Building User Micro-blog Semantic view based on topic feature extraction

Shunxiang Zhang
Anhui University of Science & Technology
Anhui, China
sxzhang@aust.edu.cn

Shiyao Zhang
Anhui University of Science & Technology
Anhui, China
yao__ing@163.com

Guangli Zhu
Anhui University of Science & Technology
Anhui, China
glzhu@aust.edu.cn

ABSTRACT
Micro-blog has become an important crowdsensing place for a lot of real-time information dissemination and discussion. This paper explores building user micro-blog semantic view based on topic feature extraction to provide theoretical support for future application such as user clustering, micro-blog topic recommendation. First, user network is built to save the user relationship on micro-blog. In the process of building user network, user authority and micro-blog topic heat are respectively considered to pick up influential users and important topic. Second, the algorithm of generating user micro-blog semantic view is proposed to represent the content of a user historical micro-blog. User topic feature vector and user topic feature matrix are used to help compute topic similarity between differential users and build topic feature graph.

CCS Concepts
• Information systems→Information integration→ Wrappers(Data Mining)

Keywords
Micro-blog, user micro-blog semantic view, topic feature graph, user authority, micro-blog topic heat.

1. INTRODUCTION
As a type of crowdsensing media, micro-blog, can be displayed in a lot of electronic terminals such as laptop, iPhone and iPad. Recently, with the rapid development of micro-blog, the number of registered users on *Sina* micro-blog has reached 500 million according to statistics. The daily active users has reached 47 million. According to the statistics of reference7 , the average number of following users is 469 from 1.18 million users. These users will generate massive micro-blog information every day. Facing these micro-blog information, whether users are interested in each of them? Whether users can distinguish and analyze the micro-blog database effectively? That is, users are facing a very serious information overload.

To solve this problem, the first task is to build user micro-blog semantic view to semantically represent the user interests. An effective and reasonable user micro-blog topic semantic view should consider following problems. (1)The topic followed by influential users are usually easy to be accepted by other users. Thus, the way to choose influential users, is an inevitable problem. (2)How to extract user interested topics and calculate similarities across users.

In our proposed algorithm, user network is a data structure to save the user relationship on micro-blog. In this process, user authority is used to select influential users and micro-blog topic heat is used to pick up those important topics. Topic Feature Graph is utilized to express topics, User Micro-blog semantic view is to help analyze whether a user is interested in a topic. User Topic Feature Vector and User Topic Feature Matrix are extracted from the users' micro-blog history. It helps compute topic similarities between differential users on the basis of user topic feature vector and user topic feature matrix.

The building process of building user micro-blog semantic view in this paper includes two levels:

(1) **Building micro-blog user network:** Micro-blog User network will be built as a directed weighted graph. If a user follows another one, it will be considered as out-degree in directed weighted graph. On the contrary, it will be considered as in-degree. Through the PageRank algorithm, the PR value of each user is got from user directed weighted graph. The PR value is defined as the user authority ρ .

(2) **Building user micro-blog semantic view:** To make word segmentation for the historical micro-blogs of the users and Micro-blog topic, segmentation word features are extracted from user historical micro-blogs and Micro-blog topics, which bring in TFG and UMSV. User micro-blog topic feature matrix generation algorithm is proposed to generate Topic Feature Vector and User Topic Feature Matrix based on TFG and UMSV.

The rest of this paper is organized as follows: Section 2 introduces related works. Section 3 introduces the method of building micro-blog user network. Section 4 presents the algorithm of building user micro-blog semantic view. Conclusions are given in Section 5.

2. RELATED WORK
Keyword extraction is the fundamental of building topic expression model, many scholars have made contribution to this aspect. Li et al proposed a new multi-strategy keyword extraction method based on tf/idf. And the specification of keywords depended on analyzing linguistic characteristics of news documents[2]. Litvak et al proposed supervised and unsupervised graph-based approaches for the cross-lingual keyword extraction , which was used in extractive summarization of text documents[3].

* Corresponding Author: Shunxiang Zhang, sxzhang@aust.edu.cn.

