A Fuzzy Comprehensive Evaluation of Public Satisfaction with Urbanization Based on Social Media Monitoring

Ruijie Wu

ABSTRACT

The paper presents a method of identifying satisfaction with Urbanization based on studying the content of comments on the social networks. The analysis of public opinions revealed areas of concern and priority areas in the public Satisfaction with Urbanization. This paper analyses the influencing factors of urbanization satisfaction. Through Latent Dirichlet Allocation (LDA) Model and Sentiment Analysis, the evaluation of urbanization satisfaction in social media environment is realized.

1. THE PROBLEM AND THE ORIGIN OF RESEARCH

Urbanization is the inevitable trend for the economic and social development of all countries and regions in the world. It is also an important symbol of the degree of modernization of a country or region. According to the survey data released by the National Bureau of Statistics of China, China's urban population accounted for 59.58% of the total population in 2018. The number of urban population exceeds the number of rural population. The acceleration of urbanization process will inevitably promote profound changes in people's mode of production, lifestyle, occupational structure and values. The Process of urbanization in china has entered a new stage.

Due to the restriction of the existing urban and rural household registration system, they have been in the edge of the city. This not only increases the difficulty of rural and urban household registration management, but also caused a series of economic and social problems.

With the development of the network society, the social medias are playing an increasingly important roles, the number of active users on social networking websites has increased exponentially. The popularity of these websites is due to the
fact that users aren’t the only content consumers, they can also generate content and collaborate with other users. In case of the occurrence of unexpected events, people can react by uploading or sharing comments, meaning that the amount of data generated and related to the event will be huge. These reactions of the users provide insight on the events.

Considering this reality, we intend to analyze the influence factors of public satisfaction with urbanization, and based on social media monitoring, construct a scientific and reasonable index system of public satisfaction with urbanization to improve the level of local government administration.

2. THE INFLUENCE FACTORS OF PUBLIC SATISFACTION

Considering the public satisfaction with urbanization is influence by many factors, we try to find out the key factors which are the topics in social media.

We collected 16799 questionnaires from Henan Social Opinion Survey Center, and cleaned up the data, got 3334 valid samples. Because the questionnaire involves 14 measurement items, we uses the principal component analysis method to reduce the dimension, and gets 6 principal components, as shown in TABLE I.

<table>
<thead>
<tr>
<th>Component</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAC1</td>
<td>Health Insurance</td>
</tr>
<tr>
<td></td>
<td>Arts, Culture and Entertainment</td>
</tr>
<tr>
<td>FAC2</td>
<td>Land Sublease</td>
</tr>
<tr>
<td></td>
<td>Percentage of Farming Income</td>
</tr>
<tr>
<td>FAC3</td>
<td>Poultry Rearing</td>
</tr>
<tr>
<td></td>
<td>Quantity of Cultivated Land</td>
</tr>
<tr>
<td></td>
<td>Assessment of Current Residential Surrounding Environment</td>
</tr>
<tr>
<td>FAC4</td>
<td>Family Size</td>
</tr>
<tr>
<td></td>
<td>Proportion of the Population Without Working Opportunities</td>
</tr>
<tr>
<td>FAC5</td>
<td>Impact of Internet on Real Life</td>
</tr>
<tr>
<td></td>
<td>Yearning for Urban Life</td>
</tr>
<tr>
<td>FAC6</td>
<td>Livestock Raising</td>
</tr>
<tr>
<td></td>
<td>Quantity of Surrounding Enterprises</td>
</tr>
</tbody>
</table>
According to the characteristics of data structure, we use the Ordered Probit Model for regression analysis. The results are shown in TABLE II.

TABLE II. PRINCIPAL COMPONENT ANALYSIS AND VARIABLE INTERPRETATION.

<table>
<thead>
<tr>
<th>Ordered probit regression</th>
<th>Number of obs = 3334</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR chi2(6) = 321.43</td>
<td></td>
</tr>
<tr>
<td>Prob &gt; chi2 = 0.000</td>
<td></td>
</tr>
<tr>
<td>Log likelihood = -4073.1997</td>
<td></td>
</tr>
<tr>
<td>Pseudo R2 = 0.0380</td>
<td></td>
</tr>
</tbody>
</table>

\[
Y = 0.11(FAC1) - 0.12(FAC2) + 0.28(FAC3) - 0.06(FAC4) - 0.10(FAC5) + 0.04(FAC6) \] [1]

Y Represents the public satisfaction with urbanization.

