Analysis of User’s Browsing Microblog

Lanying Li, Yong He and Xiaofang Li

ABSTRACT

The behavior of the users in the microblog is often influenced by factors such as age, education and so on. In this paper, the conditional Markov model in a particular situation is established to analyze the behavior of users in the microblog. The method of conditional transition probability calculation is analyzed and the conditional transfer matrix is established. The matrix is a sparse matrix which can be represented by list of 4-tuples. The method of removing the invalid data is used in data processing, which makes the data more reflect the common characteristics of users. In conditional Markov model, the scope of sample space is narrowed and characteristics are reflected accurately by the random events. Therefore, the prediction of the users' browsing microblog will be more accurate.

INTRODUCTION

Microblog is an important means for people to obtain information in modern society, and it is a shared platform for information acquisition and dissemination. By the end of 2012, Sina microblog registered users have exceeded 500 million [1]. There is a very large amount of information on the microblog, including all aspects of social life, which spread fast. Users can express their own opinions, and express
their emotions by microblog platform. The analysis of microblog data has become a very important research field [2]. Domestic and foreign scholars have conducted a lot of research on the microblog[3]. Zukerman [4] established user browsing analysis model. The user's browsing process is abstracted as a discrete Markov process. However, the model is simple and the accuracy is not high.

Browsing the microblog content, tweeting to expressing views, recommending content, turning to focus on the new microblog from friends’ recommendation, users have their own feelings and hobbies. These features are related to the age, the level of knowledge, life experience, and so on. A large number of these different groups of users showed some correlation. They have a certain similarity in the characteristics of users in microblog. The users’ next microblog can be predicted and recommended through the analysis and data mining of different user’s behavior features. It can provide decision-making services for the construction and development of the microblog site. To master the behavior process and tendency of users can serve for education and politics.

Different users content different microblog, which is affected by many factors, such as age, education, interest and so on. There is a major factor in these factors. The main influencing factor is used to analyze the behavior of users in microblog. Based on the major factor, the conditional Markov model is constructed to analyze the behavior of the users, which is the idea of this paper.

1 MARKOV MODEL

An user browses the microblog from a topic to another, which is random. A large number of users’ browsing microblog present a random feature. In order to better analyze the behavior of users in the process of microblog, the following assumption is made in this paper.

**Hypothesis 1** Assuming that the topic transfer of users in the microblog is random, and the users will concern about the next microblog only related to the current microblog and not related to the previous microblog. Such process of users in the microblog can be regarded as Markov chain.

Let $X$ be a discrete random variable. The state space is $E=\{b_1, b_2, \cdots, b_n\}$, $b_i$ represents a microblog topic, for different research, you can set different size of the content scope. The expression of $P\{X(n+1)=b_{n+1}|X(n)=b_n\}$ indicates the probability of users concern about microblog $b_{n+1}$ after microblog $b_n$. The expression of $P\{X(n+1)=b_{n+1}|X(1)=b_1,X(2)=b_2,\cdots,X(n)=b_n\}$ indicates the probability of users concern about microblog $b_{n+1}$ after microblog $<b_1, b_2, \cdots, b_n>$. The formula of $p_{ij}(n) = P\{X(n+1)=b_{n+1}|X(n)=b_n\}$ is set up, and one step transition probability matrix is $P$. 
The users can carry out a series of activities in a certain topic, such as browsing the content, tweeting to expressing views, recommending content to others, and so on. As shown in Figure 1, the circles represent the activities of the users. An user after browsing (b) or evaluating (e) in the topic 1 (T1), turns to the topic 2 (T2) (to browse (b), tweet (t) and so on), and then enters the topic 3 (T3) to conduct activities. By collecting the user's activities, we can get the corresponding data.

Ordered pair \(< b_i, b_j >\) indicates that the user concerns about the microblog \(b_j\) after browsing microblog \(b_i\). \(< b_{k1}, b_{k2}, b_{k3}, \cdots, b_{km} >\) represents the user's entire process, which can be decomposed into a number of different ordered pairs \(< b_{ki}, b_{kj} >\). All users of the microblog process are counted. The maximum likelihood estimation method is used to calculate.

\[
P_y = P\{X(j) = b_j | X(i) = b_i\} = \frac{\sum N_{ij}}{N_{ij}}, \text{ in which } N_{ij} \text{ means the count of ordered pair } <b_i , b_j> \text{ of all users.}
\]

