Statistical Methods based on Semantic Similarity of Topics Related to Microblogging

Dongqing Wu
Computation Science Department, Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong 510225,wwddqq@126.com

Fengjian Yang* and Chaolong Zhang
Guangzhou Guangdong 510225,wwddqq@126.com
2, Computation Science Department , Zhongkai University of Agriculture and Engineering, Guangzhou Guangdong 510225,yangfj88@sina.com,zhcl88@126.com

***Abstract*—Existing topic tracking methods are mostly for news and forum data, which lack of statistical methods for microblogging on relevant topics. Combined with characteristics of micro-blog information, the paper proposes a microblogging statistical methods based on semantic similarity. Firstly by building topic semantic model and then use the HowNet semantic similarity calculation of two terms, and measures the relevance of the topic and microblogging. Finally statistics method is provided on the degree of correlation. Experiments show that novel method works on the problem soundly.**

***Index Terms*—microblogging topic, semantic similarity, HowNet, LSI, statistical methods**

I. Introduction

A. Backgound

Microblogging is a broadcast medium in the form of blogging. Microblogging differs from a traditional blog in that its content is typically smaller in both actual and aggregate file size (normally 140 words limited). It allows users to exchange small elements of content such as short sentences, individual images, or video links. According to report released by CNNIC in Jan 2012, numbers of microblogging users soar to 250 million which increase almost 296% compared with one year ago in China. Among all Internet users, microblogging use rate is 48.7%. Just over a year, microblogging develop into a significant Internet application used by nearly half of Internet users in China.

Through "concern", "concerned", network is formed by microbloggers. They spread messages though commenting, concerning and forwarding. The spread of this geometric progression can achieved within a very short time and get great dissemination of results. It has beyond its technical sense as a communication tool. Furthermore it shapes and remarks ecology of the spread of public opinion. Nowadays researching around microblogging, sorting, extracting useful information, and removing false information, has become an urgent task.

B. Related Works

In the aspects of data mining of microblogging, Z. Liu, W. Yu [1] extend semantic features of words based on part-of-speech theory and Hownet. They propose feature selection method for short text. Aim at the short characteristics of microblogging text, F.Zheng[2] constructs adjacency matrix to measure the semantic similarity among words, and extracts keywords of microblogging automatically in the light of PageRank algorithm from Google. In addition, microblogging has been studied for monitoring of accidents associated with geographic information such as fire, traffic jam, and natural disaster. Some progress has been made [3-8]. As far as forum topic statistics, Y. Xi, and S. Lin propose a tracking method based on semantic similarity forum topic [9-13]. Text representation model were created through building subject word table. Then calculate the semantic similarity of the words tables of posts and topics with Hownet. This method is effective to solve the tracking problem of forum topic. Until publishing of this writing, research of microblogging statistics for specific topics have not yet reported.

In summary, the research of microblogging topic tracking and statistics has great research value and space. On one hand, it could not only help to filter unwanted information to improve the quality of content and user experience, but also play a important role in event monitoring, mining and directing public view. On the other hand, since microblogging information characteristics includes mass, fragmented and special structured as well as a lot of noise data, existing methods is difficult to extract accurate and useful information from it.

Manuscript received May 25, 2012; revised May 30, 2012; accepted June 3, 2012.
* Corresponding author, Natural Science Foundation of ZHKU.(No.G3100037)

doi:10.4304/jsw.8.1.192-199

C. Identifying Problems

When an event has occurred, institutional stakeholders need to analysis related microblogging to assess the degree of impact on the community. Currently, microblogging operators have provided some statistical services on its data. However, the whole social dimension of automatic statistics on the same topic has not been reported. How to combine the characteristics of microblogging and provide an effective statistical method for the same event from different sources microblogging remains unsettled.

II. ALGORITHM DESIGN

A. Analysis and Solutions

Chinese microblogging data, in essence, is a series of short text. Not more than 140 words of each text and may contain some special format to indicate the interaction between themes and users.

Sina microblogging for example, "@abc" represents the text mentioned user named "abc", and user "abc" will receive a reminder "who referred to me". "//XX" means that the content forward since user "XX". "#Subject#" indicates that participation in a particular topic discussion. In addition, microblogging has some additional properties, such as delivery time, the source of geographic information, the sender's information etc. Microbloggings grow very fast, and every hour up to as much as tens of thousands. Topics covered are also much dispersed. Synthesis problem itself particularity, proposed solutions is shown in Figure 1.

