Data And Content Analysis For Social Network Using LDA Text Model

Linxiang Zhang
Internation School, Beijing University of Posts and Telecommunications, Beijing 100876, China
zhanglinxiang@bupt.edu.cn

Abstract: WhatsApp is a popular cross-platform application for communicating and connecting people. In this paper, the author analyzes massive data from WhatsApp public groups and obtains so-called tacit knowledge and social complex knowledge. During the processing, the author uses LDA (Latent Dirichlet Allocation) model to analyze the content which helps us to be aware of many aspects from public groups such as social trend, sentiment analysis and other hidden information. In the end, author get the results about topic for each group and the occurrence of words for each topic.

1. Introduction
WhatsApp is an instant messaging software. We can receive message, pictures, audio files, and video information in WhatsApp, WhatsApp divided people in different groups and people can communicate with each other by expressing their own opinions. In addition, these public groups have some connections with other groups. Therefore, the WhatsApp public groups’ data involves large amount of information such as different topics people talking about and different connections between different groups.

In the project, author analyze the public groups from several aspects and divide them into following sections.

First, author analyze different content that people in different groups talk about. Publics talk about different topics in various groups. For example, persons care more about health in a group while people discuss their job in another group. Author analyze the topics people talking about in groups. In the analyze, author make data pre-preprocessing for the data at first. Data cleansing is significant since the original data is not in standard format and contains useless information. The data pre-preprocessing contains transformation, text segmentation, stemming and data reduction. In transformation part, the author makes transformation for the information so that it is uniform and can be analyzed more easily. In segmentation part, words are split into groups. In stemming part, the author removes the condition that one word contains various meanings. In data reduction step, tag which does not contains useful meanings has been removed. After these steps, author uses LDA model to analyze the content.

The LDA model is a probabilistic topic model which was rediscovered by David Blei, Andrew Ng and Michael I. Jordan in 2003 for discrete data sets(such as document sets).[1] The author has adjusted the parameters according to perplexity in the LDA model. Then the author uses LDA model to obtain the final results showing the topic for each group.
2. Background

2.1 Description of Data from WhatsApp Public groups

2.1.1 Application used for management
The data from WhatsApp public groups are imported and stored in the .db file. Author use visual database management tool, Navicat to help manage the data. Figure 1 shows using Navicat to manage the database.

2.1.2 Characteristics of the data
There are a number of distinguishing characteristics of data including scattered, colloquial, symbolic. Scattered data means that there are various kinds of message including text, graph, video and audio. The contents are not coherent because each piece of message maybe short and half-baked. Colloquial words mean that there are amount of oral words but written words. People are used to use idioms, slang and colloquial words in the daily communications. When using social software, most of people like to use emoji to show their emotions or opinions. Therefore, there are a huge number of symbolic data.

2.2 LDA model

2.2.1 LDA model representation
The latent Dirichlet allocation model is a generative statistical model which includes the text theme and words. The basic idea of this model is that each sets of observations can be explained by unobserved groups and each of which is represented as a series of words in the vocabulary. Each topic can be represented as a series of words in the vocabulary.

The process of creating LDA’s text is shown in Figure 1. The variables like the topic are implicit in the database. Arrows represents dependency relationships and constraints between entity sets. The box indicates repeated sampling and the footer represents the number of repetitions.

![Figure 1](image)

In Figure 1, \( \varphi_k \) indicates the words’ probability distribution in topic \( K \).

LDA assumes that the prior distribution of document topic is Dirichlet distribution. For any document \( m \), its topic distribution is: \( \theta_m = \text{Dirichlet}(\alpha) \). For any topic \( k \), the word distribution of \( k \) is: \( \varphi_k = \text{Dirichlet}(\beta) \). \( \eta \) is the Dirichlet hyperparameters of the distribution and \( \alpha \) represents the Dirichlet hyperparameters distributed by document topic polynomials. \( K \) is topics’ number and \( M \) is the number of documents. \( N_m \) indicates the length of the document \( m \). \( W_{m,n} \) and \( Z_{m,n} \) represents the N words and its theme in the m document. For \( nth \) word in document \( m \), we can obtain its theme number \( z_{m,n} \) distribution as: \( z_{m,n} = \text{multi}(\theta_m) \) and \( W_{m,n} = \text{multi}(\varphi_k) \)

The steps of creating LDA are as follows:
1. Repeat sampling \( \varphi_k \sim \text{Dir}(\beta) \) for \( k \) times.
2. Create the document file \( m \) through the corpus with following steps:
   I. Select \( N_m \) that obeys poisson(\( \xi \)) distribution;
   II. Select \( \theta_m \) that obeys the \( \text{Dirichlet}(\alpha) \) distribution where \( \theta_m \) is a column vector and representing the \( m \) theme’s probability of occurrence of.
III. Select the theme \( Z_{m,n} \) and \( Z_{m,n} \) obey the multinomial distribution. \( Z_{m,n} \) indicates the current choice of the theme. Select \( W_{m,n} \) which obey \( \text{Multinomial}(\varphi_{m,n}) \).

Reduplicate the step \( \parallel \) for \( M \) times. After that, author has completed the process of creating LDA.

3. Design and Implementation

3.1 Constructing the LDA Model

3.1.1 Determination of Model parameters

The determination of LDA model’s parameters mainly contains two aspects: the determination of the hyperparameters and the determination of the number of theme.