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China

DOI 10.4108/eai.18-6-2016.2264147

Christian et al presented three statistical methods to improve keyword extraction that went beyond the use of tf-idf. Unlike the classical tf-idf measure however, they took the relations and context of words into account by using the so called co-occurrence distribution. This led to an improvement over tf-idf based ranking[4]. Chen has improved the efficiency of keyword extraction in a set of relevant Chinese documents by using a new PAT-treebased approach[5]. Jiao et al proposed a method of keyword extraction based on N-gram and word co-occurrence statistical analysis [6]. Zhang K et al proposed the significance of keyword extraction with global context information' should be raised and so as the "local context information" on the basis of Support Vector Machines [7].

3. BUILDING MICRO-BLOG USER NETWORK

3.1 Basic Conceptions

Definition1: Users Graph

Relationships between the users are only in two situations, one is following, the other is followed by others in the micro-blog. Relationships between users can be unidirectional, namely the user A follows the user B, user B does not follow the user A. Relationships between users can also be bidirectional, namely the user A follows the user B, user B follows user A at the same time. If users are regarded as the nodes in the graph, then the relations between users are the edge in this picture, and the edge in the graph are directed. In this way, the users and the relationships between them can construct a directed graph, we define the graph as a user network graph G. The user network G can be described as follows:

$$G = \left\{ \left(V_i, E_j \right) \mid i = 1,2,3\ldots k,\ j = 1,2,3\ldots n \right\} \quad (1)$$

Where V_i represents the nodes of the users, E_j represents the edge in the graph, that is relationships between users, k denotes the number of the users, n represents the number of the relationship between users.

Definition2: User Authority

For a micro-blog topic, influence degrees of micro-blog users to a topic is not the same. For example, a micro-blog topic followed by user with several million fans is easy to become a hot topic, and a micro-blog topic followed by user with several fans is less likely to be a hot topic. On the Internet, the PageRank algorithm can reflect the importance of a web page. This paper considers a micro-blog user as a web page, according to the PageRank algorithm calculation method, the value of the page PR is used to calculate the user's authority $\rho(U_i)$. If the topic followed by a high authority user, the chance of obtaining other users' following to the topic is much greater than the common topics. Users authority $\rho(U_i)$ can be described as follow:

$$\rho\left(U_i\right) = \left\{ \left(U_i, \rho\right) \mid i = 1,2,3\ldots k \right\} \quad (2)$$

Where U_i represents a single user, k is the number of the users, ρ is the value of the PR for the single user.

3.2 The Structure of User Data Source

Usually, the user id, nickname, location, gender, sexual orientation, relationship status, birthday, blood type are displayed in a micro-blog user basic information page. Some basic information can be chosen and defined as a 5-tuple for each user, namely U = $\{U_{id}, U_{name}, U_{address}, U_{sex}, U_{age}\}$.Where U_{id} is micro-blog ID of user, U_{name} is micro-blog nickname of user, $U_{address}$ is user's location, U_{sex} is the sex of user , U_{age} is the age of the user. The users are linked by the relations between the users. The relationships between the users are regarded as a graph. If the user A follows the user B, then user A, B can be connected. The user A is out-degree, the user B is in-degree. A directed graph G= (V, E) is constructed, the vertex V is the set of all micro-blog users, while edge E is the link between the users. The specific situation is shown in figure 2:

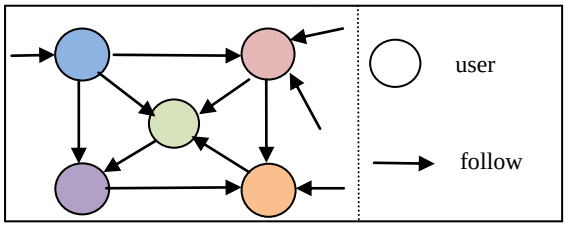

Figure. 1. User network graph

The circle represents a single user in the Figure2, and the arrows represent the relations. The user A points to the user B means A follows the user B. The user A, B, C, D, E and the relationships between them build a directed graph.