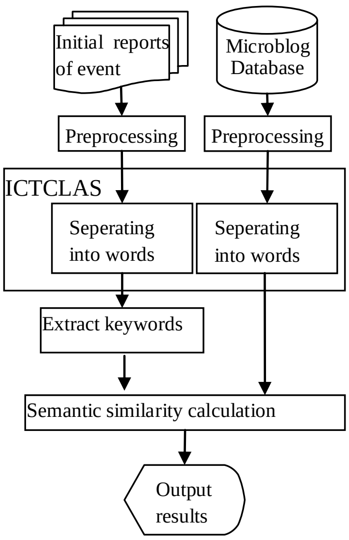

Figure1 Problem-solving roadmap

On the first branch, we obtain data from microblogging library and then preprocess it. After that, use ICTCLAS to separate words from text, then get keywords vector at last. On the other branch, the initial reports of the event are separated into words by means of ICTCLAS. Then extract keywords and form topic keywords vector. Finally, based on HowNet semantic similarity calculation, determine degree of relevance. Concrete steps are indicated below.

B. Data Preprocessing

The data pre-processing involves two steps: filtering noise and segmentation.

The filter procession is mainly for microblogging. Purpose is to eliminate the noise data as far as possible and shielding irrelevant data, filter out the useless information in the text. Microblogging text is not long, and contains a large number of specially formatted texts which would affect the following aspects of the calculation speed and accuracy without filtering. Specifically, there are three levels of filtering which should be considered.

1) Account level

For some commercial purposes, there are a large number of "zombie accounts" in microblogging. These accounts are manipulated by software and they manufacture large amount of redundancy information which must be filtered. There are characteristics provided by papers [14-19]. Firstly its attention is zero. Secondly, they repeat sending some specious, irrelevant comments (example: "true", "Do not worry", "up"), and often contain large number of advertising terms. We filtered these account through the establishment of "zombies terms" dictionary, and add a new filter words in it dynamically. Therefore establish the characteristics of library for "zombie account". When the number of behavioral characteristics of an account exceeds the threshold, microblogging from that account is completely shielded.

2) Message level

Ignore those microblogging only contain "//user". These microbloggings show user’s viewpoint of the topic, simply for forwarding (onlooking) purpose. These messages contributed to attention in some extent. But have no effect on the microblogging semantic detection. To filter such messages, we can increase the detection accuracy of keywords. For those microbloggings which include own words should be retained.

3) Content level

First consider stop words. Stop words usually have no clear meaning. The purpose of putting it in text is merely to complete a sentence. Stop words includes "you", "I", "then" high frequency words and modal particles, adverbs, prepositions, conjunctions and so on. Such as the common "," "," and "Then "etc.

After filtering, text needs to be separate into words for subsequent processing. In this paper, we use analysis system (ICTCLAS) from the Institute of Computing, Chinese Academy of Sciences. Its features include: Chinese word segmentation, POS tagging, named entity recognition, word recognition function; support user-defined dictionary. In addition, "# XXX #" format required special treatment. The format is a topic tag, which is often topic keywords, should be singled out and given higher weights.

C. Extract the Keywords of the Topic

Due to limitation of length, microblogging is difficult to describe a complete event. Therefore, the initial topic keyword extraction has to follow the traditional topic keywords extraction rule. That is, the initial full coverage of the event served as corpus. This paper extracts keywords by LSI (Latent Semantic Indexing). LSI's core operation is truncating singular value decomposition on the term-document matrix (TD). So we get the best approximation of the original term document matrix on sense of least squares. LSI could extract the implicit semantics of the document and generate vectors which express the semantics of the whole document.

If all entries from document were constituted to it, the TD matrix would be very large and sparse. Since lots of irrelevant or minimal entries have impact on the description of the document, they will cause the description of document semantic confusion and fuzzy. Therefore, in order to improve the efficiency and accuracy of the classification algorithm, screening is needed to reduce the dimension of the entry vector.

The specific steps are as follows.

1) Recommended by papers [20-25], we use N ($N = 2^k$ $(k = 0,1,2,\ldots)$) initial reports as the corpus of documents $d_j (j = 1,2,\ldots N)$. And the corpus also contains other areas of reports with ratio of 1:4 to ensure the discrimination of keywords.

