International Journal of Database Theory and Application
Vol.7, No.4 (2014), pp.11-22
http://dx.doi.org/10.14257/ijdta.2014.7.4.02

A Wrapper for Extracting Information Records from Forums based on Page Segmentation

Chunshan Li[1,2], Jingjing Chen[3,4], Dianhui Chu[5], Ge Song[1,2] , Haijun Zhang[1,2], Yunming Ye[1,2] and Jianliang Xu[3]

[1]*Shenzhen Graduate School, Harbin Institute of Technology*
[2]*Shenzhen Key Laboratory of Internet Information Collaboration*
[3]*Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, China*
[4]*School of Science and Technology, China Jiliang University, Hangzhou, China*
[5]*Department of Computer Science, Harbin Institute of Technology*
lichunshan.hit@gmail.com, jjchen@comp.hkbu.edu.hk,carroll0708@qq.com, chudianhui@hit.edu.cn, aarhzhang@gmail.com, yeyunming@hit.edu.cn, xujl@comp.hkbu.edu.hk

Abstract

Foraging information from web forums is still one of the most challenging information retrieval tasks due to various combinations of auto-generated page structural information and user-created contents. Traditional information extraction methods employ either duplicated subtree pattern detection methods, or machine learning methods. Due to the periodical update of forum templates and diversity of page contents, aforementioned approaches do not work very well on forum sites. In this paper, we present a page-segmentation based wrapper specially designed for mining data pattern of web forums, which combines a novel page segmentation algorithm and decision tree classifier together to detect the data pattern in forum. In the segmentation phase, a novel page segmentation algorithm is proposed to identify the records areas in a page, then a classifier is adopted to identify the detailed pattern of each record in the extraction phase. Extensive experiments on various types of web forums are conducted and the results conclude that our wrapper is a more generalized one which requires only few labeled training data.

Keywords: *Data Extraction; Page Segmentation; Web Structure Mining; Decision Tree*

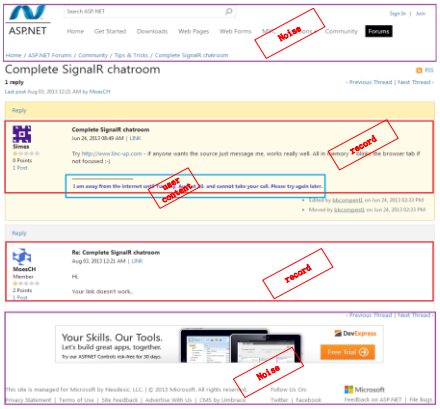
1. Introduction

The rapid growth of social media makes web forums to be an important data resource. In forums, people can create, share and discuss information freely. Consequently, it accumulates plenty of highly valuable information [1]. This phenomenon attracts considerable research efforts to extract valuable information for further analysis, *e.g.*, public opinion analysis and monitoring, forum search engine, question-answer system [2] and expertise community discovery [3].

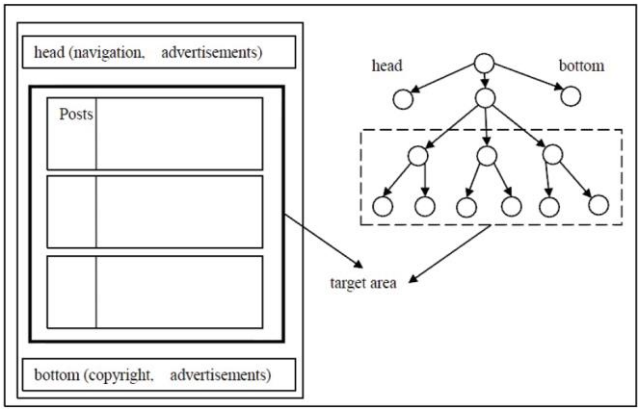
However, foraging information form web forums is still one of the most challenging information retrieval tasks. Firstly, a forum page is usually congested with many "noisy" information such as site title, banner, navigation, advertisement, decoration and *etc.*, [4] which seriously skews the performance of knowledge discovery process. Furthermore, most forum sites not only suffer from noise data, but also allow users to generate their contents in

