Modeling Research on Micro-blog Users

Chen Zhihua
Computer & Network Center
Guangdong Polytechnic Normal University
Guangzhou, China
e-mail: czh@gdin.edu.cn

***Abstract*—Through the research and analysis of information characteristics on micro-blog users, this paper defines a micro-blog users model and proposes a method for constructing the micro-blog user model which combining theme representation with vector space model representation. It also designs a learning algorithm with dynamic update user model, and proves the learning algorithm effective by experiments.**

Keywords-micro-blog;user modeling;iInterest mining

I. INTRODUCTION

Micro-blog is a micro-blogging, which is based on the information platform for sharing, communication and accessing among the user's relations. Through the WEB, WAP and various clients, Users may build the formation of individual communities and realize the instant sharing with around 140 words to update information. Owing to its characteristics of brief text spread fast, update convenient, etc.. Micro-blog becomes very popular for people to obtain the latest information, review of media information and expand personal social network. Up to February,2012, the largest domestic micro-blog site of China -- Sina micro-blog, registration has more than 300000000 users, The Number for daily publishing micro-blog reached more than 100000000, active user accounts for more than 9% among total users[1]. In such a huge information platform, users could be able to enjoy the efficient, personalized service, and more accurate advertisement content as many scholars and research focus. In personalized information services, the research on user modeling technology has become the key technology of personalized service. In this paper, through the analysis of information characteristics of micro-blog users and research on the user model construction method, a method of constructing micro-blog user model is introduced which is based on combing the method of theme representation and vector space model representation.

II. MICRO-BLOG USER MODELING AND FEATURE SELECTION

A. Information sources selection of personalized characteristic for Micro-blog users

(1)user information

Prior to using micro-blog system, people must register individual account and fill in personal information, including basic information, education information, occupation, personality label information. The basic information mainly include: real name, location, sex, birthday, e-mail etc. Nickname is the explicit marking, for users' communication, can be modified at any time. Education information include: education, school year, the school name and the school faculty. Occupation information include: a place of work, company, work time, work department or position. Personality label is described personal characteristics and hobbies with keywords, by the user's selection. According to the above user information the individual information of user features can be summarized as shown in table I.

TABLE I. EXTRACTION FEATURE OF THE USER INFORMATION

Characteristics of theme	Feature item
Basic information	{nickname, gender, age}
Education information	{degree, graduate school, learning professional}
Occupation information	{work city, work unit, post, occupation category}
Personality label	{label 1, label 2,...,label n}

(2) micro-blog information and user behavior

Micro-blog user's information includes text information, insert pictures, music, video, webpage link. For Micro-blog users, there are a lot of behaviors in the use of micro-blog process. Generally speaking, it can be divided into active and passive behavior according to the fact that user is the act or behavior of the recipient. Micro-blog users active behavior is focusing on the user published micro-blog, forwarding micro-blog, comment micro-blog, @ (and XX), personal information, label, join the micro group, participate in topics and activities. Passive behavior of Micro-blog users to be concerned about forwarded micro-blog, micro-blog comments and so on. According to the user information and activities of micro-blog extracted personalized features as shown in tableII.

TABLE II. THE USER INFORMATION AND ACTIVITIES BY MICRO-BLOG FEATURE EXTRACTION

Characteristics of theme	Feature item
Effect	{audience numbers, released micro-blog quantity, forwarded micro-blog number}
Active degree	{comment number, the number of the user to listen, forwarding number of micro-blog, topic quantity}
Interests and hobbies	{ interest 1, interest 2,..., interest n}

B. user model representation

The user model is used to describe, store and manage user characteristic information such as gender, age, education, profession, occupation, interest. User model is not a general description to user information, but a kind of algorithm oriented, with a specific data structure, formal description. User modeling is a kind of calculation process summarized from the relevant user interest and behavior information[2]. Only when the user information from the non-structure of the original form converted to the structural form understand by the computer, the user interest analysis and processing can be conducted.

User modeling determines its reflect real user information ability and computational capability, while at a certain extent it restricts the user modeling method selection. At present the common user model representations include: theme representation, key words list representation, neural network representation, Ontology of the representation and the representation based on vector space model. Theme representation expresses user models with user preference theme. If the users have the interested in such aspects recreational and scientific as basketball, the user model can be represented as {entertainment, science and technology}; for the key words list representation to the user interested information, key words to express user models, such as users of basket ball, then user interest model for {NBA, CBA, Jordan, Kobe, Yao Ming}. Representation based on neural network with stable network, Network connection weights are characteristic of network state to represent a user model; representation based on ontology display a user's field of interest, such as the Quickstep system[3]. Up to now, the most popular representation is the vector space model representation. it will be made the user model expressed as an n-dimensional characteristic vector $\{(k_1,w_1),(k_2,w_2),\ldots,(k_n,w_n)\}$, each of its one-dimensional component is consists of words and weights, the weight means the level of interest on a concept for a certain user.