2) Construction of TD matrix

Scan d_j , then separate words with ICTCLAS, and generate entry library. $\{t_i\}(i = 1,2,\ldots M)$. Statistics frequency of t_i appearing in d_j then get X' (term frequency matrix). It is generally believed that weighting function is critical to optimize the LSI. TF-IDF is a one of the widely used method. Its weight is calculated as follows.

$$x_{ij} = tf_{ij} \quad idf_i \tag{1}$$

tf_{ij} as the frequency in d_j , The formula is

$$tf_{ij} = \frac{\text{The number of occurrences } t_i \text{ in } d_j}{\text{The total number of entry in } d_j} \tag{2}$$

Formula for idf_i

$$idf_i = \log \frac{|D|}{|\{d \in D : t_i \in d\}|} \tag{3}$$

$|D|$ represents the total number of documents in corpus. $|\{d \in D : t_i \in d\}|$ represents number of documents contains t_i . From equation (2) see that t_i weight is proportional to the number of times it appears in d_j . From equation (3) we can see: low-frequency words are more important than high frequency words on the term of idf_i . Obviously it would cause some deviation. Thus equation (1) can not effectively reflect distribution, extent and characteristics of the word. Furthermore, the TF-IDF method and does not reflect the position information of the word. From different locations on the characteristics of words in the probabilistic sense to reflect on the content of the article, the weight calculation method should also be different. Feature words in a different location should be given different weights, and then multiplied by the features of word frequencies. References [26-28] propose method using the following formula:

$$tf_{ij} = \frac{\sum_{k=1}^{3} 2^{3-k} tf_{ij}'(k)}{7} \tag{4}$$

Each document is divided into three parts: head, middle and rear. Each part covers 1/3 of document. Depending on its occurrence position, weight of entry is respectively $\frac{4}{7}, \frac{2}{7}, \frac{1}{7}$. In equation (4), $tf_{ij}'(k)$ is the word t_i frequency of the Part k in d_j .

3) SVD decomposition of TD matrix

By singular value decomposition, we get the TD matrix which reflects the keywords of the text on latent semantic vector $\mathbf{Z}^{Topic} = \{Z_i^{Topic}\}$. At the same time, the weight vector of the subject headings ($\mathbf{W}^{Topic} = \{weight_i\}$) is acquired

D. Similarity Calculation Model of Topic and Microblogging

Task of this step is calculating the similarity based on $\boldsymbol{Z}^{Topic}$ getting in previous step while traversing the microblogging set. Accordingly determine whether the microblogging related to the topic.

Firstly, each microblogging is filtered and preprocessed. Secondly, all remaining entries are reserved as microblogging entry vector $\boldsymbol{Z}^{MB}$. And then calculate the similarity $Sim(\boldsymbol{Z}^{Topic}, \boldsymbol{Z}^{MB})$ between the two vectors. Take those similarities greater than the threshold l as the theme topic related microblogging.

Words structure of HowNet can be used to calculate the similarity between the two vectors. In HowNet, a word W may include a number of concept C . While C is described by sememe (the original description) [29]. Suppose the word W_1 , including the concept of group $C_1 = \{C_{11}, C_{12}, \ldots, C_{1m}\}$ word W_2 including group $C_2 = \{C_{21}, C_{22}, \ldots, C_{2n}\}$. Then the similarity between W_1 and W_2 can be measured with maximum of similarity between C_1 and C_2 .The formula as following:

$$Sim(W_1, W_2) = \max\{Sim(C_{1i}, C_{2j})\} \tag{5}$$

Inspired by reference [30], we use HowNet concept of structure and semantics of the multi-dimensional forms of expression. We calculate word similarity from three aspects. They are main sememe of concept, framework of main sememe and conceptual characterization. In this process, we distinguish between the semantic features of similarity and the similarity of syntactic features. Two conceptual similarities is calculated as follows

$$Sim(C_1,C_2) = q[b_1 ?Sim_1(C_1,C_2) \quad b_2 \text{鬚} g \ Sim_2(C_1,C_2) \quad b_3 \ Sim_3(C_1,C_2)] \qquad (6)$$

