Research on Decision Tree in Component Retrieval

Yanhua Shao, Mingsheng Zhang
College of Economics & Management
Guizhou University for Nationalities
Guiyang, China

Shengnan Xu
Center for Computer Information
Chengde Teachers' College for Nationalities
Chengde, China

***Abstract*—to retrieve and extract the most satisfying among the library of components is important in component library management system. The general component retrieval system seldom provides information about respect of reused actually. Data mining technology provides a feasible approach to above problem. In the paper, how to use the application of classification method decision-tree-based to the component reuse was discussed. In accordance with the limitation of research on traditional software component library management, we proposed the idea to apply data mining technology to the management of software component, provided auxiliary decision support to the relevant personnel of the component library. Secondly, in accordance with actual application, built an applied model of software component retrieval management by data mining technology, and analyzed the execution step of the applied model. Lastly, the model had been verified through experiment, thus the feasibility and validity of this strategy had been verified.**

Keywords-software reuse; component library; component retrieval; data mining; decision tree

I. INTRODUCTION

Software component technology is the key technology of supporting the software reuse [1, 2]. In the process of the software development based on software component, component library plays a key role. To retrieve and extract the most satisfying component among the library of components is important in component library management system. The retrieval efficiency of the component library not only includes the retrieval of the component itself, but also includes the understanding of the component. The general component retrieval system provides descriptive information about understanding the component simply, but seldom provides information about respect of reused actually. Data mining technology provides a feasible approach to above problem [3].

In the paper, how to use the application of classification method decision-tree-based to the component reuse was discussed. In accordance with the limitation of research on traditional software component library management, we proposed the idea to apply data mining technology to the management of software component, provided auxiliary decision support to the relevant personnel of the component library. Secondly, in accordance with actual application, built an applied model of software component retrieval management by data mining technology, and analyzed the execution step of the applied model. Lastly, the model had been verified through experiment, thus the feasibility and validity of this strategy had been verified.

II. LIMITATION OF COMPONENT RETRIEVAL

Component Library is an important basis for software reusing, software reuse success depends largely on the structure of component library, composition, management, etc [4,5,6]. Studies have shown: retrieve and select the components to meet the user requirement specification is the core problem of component library constructed. With the increase of components, component library continues to expand, which has brought many inconveniences for component library management. Therefore, to retrieve the components meeting client requirements has become increasingly difficult. in large software component library. In addition to a similar problem with the information systems, the component library, there are many different from the information systems. Some difficulties when users to retrieve and select the components will encounter in the following:

- In the retrieval process, there may be multiple components to meet the users search terms, how to judge accurately and select the required components from a large number of candidate components quickly and efficiently, which is a complex decision-making process, is generally dependent on the user's reuse experience and understanding of the components and subjective evaluation.
- Usually the user carry out component retrieval by expressing the features of components through the, attributes, keywords, expression of the relationship. However, the users accessing component library have different levels, which may not be familiar with the components carved face patterns. Therefore, it is difficult to understand the components.
- Perhaps the user does not have a clear objective before query, only want to retrieve the software component library to see whether there are components to can use and reuse for him or not. Therefore, it is very necessary to provide users with a certain degree of decision-making through the reuse history and the reuse experience of other users.
- The components reuse on the requirements specification, design, pattern, test plans belonging to an

978-1-4244-5934-6/10/$26.00 ©2010 IEEE

indirect reuse, the users need to analyze and understand the component at first. Moreover, in most cases, the reuse of the components is white-box reuse, i.e., component needs to be adapted according to the feedback of the components re-users. How to track the experience of component reuse and reuse of history, supporting the user select the appropriate analysis and design parts, as well as changes to the smallest component is critical.

- The components are often misused because of not understanding the functions and the lack of documentation, even the relevant components are provided by components library.

Can be seen from the above problems, component reuse historical information and the reuse feedback information is very important: On the one hand, it will not only help those who submitting component carry out component by re-engineering to improve the original components, but also develop new components by deepening the components; On the other hand, the user of component can enhance the understanding of the components and reduce the workload of re-use through feedback information from the components user. In the paper, data mining sets were established from the point of the data of component reuse history and the reuse feedback information view by attempting to use data mining technology. A number of component reuse laws were excavated by using the appropriate method of data mining, exclusion of human factors. By using this method, the reuse based on impression and experience becomes quantitative analysis of science, reuse based on isolated and scattered becomes the judgments of regularity and trends based on the scientific computation analysis.

