International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.7 (2015), pp.1-10
http://dx.doi.org/10.14257/ijmue.2015.10.7.01

Software Defect Model Based on Similarity and Association Rule

Wan Jiang Han[1,] He Yang Jiang[2], Tian Bo Lu[1], Xiao Yan Zhang[1] and Weijian Li[2]

[1]School of Software Engineering, Beijing University of Posts and Telecommunication, Beijing 100876, China

[2]International School, Beijing University of Post and Telecommunication Beijing 100876, China

hanwanjiang@bupt.edu.cn

Abstract

In order to detect defects efficiently and improve the quality of products, this paper puts forward the concept about defect classification model and defect association model by a lot of defect data. The technology of similarity is applied to defect classification model, and the idea of Knowledge Discovery in Database is applied to defect association model. Defect classification model can analyze the defect efficiently and provides the basis of solving problems quickly while defect association model can be used to detect early and prevent problem, which can make effective improvements for testing and development. This paper summed up GUI defect model based on a large number of interface defects. The model is useful to improve the accuracy of forecast and be used for test planning and implementation through the practice of several projects.

***Keywords:** defect model, association rule, defect classification, defect association, similarity*

1. Introduction

With the rapid development of information technology, the application of computer software is more and more widely. A variety of information products emerged and the requirements of product quality are also increasing. By Analyzing and studying defects, building defect classification model and defect association model can help to find solutions for similar defects quickly and efficiently, which provides a good exploration to improve the quality of product. This paper puts forward the defect classification model based on similarity techniques by a large number of defect data from testing. Meanwhile, defect association model according to the association rule theory is also proposed. Through the defect classification model and defect association model, you can classify and solve problem on time. It provides a good recommendation for product development and design. During the testing process, there is a large of defect data, it is necessary to merge the same or similar defect to the same type of defect for unify solution easily. And more defects can be found by a special defect. So, defects can be effective managed by defect classification model and defect association model [1].

Similarity measurements can also be used to define defect classification. By similar defect recognition, repeated defects or very similar defects can be removed from defect library. Similar problems recognition needs to use the knowledge of natural language management to identify the similarity of sentences, so as to achieve the purpose of classification. Defect association analysis can indicate that an emergence of defect may lead to one or the other defects to appear. Defect classification model and defect associate model can improve the quality of products and give a better way to discovery issues [2].

ISSN: 1975-0080 IJMUE

2. Related Partition and Similarity

We start with some preliminary concepts and notation. Consider a set of n elements [n]={1,2,…,n}. A partition of [n] is for nonempty sets Ci with $[n] = \bigcup_{i=1}^{k} c_i$ and $c_i \bigcap c_j = \phi$ for $i \neq j$. A partition can be expressed by means of the n n matrix with entries if both elements l1 and l2 belong to the same cluster Cj for some 1 <j< k and $R_{l_1 l_2} = 0$ otherwise. Partitions can be alternatively defined by the memberships ci where ci =j if and only if i Cj [3].

A similarity matrix (SM) S of size n is a square, symmetric and positive semi-definite matrix with entries 0 < Sij<1 such that Sii = 1 for all i $\in$ [n]. In terms of probability models on partitions, a SM can be motivated as the expected value of a (random) PM. From this point of view, positive semi-definiteness of the SM should be a consequence of the analog property of PMs, but in general, this is not always the case. To motivate the requirement we postulate that SMs should be coherent in the sense that the information they provide about the probabilistic clustering structure for [n] does indeed respect the transitivity property of the underlying partitions [3].

For the research on defect classification model, defect description language needs to be analyzed to determine similar descriptions of the defect then classify defect and form an effective defect library.

When comparing a pair of defects, a high similarity is desirable. Consider a simple experiment where a pair of defects have a high similarity. We then compare another pair of defects and find a lower similarity value. One can safely assume that the first pair of defects is more similar to each other than the second pair. So, we can define the first pair of defect the same classification [4].

Generally speaking, the classification and pattern recognition problems, as well as various methods and algorithms for their solution are well presented and discussed in the literature [5-11]. However in the most often cases the problem is viewed as an off-line classification of preliminary given data set (patterns). Then the typical task is to classify every single pattern from the given data set as belonging to one or another class. This is actually the case of supervised classification [5-12].

The research about defect similarity except recognize general words and expressions, involves lexical analysis in professional field. Therefore, we need to improve algorithm on the basis of common words similarity.