3. LATENT DIRICHLET ALLOCATION (LDA) MODEL

Faced with the complex problem of public satisfaction analysis, it is necessary to conduct a comprehensive analysis of public opinion information from multiple thematic perspectives, so as to obtain more accurate results of public opinion analysis. Therefore, this paper intends to construct a LDA (Latent Dirichlet Allocation) probabilistic topic model to identify large-scale Chinese text data.

Latent Dirichlet Allocation (LDA), a total probability generative model, is a three-tier Bayesian model. LDA use a K-dimensional latent random variable which obeys the Dirichlet distribution to represent the topic mixture ratio of the document, which simulates the generation process of the document. At present, the LDA model is one of the most popular probability topic models, and it has more comprehensive assumptions of text generation than other models.

In the process of text generation, LDA uses sampling from the Dirichlet distribution to generate a text with the specific topic multinomial distribution, where the text is usually composed of some latent topics. And then, these topics are sampled repeatedly to generate each word for the document. Thus, the latent topics
can be seen as the probability distribution of the words in the LDA model. And, each document is expressed as the random mixture of these latent topics according to the specific proportion.

Figure 1 is the graphical model representation of LDA. The boxes are “plates” that represent replicates. The outer plate represents documents, which means that the topic distribution $\theta$ is sampled repeatedly from the Dirichlet distribution for each document in the document collection. And, the inner plate represents the repeated choice of topics and words within a document, which means that words in a document are generated by repeated sampling from the topic distribution. In addition, the hollow circles represent the latent variables $\alpha$, $\beta$, $\theta$, and $z$, the filled circle represents the observed variable $w$, and the black arrow represents the conditional dependency between two variables, namely the conditional probability.

Figure 1. The graphical model representation of LDA.

The LDA model is a typical directed probability graph model with a clear hierarchical structure, such as the document collection-tier, the document-tier and the word-tier. The LDA model is determined by the parameters $(\alpha, \beta)$ of the document collection-tier, where $\alpha$ reflects the relationship between the latent topics in the document collection and $\beta$ reflects the probability distribution of all the latent topics themselves. $\alpha$ and $\beta$ are assumed to be sampled once in the process of generating a document collection. The random variable $\theta$ is the document-tier parameter, and its components stand for the proportion of various latent topics in the destination document. $\theta$ is sampled once per document. In addition, $z$ and $w$ are the word-tier parameters, where $z$ represents the proportion of the destination document assigning the latent topics to each word and $w$ is the representation of the word vector of the destination document. $z$ and $w$ are sampled once for each word in each document. In brief, for a document, its topic distribution $\theta$ is a Dirichlet prior function based on the parameter $\alpha$. For a word $w$ in a document, its topic $z$ is generated by the topic distribution $\theta$, and the word $w$ is generated by the parameter $\beta$.

The process of LDA probability topic model generating a document is described as follows [2]:
Step 1: For the topic \( j \), according to the Dirichlet distribution \( \text{Dir}(\beta) \), a word multinomial distribution vector \( \theta^{(j)} \) on \( j \) is obtained.

Step 2: According to the Poisson distribution \( \text{Poisson}(\xi) \), the number of words \( N \) in the document is gained.

Step 3: According to the Dirichlet distribution \( \text{Dir}(\alpha) \), a topic probability distribution vector \( \theta \) of the document is got, where \( \theta \) is a column vector and \( \alpha \) is a parameter of the Dirichlet distribution.

Step 4: For a word \( w_n \) of \( N \) words in the document: A topic \( k \) is chosen randomly from a multinomial distribution \( \text{Multinomial}(\theta) \) of \( \theta \). A word is selected from a multinomial conditional probability distribution \( \text{Multinomial}(\phi^{(k)}) \) of the topic \( k \), which can be seen as \( w_n \).