ISSN: 2005-4270 IJDTA

free-style. For example, posts may contain text and images or just a fragment of HTML code. Due to the diverse representation formats of posts, most existing data extraction tools do not work very well. Figure 1(a) shows an example of forum page. In Figure 1(a), page areas highlighted by red rectangles are data blocks to be extracted, areas in blue and purple rectangles are noise that should be filtered by forum wrapper.



(a) An Example Page of Forum



(b) Correspondence between Dom Tree Structure and Page Modules

To develop a generalized and accurate tool for web data extraction, many approaches have been proposed. Firstly, data extraction is thought of as duplicated subtree pattern detection problem [1]. These methods have shown the feasibility of handling post data of different layout styles, which work well on web sites generated by single template, but not on forums generated by multiple templates. Secondly, machine learning and ontology-based approaches have been used to various forums [5]. In practice, these approaches can only be used in many static web pages, but can't be adopted to current dynamic social media sites.

In this paper, we present a page-segmentation based wrapper specially designed for structural information extraction from web forums which is desired to extract generalized, robust and accurate data records. The system successfully combines page segmentation algorithm and decision tree classifier to detect the structural information in forum. In segmentation phase, a novel page segmentation algorithm was proposed to detect the record area in a page. In the extraction phase, the decision tree classifier is adopted to identify the details of each record. By doing so, the data pattern of each web forum page could be successfully parsed.

The remaining of this paper is organized as follows. Section 2 reviews some related works. We clarify some basic concepts in Section 3.1. The wrapper system is introduced in Section 3. Experiments and evaluation results are demonstrated in Section 4. At last, we conclude the paper in Section 5.

2. Related Works

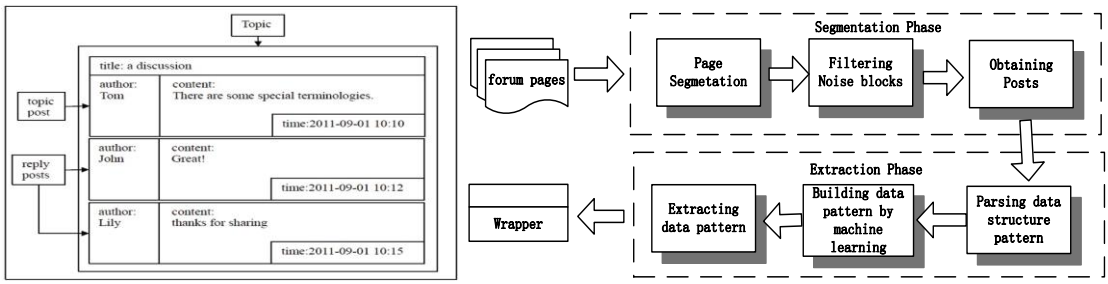
Due to its importance and the wide range of application area, data extraction techniques on web forums have attracted more and more research efforts in recent years. As discussed in the previous section, the most related approaches can be categorized into two-fold: (1) duplicated subtree pattern detection approaches, and (2) machine learning based approaches

Duplicated subtree pattern detection approaches are widely applied to extract information from web pages. Generally, they utilize the structural information of DOM tree to generate a wrapper [6, 7]. The assumption is that sub-trees with frequently repeated structures are used to carry data records. However, diverse layout styles make it difficult to identify the repeated structures in the DOM tree [6], and thus additional manual interactions are needed to assist these data extraction approaches to improve the performance [7]. Moreover, only few specific web sites are designed to have duplicated structure patterns which makes structure pattern detection based approach not to be a general one.

