Connecting Social Network with Smart Education: A Microblog-Based Experimental System

Jie Zhao, Chen Chen

Abstract— In this paper, we propose a microblog-based experimental system for enhancing smart education. In the age of social network, smart education lacks effective platform to help students practice social-network-oriented tasks, such as information pre-processing and information extraction. For this purpose, we develop a microblog-based experimental system that offers many functions to process microblog information, including microblog crawling, microblog pre-processing, and event detection from microblogs. In particular, we implement a number of state-of-the-art models and algorithms in the system, which enable students to learn existing technologies related to social-network information processing more effectively. After a brief introduction on the system framework, we present the experimental results of microblog-based event detection. Finally, we discuss the impact of our system on smart education.

Index Terms— Microblog; Smart education; Experimental system.

I. Introduction

With the rapid development of social network, microblogs become one of the major sources of getting fresh news and events [1]. Compared with traditional news media such as newspaper, Web news portals, etc., microblogs can offer more timely information because of its convenience in posting microblogs, which means that people can use many devices including PCs, smart phones, tablets, or PDA to post microblogs. Therefore, microblogs and related topics have been one of the research focuses in recent years [2].

However, current smart education systems lack the ability of providing effective experimental systems for students to learn social-network-related technologies effectively. For example, in many courses, students are not able to conduct experiments on real social-network information to learn how to utilize social-network information for decision making. Therefore, it is urgent to develop an experimental system to let students understand social network more deeply and more effectively.

Based on such background, in this paper we propose a microblog-based experimental system for enhancing the effectiveness of social-network-related smart education. We aim to construct a platform that offers students a number of functions to practice many necessary tools on microblog information, including microblog crawling, microblog preprocessing, and microblog information extraction. In summary, we make the following contributions in this paper:

(1) We propose a microblog-based experimental system for smart education in the Web 2.0 age. Our system provides a number of functions that enable students to practice many

Jie Zhao, School of Business, Anhui University, Hefei, China, Phone No. +86-551-65108040.
Chen Chen, School of Business, Anhui University, Hefei, China, Phone No. +86-551-65108040

tools necessary for processing microblog data. Therefore, students can learn quickly and effectively social-network-related knowledge via this experimental system.

(2) We present an experiment on real microblog data to show the use of the experimental system. In this experiment, we implement five classification models and run them on real microblog data to extract interested events.

(3) We present a discussion on the impact of the proposed experimental system on smart education, especially on social-network-related education.

The remainder of the paper is organized as follows. In Section 2 we discuss the framework of the microblog-based experimental system. Section 4 discusses a case study of the experimental system as well as the results. In Section 4, we give a discussion on the impact of the experimental system on smart education. In Section 5, we survey the related work. Finally, Section 6 concludes the paper.

II. Microblog-Based Experimental System for Smart Education

A. General Framework

Figure 1 shows the general framework of the microblog-based experimental system. The framework mainly consists of two modules, namely user interface, microblog crawler and preprocessing, and microblog event detection.

The user interfaces is used by students to config the experimental system and receive running results. Through the user interface, students are allowed to select microblog data sets and algorithms as well as to see different results of running different algorithms. In addition, students are allowed to view the implementations of all the involved algorithms and data structures.

The module of microblog crawler and preprocessing aims to get microblogs from microblogging service platforms and remove the microblogs that are meaningless. Microblogs can be downloaded via APIs provided by microblog platforms. For the filtering of microblogs, we employ a quality-based approach to remove the microblogs with low quality. Microblog quality evaluation can be done in terms of user aspect and content aspect. According to the user-oriented microblog quality evaluation, the user behaviors are used to tell whether a user is a spamming robot. According to the content-oriented viewpoint, the microblog messages are used in the evaluation process, and different metrics can be applied for this purpose. In addition, this module also provides NLP (Natural Language Processing) tools to perform word-splitting and POS tagging tasks.