3.3 The Acquisition of the User Authority

The relationships between the users can be obtained according to the users network graph G mentioned before. The user authority ρ can be computed by the idea of PageRank algorithm based on these relationships.

The traditional PageRank algorithm is based on web link analysis for the search results of keywords. It draws on the traditional citation analysis idea: when a web page A has a link to a web page B, B is considered as getting a contribution value from the A. The value depends on the importance of web page A itself. The more important of the web page A is, the obtained value of the web page B is higher. The calculation formula of PageRank is as follows:

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(V)}{N_v} \quad (3)$$

Where, $B(u)$ represents the web page set that points to the web page u directly (u denotes the out-degree set; N_v indicates the number of the out-chain for web page v ; P $R(v) / N_v$ refers to the value of the web page v which is average assigned to its out-chain web pages; d is the damping coefficient, whose general value is 0.85.

From the idea of PageRank algorithm, a micro-blog user is considered as a web page. The user B will get the contribution value from the user A when the user A follows the user B. The more people follow the user A, the more important the user A is, and the more contribution value user B gets. The value in the last iteration equals to the user's authority value.

The similar authority ρ calculation formula which we can get is as follows:

$$P(u) = (1-d) + d \sum_{v \in W(u)} \frac{\rho(v)}{N_v} \quad (4)$$

Where $W(u)$ represents the user set that concern the user u directly; N_v denotes the number of the out-degree from the user v; $\rho(v) / N_v$ indicates the authority value of the user v which is average assigned to the users who have concerned him; d is the damping coefficient, the specific size obtained by experiments, tentatively scheduled for 0.85.

4. BUILDING THE USER MICRO-BLOG SEMANTIC VIEW

After constructing the users' network and calculating the user authority ρ, TFG and UMSV is built based on processing micro-blog topic and user historical micro-blog. Then, the micro-blog topics for each user are presented by user topic feature vectors. So that, user topic feature matrix TFM is proposed that combines with the users and the topic features.. There are two advantages: Firstly, it's convenient to calculate the topic similarity between the users. Secondly, it's easy to recommend topics for users through the micro-blog recommendation algorithm in next step.

4.1 Basic definition

Definition 3: Topic Feature Graph (TFG)

For each micro-blog topic, many high-frequency words will appear in discussions in the user micro-blog, these high-frequency words used a certain conditions are called feature words of micro-blog topic. According to reference [8], firstly, the segmenting micro-blog participates the topic, extracts feature keywords from segmented words, calculates weight of each keyword based on td-idf, and then minings relationship between each keyword according to association rule, finally each topic will be built as a network TFG,TFG is defined as follows:

$$TFG = \{T_{tp}, E_{tp}, W_{tp}\} \quad (5)$$

Where each item of feature keyword set T_{tp} is the node of the TFG, each item of the relationship set E_{tp} is the edge of the TFG, each item of the weight set W_{tp} represents each weight on every edge.

Definition 4: User Micro-blog Semantic View (UMSV)

Similarly, according to reference [8], we segment user historical micro-blog, extract feature keywords from segmented words, calculating weight of each keyword, mining relationship between each keywords, which are based on those processes built the UMSV for each user. UMSV is defined as follows:

$$UMSV = \{T_x, E_x, W_x\} \quad (6)$$

Where each item of feature keyword set T_x represents the node of the UMSV, each item of the relationship set $E_x = \{e_{12} \ldots e_{ij} \ldots\}$ stands for the edge of the TFG, each item of the weight set $W_x = \{w_{12} \ldots w_{ij} \ldots\}$ represents each weight on each edge.

Definition 5: User Topic Feature Vector (TFV)

Because each topic list that the user is interested in will construct a list, these lists can be represented by vectors. In this paper, the vectors are recorded as the user topic feature vector TFV, TFV is described as follows:

$$TFV = \{t_1, t_2, \cdots, t_m\} (i \geq 0, m \geq 0) \quad (7)$$

Where, t_i denotes the symbol that the topic is followed. When the topic is followed, the value of the t_i is 1. Otherwise, its value is 0. m is the number of the topics that be followed by all users.

