The Domain Classification Algorithm Based on KNN in Micro-blog

Guofeng Zhu
College of Computer and Information Science & College of Software, Southwest University, Chongqing, China
zgfwm@swu.edu.cn

Zhurong Zhou*, Fengjiao Han and Zhongyun Ying
College of Computer and Information Science & College of Software, Southwest University, Chongqing, China
zhouzr@swu.edu.cn

Abstract—**The Influence of Micro-blog User is essentially the interactions of user to user. With the greater interactions of one user to another user, the influence of the user will be bigger. The existing researches about micro-blog influence are mainly aimed at Twitter and don't consider comment function, the characteristic of interdisciplinary and intersectional domain about the micro-blog. In order to fully consider those factors, we proposed the domain classification algorithm based on KNN in micro-blog. This algorithm classifies all micro-blogs into 15 types depend on the Open Directory Project. It determines the domain of miro-blog according to the degree of similarity between the content of micro-blog and domain ontology of Open Directory Project. Finally we get the influence of user in every domain by quantizing some factors, for example, the degree of similarity between the content of micro-blog and every domain, the number of fans, the number of retweeting, the number of commented, the login and registration time. The experiments show that the proposed method can get more accurate and practicability result than traditional ones, which fully considers the characteristic of interdisciplinary and intersectional domain.**

Keywords-Micro-blog domain classification algorithm; Influence of micro-blog users; KNN algorithm

I. INTRODUCTION

With the development of Web2.0, micro-blog has become the popular media interactive platform as an important Social Network Service. It makes a great difference on society and becomes the important platform for people communicating and sharing information because of the varieties of content and the quickly spreads [3][6]. By the end of 2011, the number of micro-blog users in China has grown up to 250 million[2]. It becomes a great source of promotion [1][9]. With the greater influence of micro-blog user who is as the basic of relation web about micro-blog, he has greater influence on spreading information. Consequently, many researchers have turned to the study of the influence about the micro-blog users [8] [9]. Most of them considering many parameters, for example, the number of fans, retweeting, replying and the quality of fans, those researchers get the influence of micro-blog user by the page rank algorithm and the method based on the value of behavior about user [7] [10] [11] [12]. But they neglected the characteristic of interdisciplinary and intersectional domain about the micro-blog, and can't distinguish the influence on the basis of domain. Hence, we proposed the domain classification algorithm based on KNN in micro-blog. This algorithm classifies all micro-blogs into 15 types depend on the Open Directory Project. It determines the domain of miro-blog according to the degree of similarity between the content of micro-blog and domain ontology of Open Directory Project. Finally we get the influence of user in every domain by quantizing some factors. As compared with traditional method, it fully considers the characteristic of interdisciplinary and intersectional domain, and has good practicability.

Based on the above ideas, the first part of the paper introduces the development and social influence about micro-blog; the second part introduces and analyses the relevant theories about field classification and methods of calculating influence; the third part proposes the Micro-blog Domain Classification Algorithm which includes the problem analysis of users influence, relevant definition, the overall idea of domain classification in micro-blog, and the algorithm description; the forth part shows the experimental results and indicates the method which proposed in this paper is practical and usable; the fifth part gives a summary of this paper.

II. RELATED RESEARCH

At home and abroad, there are many methods about calculating the influence of micro-blog user[16], for example, the calculating method based on page rank algorithm[7], the method based on the value of behavior about user [10], the method based on page rank algorithm and the value of behavior about user [11], the method based on URL tracking[12]. Those four methods consider the parameters about the number of fans, retweeting, replying and the quality of fans. Those can effectively calculate the influence of Twitter users. At present, the influence list of Sina micro-blog also has been applied micro-blog user influence which considers the liveness, transmissibility, coverage of users and decides the domain of users according to the field when users registered and labeled.