The module of microblog event detection is to extract events from the microblogs downloaded by the crawling module. It is possible to use some rule-based approaches to

detect microblog event. However, due to the large data volume of microblogs, it is more popular to apply a machine-learning method to extract events from microblogs. This can be done by employing some classification models, such as SVM and Naïve Bayes. In this module, we implement a number of existing classification models that offers students more choices when they are learning related social network knowledge.
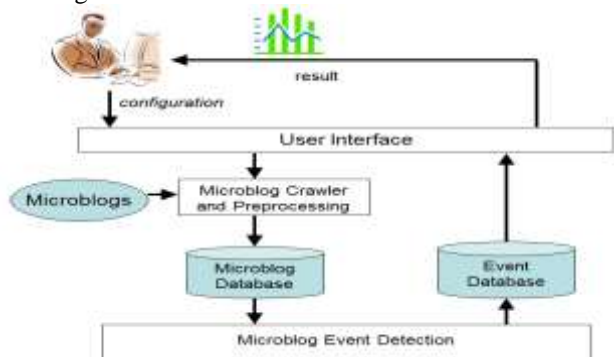

Fig.1. General framework of the microblog-based experimental system

B. Classification Models Implemented in the System

The microblog event detection module involves classification models, which is the most complex part in the system. Classification methods are very popular in data mining related researches. According to the task of event detection from microblogs, we can also transform the event detection task into a classification problem. In general, all the microblogs can be regarded as two classes, namely event microblogs or non-event microblogs. Therefore, the objective of event detection from microblogs is to determine whether a microblog is within the class of event microblogs. This can be done by classification methods, through which all the microblogs can be classified into two types automatically.

As classification methods used in real applications have to be mature and well-demonstrated, people usually utilize some existing classification methods in different applications. Hence, in the event detection from microblogs, we also plan to use some mature classification methods.

So far, there are a lot of mature classification methods proposed [12, 13, 15], such as SVM and decision tree. In this section, we will discuss some major classification methods. In the next section, we are going to test the performance of these different classification methods, and try to reveal which one has the best performance in event detection from microblogs.

(1) SVM

Support Vector Machine (SVM, also Support Vector Networks) is a typical classification method, which is a supervised machine-learning model with associated learning algorithms that analyze data and recognize patterns. It can also be used for regression analysis. Given a training set that contains a number of examples as well as their categories, an SVM-based classification method can build a model that automatically determine the right category of each new example. SVM, as a non-probabilistic binary linear classifier, first represents the examples as points in space, and then divides all the examples by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

(2) Decision Tree

A Decision Tree is a tree-like structure. Each internal node in a decision tree represents a conditional test on an attribute, each branch represents the testing answer, and each leaf node represents a class label. The paths from the root to leaf represent classification rules.

Decision trees are commonly used in many applications to perform classifications. There are various implementations of decision trees, among which the most influential one is the C4.5 algorithm (thus we can call this algorithm Decision Tree C4.5). C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [13]. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training examples according to the concept of information entropy. The training examples are a set of samples that have already been classified (with known class labels). Each sample consists of a n-dimensional vector, where each element in the vector represents a feature of the sample, as well as the corresponding class.

(3) Naïve Bayes

Naïve Bayes is another popular classification method. Naive Bayes classifiers are actually simple probabilistic ones that are based on the Bayes' theorem, assuming that all the features are independence. Naive There are a lot of variations of Bayes models, such as simple Bayes and independence Bayes, which are also well-known in the research community. The same property of all the Bayes models is that they are all based on Bayes' theorem.

Naïve Bayes has been studied extensively since the 1950s. It was first introduced to the area of information retrieval in the early 1960s. Now it has been a very popular method for many issues such as text categorization, document classification, etc. When used for these issues, documents are treated as basic data, and features about word, grammar, and contents are usually used as attributes. With appropriate preprocessing, Naïve Bayes can provide good performance and is competitive with some recently-advanced methods such as SVM.

Basically, Naive Bayes classifiers have the assumption that the value of a specific feature (attribute) is unrelated to the presence or absence of any other feature, given the class variable. For example, a fruit might be regarded as an orange in case that it is red or orange, round, and about 3" in diameter. According to Naïve Bayes classifiers, all the features considered contribute independently to the probability that this fruit is an orange, regardless of the presence or absence of the other features.

Despite the simple assumptions, Naïve Bayes classifiers work very well in many real applications. A comprehensive comparison with other classification algorithms in 2006 showed that Bayes classification is outperformed by other approaches, such as boosted trees or random forests. Another advantage of Naive Bayes is that it only requires a small number of training samples to estimate the parameters that are required by the classification procedure.