In formula (6), C_1,C_2 are the two concepts involved in similarity calculation. q is the sign of coefficient. If C_1,C_2 were antonymy, q sets to -1, otherwise 1. $Sim_1(C_1,C_2)$ calculates similarity of sememes. $Sim_2(C_1,C_2)$ is calculation of the semantic tree similarity. $Sim_3(C_1,C_2)$ is righteousness of the similarity of sememes framework. $b_1 \sim b_3$ are weights of corresponding similarity, which satisfied with $\overset{3}{\underset{i=1}{\text{å}}} b_i = 1$ and $b_3 \# b_1 \quad b_2$. Their formula is provided in paper [31-34]. g is regulatory factor of roles and characteristics. If Characterization of C_1,C_2 existed common event E , and these two concepts belong to different roles dependent on E , then $Sim_2(C_1,C_2)$ should be multiplied by 0.5 or 1 otherwise.

Assume there are M entries in $\boldsymbol{Z}^{Topic}$, and N entries in Z^{MB} . After calculate the similarity of the two entry, similarity matrix $\{s_{ij}\}_{M´N}$ are created in which $s_{ij} = sim(Z_i^{Topic}, Z_j^{MB})$. It is topic keywords and mircoblogging keywords similarity matrix. Accordingly, we can calculate the overall similarity of the microblogging and topic by the following formula.

$$Sim(\boldsymbol{Z}^{Topic}, \boldsymbol{Z}^{MB}) = \sum_{i=1}^{M} weight_i \cdot \max\{Sim(Z_i^{Topic}, Z^{MB})\} \qquad (7)$$

Where $weight_i$ is the weight of $\boldsymbol{Z}_i^{Topic}$, $\max\{Sim(Z_i^{Topic}, \boldsymbol{Z}^{MB})\}$ as the maximum similarity of all term in Z_i^{Topic} and $\boldsymbol{Z}^{MB}$. Formula (7) shows that the more similar between mircoblogging term and topic term of great weight, the greater contribution it will be for overall.

III. Experiment

A. Data Sources Design

We selected topic which has great impact on society recently. Experimental data are as follows:

1) Initial news reports were manually selected from the mainstream media as analysis corpus for topic.

2) From open platforms of Sina, Netease and Tencent microblogging, collect 100,000 , a total of 300,000 of the original microblogging data generated within 24 hours after the event happened.

B. Methods of Evaluation

We use the traditional metrics tracking system evaluation. They are Recall rate, Precision rate, and harmonic mean F. These formulas are as follows

Recall rate

$$r = \frac{|R(Q) \text{ Ç } R^*(Q)|}{|R^*(Q)|} \qquad (8)$$

Precision rate

$$p = \frac{|R(Q) \text{ Ç } R^*(Q)|}{|R(Q)|} \qquad (9)$$

harmonic mean F

$$F = \frac{2rp}{r + p} \qquad (10)$$

Formula (8), (9) Q represents a query. $R(Q)$ is the number of microblogging. $R^*(Q)$ represents all the microblogging related with topic. Define symbol$|\cdot|$ for the number of elements in the collection.

C. Experimental Results

The proposed method compared with two other traditional topics of statistical methods, the results shown in Figure 2-4.