At present, there are many algorithms for classifying the domain[14], for example, Naive Bayes Classify Algorithm, Rocchio Algorithm, Decision Tree Algorithm, KNN Algorithm etc. The Naive Bayes Classify Algorithm is easy to realize, and the cost of space-time is small in classification process. But the deficiency is that the eigenvalues based on text is independent of each other. This observation is equivalent to saying that one word can't be influenced by another word. So, it is obviously wrong and isn't applicable in classification of the micro-blog field. Rocchio Algorithm is common used to the cases when

978-1-4673-5000-6/13/$31.00 ©2013 IEEE

the between class distance is larger and in class distance is smaller, and often used to the benchmark system which calculates the performance of classification system. In addition, the algorithm effect is bad and rarely used to solve the specific classification problem. Decision Tree Classification is not suitable for micro-blog field classification because it is easy to cause the excessive adapting problem when a text set is very large, rule base will become very large and the data sensitivity will be enhanced [13]. However, KNN Algorithm is an unsupervised learning method, and do not need the training material. Especially, its implementation is simple and has high classification accuracy. So it is widely used in Chinese text automatic classification [4][5].Therefore, this paper using KNN Algorithm to classify the micro-blog field.

III. MICRO-BLOG DOMAIN CLASSIFICATION ALGORITHM

A. Problem Analysis

The existing researches about micro-blog influence are mainly aimed at Twitter. Analytically, the existing researches have the following questions specific to applications of the micro-blog.

Question 1: The Interdisciplinarity of Micro-blog.

The problem of interdisciplinarity can be formalized as follows. Given a set Ω_P, there are fifteen domains, P_1, P_2, …, P_i, …P_{15} that corresponding to the fifteen domain in Open Directory Project*[in definition 1]*. (*The symbol of* $\rightarrow$ *expresses the domain of former is subject to the latter and the symbol of* $\nrightarrow$ *expresses the domain of former isn't subject to the latter.*) Supposing $User_i \rightarrow P_j$, that is to say the domain of $User_i$ engaged in is subject the domain P_j (*j is the integer from 1 to 15*). Given a set Ω_w, there are m micro-blogs, Wb_1, Wb_2, …, Wb_i, …Wb_m that are published by $User_i$, if $\exists Wb_i \nrightarrow P_j$, that is to say the domain of the Wb_i about $User_i$ isn't subject to the domain P_j, there is the characteristic of interdisciplinary domain about the micro-blog.

Question 2: The intersectional domain about the micro-blog.

The problem of intersectional domain can be formalized as follows. Given the domain of $User_i$ registered is Pj which is record as $User_i \rightarrow P_j$ (*j is the integer from 1 to 15*). If $\exists\ W_i \rightarrow P_k \cap W_i \rightarrow P_m (m \neq k)$ in the collection of Ω_w and Ω_P, there is intersectional domain about the micro-blog.

To solve the problem of interdisciplinary and intersectional domain about the micro-blog, this paper proposed the Field Classification Algorithm based on KNN. Firstly, this algorithm analyses the Content of micro-blog, and calculates the similarity of the content and every domain. Secondly, it distributes the micro-blog to the largest similarity domain and divides all the micro-blog Ω_w into several subsets $\Omega_{w'}$ according to Ω_P. Therefore, this paper has solved the problem of interdisciplinary and intersectional domain about the micro-blog.

B. Formalization Definitions

Definition 1: the domain of micro-blog. In this paper, the micro-blog has been divided into 15 categories according to the ODP system which is the world's largest open classification catalogue [15]. Equivalently, $\Omega_P=\{P_1, P_2, \cdots, P_j, \cdots P_{15}\}$.

Definition 2: the influence of micro-blog user. From literature [7],it is known that the influence of micro-blog user is essentially the interaction between users, the greater effect of a user to other users, the influence of the user is greater.

Definition 3: Micro-blog. Micro-blog is the text of user published. Given a micro-blog Wb_i, it is consist of *W _ID, U_ID ,Content, Zf, IWc, P, Tp.*

W _ID is the identification number of micro-blog. *U_ID* is the identification number of user. *Content* is the content of micro-blog text. *Zf* is the number of retwitting. *IWc* is the number of comments. *P* is the domain of micro-blog belonged. *Tp* is the degree of correlation between the content of micro-blog with the domain.

Definition 4: Distinguishing the domain of micro-blog. All the micro-blog Ω_w has been distinguished into several domains $\Omega_{P'}$ according to the content of micro-blog. It can be formalized as follows.

$\Omega_w=\{ Wb_1, Wb_2, \cdots, Wb_i, \cdots Wb_m \}, \Omega_{P'}=\{P_i, P_j, \cdots P_k\}, \Omega_{P'} \in \Omega_P, \forall P_j, \exists \Omega_{w'} \in \Omega_w, \forall Wb_i \in \Omega_{w'}, Wb_i \in P_j$.