From the perspective view of machine learning, Hidden Markov model-based information extraction technique is thought to be a more generalized approach for data extraction [8]. The proposed approaches adopt Hidden Markov model to analyze forum pages, and extract data records. To train an accurate HMM, it requires a lot of efforts to prepare the training data as well as a longer learning process of the model parameters. Feng *et al.*, [9] assumed that there always exist 12 fields in a record. Then, they employed SVM to classify DOM nodes into the predefined 12 categories. It is obvious that the predefined categories strongly restrict the model flexibility.

3. Extraction for Web Forum Data

In this section, we first describe some basic concepts used in this paper, then present the architecture of our system. The last part of this section we focus on how to find all data records in forum pages and present a new method for representing and generating the forum wrapper.



(a) (b)

Figure 2. (a) Concepts in Post Page. (b) System Architecture

3.1. Concept Definition

To clearly formulate the problem as well as to facilitate the following discussions, we quickly review some basic concepts used in this paper.

Forum: an open public discussion space in the Internet, where personal opinions on a particular topic could be shared and exchanged. In a forum, users can publish posts, read and comment on posts generated by other forum users.

Post: a record in a forum and it usually represents user's perspectives to a specific topic. Post page: an independent page in a forum, which usually represents a discussion among forum members. It may consist of several posts.

Topic post: the first post in a post page and topic post usually represents main issue in a discussion. reply post: stands for the follow-up posts replied by other users. Figure 2(a) shows

an example of post page, in which post, topic post, reply posts are illustrated. In each post, author, time, and contents will be extracted, which are the main targets of proposed wrapper.

3.2. System Architecture

Our system incorporates two main phases, as depicted in Figure 2(b): (1) segmentation phase, and (2) extraction phase. The goal of the first step is to automatically find each independent post. In practice, the first step will segment a forum page into several independent post modules. After page segmentation, structured data with explicit semantic details could be identified in those posts. In Figure 2(b), each step is denoted by a dash-lined box.

3.3. Segmentation for Posts Extraction

In a forum, post page not only contains a number of posts, but also contains some noise data. Therefore, it is necessary to segment post pages into independent semantic modules, then the post area and noise area could be separated. Inspired by the spirit of traditional segmentation methods, we assume posts are generated in such a way that contents with independent topics are encapsulated in close sub-tree of DOM tree having similar layout styles.

Figure 1(b) shows a part of our assumption. In left of Figure 1(b), post areas are highlighted in bold rectangle, and its corresponding abstracted DOM tree model can be seen at the right of the Figure 1(b). It can be seen that there does exist a mapping relationship between post modules and sub-trees of DOM tree. Users can analyze posts via features of DOM tree. Another part of our assumption is one post is generated by one user in his/her own style, which make it have clear visual clues. Figure 1(a) indicates 2 data records in a post page. As can be seen, they have similar but different visual styles, e.g. module size and text font.

Based on these assumptions, a set of heuristic rules can be generated to judge whether a DOM tree node is a post node or not. If a node is post node, then this post is to be extracted. In order to exactly detect post, a novel page segmentation algorithm is proposed by considering the following aspects: 1) DOM tree structure; 2) visual style; 3) additional information from users. The heuristic rules are listed in Table 1.

Table 1. Heuristic Rules in Post Extraction Phase

DOM tree rules	rule 1	In a page, post nodes have the specific type, they are either “DIV” node, or “TABLE” node
	rule 2	Nodes in subtree of post node have similar type, such as “A”, “UL” and “SPAN”. If all types appearing in node 1 is the same as types in node 2, node 1 and 2 are similar.
visual rules	rule 3	The width of post module must be larger than 480 pixel, the height of post module must be large than 20 pixel
	rule 4	The width difference between two posts must be small than 20 pixel
additional rule	rule 5	Post modules must have a time stamp

3.4. Classification Based Records Extraction

After post page segmentation, all posts contained in forum page are already extracted. Then posts will be parsed into records. According to post's semantic detail, a record contains 4 feature fields e.g. "author", "time", "content" and "additional information". In this phase, we first extract all text nodes in posts. Then, the record extraction task is mapped to the classification problem, in which text nodes in posts can be viewed as samples and their labels are the above 4 feature fields. In the following, our system summarizes 8 attributes for samples which is very related to classification label. Those attributes will be defined in Table 2 which is the basis of the following classification task.