1. The determination of the hyperparameters:

The LDA model has two sets of pre-settings. They are per-document topic proportions from \( \text{Dirichlet}(\alpha) \) and per-word topic assignment from \( \text{Dirichlet}(\beta) \). Though the values of \( \alpha \) and \( \beta \) will affect the proportions of the topic and the words, using different topic and the words will not influence the result. Therefore, it is possible to assume the values of \( \alpha \) and \( \beta \) are 50/k and 0.01. Then we use the Smoothing method to deal with the data. [2]

2. The determination of the number of themes:

The number of theme \( K \) has an influence on the derivation of model and the value of the hyperparameters. In text mining, the researchers usually observe the effect by debugging from time to time or enumerating. For example, researchers observe the quality of the high probability theme words. Extending the topic model using non-parametric Bayesian method is a most common method since the size of the model can be adapted with the change of trend of the data in the model. The parameters also can be chosen by the trend and proportions of data.

3.1.2 Model Solution

It is an complex optimization problem to obtain the solution, \( \bar{\theta} \) and \( \bar{\alpha} \). Usually, researchers could not obtain accurate results. In common, there are two ways to solve the LDA model: the first one is based on Gibbs sampling algorithm [3], and the second one is based on variational inference EM algorithm. Gibbs's sampling algorithm is most widely used in LDA models since it is simple and effective. And in the LDA based on Gibbs sampling algorithm, the collapse Gibbs sampling [4] method is most widely used. The purpose of collapse Gibbs sampling method is to obtain the overall probability distribution of \( Z_{m,n} \) and \( W_{m,n} \) corresponding to \( \bar{Z}, \bar{W} \) which means the distribution of document themes and the distribution of theme words. If \( Z_{m,n} \) and \( W_{m,n} \) are determined, the value of \( \theta_m \) and \( \alpha_k \) can be obtained after knowing the statistical frequency of words.

3.2 Content Analysis

In disparate groups, persons discuss about disparate topics. Different topics represent different contents which can help researchers to know the trend of contents. It is significant to summarize the theme from the content. Author does this by following steps.

3.2.1 Data preprocessing

Author need to do some processing of data before major processing to facilitate the operation of a computer. In the real world, raw data is generally incomplete and inconsistent. First, the author translates all the other languages in to English and obtain the words from text segmentation with one form. Then, the author use python to remove the stop words.
3.2.2 LDA model implementation

Input: word vector w, super parameters α and β, number of topic k. Pseudo-code:

```
for all document m do:
    for all words n in document m do:
        sample topic index z(m,n) =k~Mul(1/k)
        n(m,k)+=1; n(k,n)+=1; n(m)+=1; n(k)+=1;

while not finished do:
    for all document m do:
        for all words n in document m do:
            n(m,k)-=1; n(k,n)-=1; n(m)-=1; n(k)-=1;
            sample topic index k~p(z|z(not n),w);
            n(m,k)+=1; n(k,n)+=1; n(m)+=1; n(k)+=1;

If converged then:
    compute θ;
    compute ψ;
```

In this model, the author determines $α=50/K$ and $β = 0.01$. For the parameter K, the author uses perplexity to adjust it. The value of perplexity varies with the number of topic.

![Figure 2 Perplexity](image)

Using the LDA model and the Gibbs sampling algorithm, the topic of group can be obtained after constructing the input vectors. From Figure 3 we can obtain the best topic number since the value of perplexity in this point is the least. Combining with the input vector and Gibbs sampling, the author obtains the Probability of key words in topics and the probability distribution of the potential topic for the groups.

| Topic 1 Key words | probability | Topic 2 Key words | probability | Topic 3 Key words | probability |
|-------------------|-------------|-------------------|-------------|-------------------|-------------|
| sale              | 0.000162    | video             | 0.000161    | job               | 0.030640    |
| sage              | 0.000162    | Success           | 0.000163    | company           | 0.019451    |
| suppose           | 0.000321    | beat              | 0.000161    | salary            | 0.018848    |
4. Results and Discussion

4.1 Result of Content Analysis

In content analysis, the author analyzes the message and tries to find the content users talk about. In this analysis, the author makes data preprocessing at first and builds a LDA model to analyze the topic that users talk about. Figure 4 shows the preprocessed data. Figure 5 shows the results from the LDA model.

In LDA model’s results, we can get to know the possible topic for each group.

5. Conclusion

In this project, the author analyzes WhatsApp public groups from the content aspect. The author tries to find what users talk about in each group. First, the author makes preprocessing for the message, there are four steps containing translation, text segmentation, stem process, and removing stop words. Then the author tries to build a model to analyze the content and selects LDA model to do the analysis. After that,
the author adjusts the parameters in LDA model and starts the analysis to get the final results about
possibility of topic for each group and the possibility of word for each topic.

By analyzing the data, the author can discover a great deal of hidden messages about contents. Also,
there are many other aspects such as sentiment and popularity to be analyzed. Therefore, the author will
also try to analyze the other aspects to find more information in the future.

References
[1] Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. "Latent
Dirichlet Allocation". Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
[2] LI Xiaoke, ZHAO Zijuan, GUO Qiang, LIU Jianguo, LI Rende. On the Theme of Network Science
Conference via Text Mining. Complex Systems and Complexity Science, 2018, 15(3): 9-9.
[3] Baker, L. D. & McCallum, A. K. (1998). Distributional clustering of words for text classification.
In Proceedings of the 21st ACM-SIGIR International Conference on Research and Development
in Information Retrieval (SIGIR).
[4] Iyer, R. & Ostendorf, M. (1996) in Proceedings of the International Conference on Spoken Language
Processing (Applied Science & Engineering Laboratories, Alfred I. duPont Inst., Wilmington,
DE), Vol 1., pp. 236-239.