(4) Backpropagation Artificial Neural Networks

Artificial Neural Networks (ANNs) are computational models which have been applied in machine learning and patter recognition. Artificial neural networks are generally presented as systems of interconnected neurons which can compute values from inputs. Similar to other machine learning models, artificial neutral networks are also based on

training and prediction, i.e., they are supervised models as SVM and Naïve Bayes.

The most influential ANN model is the Backpropagation Artificial Neural Networks (BP ANNs). The term "backpropagation" refers to "backward propagation of errors". The backpropagation is a very common method used in training artificial neural networks. Backpropagation requires an expected output for every input value so that we can calculate the loss function gradient.

(5) KNN

The K-Nearest Neighbors algorithm (KNN) is also widely used as a classification model. KNN is a non-parametric method that needs not to determine parameters for classification. According to KNN, the input consists of K closest training samples, and the output is a class label. The category of an inputted object, e.g., a document, will be determined by the votes from its neighbors, and the object will be labeled as the class that is most common among its K nearest neighbors.

KNN is a simple but effective classification model and can be used for various applications such as document classification and location recommendation. In some cases, we can assume that the contributions of neighbors are not the same. For instance, a closer neighbor may contribute more than other neighbors far away from the object.

III. A CASE STUDY OF THE SYSTEM

In this section, we discuss a case study of using the microblog-based experimental system to show how this system can support smart education.

A. Settings

In this case study, students are allowed to use a data set downloaded via APIs from Sina Weibo (http://weibo.com/). It is also possible for students to use a crawler to manually get tailor-made data set [19]. This data set consists of 1,02 million microblogs which cover a time range from April 6, 2013 to April 8, 2013. Then, students are required to preprocess these original microblogs to remove some spams. We offer ICTCLAS (http://ictclas.org), which is the most widely used tool for word segmentation in Chinese, for students to divide words and perform POS tagging. After that, students can use the filtering module of the system to remove the microblogs without any person names, location names, organization names, and time words. Finally, we urge students to select 5,000 microblogs as the final data set to perform microblog event extraction.

In order to determine the correct events, the original 5,000 microblogs are manually annotated. We invite three students to perform the annotation. If a microblog is recognized as an event by at least two students, this microblog is annotated as containing a correct event. Finally, there are 2,704 microblogs that are recognized as event microblogs, and 2,296 ones are non-event microblogs. The 2,794 microblogs are labeled as the relevant set.

Four different metrics are provided for students to measure the performance of various methods including (1) SVM, (2) Decision Tree C4.5, (3) Naïve Bayes, (4) BP ANN, and (5) KNN. These metrics include precision, recall, F-measure, and AUC [12], which are all commonly used in information extraction and retrieval. Suppose that TP is the number of correctly detected events, FN is the number of events that are missed in the results, FP is the number of non-events in the result, the precision of event detection is defined as $P = TP/(TP+FP)$ and the recall is $R = TP/(TP+FN)$. F-measure is a combination of precision and recall, which is defined as $2PR/(P+R)$. AUC is another metric for evaluating classification performance, which refers to the area under the ROC curve of the result [12].

B. Results

In order to conduct extensive comparisons on the five classification methods, we first need to determine the features used for event detection.

Previous studies on microblogs have considered some unique features of microblogs [16], e.g., authority information and some unique symbols of microblogs. Liu et al. [17] proposed to consider the POS tagging feature. They also propose to use some semantic dictionary such as HowNet to expand the textual words in microblogs so that new features can be added [18].

Based on existing work [16-18] and the properties of microblogs, we finally select 17 features for students to form the entire feature set. Further, we select a sub-set of features from the 17 features to form another feature set, which will also be used in the experiments. As a consequence, all the features are shown in Table 1 and 2. Note that the first feature set contains all the 17 features, while the second feature set contains the features numbered 2, 3, 4, 5, 6, 8, and 16. The second feature set is based on the view of news reports, in which the features are very popular in traditional news event detection.

Table 2 shows the results of event detection, where the first feature set is used. As Table 2 shows, Decision Tree C4.5 gets the highest precision and the AUC value, but its F-measure is a bit lower than BP ANN. BP ANN has the best results in terms of recall and F-measure, but it has too high cost in run time. Especially, the run time of BP ANN is 30 times more than that of Decision Tree C4.5. In summary, Decision Tree C4.5 has the best overall performance when all the 17 features are used in classification.