III. THE APPLICATION OF DECISION TREE TO COMPONENT RETRIEVAL

The mining data sets were established based on the tables of the reuse historical information sheet and feedback information of components by using data mining technique. A number of component reuse rules were excavated from the tables, which provide a supplementary decision support for the relevant personnel of component library.

A. Decision Tree Method

The decision tree classification is inductive learning algorithm based on examples [7]. The classification rules denoting by decision tree were extracted from a group of instances of no order, no rules. The decision tree classification adopts recursive method, in the tree's internal nodes to compare the properties and to determine the branch based on different node property values, to reach a conclusion in the tree's leaf nodes. Therefore, a path from the root to a leaf node corresponds to a conjunctive rule. The decision tree corresponds to a group of the disjunctive expression rules [8, 9].

The decision tree classification is composed of two steps [10, 11]. One step is to build a decision tree model, which is the process to get knowledge from data and learn. The other step is to classify the unknown data samples through using the decision tree model built by the first step. In accordance with the unknown data samples, the attribute value is tested in turn from the root node to certain leaf node, thereby the classification which the data samples belonged to is found.

B. Model of Component Retrieval Based on Decision Tree

The feedback information of reuse is submitted to component feedback information table of the component library, which is verified. At the same time, the reuse history information is also conserved in the component library. The data contain a large number of latency useful knowledge which is hidden and unknown in advance. The mining data set is established based on the reuse historical information sheets and the feedback information table of the component library by using data mining technique. The component reuse rules are excavated from the mining data set, which is presented to users in the form of visualizing. The component user was able to understand the components from other component user point of view, therefore, to understand the practical application information of the components concerned in. In addition, the component library manager can also manage the components by using the reuse regulations of components, the components for the need to improve, pay it back to the component producers in order to develop higher-quality components. It is necessary to consider the appropriate strategy in the component library classification to increase its weight for a component to be reused excellently in the actual, so that it has a higher hit rate in the retrieval.

The model of component library management based on decision tree is presented in Fig. 1.

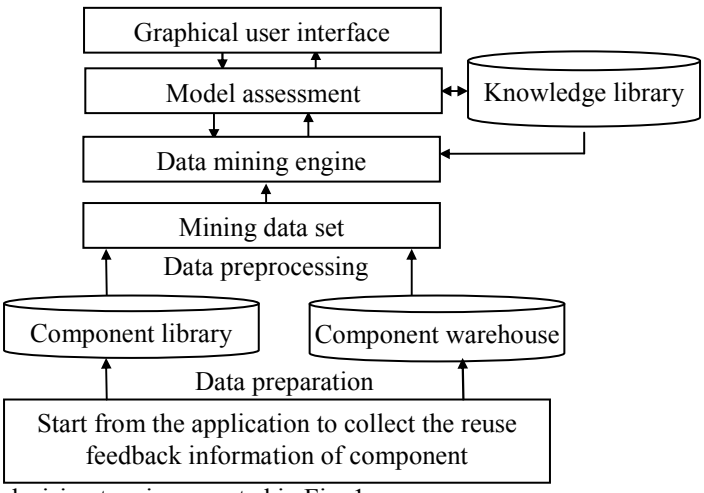

Figure 1. model of component library management based on decision tree

C. Working Process of the Model

1) Data preparation

Data preparation is to collect data and establish a rational database schema for data entry on the basis of data mining needs. In the paper, the original data comes mainly from the table of component reuse and reuse historical information as required by the knowledge of data mining. Component reuse historical information exists in the component reuse historical information table of the component library, which includes the

Stadholder Fundation of Guizhou Province(Grant No.2009-111);
Science and Technology Funds of Guizhou Province(Grant No.[2009]2125).

description information of component and the reuse number of components, etc. The reuse feedback information of component was stored in the feedback information form of component, which includes the main areas of application, reuse methods, reuse environment, the evaluation of the component.