Identification of general words needs to calculate primary similarity of a sentence. For sentence similarity computing, the general steps are as follows:

Step 1 Segmentation management to statement

Step 2 Phrase management to statement, calculate similarity about semantic

Step 3 Import syntactic rule, analysis sentence structure

Step 4 Calculate similarities about the semantic of statement

Step 5 For a field composed by a complex sentence, calculating complex sentences or sentence group semantic similarity

For the segmentation management, there are many sophisticate systems such as segmentation system of information retrieval lab. For the management of the phrase should focus on the management of structural analysis. Then calculate similarity of the phrase semantic.

General statements such as general Chinese area of information management, there are many methods of comparing text similarity. The ideas worth learning. Its analysis and introducing word recognition in defects professional area will make us to find a suitable defect similarity comparison method. So that helps to classify the problem. The followings are some similarity methods.

Boolean model is used to calculate the similarity of statement with fast speed and high efficiency. But its calculated result is only two values, either the same or different. For some semantically similar words, Boolean model may give wrong judgment [13].

TF*IDF used in the field of information retrieval, the method is a statistical method, only the number of words in the sentence that contains a lot of relate words may be repeated, the effects of this statistical method can be reflected. TF*IDF methods only consider the words statistical properties in the context, without considering the semantic information of the word itself. It also has some limitations [14].

Vector space model is expressed by sequences statements and sentence similarity is calculated by calculating the similarity of lexical between sequences, which is based on their semantic similarity model. It is a good response to the sentence semantic information [15-17].

In this paper, defect similarity computing model based on vector space model was proposed.

3. Related Association Rules

Association rule mining, as an important data mining method, can effectively discover relationships between properties derived from the spectra of large celestial bodies and has great potential in studying the origin and evolution of the universe.

Through a large number of test data, defect distribution can be summed up. We applied association rules of data mining, and studied the relative defects, which can analyze efficiently and prevent product defects, then improve product quality.

Association rules reflect the relationship among data or is a study whether the generation of a data can speculate the generation of another data. Data association reflect the relationship in Database. The association degree can be expressed through support and confidence. Thus, finding the relationship among data in the transaction database is the mining purposes of association rule [8-10].

Association rules in a transaction database are defined as follows [18-22].

Definition 1: let $I = \{i_1, i_2, \cdots, i_n\}$ be a collection of items, the transaction database $D = \{t_1, t_2, \cdots, t_n\}$ is a series of transactions with a unique identifier TID. Each transaction corresponds to a subset of I. The collection of items called itemsets.

Definition 2: Suppose $I_1 \subseteq I$, the support of item set I_1 in the data set D is the percentage of I from transactions that contain I_1, that means, $\text{support}(I_1) = \|\{t \in D | I \subseteq t\}\| / \|D\|$

Definition 3: For item set I and transaction database D, all meet user-specified Minsupport item sets in T, which is greater than or equal to Minsupport nonempty subset of I, is the frequent item sets or large item sets. Frequent item sets that do not include other elements from frequent item set is called maximum frequent item sets or maximum item set.

Definition 4: In I and D, the definition of association rules, such as $I_1 \Rightarrow I_2$ can be presented as credibility or confidence. The so-called of rules' confidence is the ratio of number of transactions that include I_1, I_2 and the number of transactions that include I_1 .That is, $\text{confidence}(I_1 \Rightarrow I_2) = \text{support}(I_1 \cup I_2) / \text{support}(I_1)$ .Where $I_1, I_2 \subseteq I, I_1 \cap I_2 = \phi$.

Definition 5: Strong association rules is association rules with D on I which meet minimum support and minimum confidence. Commonly association rules generally refers to the strong association rules as defined above.

4. Defect Classification and Association Model

Now, we study a new defect model, which is derived from the defect analysis and summary. This defect model plays an important role in the discovery of more defects. The defect model proposes efficient improvement to the development and testing process, and improve the product quality much better. Defect model is concerned with applying the similarity theory and the idea of the Knowledge Discovery in Database. It is the result of summing up defect distribution, using association rules and analyzing relationship, which can be used as the theoretical basis for the development improvement [23-24].

4.1. Defect Classification Model Definition

Defect similarity calculation model is based on vector space model shown in Figure 1.