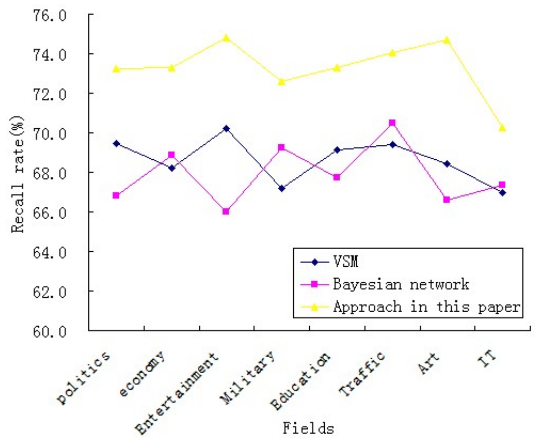

Figure 2 Comparison of recall rates

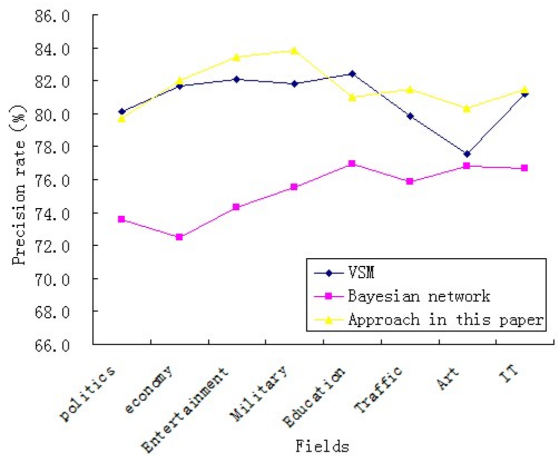

Figure 3 Comparison of precision rates

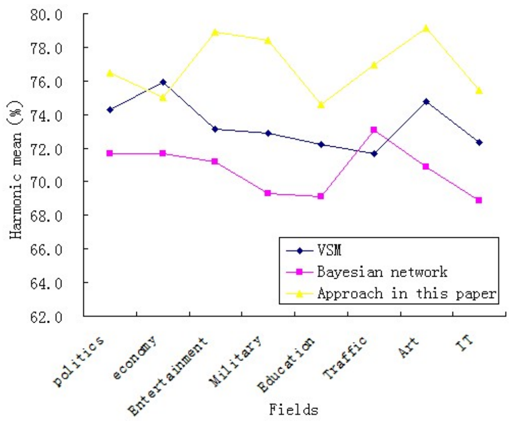

Figure4 Comparison of harmonic mean

TABLE I.
SUMMARY OF EXPERIMENTAL RESULTS

No.	Methods	Avg. Recall rate （%）	Precision rate （%）	Harmonic mean （%）
Ⅰ	VSM-based approach	69	80	74
Ⅱ	Bayesian network-based approach	67	75	71
Ⅲ	Approach in this paper	73	82	77

Though these data, we could summary it in Table Ⅰ.

Recall rate of approach Ⅰ is relatively high, but compared with the approach of this paper is still a bit weak. In the aspect of accuracy, the Bayesian network approach is not very stable on the performance and relatively dependence on the data. After the analysis of experimental results, we found that the reason for low precision rate using Bayesian networks-based approach lied in a large number of irrelevant features in the microblogging data which supposed to be screened. In addition, the contrast among the harmonic mean, this method is superior to the other two traditional methods.

D. Influence of Time Window Length

Different length of time window was trying to compare test results which were shown in Figure 5.

When time window is small, it showed that result were vulnerable to the interference of the noise data. At the same time, precision and recall rates are low. When time window is large, the selected keywords are more accurate, but due to the larger particle size and some topics had been omitted. For example, in one hour window of time, words like "Good Night" and "lunch" in a specific time of day appear frequently, might be mistaken as topics. If larger time window was selected and the topics were gotten shorted attention, then the number of keywords and other words would trend to be in line, which means that the length of window is not big enough.

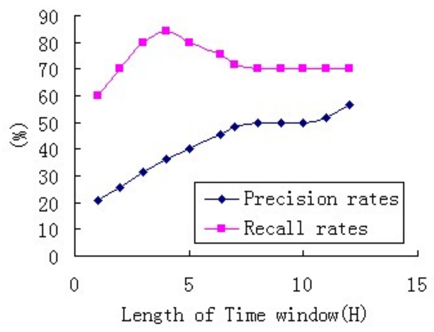

Figure5 Influence of time window length

E. Influence of Time Window Length

Using three hours time window for detection with different thresholds l , we get the results shown in Figure 6.

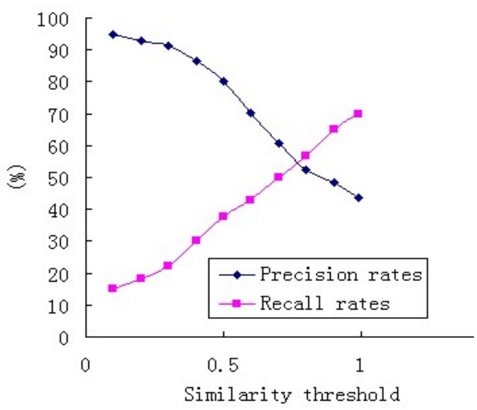

Figure 6 Influence of similarity threshold l

It shows that microblogging and topic similarity threshold which affect query results. When T value near 1, the recall rate is high, but it also get a lot of unrelated results. Obviously, the more cautious judgment of the keyword in the larger threshold, while reducing the noise, but increases the risk of missing news topics.