2) Data preprocessing

Data pre-processing is an important step in the data mining process, the mining data sets is got through data preprocessing. In the paper, in accordance with the knowledge excavated for the data mining, the original data table is established by extracting relevant data from the component reuse historical information table and the reuse feedback form in the component library, and then the mining data sets are established by dealing with the original data table.

3) Data mining

The data mining process is the basic components of data mining system. The data mining process consists of a set of functional modules, which should include data mining, knowledge query, rules expression, rule storage in general. The mining task of this paper is with a clear classification. Among the classification algorithm, decision tree algorithm has the characteristic of least impacting by the record, the model easy to understand, easy to train, easy to implement, and it is a very good algorithm to deal with text data. Therefore, decision-tree algorithm is selected to mine the reuse rules of component in the paper. At the same time, due to the continuity of the data, the see 5 algorithm is used to implement the component reuse rule mining in this paper.

4) Model assessment

The validation and evaluation of the results is indispensable in the data mining process. Which is an iterative experimentation process, the other sample of component library can be used to verify, you can also select a new sample set to verify until the user has the satisfaction with the results. The holdout and cross-validation are two commonly technical methods of assessment of classification predictive accuracy. Usually holdout evaluation methods commonly used in the initial pilot of the occasion, so holdout evaluation method is adopt by this paper to carry out model assessment.

IV. APPLICATION EXAMPLE

A simulation component library is established in the paper, and a set of experimental data are given to simulate reusable component library platform. The properties of each component including component number, version number, type, creation tool, type, function, applications, reuse frequency, reuse methods, reuse environment are extracted from the component reuse historical information table and feedback form of the component library. In view of this article is only the initial pilot study, the classification by the task of excavation, the "valuation" will be considered as the label attribute, and the label attribute will be classified as poor, in general, good, excellent. The link between reuse attributes and tag attributes were set up by see 5 algorithms [12], which prove the feasibility and effectiveness of the strategy. Mining data set and then converted into the data format suited see 5 algorithm, the data set is composed of 258 samples. Each sample contains 11 attributes, as shown in Fig. 2.

```
3348, 1.0, Windows, Delphi,VCL/CLX,other,
-,t,627, MARKET, super
3428,3.1,Windows,VC++,other,code,Message
Queueing,t,200,Communication,super
3443,1.0, Windows, PB,VCL/CLX,code,
Print and Preview, t, 103, GOVERMENT, good
        ……          ……
3345,1.0, -,VC++,Class,code,-,t,103,FINANCE,good
3422,3.0,Windows,VC++,COM/DCOM,code,Product
Suites,f,12,other,general
3346,1.0,Windows,Delphi,VCL/CLX,code,Print and
Preview,t,18,GOVERMENT,general
        ……          ……
```

Figure 2. mining data set suit to see 5

Applying see 5 algorithm to mining data set, training data were set up by extracting randomly 67% of the mining data set (the rest of the data is as a test data set). Reuse rules are derived from the training data. The results are as follows:

```
See5 [Release 2.01]     Mon Jan 25 19:13:18 2010
  Options:  Rule-based classifiers

        Use 67% of data for training

        Class specified by attribute `valuation'

Read 173 cases (11 attributes) from component.data

Rules:

Rule 1: (47/1, lift 3.5)
        Reuse Num <= 8
        -> class less  [0.959]

Rule 2: (21/1, lift 6.9)
        Reuse Num > 8
        Reuse Num <= 18
        -> class general  [0.913]

Rule 3: (105/23, lift 1.6)
        Reuse Num > 18
        -> class good  [0.776]

Rule 4: (16, lift 7.8)
        level = code
        Reuse Num > 120
        -> class super  [0.944]

Rule 5: (8, lift 7.4)
        Reuse Num > 269
        -> class super  [0.900]
```

Default class: good

Evaluation on training data (173 cases):

```
        Rules
----------------
 No     Errors
  5    7( 4.0%)   <<

 (a)   (b)   (c)   (d)    <-classified as
----  ----  ----  ----
 46     1           (a): class less
       20     3     (b): class general
             82     (c): class good
  1          2    18   (d): class super
```

Evaluation on test data (85 cases):

```
        Rules
----------------
 No     Errors
  5    6( 7.1%)   <<
 (a)   (b)   (c)   (d)    <-classified as
----  ----  ----  ----
 14                (a): class less
       7     3     (b): class general
       1    45     (c): class good
             2    13   (d): class super
```

Experimental results show that: before the use of see 5 algorithms to carry out excavation, we choose to generate rule sets, use discrete attributes tests to put out the branches of the decision tree, select 67% of the samples for training, the use of rough threshold value. Data were classified into different types through the 'valuation' properties.