As for the information characteristics of the micro-blog user, the topic representation based on vector space model will be adopted in this paper. The user model design for the structure can be divided into two layers, the first one is a characteristic theme, using the topic representation, expressed as UM= $\{ u_1,u_2,\ldots,u_n \}$; the second one is a characteristic theme of each feature, for a fixed feature, use the subject representation, such as basic user information corresponding to the feature, said ui=$\{t_1,t_2,\ldots,t_n \}$, and the feature of interest subject corresponds to the feature is composed of micro-blog text mining and extraction of key words, and each key words weight is not the same, so the use is based on vector space model representation, expressed as $W=\{(k_1,w_1),(k_2,w_2)\ldots,(k_n,w_n)\}$. The user model structure diagram is shown in figure 1.

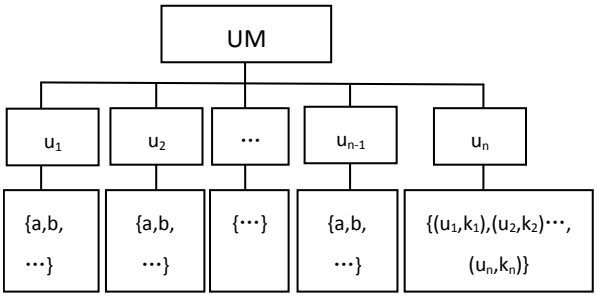

Figure 1. The user model structure

Micro-blog user model UM=$\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$={basic information, education information, occupation information, personalized labels, influence, active degree, hobbies}. Each topic contains features as follows:

$u_1=\{x_{11},x_{12},x_{13}\}$={nickname, gender, age};

$u_2=\{x_{21},x_{22},x_{23}\}$={degree, a graduate school, learning professional};

$u_3=\{x_{31},x_{32},x_{33},x_{34}\}$={work city, work unit, post, occupation category};

$u_4=\{x_{41},x_{42},\ldots,x_{4n}\}$={label 1, label 2,...,label n}

$u_5=\{x_{51},x_{52},x_{53}\}$= {audience numbers, released micro-blog quantity, is forwarded micro-blog number};

$u_6=\{x_{61},x_{62},x_{63},x_{64}\}$={comment number, the number of the user to listen, forwarding number of micro-blog, topic quantity};

$u_7=\{(k_1,w_1),(k_2,w_2),\ldots,(k_n,w_{n)}\}$= { (interest 1,interest1's weight), (interest 2,interest2's weight),..., (interest n, interest n's weight)}.

C. micro-blog users model construct and update

As the users' first login micro-blog platform, they establish user model and fill in the registration information. While the materials about the user's basic information, education, occupation and personality label feature etc. can be extracted from the registration information because of no recording of the published micro-blog, forwarding micro-blog, the influence of active. Only the above three items initialized to null will system be collected the relevant information in the process of using the micro-blog activities for the further improvement of the user model.

The model of interest eigenvector collection and extraction is the most important part of the study; the main steps of the algorithm are as follows:

(1) Collection of interest in text set, the main collection released by the user, forwarding the micro-blog text and topic text;

(2) The interest of text segmentation, POS tagging information obtained with a list of words;

(3) The noise words for initial filtering, i.e. remove low-frequency words and high frequency words and other words;

(4) the part-of-speech tagging information will be nouns using TFIDF, non-nouns using word frequency method for feature selection, and in accordance with the nouns extracted from 12%, 8% is not a noun extraction ratio determined threshold, and then get the keyword feature vector;

(5) feature words and their weights write into topic feature set for user model hobby . The feature weight is calculated as follows:

Set the collected text set as a total of N, N set to N={ $d_1,d_2,\ldots,d_n$ }, located from the document Di get words set to d_i={ $k_1,k_2,\ldots,k_n$ }, then the feature weights calculation formula for w_{ij}:

$$w_{ij} = t_{fij} \times id_{fj} = tf_{ij} \times \log(N / n_j) \quad (2.1)$$

In which:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (2.2)$$

$$idf_i = \log \frac{|N|}{|\{j : k_i \in d_j\}|} \quad (2.3)$$

In formula (2.2), n_{ij} is the number of the word occurrences in the document d_j, and the fraction is the number of occurrences of all words in document d_j.

As the characteristics of the user's interest is dispersed in the interest text set , therefore interest model improvement depends on the user interest in text set quantity. As for less use micro-blog users will cause the interest model accuracy and completeness due to their less interest text. In addition, for users, there are short-term interest and long-term interest, so the use of the above learning algorithm the regular collection and extraction of user interest key words in order to constantly improve the user model.