The greater l the higher the recall rate will be. At the same time the precision rate will rise. The smaller l the higher the recall rate will be. At the same time the precision rate will reduce. Author has tried different whose interval is 0.05 in the range of 0.1 to 1.0. In this experiment, set to 0.765 to get the best harmonic mean and the lowest sampling error rates. Author believes that the better threshold value selection method should be adaptive. It is the direction of future efforts.

Ⅳ. CONCLUSION

Combined with the characteristics of microblogging text, this paper proposed a statistical tracking method for a topic related to microblogging based on HowNet.

Microblogging data is preprocessed on the basis of the text. Then use HowNet concept of structure and semantics of the multi-dimensional forms of expression to calculate the similarity of keywords vector. Thus judge whether the microblogging and specified topic are related. The experimental results show that the proposed method is better than traditional statistical methods in precision and recall rate.

Due to the diversity of microblogging text expression, a lot of Internet language is not in line with the Chinese language specification. Informal language is popular in microblogging. Existence of non-standard phrases brings many difficulties. At the same time, the experimental results reveal that there is still considerable room for improvement. Further improvement of the algorithm speed and accuracy, and enhancement of stability, is the focus of future research work.

ACKNOWLEDGMENT

The authors wish to thank Computational Science Lab of ZHKU.

This work was supported in part by a grant from Natural Science Foundation of ZHKU.(No.G3100037) .

REFERENCES

[1] Z. Liu, W. Yu, W. Chen, et al. "Short Text Feature Selection for Micro-Blog Mining", Proceedings of CiSE2010.CiSE 2010, Wuhan,China.Dec.2010.Wuhan,China:IEEE,pp.201-204, 2010.

[2] A. Passant, J. Breslin, and S. Decker, "Rethinking microblogging: open, distributed, semantic," Web Engineering, pp. 263-277, 2010.

[3] F. Abel, Q. Gao, G. J. Houben, and K. Tao, "Semantic enrichment of twitter posts for user profile construction on the social web", The Semanic Web: Research and Applications, pp. 375-389, 2011.

[4] S. J. Barnes and M. BöHRINGER, "MODELING USE CONTINUANCE BEHAVIOR IN MICROBLOGGING SERVICES: THE CASE OF TWITTER," Journal of Computer Information Systems, vol. 51, p. 1, 2011.

[5] Zheng, D.Miao, Z.Zhang,et al. "News Topic Detection Approach on Chinese Microblog". Computer Science,2012,vol.39,no.1,pp.138-141

[6] M. Cutler, Y. Shih, and W. Meng, "Using the structure of HTML documents to improve retrieval", Proceedings of USENIX Symposium on Internet Technologies and Systems 1997, Monterey, California: USENIX Association.pp. 241-251,1997.

[7] J.Lan,H.Shi,X.Li. "Associative Web Document Classification Based on Word Mixed Weight", Computer Science, 2011,vol.38,no.3,pp. 187-190.

[8] B.Ge,F.Li,S.Guo,et al. "Word's semantic similarity computation method based on Hownet". Application Research of Computers, 2010,vol.27,no.9,pp. 3329-3333.

[9] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", Proceedings of the 19th International Conference on World Wide Web,2010.Raleigh,North Carolina: ACM Press,pp.851-861,2010.

[10] L.Zhang,C.Yin,J.Chen. "Chinese Word Similarity Computing Based on Semantic Tree", Journal of Chinese Information Processing, 2010,vol.24,no.6,pp. 23-30.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.

[11] M. Böhringer and A. Richter, "Adopting Enterprise 2.0: A Case Study on Microblogging," Mensch und Computer 2009: Grenzenlos frei, pp. 293-302, 2009.

[12] J. G. Breslin, D. O'Sullivan, A. Passant, and L. Vasiliu, "Semantic Web computing in industry," Computers in Industry, vol. 61, pp. 729-741, 2010.

[13] Y.Xi,S.Lin,B.Li,et al. "Method for BBS topic tracking based on semantic similarity", Journal of Computer Applications, 2011,vol.31,no.1,pp.93-96.

[14] S. Castano, A. Ferrara, and S. Montanelli, "P2P Semantic Coordination for Collective Knowledge Organization," Collaborative search and communities of interest: trends in knowledge sharing and assessment, 2010.

[15] F. Cena, A. Dattolo, E. De Luca, P. Lops, T. Plumbaum, and J. Vassileva, "Semantic Adaptive Social Web," Advances in User Modeling, pp. 176-180, 2012.