Table 1. Features used for classification

No.	Features
1	*text length*
2	*number of person names*
3	*number of location names*
4	*number of time words*
5	*number of organizations*
6	*number of verbs*
7	*has Web links or not*
8	*number of other nouns*
9	*number of stop words*
10	*number of words after segmentation*
11	*number of total words*
12	*frequency of stop words*
13	*containing " 【】 " or not*
14	*containing "#" or not*
15	*number of punctuations*
16	*number of all the nouns*
17	*number of followers.*

Table 2. Results of event detection on the first feature set (17 features)

	Precision	Recall	F-Measure	AUC	Run Time (ms)
SVM	0.624	0.769	0.689	0.611	95080
Decision Tree C4.5	0.705	0.74	0.722	0.706	5456
Naïve Bayes	0.607	0.776	0.681	0.661	417
BP ANN	0.685	0.779	0.729	0.721	173428
KNN	0.654	0.643	0.649	0.642	4294

Table 3. Results of event detection on the second feature set (7 features)

	Precision	Recall	F-Measure	AUC	Run Time (ms)
SVM	0.698	0.761	0.728	0.686	53566
Decision Tree C4.5	0.697	0.793	0.742	0.713	2732
Naïve Bayes	0.638	0.767	0.696	0.682	372
BP ANN	0.677	0.774	0.722	0.701	98316
KNN	0.687	0.679	0.683	0.685	6654

When the second feature set is used, we get the results shown in Table 3. All the classification results are based on a 10-fold cross-validation method, and we do not divide the training set and testing set. As Table 3 shows, SVM has the highest precision. Decision Tree C4.5 gets the second highest precision but which is pretty close to the precision of SVM. In addition, Decision Tree C4.5 has the highest recall and F-measure. Regarding time performance, Decision Tree C4.5 is worse than Naïve Bayes, but it is better than all the other three methods. Although Naïve Bayes is very time efficient, it is not a good model for event detection from microblogs because of its poor performance w.r.t. precision and recall.

Based on all the results shown in Table 2 and 3, we can see that Decision Tree C4.5 has the best overall performance in event detection from microblogs. Therefore, it is more appropriate to use this classification method as the base model in microblog-based event detection. Also the BP ANN and SVM methods can be considered in case that the time performance is not much critical for applications. If time performance is the most critical issue, we can consider using the Naïve Bayes method.

IV. IMPACT OF THE SYSTEM ON SMART EDUCATION

Social network and Web 2.0 has been a hot topic in smart education. However, many courses in smart education lack effective experimental systems to offer students deeply understanding on the underlying knowledge and tools. Thus, the proposed experimental system in this paper can be helpful in many aspects in smart education.

First, the proposed experimental system can help students understand the basic NLP processing techniques and tools. Our system implemented a lot of text processing tools that are commonly used in social-network information processing. Students can config and use these tools to process real microblog data to see the effectiveness and efficiency of these tools. Further, they can use these tools in any other applications.

Second, we provide real microblog data in the proposed experimental system. These data can be used by students to perform many tasks like data mining, information extraction, and decision making.

Third, machine learning technologies have been a focus in many smart courses. Our system integrated five machine learning models that are commonly used in many applications. After a study on the theoretical aspects of these models, students can use the proposed system to quickly learn the real effectiveness of these models. Moreover, they can compare the performance of different models on real data set via our system. This is more efficient for students to grasp machine learning knowledge.

Finally, the proposed system is open-sourced. Thus, students can extend their own designs or improvements on the provided modules in the system. Specially, our system can be used to compare a new approach with existing ones.

V. RELATED WORK

The special characteristics of microblog bring new opportunities and challenges for event detection on the microblog platforms. Those opportunities can be summarized as follows [1]:

(1) As microblogs have the real-time property, events detected from microblog messages are usually more fresh and useful than those from traditional information sources such as web pages. Users are enabled to post event information at a very early stage when the event occurs. This behavior is usually faster than other media such as news portal website or newspaper.

(2) As there are a great number of active users on the microblog platform, who may distribute among a large scope of geographical area, we are able to detect more useful events from microblogs. On the contrast, traditional events detection in new portal websites depends on the engagement of news workers.