Table 2. Data Features in Classification

Id	definition	description	Id	definition	description
1	$width = \begin{cases} 1 & v<200 \\ 2 & 500>v\geq 200 \\ 3 & v\geq 500 \end{cases}$	width of node v	5	$parent = \begin{cases} 1 & v='A' \\ 2 & v='H' \\ 3 & v='SPAN' \\ 4 & v='DIV' \\ 5 & v='B' \\ 6 & v='FONT' \\ 7 & other \end{cases}$	type of parent or grandparent
2	$height = \begin{cases} 1 & v<40 \\ 2 & 400>v\geq 40 \\ 3 & v\geq 400 \end{cases}$	height of v	6	$category = \begin{cases} 1 & v='author' \\ 2 & v='time' \\ 3 & v='content' \\ 4 & v='additional\ information' \end{cases}$	categories of v
3	$class = \begin{cases} 1 & v='class' \\ 2 & other \end{cases}$	parent of v has the 'class' attribute	7	$space = \begin{cases} 1 & v='space' \\ 2 & other \end{cases}$	text in v contains space.
4	$time = \begin{cases} 1 & v='time\ stamp' \\ 2 & other \end{cases}$	text in v contains time stamp	8	Integer	depth of v

Decision tree algorithm (DT) [10] is adopted in this phase to classify each text node into a predefined category. Specially, C4.5 algorithm is adopted in this paper. A Decision tree builds decision tree from a set of training samples, using the criteria of normalized information gain. The training data is a set $S = \{s_1, s_2, \cdots, s_n\}$ of training data. Each data sample is a vector $s_i = (x_1, x_2, \cdots)$ where $x_1, x_2, \cdots$ represents attributes or features of the sample. The training data is augmented with a vector $C = (c_1, c_2, \cdots)$ where $c_1, c_2, \cdots$ represents the class to which each sample belongs.

At each inner node of the tree, C4.5 chooses one attribute of the data that most effectively splits the set of samples. The criterion is the normalized information gain radio that results from choosing an attribute to split the samples set. The attribute with the highest normalized information gain radio is chosen to finally split the tree.

The equation 1, 2, 3, 4 and 5 illustrate how to compute the information gain radio of attributes. First, the entropy of sample set D with category information is calculated by

Equation 1. Second, the expectation of samples will be computed, where value of the attributes α is i. Third, normalized information gain of the attribute α can be obtained by Equation 3. Fourth, the entropy of sample set with split attribute α is calculated using Equation 4. Finally, the information gain radio of the attribute α can be computed by Equation 5.

$$\text{entropy}(D) = -\sum_{k=1}^{M} \frac{|D_k|}{D} \times \log_2\left(\frac{|D_k|}{D}\right) \quad (1)$$

where D is the training set, D_k is the training data set of class k and m is the number of categories.

$$\text{expectation}_{\alpha}(D) = -\sum_{i=1}^{A} \frac{|D_{i\alpha}|}{D} \times \text{entropy}(D_{i\alpha}) \quad (2)$$

where α represents an attribute, A is the number of possible values of attribute α and $|D_{i\alpha}|$ indicates the number of samples in which the value of attribute α equals to i.

$$gain_{\alpha}(D) = \text{entropy}(D) - \text{expectation}_{\alpha}(D) \quad (3)$$

$$splitEntropy_{\alpha}(D) = -\sum_{i=1}^{A} \frac{|D_{i\alpha}|}{|D|} \times \log_2 \frac{|D_{i\alpha}|}{|D|} \quad (4)$$

$$gainRateo_{\alpha}(D) = \frac{gain_{\alpha}(D)}{splitEntropy_{\alpha}(D)} \quad (5)$$

After the classification, all text nodes in post are associated with semantic information types. Actually, the category of "additional information" has little explicit semantic meaning and will be ignored. Finally, the obtained target records are labeled with "author", "time" and "content".