Hypothesis \(v(i)\) that indicates the user's original concern of the microblog is \(b_i\). The i dimension of vector \(v(i) = (0,0,\cdots,1,\cdots,0)\) is 1. The other dimensions are 0. Let PT be multiplication of \(v(i)\) and Pk. \(P^T = v(i)P^k\). According to the properties of Markov chains, the microblog of the maximum probability value in PT is most likely to be concerned after k times. In particular, when k=1, the microblog of the maximum probability value in PT will most likely to be browsed in the next time.
2 CONDITIONAL MARKOV MODEL

Users concern about a certain microblog which are influenced by factors, such as age, gender, education, working background, personal hobbies and other factors. There is a major factor in these factors. The main influencing factor is selected to analyze the behavior of users. Based on this factor the characteristics of users in microblog are discussed, which is the idea of conditional probability. The conditional Markov chain was constructed in the paper. Conditional Markov chain narrows the scope of the sample space, and users’ behavior in the microblog is able to reflect the original features of the sample. So it is more accurate. In addition, conditional Markov model is used to partition the sample space, and so the complexity of the problem is reduced. According to the characteristics of users’ information, such as age, education, etc., and users can be divided into different groups. For different characteristics of the microblog site, the different factor is selected to classify user groups. In order to discuss the conditional Markov chain, the following assumption is made in this paper.

**Hypothesis 2** Assuming users in the process of microblog to meet the hypothesis 1, setting $H$ is the main factor. $h_1, h_2, h_3, \cdots, h_m$ are mutually incompatible $m$ partition in $H$. That is the function of $H\{h_1, h_2, h_3, \cdots, h_m\}$.

For $\forall i, j \in m$, when $i$ is not equal to $j$ ($i \neq j$), that is $h_i \cap h_j = \phi$.

**Definition 1** The formula of $p_y(n | h_k) = P\{X(n+1) = h_{n+1} | X(n) = h_n, H = h_k\}$ is one step conditional transition probability. It is probability of ordered pair $< b_i, b_j >$ of all users in $h_k$. It means that the probability of concerning about the microblog $b_j$ after having concerned microblog $b_i$ by users of $h_k$ sample space.

Under the division on the condition of $H$, the conditional transition matrix (1) is established.

$$
P_k(n | h_k) = (p_y(n | h_k)) = \begin{bmatrix}
p_{11}(n | h_k) & p_{12}(n | h_k) & \cdots & p_{1m}(n | h_k) 

p_{21}(n | h_k) & p_{22}(n | h_k) & \cdots & p_{2m}(n | h_k) 

\cdots & \cdots & \cdots & \cdots 

p_{m1}(n | h_k) & p_{m2}(n | h_k) & \cdots & p_{mm}(n | h_k)
\end{bmatrix}
$$

(1)

According to the hypothesis 2, we can see that there are $m$ different conditional transfer matrices for users’ space with $m$ partitions.

The vector $w(i, h_k) = (0, 0, \cdots, 1, \cdots, 0)$ indicates the $i$ dimension is 1, and the other dimensions are 0. The formula (2) is established.

$$
P^T (h_k) = w(i, h_k) P_k(n | h_k)
$$

(2)
The microblog of the maximum probability value in $P^T(h_k)$ is the microblog which can be concerned about most likely by users in hk in the next time.

$$P^T_u(h_k) = w(i, h_k) [P_i(n \mid h_k)\cdot P_k(n \mid h_k)\cdots \cdot P_k(n \mid h_k)]$$ \hspace{1cm} (3)

In which $[P_i(n \mid h_k)\cdot P_k(n \mid h_k)\cdots \cdot P_k(n \mid h_k)]$ indicates the multiplication of u conditional transfer matrixes. The microblog of the maximum probability value in $P^T_u(h_k)$ is the most likely to be browsed in the next u times. After dividing users according to the factor h, all users concerned about microblog sequence are counted respectively in the range of hk (k=1, 2, 3…m). By using the maximum likelihood estimation method, we can get the formula (4).

$$p_q(n \mid h_k) = P\{X(n+1) = b_{u+1} \mid X(n) = b_n, H = h_k\} = \frac{N_q(h_k)}{\sum_{j=1} N_q(h_k)}$$ \hspace{1cm} (4)

In formula (4) $N_q(h_k)$ means the number of ordered pair < bi, bj> of users in hk. On the basis of the previous analysis, the conditional transition matrix is given and the user's behavior in microblog is analyzed in the following.

**Algorithm 1** Constructing conditional transition matrix

**Input** Users sequence in the microblog of hk (k=1, 2, 3…m).

**Output** m conditional transition matrixes.

**Process**

For i=1 to m

- For j=1 to n

  - To calculate the one step conditional transition probability of class K users by the formula (4);

  - To construct the conditional transition matrix of class K users by the formula (1);

For different types of users, according to the current state of the user, the next focused on microblog can be analyzed by the formula (2), the next u times concentrated microblog can be analyzed by the formula (3).