[16] D. N. Crowley, A. Passant, and J. G. Breslin, "Short Paper: Annotating Microblog Posts with Sensor Data for Emergency Reporting Applications Semantic Sensor Networks 2011 Semantic Sensor Networks 2011," Semantic Sensor Networks 2011, pp. 84-89, 2011.

[17] W.Qin. "Microblogging vs. Zombies", Computer Fan, 2011,no.22,pp.10,.

[18] M. Hepp, "HyperTwitter: collaborative knowledge engineering via twitter messages," Knowledge Engineering and Management by the Masses, pp. 451-461, 2010.

[19] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: An analysis of a microblogging community," Advances in Web Mining and Web Usage Analysis, pp. 118-138, 2009.

[20] A. Joly, P. Maret, and J. Daigremont, "Between social awareness and productivity: Results of a survey about real-time microblogging," First Monday, vol. 15, 2010.

[21] A. N. Langville and C. D. Meyer, Google page rank and beyond: Princeton Univ Pr, 2006.

[22] Zelikovitz S,Transductive M F. "Learning for Short-Text Classification Problem using Latent Semantic Intexing International".Journal of Pattern Recognition and Artificial Intelligence, vol.19,no.2,pp.143-163,2005

[23] C. H. Lee and C. H. Wu, "A Self-adaptive Clustering Scheme with a Time-Decay Function for Microblogging Text Mining," Future Information Technology, pp. 62-71, 2011.

[24] T. Li, "Using Microblogging for the PKM in the Web2. 0 Environment," Emerging Computation and Information teChnologies for Education, pp. 519-526, 2012.

[25] A. Mazarakis, S. Braun, and V. Zacharias, "Feedback in social semantic applications," International Journal of Knowledge Engineering and Data Mining, vol. 1, pp. 291-302, 2011.

[26] M. Okazaki and Y. Matsuo, "Semantic twitter: analyzing tweets for real-time event notification," Recent Trends and Developments in Social Software, pp. 63-74, 2011.

[27] S. Panigrahi and S. Biswas, "Next Generation Semantic Web and Its Application," IJCSI International Journal of Computer Science Issues, vol. 8, 2011.

[28] A. Passant, J. Breslin, and S. Decker, "Open, distributed and semantic microblogging with SMOB," Web Engineering, pp. 494-497, 2010.
[29] A. Passant, S. Kinsella, U. Bojars, J. G. Breslin, and S. Decker, "Understanding Online Communities by Using Semantic Web Technologies," Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, vol. 1, p. 429, 2010.
[30] V. Penela, G. Álvaro, C. Ruiz, C. Córdoba, F. Carbone, M. Castagnone, J. Gómez-Pérez, and J. Contreras, "miKrow: semantic intra-enterprise micro-knowledge management system," The Semanic Web: Research and Applications, pp. 154-168, 2011.
[31] T. Ushiama and T. Eguchi, "An information recommendation agent on microblogging service," Agent and Multi-Agent Systems: Technologies and Applications, pp. 573-582, 2011.
[32] C. Zheng-yan, "Short Message Classification of Microblogging Based on Semantic [J]," Modern Computer, vol. 8, 2010.
[33] Gong, S. "A Personalized Recommendation Algorithm on Integration of Item Semantic Similarity and Item Rating Similarity [J] " Journal of Computers, vol.6, no. 5, pp.1047-1054, May 2011.
[34] Huang, G., S. Wang, et al. "Query expansion based on associated semantic space". Journal of Computers, vol.6, no. 2, pp.172-177, February 2011.

Dongqing Wu was born in Guangxi, China in Sep. 15 1974. He received his B.S. from Beijing Jiaotong University, Beijing, China in 1996. In 2005, he received his Master degree in Computer graphics from Institute of industrial design and graphics, South China University of Technology, Guangzhou, China. His major field of study covers pattern recognition, enterprise application techniques.

He joined Institute of Design, Liuzhou Railway Bureau, Liuzhou, China after graduated from BJU in 1996. He studied for M.S. Degree in SCUT from 2002 to 2005. During the period, he served as software engineer in Guangzhou Century Bridge Technology Co., Ltd., Guangzhou, China. In 2005, he joined Zhongkai University of Agriculture and Engineering with years of industrial experience in July 2005. He is the author or coauthor of more than 20 journal papers. Seven out of them are indexed by SCI and EI. His works is showed as below.