4. Experiments and Evaluation Results

To evaluate the effectiveness of our wrapper, extensive experiments are performed on various data sets. The state-of-the-art segmentation algorithm PS-STS [11] and classification algorithm native Bayes(NB) are adopted for the comparison. Experimental results show that our wrapper system can acquire a better segmentation and extraction results than PS-STS and NB.

The PS-STS algorithm [11] can split a post page based on similarity of the DOM subtree structure. Equation 6 shows the similarity function for PS-STS. To acquire best performance, the parameter of PS-STS will be initialized as N = 3, ω_i= {0.6, 0.3, 0.1}.

$$Sim(x, y) = \sum_{i=1}^{N} w_i \sum_{j=1}^{M_i} \frac{1}{M_i} S_{ij} \quad (6)$$

where x, y are two subtrees for comparison, N is the maximum depth of subtree for similarity computation, which means the similarity function only considers nodes in top N layers of x, y. w_i is weight that indicates the layer i's influence on the calculation of similarity, Mi is the number of nodes in layer i, and S_{ij} = 1 if node i and j are of the same type or attributes, and 0 otherwise.

4.1. Experiment Design

As there is no benchmark data set, our testing data set is built as follows. To show the robustness, two types data will be used to estimate our system. (1) post pages and posts are generated automatically by forum generating program; (2) post pages are crawled from famous web forums. Table 3 shows the details of web forums used in the experiments. Semantic modules in each page and data records are manually labeled. If the results extracted from our approach or two baseline algorithms equals to the manual label, then the result is a positive one, and vice verse.

Table 3. Data Set Using Experiment

Id	Forum Site	Description	Id	Forum Site	Description
1	PhpBB	forum creating program	8	www.xcar.com.cn	car clubs
2	Discuz	forum creating program	9	www.xici.net	citizen clubs
3	PhpWind	forum creating program	10	bbs.163.com	general forum
4	IPB	forum creating program	11	www.19lou.com	travel information
5	LeadBBS	forum creating program	12	club.pchome.net	mobile phones
6	tieba.baidu.com	general forum	13	bbs.qq.com	general forum
7	www.tianya.cn	social events discussion	14	club.kdnet.net	general forum

4.2. Page Segmentation Comparison Results

To evaluate the performance, three commonly used criteria are adopted in the experiments which are: Precision (P), Recall (R) and F-score. Let S denote the semantic module set acquired by FPPS/PS-STS, M denote the semantic module set which is manually labeled as the ground truth label. Then Precision, Recall and Fscore can be defined as Equation 7.

$$P = \frac{|S \cap M|}{|S|} \qquad R = \frac{|S \cap M|}{|M|} \qquad Fscore = \frac{2P * R}{P + R} \tag{7}$$

Firstly, the experiment is performed on dataset forum 1-5. Table 4 shows the segmentation results of FPPS algorithm. It can be seen that our wrapper can detect all semantic modules which match the manual labels.

Table 4. Results Illustration in Machine Data

Forum Id	manual	FPPS
1	2359	**2359**
2	3743	**3743**
3	2728	**2728**
4	1457	**1457**
5	2003	**2003**

Secondly, the comparison experiment is performed on dataset forum 6-10. It can be seen from Table 5 that FPPS outperformed PS-STS in term of Precision, Recall and F-score. Possible reason is that the PS-STS algorithm only considers the subtree structure and ignores the information hidden in visual clues and post content. For example, there exist several post pages that only have one post in data set. In this case, the PS-STS is invalid, whereas FPPS works well in these two scenario.