(3) As microblog has the social network property, we are able to detect some special events from microblogs, such as events in a specific domain, events in a specific group, or events in a specific area.

Microblog and its related researches have been a hot topic in recent years, due to the large number of users and the large amount of data in microblogging platforms. Consequently, microblogging platforms are becoming new social interaction places in the Web age.

Compared with traditional Web systems, microblogging platforms have the following unique features:

(1) Microblogs are real-time information. This is mainly because users can post new short microblogs about a certain event easily and conveniently if they are at the event place and happen to know the event. On the contrary, Web pages have to be carefully prepared and reviewed which is more time-consuming. Therefore, the events detected from microblogs are more fresh and valuable for users than traditional Web systems.

(2) Microblogs usually contain more new events than traditional Web pages because of the large number of microblogging users living in a wide geographical range.

(3) Microblogs contain more social-interaction features than traditional Web pages. Thus it is possible to detect more specific events from microblogs, such as events about a specific domain, events in a specific place, etc.

Detecting events from microblogs is somehow similar with some previous work on TDT (Topic Detection and Tracking) [8]. These works on TDT typically employ some machine learning models, which are not suitable for microblogging platforms. This is because microblogs are very short (< 140

characters) and machine learning models perform poorly on short texts [8]. There are also some existing works focusing on microblog-based event detection [5-7]. However, they only use a term-frequency-based approach, which simply use the hot words to represent events [7]. This approach ignores the other features of events, such as the happening time of events, places of events, etc., and thus cannot form completed description about events. For instance, if we simply use a hot word "Olympic Game" and do not consider other information such as the year of an Olympic Game, we are not able to get the exact and detailed information about Olympic Game. On the other hand, many applications in emergency treatment and public security management have urgent needs on obtaining the details about events.

One major issue in microblog-related researches is how to effectively extract events from microblogs [3]. Previous studies have shown that public emergent events have very different ways when spreading on microblogging platforms [4]. There are some existing works focusing on analyzing the new features of microblogs and further presenting new approaches for event detection from microblogs. Basically, these works can be classified into three types: specific event detection, specific person's detection, and integrated event detection. For specific event detection [5], people usually use some specific keywords in microblogs such as earthquake and tornado to find new events. As microblog users can post information very quickly when these events happen, it is possible to detect such events timely. For specific person's detection [6, 9], people use the name of a specific person to search his or her relevant events in microblogs. For integrated event detection, a very popular method is to detect those hot words in microblogs [7]. Some researchers also utilize the location information embedded in microblogs, i.e., Geo-tags, to find local events from microblogs [10, 11].

Presently, microblog-based event detection is commonly based on some classification models. A classification model aims to determine the exact class of a given object given that there are some pre-defined classes. The input of a classification task is a set of records, each of which is a tuple (x, y). Here, x is a feature set, and y is a set of target classes. A classification model is to obtain a target function f mapping x to y through a learning stage [12]. Classification models usually utilize some learning algorithms to determine a final function that can describe the relationship between x and y perfectly. This function is not only used to explain the relationship between x and y in the given data set, but also used to predict the possible y of a new x input [13]. Thus, we usually divide the entire data set into a training set and a testing set. The training set is used to determine the right function f between x and y, and the testing set is used for performance evaluation.

There are a lot of classification models that have been proposed and widely used so far, including the decision tree C4.5 model, the BP neutral network model, the SVM (Support Vector Machine) model, the KNN (K nearest neighbors) model, etc. [12, 13]. A classification method discovers some rules from certain training sets with knows classes, and then use these rules to predict new classes [12, 13]. Classification as a common machine-learning approach has been widely used in many areas including document classification, information retrieval, intrusion detection, object recognition, information extraction, etc. However, regarding the issue of event extraction from microblogs, it still remains unknown which classification models will have the best performance. Thus, it is necessary to study the applicability of existing classification methods on event detection from microblogs.

When dividing the entire data set into a training set and a testing set, if the features of these two sets vary a lot, we can imply that the function learned from the training set will not suit for the testing set. Consequently, we will get poor performance. Therefore, in real applications we usually use cross-validation to measure the performance of a classification model. Cross-validation is a popular method in the areas of statistics and machine learning [14]. In practical experiments, the most commonly used cross-validation method is the k-Fold Cross-Validation. According to this method, the entire data set is partitioned into k groups. Each group is taken as a testing set and the remaining groups are used as the training set. Thus, we will get k functions as well as k precision results. Finally, the mean of the k results is used as the final precision metric to evaluate the overall performance of the classification model.