C. Overall Idea

As shown the dashed part in figure 1, the existing research didn't consider the character of interdisciplinarity about micro-blog and classified all micro-blogs(Ω_w) into the domain of P_j which the user engaged in. It interferes with calculating the value of influence in domain of P_j. Therefore, we divide all micro-blogs that $User_i$ publishes into 15 domains(P_1, P_2, P_3, P_4, …, P_{15}) according to the Open Diretory Project as shown in figure 1. This method classifies the domain of miro-blog using the Domain Classification Algorithm according to the content of micro-blog and gets the influence of user in every domain by quantification parameters in paper.

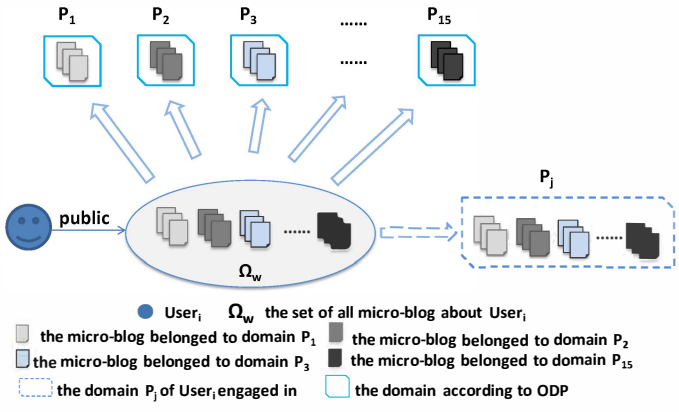

Figure 1. the character of interdisciplinarity about micro-blog

In this paper, all information about micro-blog user has been divided into the two parts---- the set of micro-blog user and the set of micro-blog.

The two parts are respectively formalized to *UserSet* and *WbSets*. *UserSet* is the collection of micro-blog users' essential information which contains the number of fans, the number of retweeting, the login and registration time. *WbSets* is the collection of all micro-blogs and the composition is as follows.

$WbSets=\{WbSet_1, WbSet_2, WbSet_3, \cdots WbSet_i \cdots WbSet_n\}$,

$WbSet_i$ is the collection of micro-blog about $User_i$,

n is the total of micro-blog users,

$WbSet_i=\{Wb_1, Wb_2, Wb_3, \cdots Wb_j \cdots Wb_m\}$,

Wb_j is the first j micro-blog of $User_i$, the definition of Wb_j shown in definition 3

m is the total of micro-blogs that $User_i$ published.

The domain classification model has been shown in figure 2 which contains three basic steps.

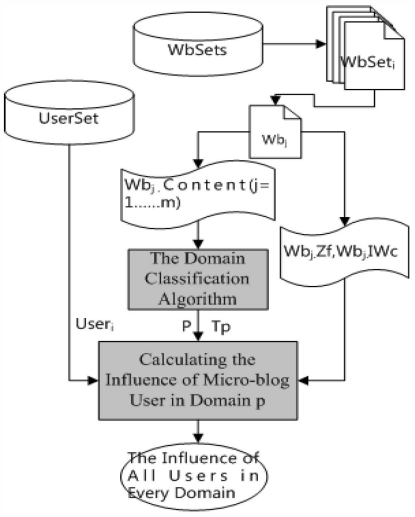

Figure 2. system flowchart

- Step 1, obtaining the information about micro-blog user.Get the set of UserSet and WbSet through the API technology[6].
- Step 2, classifying every Web_i of *WbSet* into their respective fields. Assign the micro-blog Web_i to the most similar field p and get the value of Tp -- degree of correlation between the content of micro-blog with the domain by domain classification algorithm according the *content* of Wb_i
- Step 3, calculating the influence of micro-blog user in domain *P*. Quantification parameters about the total number of micro-blogs in field *P*, the number of retweeting, the login and registration time, the number of fans about user.

This paper mainly studies the field classification algorithm (Step 2), and shows in Section 3.4.