Table 5. Page Segmentation Comparison Result

Forum Id	Precision		Recall		Fscore	
	FPPS	PS-STS	FPPS	PS-STS	FPPS	PS-STS
6	**1.000**	0.835	**1.000**	0.914	**1.000**	0.842
7	**0.974**	0.861	**0.974**	0.929	**0.974**	0.963
8	**1.000**	0.955	**1.000**	0.947	**1.000**	0.947
9	**1.000**	0.943	**1.000**	0.833	**1.000**	0.833
10	0.945	**1.000**	**1.000**	0.856	**0.972**	0.856

4.3. Data Extraction Comparison Results

In this experiment, our focus is to evaluate the performance of information extracted. The native Bayes classifier (NB) is chose for comparison as: (1) naive Bayes classifiers works quite well in many complex real-world situations; (2) it only requires a small amount of training data to estimate the parameters.

The experimental data are forum 11-14. It contains 9383 samples, in which 128 samples will be used for training, others will be used for test. All samples will be generated based on FPPS.

Table 6 reports the comparison results from decision tree classifier (DT) and naive Bayes classifier. As can be seen, both two algorithms work well on data extraction task. The performance of NB is quite close to DT in categories of “author” and “time”. However, the NB classifier confuses the samples in categories of “content” and “additional information”. The recall values of the two categories are poor while the DT obtains a higher value. It is clear that DT outperforms NB in this setting.

Table 6. Comparison Results on Data Extraction

Category	Precision		Recall		Fscore	
	DT	NB	DT	NB	DT	NB
Additional information	**0.935**	0.857	**0.993**	0.443	**0.997**	0.614
author	0.985	**0.988**	**0.994**	0.994	**0.970**	0.991
time	**0.993**	0.927	**1.000**	0.997	**1.000**	0.998
content	**0.996**	0.545	**1.000**	0.856	**0.998**	0.705

4.4. Time Consume Comparison Results

It is also noticed that DT classifier is a more efficient approach and is more suitable in the real world application where a large scale data set is commonly used. Table 7 shows the time spent by DT and NB classifier from 1000 samples to 9383 samples. DT classifier is 2 times faster than NB. Possible reason is that the average height of DT classifier is 3.2. It means that all test samples will be assigned with a label after only 4 comparisons.

Table 7. Time Consumption Comparison (Second)

	2000	4000	6000	8000	9200
DT	**3.87**	**6.04**	**9.75**	**12.2**	**14.59**
NB	5.7	12.6	18.24	22.08	27.11

5. Conclusions

In this paper, we have presented a novel page segmentation based wrapper system to extract structural data from post page of forums. Most existing methods either depends on simple heuristics rules or learning algorithms with a predefined data template. Facing with complex layouts and diverse user created posts, the performances of those approaches are poor. By combining the DOM-tree features and visual clues, our proposed page segmentation method can successfully extract post modules. And then our system extracted structured data using decision tree algorithm. Experimental results show that our wrapper is efficient and effective to extract web forum data. Currently, we are investigating how to extend our system to other types of web data.

Acknowledgements

This work is supported in part by NSFC under Grant no.61300209, no.61370213, National Commonweal Technology R&D Program of AQSIQ China under Grant No.201310087, Shenzhen Science and Technology Program under Grant No.CXZZ20130319100919673, Shenzhen Foundation Research Fund under Grant no. JCY20120613115205826, the Shenzhen Strategic Emerging Industries Program under Grant no.ZDSY20120613125016389, National Key Technology R&D Program No.2012BAH10F03, No.2013BAH17F00 and science and technology development of Shandong Province No.2010GZX20126, 2010GGX10116.