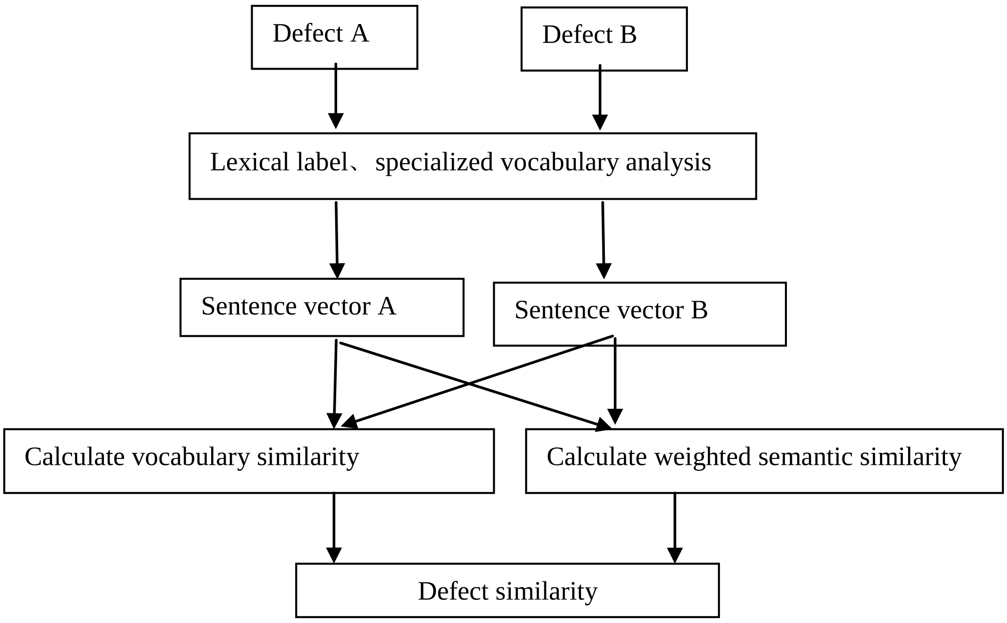
Defect A
Defect B
Lexical label、 specialized vocabulary analysis
Sentence vector A
Sentence vector B
Calculate vocabulary similarity
Calculate weighted semantic similarity
Defect similarity

Figure 1. Defect Similarity Calculation Model Based on the Vector Space

Firstly, the similarity between defect A and defect B is defined as a value in [0-1], where 0 represents no similarity, 1 represents completely similar, the larger value indicates the more similar they are.

A sentence that states defect issue can be expressed as a vector $T =< T_1, T_2, \ldots, T_n >$,among which T_i represents a single word after separate words management in defect sentences, these words are mainly nouns, verbs, adjectives and numerals, the sentence’s semantic information is described mainly by the part of speech of these types of word.

The main idea of word-based similarity model is: First calculate the words’ semantic similarity in the statements, and then through semantic similarity calculate defect statement similarity, so that the rich semantic information can be taken into account. The similarity between sentence T and sentence $T^{'}$ can be obtained by their similarity matrix $M_{TT^{'}}$, as shown in Eq. (1).

$$M_{TT'} = \begin{bmatrix} s(x_1, y_1) & s(x_1, y_2) & \cdots & s(x_1, y_m) \\ s(x_2, y_1) & s(x_2, y_2) & \cdots & s(x_2, y_m) \\ \vdots & \vdots & & \vdots \\ s(x_n, y_1) & s(x_n, y_2) & \cdots & s(x_n, y_m) \end{bmatrix} \tag{1}$$

Where, $s(x_i, y_i)$ represents the semantic similarity of word x_i and word y_i, each row of the matrix represents the semantic similarity of a word in sentence T and each word in sentence T' .Suppose $w_1, w_2, \cdots, w_n$ respectively denotes the weights of $x_1, x_2 \cdots, x_n$ in sentence T, take the maximum value of each row or each column in the matrix, that is seeking the maximum semantic similarity of a word in sentence T and each word in sentence T'. Will matrix $M_{TT'}$ compress to one-dimensional, and then weighted averages to these maximum, as shown in Eq. (2). So we can get the weighted semantic similarity $X_{TT'}$ between sentence T and sentence T' .

$$X_{TT'} = \frac{\sum_{i=1}^{n}(w_i(\max(s(x_i, y_1), s(x_i, y_2), \cdots s(x_i, y_m))))}{\sum_{i=1}^{n} w_i} \tag{2}$$

According to the given similarity, we can determine similar types of defects, thereby to classify the defects.