D. Algorithm Description

To get the influence of users in every field, we use the domain classification algorithm based on KNN to distinguish the domain of micro-blog. KNN algorithm is a learning method without supervised. It gets k texts which are the most similar feature space text by comparing the target text with the text in feature space set using the similar function, and judges the category of new text according to the k texts using the class attribute function as shown in formula 2. The similar function is as shown in formula1.

$$Sim(d_i, d_j) = \frac{\sum_{k-1}^{m} W_{ik} * W_{jk}}{\sqrt{(\sum_{k-1}^{M} W_{ik}^2)(\sum_{k-1}^{M} W_{jk}^2)}} \quad (1)$$

In this fomula, *Sim* is the similarity of d_i and d_j , and d_i is the target text, and d_j is the feature space text.

$$T(Content, p_j) = \sum_{\bar{d}_i \in KNN} Sim(Content, \bar{d}_i) y(\bar{d}_i, p_j) \quad (2)$$

T is the weight of *Content* belonged to P_j. *Content* is the content of new micro-blog, $y(\bar{d}_i, p_j)$ is the field categorical attributes function, if $\bar{d}_i$ belong to *Pj*, the function value is 1, else the function value is 0.

From the above analysis, *the domain classification algorithm* such as algorithm 1 shows.

Algorithm 1: The Domain Classification Algorithm Based on KNN

Input: WbSets; the set of field text feature space; Mp ---the number of micro-blog in domain p (Initial value is 0).

1) For i=1 to n /n is the total number of users
2) For j=1 to m //m is the total number of micro-blog about user
3) Extracting the feature vector about micro-blog;
4) Selecting the k most similar text with the new content about micro-blog using function Sim(di,dj) in the set of field text feature space;
5) For l=1 to 15
6) Calculating the degree of correlation between k neighbors and every field by the function of T(Content,Pl);
7) End for
8) Comparing the weight of domain and assigning the content of micro-blog to the domain p which the degree is the biggest;
9) Tp = Max（T（Content，Pl））;
10) Mp++; // AddSelf the number of micro-blog in domain p
11) End for
12) End for

Output: p—the domain of micro-blog belonging to; Tp----the degree of correlation between Wbi and domain p.

IV. EXPERIMENTS

This paper uses the application program interface (API) of Sina Micro-blog to excavate 680000 users that 0.9% of them have no fans and followers,6.8% of them whose micro-blog number is less than ten, 92.3% of them are with high degree of active, the number of business users through the Sina certification is 1023.This paper does research into 10 users who have registered as business fields in Sina micro-blog, calculates and compares the influence under the situation of considering the characteristic of intersectional field and doing not.

When use the traditional method, we don't consider the characteristic of intersectional field about users, the user's all micro-blogs are regard as belonging to the business field which the user registered as. The influence diagram is showed in figure 3.

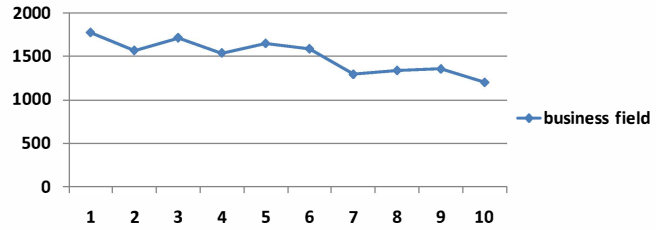

Figure 3. the influence of 10 users in business field which the user registered as

Through analysing the experiment results, the traditional methods obviously is wrong which regardless whether if the micro-blog is related to the business field. It leads to the value of user influence are only in the commercial field.

When consider the characteristic of intersectional field about users and divide micro-blog fields into 15 categories according to the ODP system, we calculate the influence of 10 users in every field, and list the influence of them in business, IT, sports, and social culture field. The concrete data diagram is as shown in figure 4.

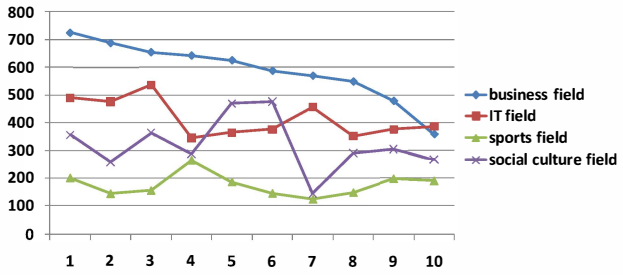

Figure 4. the influence of 10 users in every field

Through analyzing the experiment results, when consider the characteristic of intersectional field about users, the users have influence on every fields. As for the user 3, he has influence on IT, sports, and social culture field in addition to the business field when considering the characteristic of intersectional field. However, when using the traditional methods, the user 3 just has influence on the business field. Hence, the traditional methods are not in conformity with the common sense.