Definition 6: micro-blog topic Heat(φ)

For a micro-blog topic, mentioned before, the probability of a topic which is followed by users, with high user authority degree be followed by other users is greater than a topic which is followed by users with authority degree. But for a micro-blog topic, when the heat of a topic is high, from another perspective, when the micro-blog topic becomes hot, the probability of other users following it will be greatly increased. In this paper, we define a micro-blog topic heat as φ, the topic of micro-blog heat φ can be described as follows:

$$\varphi = R / 100000 \quad (8)$$

Where, R denotes the current topic of search results returned by micro-blog search results, the search results of a topic can have very obvious reaction of the current heat of a topic .It impacts much for the future topic recommendation.

Definition 6: User Topic Feature Matrix(TFM)

According to definition 1, we get the user topic feature vector TFV, assuming that the number of users is n, then n-topic feature vectors constitute user topic eigenvectors TFM, the length of single user topic feature vector is m, the size of user topics eigenvectors TFM is n*m, the topic of user feature vector can be described as follows:

$$\lambda = a\rho + b\phi \quad (9)$$

Where a and b are two adjustable constants. Table 1 gives some examples of the topic matrix.

Table 1. user topic feature matrix

User/topic	Kobe retired	Curry MVP
User A	1	0
User B	0	0
User C	1	1

4.2 Acquisition of user topic matrix TFG

For the micro-blogs from the same topic, we segment it and extract feature keywords from segmented words, then calculate the weight of each keyword based on td-idf, and then through association rule mining relationship between each keyword, build each topic as a network TFG. Finally, we will acquire a set of TFG. For the historical micro-blog from a single user, we do the same things on them as what we do on micro-blogs. Then user historical micro-blog will be built as a network UMSV. Finally we will acquire a set of UMSV.

Figure 3 shows some examples about TFG and Figure 4 shows about UMSV.Based on TFG for each topic and UMSV for each user, this paper calculates matching degree for each TFG and each UMSV. When matching degree is greater than the threshold ϵ, we believe that the user is interested in this topic.

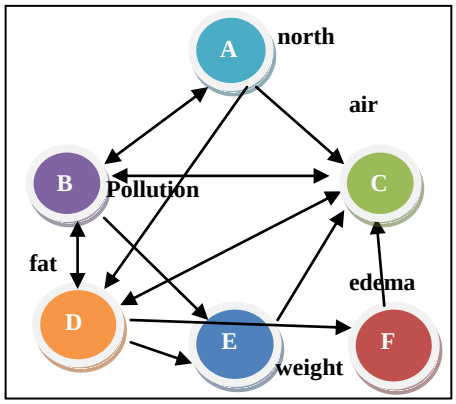

Figure 3 Topic Feature Graph

Figure 3 shows an example for TFG, node A to F represent some keywords like 'fat' and 'air'. These keywords and relationships constitute a Topic Feature Graph about "Air pollution can lead to fat".

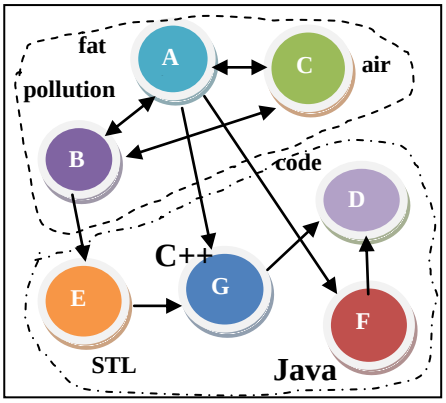

Figure 4 User Micro-blog semantic view

Figure 4 shows an example for UMSV, which contains two topics, that are, "Air pollution can lead to fat" and "two main programming languages". From this User Micro-blog Feature Graph, we can find the user is interested in the semantic "Air pollution->fat" and "C++->STL(Standard Template Library)".

In the following algorithm, step 1 build user topic characteristic matrix TFM. Step 3 to step 9 are used to initialize the user topic characteristic matrix TFM. Step 14 is used to acquire the marching degree for each TFM and each UMTF. Step 14 to 17 are used to realize matching. If the matching degree is greater than ϵ, then we set the value of this topic as 1, which means the user is interested in this topic. Step 12 to 18 are used to find each user's followed topic and we can set its value as 1.