[1] D. Wu, F. Yang, and C. Zhang, “Implementation of Combined Query Component in Enterprise Application,” Modern Computer. Guangzhou, vol. 7, pp. 7-11, April 2011.

[2] D. Wu, F. Yang, and Z. Wu et al, “A Motor Control Method for Vehicles Brake Test Bed Based on Kalman Filter,” Computer Measurement & Control. Beijing, vol.19, pp.869-871, April 2011.

[3] D. Wu, F. Yang, and C. Zhang, “TrueType font s automatic character embroidery knit algorithm,” Computer Engineering and Design. Beijing, vol.29, pp. 513-515, Jan. 2011.

Mr. Wu is currently a member of China Computer Federation. He has more than 10 years experience in the areas of text information processing and software engineering. He has received over $100 thousand research grants since 2005 and collaborated with many experts from academia, national laboratories, and industries. She has been developing state-of-the-art text recognitions models for various industrial applications. Since 2007, Mr. Wu has being leading a research team to help the government improve its public opinion monitoring system. Recently, Mr. Wu has established a new model to analyze vast amounts of microblogging, which aim is extracting useful information from it.

Prof. Fengjian Yang was born in Hunan, China in Nov. 1954. He received his B.S. from Hunan Institute of Education, Xiangtan, China in 1985. In 1987, he received Master degree in Computational Mathematics from School of Mathematical Sciences, Xiangtan University, Xiangtan, China. His major field of study covers algorithm and stability, Dynamics of Neural Networks.

He joined Hunan Institute of Engineering, Xiangtan, China after graduated from XTU. in 1987. During the period, he served as associate professor for ten years. In 1997, he joined Zhongkai University of Agriculture and Engineering with years of academic experience. He is the author or coauthor of more than 50 journal papers. Twenty out of them are indexed by SCI and EI. His works is showed as below.

[1] Y. Fengjian and L. Jiechun, "Positive solution of even order nonlinear neutral difference equations with variable delay," Journal of Systems Science and Mathematical Sciences, 2002.

[2] F. Yang, C. Zhang, and D. Wu, "Global stability analysis of impulsive BAM type Cohen–Grossberg neural networks with delays," Applied Mathematics and Computation, vol. 186, pp. 932-940, 2007.

[3] F. Yang, C. Zhang, D. Wu, and X. Hu, "Uniformly stability of impulsive BAM neural networks with delays," Applied Mathematics and Computation, vol. 186, pp. 23-27, 2006.

Prof. Yang is currently an executive director of Guangdong Province Industrial and Applied Mathematics, Guangzhou Industrial and Applied Mathematics, Guangdong scene Statistical Society. He has more than 20 years experience in the areas of text information processing and algorithm stability. He has received over $100 thousand research grants since 1990 and collaborated with many experts from academia, national laboratories, and industries. Prof. Yang has been leading a research team to explore the stability of neural networks.

Chaolong Zhang was born in Hunan province, China, in 1974. He received his M.S.degree in Foundations of Mathematics from South China Normal University,Guangzhou,China,in2004.He is the author and coauthor of more than 46 journal papers. His works is showed as below.

[1] Z. Chaolong, Y. Fengjian, and H. Xiaojian, "Global Exponential Stability of BAM Neural Networks with Varying. Coefficient and Impulses," Journal of Biomathematics, vol. 22, pp. 395-402, 2007.

[2] C. Zhang and W. Feng, "Oscillation for higher order nonlinear ordinary differential equations with impulses,"

Electronic Journal of Differential Equations, vol. 2006, pp. 1-12, 2006.

[3] C. Zhang, W. Feng, J. Yang, and M. Huang, "Oscillations of second order impulses nonlinear FDE with forcing term," Applied Mathematics and Computation, vol. 198, pp. 271-279, 2008.

His current research interests include stability theory, artificial neural networks, impulsive functional differential systems, stochastic functional differential systems.

Mr. Zhang is currently an executive director of Guangdong Province Industrial and Applied Mathematics, Guangzhou Industrial and Applied Mathematics, Guangdong scene Statistical Society. He has more than 15 years experience in the areas of text information processing and algorithm stability. He has received over $50 thousand research grants since 1997 and collaborated with many experts from academia, national laboratories, and industries. He has been leading a research team to explore dynamics of neural networks.