When data mining, read 173 samples for training, resulted in a decision tree and a set of rules set (including 5 rules). The following typified the rules.

Rule 1: If (Reuse Num <=8) Then (valuation = less);

Rule 2: If (8< Reuse Num <=18 Then (valuation = general);

Rule 3: If (Reuse Num >18) Then (valuation =good);

Rule 4: If (Reuse Num >=120 and level=code

Then (valuation =super).

Rule 5: If (Reuse Num >269 Then (valuation =super).

From the excavating rules can be seen, see 5 considers reuse number, component creating tool and component-level, functions and other features as major factor to determine reuse good or bad, which is component reuse, managers, producers are concerned about is the same. Thus the feasibility and validity of this strategy had been verified.

V. CONCLUSION

Component Library system as the core of software reuse systems, understanding of the retrieval component is the key. In this paper, the method that how to introduce the decision tree classification into the component library management was discussed, and the feasibility and effectiveness of such method were validated, in order to provide technical support to reuse success. Therefore, the efficiency of software reuse of was enhanced. However, there are still a lot to be desired in the paper: the prediction accuracy of the component reuse by using data mining algorithm is not ideal. In addition, the reuse feedback information is generally semi-structured text. However, data mining algorithms are mostly carried out in the structured data, and how converted the semi-structured text to the structure data is to be further addressed.

ACKNOWLEDGMENT

This paper is supported by College of Economics and Management of the Guizhou University for Nationalities. At the same time, I would like to express my gratitude to my family.

REFERENCES

[1] LI Xiao-li; LIU Chao; JIN Mao-zhong; GAO Zhong-yi. Software Component Reusability Quality Metrics. Application Research of Computers, Vol. 24, No. 6, 2007. pp. 280-283

[2] Fuqing Yang, "Software Reuse and Its Correlated Techniques," in Computer Science, Vol. 26, No. 5. 1999, pp. 1-4

[3] Zhu Ming, Data Mining. He Fei: University of Science and Technology of China Press, 2002, pp. 1-9

[4] LI Zhao-hui, MAO Hai-jun. Frame representation and retrieval of reusable software component, Application Research of Computers, Vol. 39, No. 2. 2008, pp. 129-131

[5] YAO Quan-zhu; MENG Li; CUI Du-wu. Component retrieval method based on case-based reasoning and XML. Journal of Computer Applications, Vol. 27, No. 7. 2007, pp. 1711-1714

[6] SHENG Yi-fang ZHANG Wei-shi ZHANG Xiu-guo SHI Jin-yu. Research on Transformation Mechanism of Component Retrieval Conditions in Multi-Library, Computer Engineering and Applications, No. 25, 2006. pp. 23-26

[7] Hu Xiaohua, Nick Cercone. Discovery of Decision Rules in Relational Databases: A Rough Set Approach. Proceedings of the third international conference on Information and knowledge management. CIKM 1994: 392-400

[8] Jiawei Han, Micheline Kamber. Data Mining: concept and techniques, Morgan Kaufmann, 2001

[9] Zhong-zhi Shi. Knowledge Discovery. Beijing: Tsinghua University Press, 2002

[10] Quinlan J. R. C4.5: Programs for Machine Learning [M]. San Francisco: Morgan Kauffman. 1993

[11] Theodoros Evgeniou, Massimiliano Pontil, Tomaso Poggio. Statistical Learning Theory: A Primer, International Journal of Computer Vision. Vol. 38, No. 1. 2000

[12] J. R. Quinlan, Bagging, Boosting, and C4.5, In Proceedings AAAI96 Fourteenth National Conference on Artificial Intelligence, Portland, Vol. 1, 1996