As shown in Figure 1, the distribution of the hyper-parameters \( \alpha \) and \( \beta \) can control the topic distribution \( \theta \) and the word distribution on topics \( \phi \) by the Dirichlet distribution \( \text{Dir}(\alpha) \) and \( \text{Dir}(\beta) \)[2]. And then, \( \theta \) and \( \phi \) codetermine each word in the document. For the document \( d \), the joint distribution of all the known and latent variables is as follows:

\[
p(w_d, z_d, \theta, \phi | \alpha, \beta) = p(\phi | \beta) \prod_{n=1}^{N_d} p(w_{d,n} | \phi^{(z_{d,n})}) p(z_{d,n} | \theta_d) p(\theta_d | \alpha)
\]

(2)

After eliminating the variables \( \theta_d, \phi, z_d \), the probability distribution of \( w_d \) is obtained as below:

\[
p(w_d | \alpha, \beta) = \int \int p(\theta_d | \alpha) p(\phi | \beta) \prod_{n=1}^{N_d} p(w_{d,n} | \phi^{(z_{d,n})}) d\phi d\theta_d
\]

(3)

Therefore, for the whole text set \( D \), its corresponding probability distribution is as follows:

\[
p(D | \alpha, \beta) = \prod_{d=1}^{|D|} p(w_d | \alpha, \beta) = \prod_{d=1}^{|D|} \int \int p(\theta_d | \alpha) p(\phi | \beta) \prod_{n=1}^{N_d} p(w_{d,n} | \theta_d, \phi^{(z_{d,n})}) d\phi d\theta_d
\]

(4)

In order to obtain the probability topic distribution of text, this thesis does not directly calculate the word distribution on topic \( \phi \) and the topic distribution on document \( \theta \). Instead, according to the visible word sequence in document, the posterior probability \( p(w | z) \) namely probability of word \( w \) giving topic \( z \) can be obtained. At last, using the Gibbs sampling algorithm can indirectly gain the value of \( \phi \) and \( \theta \).
4. SENTIMENT ANALYSIS

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information. Sentiment analysis is widely applied to voice of the customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation, affective state or the intended emotional communication. [3]

We artificially annotate the text information of public sentiment, and construct three types of topic feature thesaurus. The first type of topic feature thesaurus is clustering according to the summary results. Each thesaurus represents one category. The thesaurus contains all the topic feature words belonging to this category labeled manually, which are used to extract the topic words. Finally, the topic words are classified and archived according to the content of the thesaurus. [4]

The second type is the emotional polarity vocabulary. According to the co-occurrence probability of emotional words in the dictionary and corpus, the emotional vocabulary of text commentary is constructed. The positive and negative values are used to express the positive and negative emotional words, i.e. the extreme values. "1" represents positive words, including positive evaluation words and positive emotional words; and "0" represents neutral words. "-1" represents negative words, including negative evaluation words and negative emotional words. [5]

The third type of vocabulary, polarity intensity vocabulary, is used to determine polarity intensity. This vocabulary is composed of emotions words. [6]

According to the users’ comments of public sentiment in social media, emotional distribution include positive emotions M, neutral emotions K and negative emotions N. Define the sum of the three as emotional polarity $E$, such as Formula 1:

$$E_0 = M_0 \times 1 + N_0 \times (-1) + K_0 \times 0$$  \hspace{1cm} (5)

$M_0$ represents the proportion of positive emotions $M$ in all emotional words in public sentiment, $N_0$ represents the proportion of negative emotions $N$ in all emotional words in public sentiment, and $K_0$ represents the proportion of neutral emotions $K$ in all emotional words in public sentiment.

$$E_f = 5 \times A_x + 4 \times B_x + 3 \times C_x + 2 \times D_x + 1 \times E_x$$  \hspace{1cm} (6)
$E_f$ stands for the emotional degree of each comment content, $A_x$ stands for the percentage of emotional words with the strongest intensity, i.e. grade 5, $B_x$ stands for the percentage of emotional words with the strongest intensity, i.e. grade 4, and $C_x$ stands for the percentage of emotional words with the strongest intensity, i.e. grade 3, in this user comment. In contrast, $D_x$ represents the percentage of emotional words with weaker intensity, i.e. grade 2, in this user comment, $E_x$ represents the percentage of emotional words with weakest intensity, i.e. grade 1, in this user comment. $A_x$, $B_x$, $C_x$, $D_x$, $E_x$ may be negative, representing negative emotions, subtracting from the formula.

ACKNOWLEDGEMENTS

This paper is sponsored by “Young Key Teachers Training Program of Henan Higher Education Institutions (2017GJS256)” and “Henan soft science research project (182400410289)”.

REFERENCES