4.2. Defect Association Model

Relative analysis informs that a happening of one defect may incur another or many other defects occur. For example, the problem of 'interface displaying' is associated with 'interface indication problems',' consistency problems' and 'boundary problems'.

In order to confirm the defect association model and quantify the defect classes, the test object's version is defined as transaction set T and the defect category set is defined as set I at first. Then the association defect each defect model corresponds is determined. After the transaction database D is generated and the minimum support degree and minimum confidence threshold value are given, by applying the Apriori algorithm the frequent item set and the association rule will be obtained at last.

We focus on the strong association rules which set D meets the rules for minimum support degree and minimum confidence degree of set I.

Association Rule Mining is searching for a process satisfying the minimum support degree and minimum confidence degree. The support degree and confidence degree are the ones that are useful and meaningful.

This paper will quantify different versions of each test object, and finally get the transaction database D. Then define two thresholds as follows: Minimum support degree (min_sup) and Minimum confidence coefficient (min_conf).

Confidence degree measures the rule's intensity while the support degree measures the turn up frequency of the rule. A larger confidence degree and a smaller support degree can be applied to typical cases.

Thus, steps to determine association model are as follows:

Step 1 Determine the association rule X in defect classification database

Step 2 Ensure a condition that X.support>=min_sup

Step 3 Ensure a condition that X.confidence>=mini_conf

At last obtain the defect set's strong relation set which is the defect association model we want.

5. Experimental Results

In this section, we will show an experiment and its results. This experiment applied this model discussed above, based on a typical GUI's test data, concluded a GUI defect model [17].

5.1. Definition of Defect Classifying

This step mainly focuses on settling and analyzing the data source. These data, which is the defect set, are the total defects by using several methods such as testing. By fully analyzing the description of these defects, extracting the useful information, classifying and building up the defect model species, calculating the similarity of each two defects according to the Eq. (1) and Eq. (2). First define the minimum similarity $\mathrm{Sim}_{\mathrm{Min}}$ and combine two defects which similarity is larger than $\mathrm{Sim}_{\mathrm{Min}}$,then get a minimum defect classification set , finally, construct a preliminary frame of model [25].

This piece matched the defect similarities according to the defect data of GUI test result. This defect set is $D = \{d_1, d_2, \cdots, d_n\}$,where n=50. By applying Eq. (1) and Eq. (2), the similarity is calculated. we defined $\mathrm{Sim}_{\mathrm{Min}}$ as 80% in this model. Then the defects, whose similarity above 80% is classified to one set, formed a minimum defect classification set. After classified the problem types and problem distribution, calculated the similarity degree, a GUI defect classification table is showed as Table 1.

Table 1. GUI Defect Classification

Term	Defect classification
1	Interface display problem
2	Interface indication problem
3	Wrong character problem
4	Punctuation format problem
5	Inconformity semantic expression
6	Unreadable codes
7	Messy versions
8	Inconformity pictures
9	Inappropriate learning problem
10	Sequencing problem
11	Boundary problem

5.2. Defect Association Model

The process to determine the defect association model is a process of data mining. Set the product version as transaction set T, quantized defect as item set I and construct a transaction database D. After the threshold values of minimum support and minimum confidence support are given, frequent item sets and association rule will be generated by applying Apriori arithmetic. We focus on the strong association rule which is the rule satisfying the minimum support and minimum confidence coefficient D on I.

This paper quantized 26 versions of the test object and obtained transaction database D. Two threshold values are defined as follows according to previous experiences.

- Minimum support degree (min_sup) = 38%
- Minimum confidence coefficient degree (min_conf) = 68%

37 rules are obtained by using Apriori arithmetic. In order to facilitate the analysis results, the digit results were turned to be specific problems. Then based on the constraints such as a strong association rule should include a minimum support degree which is 38% and a minimum confidence degree which is 68%, decide whether the output association rule is strong association rule, result is showed in Figure 2 [26].