Algorithm 1:User Micro-blog semantic view generation

Input: topic feature graph set TFG_Set={TFG_1,TFG_2,…,TFG_i,…, TFG_m}, the number of topics m, the number of users n, the User Micro-blog semantic view list set UMSV_Set={$UMSV_1$, $UMSV_2$, …, $UMSV_i$, …, $UMSV_n$}

Output:User topic characteristic matrix TFM

```
1：int TFM[n][m]
2：i=0,j=0
3 ：while(i<n)
4：    while(j<m)
5:        TFM[i][j] = 0
6：       j++;
7:    endwhile;
8：     i++;
9：endwhile
10: i=0;
11: int k
12：for UMSV in UMSV_Set
13：   for TFG in TFG_Set
14：    k = Match(TFG, UFMG);
15     if k>ϵ
16：       TFM[i][j] = 1;
17     endif
18:    j++
19:   endfor
21：  i++;
22：endfor
23：  end
```

The complex degree of this algorithm is mainly constituted by two double loops, the time complexity of first loop (steps 3 to 9) is O (n * m) and a time complexity of second loop[(step 12 to 22) is O (n*m*m) ,the time complexity the algorithm is O (n*m*m).The algorithm is mainly used for the n*m two-dimensional matrix to storage TFM, so the space complexity of the algorithm is O (n*m)

5. CONCLUSIONS

By using different electronic terminals, micro-blog has become an important crowdsensing place. This paper has built user micro-blog semantic view based on topic feature extraction to provide theoretical support for micro-blog topic recommendation.

Firstly, we have built user network to save the user relationship on micro-blog. In the process of building user network, user authority and micro-blog topic heat are respectively considered to pick up influential users and important topic.

Secondly, we have proposed the algorithm of generating user micro-blog semantic view which is proposed to represent the

content of user historical micro-blog. User topic feature vector and user topic feature matrix are used to help compute topic similarity between differential users and building topic feature graph.

Our future works will include exploring user clustering and the way to realize micro-blog topic recommendation system.

6. ACKNOWLEDGMENTS

This Research work was supported in part by the Natural Science Foundation of Anhui Province Universities (No. KJ2015A111), the Opening Project of Shanghai Key Laboratory of Integrate Administration Technologies for Information Security (Grant No.AGK2013002) in part by the National Science Foundation of China under (Grant No. 61300202).

7. REFERENCES

[1] Mu, F.N. Research on recommendation diversity for micro-blog users [D]. Harbin Institute of Technology, 2013

[2] Li, J.Z., Fan, Q., Zhang, K. Keyword Extraction Based on tf/idf for Chinese News Document . Wuhan University Journal of Natural Sciences, 2007, 12(5):917-921.

[3] Litvak , M., Last, M. Graph-based keyword extraction for single-document summarization. Mmies 08 Workshop on Multi-source Multilingual Information Extraction & Summar. 2008:17--24.

[4] Wartena, C., Brussee, R., Slakhorst, W. Keyword Extraction Using Word Co-occurrence. Proceedings of the 2010 Workshops on Database and Expert Systems Applications. IEEE Computer Society, 2010:54-58.

[5] Chen, L. F. PAT-tree-based keyword extraction for Chinese information retrieval. ACM SIGIR Forum. Association for Computing Machinery, 1989:221–222.

[6] Jiao, H., Liu, Q., Jia, H.B. Chinese Keyword Extraction Based on N-Gram and Word Co-occurrence. Computational Intelligence and Security Workshops, 2007. CISW 2007. 2007:152-155.

[7] Zhang, K., Xu. H., Tang. J. et al. Keyword Extraction Using Support Vector Machine . Lecture Notes in Computer Science, 2006, 4016:85-96.

[8] Luo, X.F., Fang, N. et al.: Semantic representation of scientific documents for the e-science Knowledge Grid. Concurrency and Computation: Practice and Experience, 20, 839--862 (2008).

.