And as in the time of 2012 European Cup, the users are engaged in all walks of life and love football who share the micro-blog related to football. Figure 5 shows the 10 users' influence in every field in the time of 2012 European Cup. It indicates that the users published quite many micro-blogs related to the sport even more than in the business field during the European Cup.

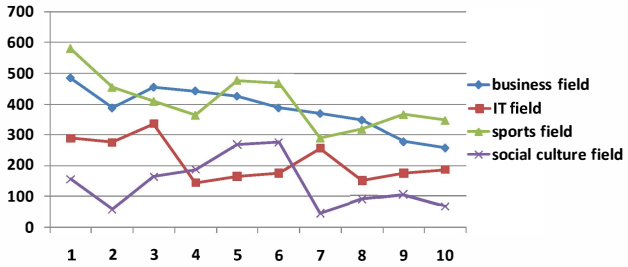

Figure 5. the influence of 10 users in every field during the European Cup

To sum up, micro-blog users have influence on all walks of life and not only limit to the field engaged in. Even in some period of time, the users have higher influence on the other areas than the field engaged in.

V. CONCLUSION AND PERSPECTIVES

In this paper, we propose the Domain Classification Algorithm Based on KNN in Micro-blog. This algorithm full considers the characteristic of intersectional and interdisciplinarity about field in this paper. In conclusion this method can clearly estimate the influence about users in all walks of life. It is easy to get the latest dynamic about all fields and realize the intelligent recommend function about every field effectively too. In addition it is valuable to cognize user's peciality and determine the key network marketing. Certainly this paper also has the following limitations: the efficiency of classifying isn't enough high and don't give enough weight to the field that users engaged in. In the future work, we will focus on improving the efficiency of the field classification and full considering the weight of field that the user engaged in.

VI. ACKNOWLEDGMENT

The authors would like to thank the colleagues and schoolmates for supplying great support for this paper. And we all wish to acknowledge Institute of Computer and Information Science, Southwest University, Chongqing, China. This work was supported in part by a grant from them.

REFERENCES

[1] SUN Sheng-Ping. Research on Chinese Micro-Blog Hot Topic Detection and Tracking. Beijing Jiaotong University.2011

[2] XU Zhi-Kai.Research and Implementation on Key Techniques of Online Public Opinion Analysis. Harbin Institute of Technology.2011

[3] XU Dong. Worries and Thinking of the Right of Discourse in Network. Sichuan University.2007

[4] ZHANG Ning, JIA Ziyan, SHI Zhongzhi. Text Categorization with KNN Algorithm. Computer Engineering. 2005 ,31(8).171-172

[5] ZHANG Zhu-Ying, HUANG Yu-Long ,WANG Han-Hu. A New KNN Classification Approach. Computer Science.2008,35(3).170-172

[6] LI Jun, CHEN Zhen, HUANG Ji-Wei. Micro-blog Impact Evaluation Study. Netinfo Security.2012(3).10-13

[7] LIU Yao-Ting. Research on Social Network Structure [D].Zhejiang University,2008. 6 .

[8] Shaozhi Ye, S. Felix Wu. Measuring Message Propagation and Social Influence on Twitter.com[C]. SocInfo 2010. 223-228.

[9] Louis Yu et al., What Trends in Chinese Social Media[C]. SNAKDD' 2011. 2-4.

[10] Meeyoung Cha et al., Measuring User Influence in Twitter: The Million Follower Fallacy[C]. AAAI 2010. 11-13.

[11] Yuto Yamaguchi et al., TURank: Twitter User Ranking Based on User-Tweet Graph Analysis[C]. WISE 2010. 243-246.

[12] Eytan Bakshy et al., Everyone's an Influencer: Quantifying Influence on Twitter[C]. WSDM 2011. 67-69.

[13] ZHANG Zheng-Jie, WANG Zi-Qiang. The Summarize about Text Classification Algorithm. Computer Knowledge and Technology.2012,8(4).825-828

[14] WU Chun-ming, XIE De-ti . Research on Deep Web Classification Based on Domain Feature Text. Computer Science.2012, 39(4).177-180

[15] Lu Xiaoxi. Analysis on Classification System of ODP. The Library Journal of Shandong.2009(1).62-65

[16] XIAO Yu, XU Wei, SHANG Zhao-xi. Analysis on Algorithms of Identifying Regional Influential Users in Micro-blogging. Computer Science. 2012, 39(9).38-42