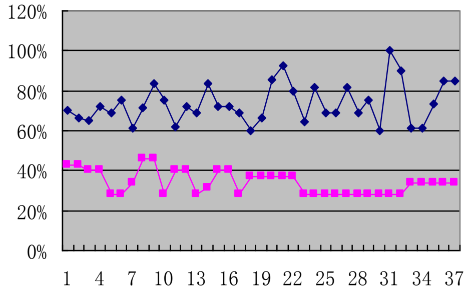

Figure 2. Result on Association Rule

The horizontal axis represents the quantized rule while the vertical axis represents confidence degree or support degree. In Figure 2, 22 strong association rules are obtained by calculating the output satisfying min_sup and min_conf. For the association rule of ’interface display problem’ =>’interface indication problem’, it’s confidence is 70%, the support degree is 43%, which indicates the probability that both ‘interface indication problem’ and ‘interface indication problem’ occurs is 70%, while the probability is 43% that ‘interface indication problem’ may occur when ‘interface display problem’ occurs. Thus we mainly focus on these associations.

The confidence degree of the 31st association rule in Figure 2 is 100% and the support degree does not satisfy the minimum support degree, hence it is not a strong association rule. But when precondition occurs, the consequence rule always appears. It’s appropriate to focus more on consequence rule when precondition occurs. Despite it is not a strong association rule and the support degree doesn’t reach a defined minimum support degree, but a high confidence coefficient degree can also reflect an important association problem.

According to the research on frequent set and strong association rule above, defect types and descriptions for these types are concluded into a corresponding table which forms a defect association model, as in Table 2, which list the defect association with user’s interface. Eligible defects and its associations will be listed in the model and defect distribution and defect association can be known from this.

Table 2. GUI Defect Association Model

Defect types	Defect associations
Interface display problem	Interface indication problem, consistency problem, boundary problem
Interface indication problem	Interface display problem, consistency problem, boundary problem, messy codes
Messy Versions	Interface display problem

Graphical inconsistency	Interface indication problem, consistency problem
Learnability problem	Interface indication problem, consistency problem, boundary problem
Boundary problem	Interface display problem, Interface indication problem, Consistency problem
Consistency problem	Interface display problem, Interface indication problem, Boundary problem, Graphical inconsistency

Association analysis indicates that a problem may lead to one or the other problems to appear. Such as 'interface display problem' associates with 'interface indication problem', 'consistency problem',' boundary problems' etc.

6. Conclusions

This study based on vector space similarity calculation model and the association rules technic for data mining, by researching large amount of defect data, proposes a defect classifying method and a defect association model. In this way, not only the defect analyzing efficiency is improved but also based on defect association model, pertinent suggestions for software testing and designing will be presented. Further study will continuously work on specializing the similarity of terminologies that describe defects and keep on researching other defect associations as well.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61170273).

References

[1] Glenford J. M., Tom B., Todd M. T. and Corey S., "The Art of Software Testing," **(2004)**.

[2] Bender, "Requirements Based Testing Process Overview," **(2009)**.

[3] Carlos A. N. and Fernando A. Q., "Similarity analysis in Bayesian random partition models," Computational Statistics and Data Analysis, vol. 55, **(2011)**, pp. 97-109.

[4] Brian S. M. and Spiros M., "Comparing the Decompositions Produced by Software Clustering Algorithms using Similarity Measurements."

[5] J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," New York: Plenum Press, **(1981)**.

[6] A. Jain, "Data Clustering: 50 years beyond K-means", Pattern Recognition Letters, vol. 31, **(2010)**, pp. 651-666.

[7] S. Ozawa, S. Pang and N. Kasabov, "An Incremental Principal Component Analysis for Chunk Data", Proc. of the 2006 IEEE Int. Conference on Fuzzy Systems, FUZZ-IEEE 2006, Vancouver, July **(2006)**, pp. 10493-10500.

[8] X. Zhou and P. Angelov, "Real-Time Joint Landmark Recognition and Classifier Generation by an Evolving Fuzzy System," Proc. of the 2006 IEEE Int. Conference on Fuzzy Systems, FUZZ-IEEE 2006, Vancouver, July, **(2006)**, pp. 6314-6321.

[9] S. Soltic, S. Wysocki and N. Kasabov, "Evolving Spiking Neural Networks for Taste Recognition," Proc. of the 2008 IEEE Int. Conference on Fuzzy Systems, FUZZ-IEEE 2008, Hong Kong, June **(2008)**, pp. 2092-2098.

[10] S. Pang and N. Kasabov, "r-SVMT: Discovering the Knowledge of Association Rule over SVM Classification Trees," Proc. Of the 2008 IEEE Int. Conference on Fuzzy Systems, FUZZ-IEEE 2008, Hong Kong, June **(2008)**, pp. 2487-2494.

