m$ with time stamp $t_m$ under theme $k$, and $n_{k,j}$ denotes the number of Microblogs with theme $k$ in the $l$ interval.

### 4. Experimental Results and Analysis

#### 4.1. Experimental Corpus and Settings

This paper uses data from sina Microblog, which collects 2936 popular Microblog posts from December 1 to 18, 2017 in chronological order. The themes of ordinary Microblogs are not clear enough, and popular Microblogs are more suitable for theme mining. The content of Microblogs involves four themes: national treasure, world Internet Conference, Fanghua and Jiangge case.

After preprocessing operations such as Chinese word segmentation and removing stop words, the time stamp of Microblog is converted. The threshold value $\epsilon$ is set to 0.25, $\mu$ is set to 0.2, and the number $K$ of themes is set to 4. The time interval is 24 hours, $L=18$.

#### 4.2 Evaluating Indicator

$KL$(Kullback Leibler) distance, also known as relative entropy, is used to measure the difference between probability distributions. It can be used to calculate the similarity between themes. The theme model with far distance has higher theme discrimination, the better the model. $KL$ distance is defined as formula (5).

$$D_{KL}(P, Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

(5)

$P, Q$ is the distribution to be measured

$KL$ distance is asymmetric, but the similarity between themes should be symmetric. Therefore, $JS$ distance is used to measure the similarity between themes. $JS$ distance is a variant of $KL$ divergence, and $JS$ distance is shown in formula (6).

$$JS(P, Q) = \frac{1}{2} KL(P, \frac{P+Q}{2}) + \frac{1}{2} KL(Q, \frac{P+Q}{2})$$

(6)

#### 4.3 Performance Comparison Experiment Analysis

Microblog data without balanced processing has the phenomenon of mixed subject words. The theme model is built for the data without balance processing, and the distribution of subject words is calculated. The words with the top 8 extraction probability are shown in Table 1.
Table 1. Unbalanced word distribution

| Theme 1       | Theme 2       | Theme 3       | Theme 4       |
|--------------|--------------|--------------|--------------|
| Country      | Internet     | Fanghua      | Liu Xin      |
| treasure     | world        | film         | Jiang Ge     |
| story        | Congress      | Li           | mother       |
| national treasure | China     | Liu Feng     | Feng Xiaogang |
| program      | economy      | say          | Chen Shifeng |
| culture      | Wuzhen       | youth        | jiangge case |
| museum       | development   | he Xiaoping  | Miao Miao    |
| cultural relic | network   | years        | kindhearted  |

It can be seen from table 1 that the Microblog words about Fanghua and jiangge case are mixed in theme 4, while the subject words related to Fanghua exist in theme 3 alone. After the unbalanced treatment in this paper, 420 Microblogs are eliminated. The words with the top 8 probability in the word selection distribution are shown in Table 2.

Table 2. Balanced word distribution

| Theme 1       | Theme 2       | Theme 3       | Theme 4       |
|--------------|--------------|--------------|--------------|
| Country      | Internet     | Fanghua      | Jiang Ge     |
| treasure     | world        | film         | Liu Xin      |
| story        | Congress      | Feng Xiaogang| attentions    |
| national treasure | China     | He Xiaoping  | mother       |
| program      | economy      | dance        | Chen Shifeng |
| culture      | Wuzhen       | youth        | Understand   |
| protect      | development   | kindhearted  | Japan        |
| cultural relic | network   | years        | hurt         |

From the table data, it can be seen that the first eight subject words in the balanced model are no longer mixed. This experiment also compares the JS distance between different themes before and after the balance treatment, and further evaluated the model. The calculation results are shown in Figure 2.

![Figure 2. Comparison of JS values](image)

T1, T2, T3 and T4 represent four themes. From the results of the table, we can see that the JS value between different themes after balanced treatment tends to be a fixed value, and the JS value between Theme 3 and theme 4 increases relatively obviously. It is proved quantitatively that the distance...
between Theme 3 and theme 4 in the balanced model becomes longer, and the difference between themes increases.

4.4. Comparative Experimental Analysis of Theme Evolution
In this paper, every 24 hours is taken as a time interval. From the table content, we can see the hot events occurred from December 1 to 18. The distribution of theme intensity is shown in Figure 3, 4 and 5.

**Figure 3.** Theme evolution law

**Figure 4.** Theme evolution law

**Figure 5.** Theme evolution law

From the overall trend, CCTV variety show national treasure will be broadcast on December 3 with weekly issue. As can be seen from the figure, the theme intensity of relevant national treasure is higher.
on the 3rd, 11th and 17th, reflecting the broadcasting cycle. The theme intensity of the Internet Conference increases to the peak on the 5th, and then slowly declines. The movie "Fanghua" was launched on December 15. Its popularity increased significantly, and the overall theme intensity of Fanghua was higher than that of Jiangge case, indicating that the theme of Fanghua was more popular in this period of time. The curve in the figure better reflects the evolution law of the four major events.

5. Conclusion
In this paper, aiming at the sparse text features and strong real-time characteristics of Microblog, we use the features of theme distribution to balance the processing, and then get a more accurate theme model. The model introduces time variables to calculate the theme intensity of different time intervals, which effectively reflects the evolution trend of themes.

6. References
[1] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 35(2): 328-332.
[2] WANG Han, XIA Hong-bin. Collaborative Filtering Recommendation Algorithm Mixing LDA Model and List-wise Model[J]. Computer Science, Sep. 2019, 46(9): 216-222.
[3] Wang C, Blei D, Heckerman D. Continuous time dynamic theme models[C]. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence. Corvallis: AUAI Press, 2008: 579-586.
[4] Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using theme models[C]. Proceedings of the 33rd European Conference on Advances in Information Retrieval. Berlin: Springer-Verlag, 2011: 338-349.
[5] Quercia D, Askham H, Crowcroft J. TweetLDA: supervised theme classification and link prediction in Twitter[J]. ACM Web Science Conference, 2012, 12 (68): 247-250.
[6] Shi Qing-wei, Liu Yu-shi, Zhang Feng-tian. Biterm theme evolution model of Microblog[J]. Journal of Computer Application, 2017, 37(5): 1407-1412.
[7] Liu Bing-yu, Wang Cui-rong, Wang Cong, et al. Microblog community discovery algorithm based on dynamic theme model with multidimensional data fusion[J]. Journal of Software, 2017, 28(2): 246-261.
[8] He H, Bai Y, Garcia EA, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]. Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, 2008, 1322-1328.
[9] Barua S, Islam M M, Yao X, et al. MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(2): 405-425.
[10] Wilson D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems, Man and Cybernetics, 1972, 2(3): 408-421.
[11] Jorma L. Improving Identification of Difficult Small Classes by Balancing Class Distribution[C]. 8th Conference on Artificial Intelligence in Medicine, Cascais: Silvana Quagliani, 2001: 63-66.
[12] Tao Xin-Min, Zhang D-X, Fu D-D, et al. The SVM classifier for unbalanced data based on spectrum cluster-based under-sampling approaches[J]. Control and Decision, 2012, 27(12): 1761-1768.
[13] Porteous I, Newman D, Ihler A, et al. Fast collapsed gibbs sampling for latent dirichlet allocation[C]. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM, 2008: 569-577.