**Algorithm 2** Users’ browsing analysis

**Input** The state vector of the users, and the value of the users belong to the category of hk.

**Output** The microblog which the user will pay attention to in the next time, the microblog which the user will concern about in the next u times.

89
**Process**

Step 1, to find the conditional transfer matrix according to the classification of the value of $h_k$;

Step 2, to calculate the next possible concerned microblog by using the formula (2);

Step 3, to find out the microblog of the maximum probability value of the vector $P^T(h_k)$;

Step 4, to calculate the microblog in the next $u$ times that may be concerned by using the formula (3);

Step 5, to find out the microblog of the maximum probability value of the vector $P_u^T(h_k)$;

---

**3DATA PROCESSING AND STORAGE**

**3.1 Data Processing**

In the actual, users can browse a microblog at will. This is purely accidental. It does not represent the tendency of the majority users’ behavior. Therefore, the invalid data which deviate from the data of the majority users’ behavior should be removed. As shown in Figure 2, after concerning about the microblog $b_m$ by the users, the times of paying attention to the microblog $b_i$ is $N_i$, and the times of paying attention to the microblog $b_j$ is $N_j$. The times of paying attention to the microblog $b_n$ after $b_i$ is $M_i$. The times of paying attention to the microblog $b_n$ after $b_j$ is $M_j$.

In this paper, the formula (5) is used to estimate the degree of deviation of the data.

$$H = \frac{(N_i + M_i) - (N_j + M_j)}{(N_i + M_i)}$$  \hspace{1cm} (5)

Set a threshold $h_0$. The activity of $b_i$ is very small when $H > h_0$. It is the user's casual behavior, which is the invalid data and does not represent the characteristics of users. Therefore, it should be removed.
3.2 Data Storage Analysis

In the actual microblog site, the number of the microblog topic is quite large and the number of users is large. In fact, the number of concerned microblog is very limited. Especially under condition of partitioned users, the specific users are often focused on the topic of a few kinds of microblog, so there is a considerable number of microblog without users’ browsing. As a result the probability is 0. According to the formula (4), there are many elements in the conditional transition matrix to be zero, which is a sparse matrix. The zero elements in the matrix are not stored, which compress the matrix storage space. It can be represented by list of 3-tuples or cross linked list. Due to $h_k$ for the classification of users, list of 4-tuples is used to storage in the conditional Markov model.

$$( h_k, i, j, p )$$

In which $h_k$ represents the interval value of the users’ division, $i$ is the row index of the conditional transition matrix, $j$ is the column index of the conditional transition matrix, $p$ is the probability value of the conditional transition matrix.

In addition, in the actual site, the number of accessing microblog is rare, and its value in the calculation of the probability is very small. For convenient handling, this rarely visited microblog is regarded as zero. This further increases the zero elements of the conditional transition matrix. The conditional Markov model is $m$ times of Markov model because of the need to store $m$ conditional transition matrixs, that is $O(mn^2)$. In this paper, the list of 4-tuples is used to store only nonzero elements. The storage space of the conditional Markov model will not be much.

4 ANALYSIS OF COMPLEXITY

Markov model and conditional Markov model need to scan all users’ browsing the microblog sequence to establish the transfer matrix. Conditional Markov model is needed to scan the different subsets. As a result, the counts of scanning in two
algorithms are the same. Just the number of the matrixes is different. Only one matrix is built in Markov model, while m matrixes are built in conditional Markov model.

The different conditions of the transfer matrix are searched in conditional Markov model for different users. The interval divisions of users are few. So the number of conditional transfer matrix is not much. After finding the corresponding conditional transition matrix, according to the formula (2) or formula (3), the computational complexity of the two algorithm is the same.

CONCLUSIONS

Markov model is an effective way to analyze the behavior of users. In this paper, a major one of many factors, such as age, education, gender, personal preferences, is selected to analyze the conditional probability of user's attention to microblog. The conditional Markov model is established to analyze the behavior of users. Compared with the Markov model, the conditional Markov model reduced the scope of the sample space. It is the probability of occurrence under certain conditions. It is better on reflecting the user's browsing characteristics and the analysis is more accurate. In addition, the method of data miscellaneous is used in the process of data. The data is able to reflect the common characteristics of groups. List of 4-tuples is used to store non zero elements in the conditional Markov model. Compared with the Markov model, the storage space is not increased. The computational complexity is similar.

ACKNOWLEDGEMENTS

This work was financially supported by the Education Department of Sichuan Province (16ZB0404) and the Engineering & Technical College of Chengdu University of Technology (C122015008).

REFERENCES