[11] M. Svensson and S. Byttner, "Self-organizing maps for automatic fault detection in a vehicle cooling system," Proc. of the 4th Int. IEEE Conference on Intelligent Systems, IS 2008, Varna, Bulgaria, September, **(2008)**, pp. 24-8 – 24-12.

[12] Stefan B., Magnus S. and Gancho V., "Incremental Classification of Process Data for Anomaly Detection Based on Similarity Analysis, IEEE, **(2011)**.

[13] Yi G., "Quantifying semantic similarity of Chinese words from HowNet," In 2002 International Conference on Machine Learning and Cybernetics, **(2002)**, pp. 234-239.

[14] Zheng T. Y. and Lei H., "Similarity Computation of Chinese Question Based on Chunk," In 2006 International Conference on Machine Learning and Cybernetics, **(2006)**, pp. 17–22.

[15] Gan K. W. and Wong P. W., “Annotation information structures in Chinese texts using How net,” In Second Chinese Language Processing Workshop Hong Kong, **(2000)**, pp. 85-92.

[16] Sergei N., Constantine D., Dean J. G., “Two approaches to Matching in Example-based Machine Translation,” In Proceedings of the fifth International Conference on Theoretical and Methodological in Machine Translation of Natural Languages, **(1993)**, pp. 45-57.

[17] Wanjiang H., “Study on the defect Classification model,” Applied Mechanics and Materials vol. 475-476 513-517, **(2014)**, pp. 4008-4011.

[18] Long H. and Lee H. Y. A., “New Visualization Technique for Knowledge Discovery in OLAP,” Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining, **(1997)**.

[19] Jaiwei H. and Michelinne K., “Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Second Edition, **(2006)**.

[20] Margaret H. D., “Data Mining: Introductory and Advanced Topics,” Pearson Education, **(2007)**.

[21] Dion H. G. and Rebecca P. A., “An Introduction to Association rule mining: An application in counseling and help-seeking behavior of adolescents,” Behavior Research Methods, vol. 39 no. 2, **(2007)**, pp. 259-266.

[22] Li C. and Fan M., “Generating association rules based on threaded frequent pattern tree,” Computer Eng. Appl. (in Chinese), vol. 4, **(2004)**, pp. 188-192.

[23] Jiang H. C., “Stellar spectra association rule mining method based on the weighted frequent pattern tree,” Research in Astron. Astrophys, vol. 13 no. 3, **(2013)**, pp. 334–342.

[24] Jiang Y., Li M. and Zhou Z. H., “Software defect detection with Rocus,” JOURNAL OF Computer Science and Technology, vol. 26 no. 2, **(2011)**, pp. 328-342.

[25] Zhao Q., Zheng J. and Li J., “Software Reliability Modeling with Testing-Effort Function and Imperfect Debugging,” TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 10 no. 8, **(2012)**, pp. 1992-1998.

[26] Wanjiang H., Tianbo L. and Sun Y., “Research on the Problem Model of GUI based on Knowledge Discovery in Database,” Proceedings of the 2013 International Conference on Software Engineering and Computer Science, **(2013)**, pp. 5-9.

Authors

Wan-Jiang Han, was born in Hei Long Jiang province, China, 1967. She received her Bachelor Degree in Computer Science from Hei Long Jiang University in 1989 and her Master Degree in Automation from Harbin Institute of Technology in 1992.

She is an assistant professor in School Of Software Engineering, Beijing University of Posts and Telecommunication, China. Her technical interests include software project management and software process improvement.

Tian-Bo Lu, was born in Guizhou Province, China, 1977. He received his Master Degree in computer science from Wuhan University in 2003 and his PH.D Degree in computer science from the Institute of Computing Technology of the Chinese Academy of Sciences in 2006.

He is an Associate professor in School of Software Engineering, Beijing University of Posts and Telecommunications, China. His technical interests include information and network security, trusted software and P2P computing.

Xiao-Yan Zhang was born in Shandong Province, China, 1973. She received her Master Degree in Computer Application in 1997 and her PH.D Degree in Communication and information system from Beijing University of Posts and Telecommunication, China, in 2011.

She is an Associate professor in School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing, China.

Her technical interests include software cost estimation and software process improvement.