VI. CONCLUSION

In this paper, we presented an experimental system for smart education in the age. Specially, we use the microblog data as the social network information source, and implement a number of text processing tools as well as classification models in the system, which allows students to perform various social-network-related experiments and fasten the knowledge-learning process. We discussed the general framework of the proposed system and explained the functions as well as the modules of the system. Then, we presented a case study of the system. In addition, we briefly discussed the impact of the proposed system on smart education.

ACKNOWLEDGMENT

REFERENCES

[1] P. Jin, L. Mu, L. Zheng, J. Zhao, L. Yue. News Feature Extraction for Events on Social Network Platforms. *Proc. of WWW*, 2017, pp. 69-78
[2] L. Zheng, P. Jin, J. Zhao, L. Yue. A Fine-Grained Approach for Extracting Events on Microblogs, In *Proc. of DEXA*, LNCS 8644, 2014, pp. 275-283
[3] Y. Xie, *The Report on Chinese Social Opinion and Crisis Management*, Social Science Literature Press, 2012
[4] J. Zhao, P. Jin, G. Huang. A Survey on Detecting Public Emergencies from Web Pages, *Advances in Information Sciences and Service Sciences*, 2011, 3(3): 56-63
[5] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *Proc. of WWW*, Raleigh NC, USA, 2010, pp. 851-860
[6] A.M. Popescu, et al. Extracting Events and Event Descriptions from Twitter. In *Proc. of WWW*, Hyderabad, India, 2011, pp. 105-106
[7] R. Long, H. Wang, Y. Chen, et al., Towards Effective Event Detection, Tracking and Summarization on Microblog Data. In *Proc. of WAIM*, Wuhan, China, 2011, pp. 652-663
[8] J. Zhao, X. Li, P. Jin, A Time-Enhanced Topic Clustering Approach for News Web Search, *International Journal of Database Theory and Application*, 2012, 5(4): 1-10

[9] A.M. Popescu, M. Pennacchiotti. Detecting Controversial Events from Twitter. In *Proc. of CIKM*, Toronto, Canada, 2010, pp. 1873-1876

[10] K. Watanabe, M. Ochi, M. Okabe, et al., Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs. In *Proc. of CIKM*, Glasgow, UK, 2011, pp. 2541-2544

[11] J. Zhao, P. Jin, Q. Zhang, R. Wen. Exploiting location information for Web search. *Computers in Human Behavior*, 2014, 30: 378-388

[12] A. P. Field, *Discovering Statistics Using SPSS (Introducing Statistical Methods S.)* (2nd Edition), Sage Publications Ltd, 2005.

[13] J. Zhao, X. Wang, P. Jin. Feature Selection for Event Discovery in Social Media: A Comparative Study. Computers in Human Behavior, 2015, 51: 903-909

[14] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proc. of IJCAI*, Quebec, Canada, 1995, pp. 1137-1145

[15] S. Lin, P. Jin, X. Zhao, L. Yue, Exploiting Temporal Information in Web Search, *Expert Systems with Applications*, 2014, 41(2): 331-341

[16] B. Sriram, et al., Short Text Classification in Twitter to Improve Information Filtering. In *Proc. of SIGIR*, Geneva, Switzerland, 2010, pp. 841-842

[17] Z. Liu, et al., Short Text Feature Selection for Micro-blog Mining. In *Proc. of CiSE*, Wuhan, China, 2010, pp.1-4

[18] Keenage.com, HowNet, In http://www.keenage.com/, Accessed 10 September, 2017

[19] Y. Sun, P. Jin, L. Yue. A Framework of A Hybrid Focused Web Crawler, In *Proc. of the Second International Conference on Future Generation Communication and Networking Symposia (FGCNS'08)*, 2008. Volume 2, pp. 50-53

Jie Zhao received her PhD. degree from University of Science and Technology of China in 2014. Now she is a full professor in the School of Business, Anhui University. In recent years, her research interests focused on social network and Web information management.

Chen Chen is currently a master student in the School of Business, Anhui University. His research interests are microblogging services and Web information extraction.