References

[1] J. Yang, R. Cai, Y. Wang and J. Zhu, "Incorporating site-level knowledge to extract structured data from web forums", WWW, **(2009)**, pp. 181–190.
[2] G. Cong, L. Wang, C. Lin and Y. Song, "Finding question-answer pairs from online forums", SIGIR, **(2008)**, pp. 467–474.
[3] J. Zhang, M. Ackerman and L. Adamic, "Expertise networks in online communities: structure and algorithms", WWW, **(2007)**, pp. 221–230.
[4] D. Cai, S. Yu, J. Wen and W. Ma, "Block-based web search", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, **(2004)**, pp. 456–463.
[5] Y. Xiudan and Z. Yuanyuan, "Ontology-based information extraction system in e-commerce websites", Proceedings of 2011 International Conference on Control, Automation and Systems Engineering (CASE), **(2011)**, pp. 1–4.
[6] N. Kushmerick, "Wrapper induction: Efficiency and expressiveness", Artificial Intelligence, vol. 118, no. 1, **(2000)**, pp. 15–68.

[7] S. Zheng, R. Song, J. Wen and D. Wu, "Joint optimization of wrapper generation and template detection", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, **(2007)**, pp. 894–902.
[8] F. Barros, E. Silva, R. Prud'encio, A. Nascimento, *et al.*, Hidden markov models and text classifiers for information extraction on semi-structured texts, in Proceedings of the Eighth International Conference on Hybrid Intelligent Systems, **(2008),** pp. 417–422.
[9] F. J, H. P, G. M., A learning approach to discovering web page semantic structures, In Proceedings of the Eighth International Conference on Document Analysis and Recognition, **(2005),** pp.1055-1059.
[10] L. Breiman, J. Friedman, C. Stone, R. Olshen, Classification and regression trees, Chapman & Hall/CRC **(1984)**.
[11] Y. Wang, B. Li, C. Lin, Data extraction from web forums based on similarity of page layout, In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, **(2009),** pp. 1–5.

Authors

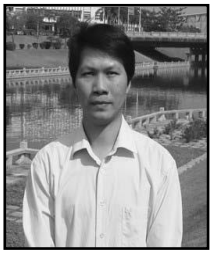
Chunshan Li, received the Master degree in Harbin Institute of Technology. He is now a PHD candidate in the Shenzhen Graduate School, Harbin Institute of Technology. His research interests include data mining, text mining, social computing and topic model.

Jingjing Chen, currently is the Ph. D student from the Department of Computer Science, Hong Kong Baptist University. He is also a lecturer of China Jiliang University. His research interests include educational data mining, e-learning technology and computer supported collaborative work.

Dianhui Chu, is Ph.D. candidate and associate professor of Harbin Institute of Technology. His research interests are within the domain of service computing, service engineering, and software architecture.

Ge Song, received the M.S. degree in Computer Science from Guizhou University in China in 2010. She is currently working towards the Ph.D. degree in the Department of Computer Science, Harbin Institute of Technology. Her research interests include textual classification, especially textual stream classification based on ensemble model.

Haijun Zhang, received the B.Eng and Master's degrees from the Northeastern University, China in 2004 and 2007, respectively, and the PhD degree from City University of Hong Kong in 2010. His research interests include multimedia data mining, machine learning, pattern recognition, evolutionary computing, and communication networks.



Yunming Ye, received the Ph.D. degree in Computer Science from Shanghai Jiao Tong University. He is now a professor in the Shenzhen Graduate School, Harbin Institute of Technology. His research interests include data mining, text mining, and ensemble learning algorithms.

Jianliang Xu, received the BEng degree in computer science and engineering from Zhejiang University, Hangzhou, China, in 1998 and the PhD degree in computer science from the Hong Kong University of Science and Technology in 2002. He is a professor in the Department of Computer Science, Hong Kong Baptist University. He is a member of the Database Group at Hong Kong Baptist University. His research interests include data management, mobile/pervasive computing, wireless sensor networks, and distributed systems. He has published more than 100 technical papers in these areas. He is a senior member of the IEEE.