III. EXPERIMENTAL MODELING AND MODEL EVALUATION

Micro-blog user modeling is the most complex and important part. Therefore this section primarily test and verify the user interest model. Interest model in the experiment are generally calculated by the Chinese lexical analysis system ICTCLAS that invented by the Chinese Academy of Sciences and on the basis of which conducted text segmentation and POS tagging. Evaluation indicators are used widely in the field of data mining and information retrieval using precision and recall.

$$\text{Precision} = \frac{\text{The number of correct key words in the interest model}}{\text{The number of interest key word of the user}} \quad (3.1)$$

$$\text{Recall} = \frac{\text{The number of correct key words in the interest model}}{\text{The number of key word in the interest model}} \quad (3.2)$$

Considering the current usage status of micro-blog users are mainly young people, The 50 students will be given Sina micro-blog accounts those who in our university selected from three grades and six schools as computer software engineering, electronic information engineering, network engineering, English translation, Chinese language and literature, management. A total of 300 micro-blog users information will be taken as user modeling test object.

Through API provided by Sina micro-blog, using tools reptiles to obtain the 300 micro-blog users to publish, forwarding the micro-blog and attention topic text information, through the text segmentation, in accordance with the nouns extracted from 12%, 8% is not a noun extraction of keywords, and then according to the formula of calculating the weight value of 2.1 character words, that the user's interest model. In addition, through the questionnaire survey, the 300 micro-blog user interest model is taken as a test of the interest model. Taking into account the micro-blog users to publish, transmit and concerns of the micro-blog information (collectively referred to as interest in text) too small to accurate and comprehensive access the features of word, so users are divided into three categories, tables were calculated for these three types of user interest model precision and recall, results as shown in table III.

TABLE III. TABLE 3.1 THREE TYPES OF USER INTEREST MODEL EVALUATION INDEX DATA TABLE

User classification	Classification of terms	User number	Average precision	Average recall
Class A	text number≤10	21	0.12	0.13
Class B	10 < Interest in text ≤30	46	0.35	0.41
Class C	Interest in text >30	233	0.86	0.84

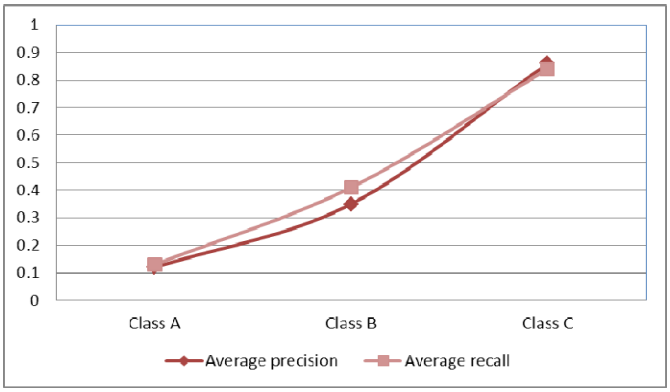

Figure 2. Three types of user interest model evaluation of line chart

Through the above analysis of data obtained, micro-blog user interest model depends on the user interest in the amount of text, text with interest in quantity. Average precision and recall is more and more high. It reaches 0.86 and 0.84 for the user interest text number over 30, which proved this method can accurately find the user interest class.

IV. CONCLUSIONS

This paper analyzes the micro-blog users information features, and explores the micro-blog user modeling and feature selection techniques, on the basis of which introduces a theme representation based on vector space model representation method of constructing micro-blog user model. In addition, dynamic learning algorithm , dynamic update user interest model is proposed, Finally, experimental results proved the validity and accuracy of the learning algorithm.

REFERENCES

[1] Sina science and technology. Results [EB/OL]. 2012, 2

[2] Middleton S E,Shadbolt N R.Ontological User Profiling in Recommender Systems[J].ACM Transactions on Information System,2004,22(1):54-88.

[3] Christian S, Emmanuel F.Soundspotter－A Prototype System for Content-based Audio Retrieval[C]//Proc.of Digital Aodio Effects.2002.

[4] Huang He, Huang Hai, Wang Rujing. FCA-Based Web User Profile Ming for Topics of Interest[C]. Proceedings of the 2007 IEEE International Conference on Integration Technology. Shenzhen, China. 2007:20-24

[5] Pazzani M, Billsus D. Learning and Revising User Profiles: The identification of interesting Web sites[J]. Machine Learning27. 1997:313-331

[6] Fragoudis D. User Modeling in Information Discovery: An overview[C]. Proceedings of Advanced Course on Artificial Intelligence, 1999(ACAI99), July 1999, Greece