International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.3, Issue No.12, pp : 757-760 01 Dec. 2014

Research on the Clustering Algorithm of Component based on the Grade Strategy

G.Vamshi Krishna*[1], Dr.P.Niranjan*[2] , Dr. P.Shireesha *[3]
Department of computer science, Engineering, Kakatiya Institute of technology and science, Warangal.

Abstract: The rapid development in the software component technology increases the number of components, reasonable component classification is the foundation to achieve effective retrieval. The former methods like faceted classification and full text retrieval methods and some traditional methods have always some subjective factors to achieve it. The user is not able to satisfy with these methods. So from the point of user requirements, grade strategy is introduced, which gives each facet different weight, and the similarity between the components. A component clustering algorithm based on the grade strategy is proposed. They prove that component clustering algorithm based on the grade strategy makes the component clustering result more humanized and better serve user requirements.

Keywords- **software reuse, component clustering, faceted classification, grade strategy.**

I. INTRODUCTION

The gradual development in the recent 20 years of object-oriented technology and software component technology, software reuse is regarded as an effective way to resolve software crisis and to improve the efficiency and quality of software production. The rapid development in software component technology as one of the key factors to achieve software reuse. And several mature software component-models have been appeared, such as COM, EJB and Web Services.

Managing the components effectively, necessarily need to establish component library. In the component library system, component classification and component retrieval are the Two problems. Component classification is the foundation and the key to retrieve component high efficiently and rapidly. Currently, there are many component classification methods. According to component representation, these classification methods have been divided into three sorts: artificial intelligence methods, hypertext browsing methods and information science methods by W.Frakes. And among information science methods, faceted classification method are used most widely this method has comparatively obviously subjective factors, because the definition of each facet and facet's value depends on expert experiences. As a result, it may prevent retrieving components from different component libraries, and can restrict the description of component facets' values and component retrieval conditions. A method of combining faceted classification with full-text retrieval is used to represent components. Based on this component representation, from the user's needs, grade strategy is introduced to obtain better similarity between components and then cluster components. By these, better clustering is achieved; and the classification result is more reasonable and closer to user requirements.

I. REPRESENTING COMPONETS AS FACETED CLASSIFICATION AND FULL TEXT RETRIEVAL COMBINDLY

Full-text retrieval method and faceted classification can avoid the subjective factors which are used the semantic analysis technology, and then extracts the keywords of text, and performs self-organizing clustering. When there are many types of text content, which are using semantic analysis technology being still in the phase of research, the efficiency of the method is very bad and the retrieval speed may be decreased. So the method uses combining faceted classification with full-text retrieval is used, which can lessen the subjectivity of faceted classification and can to some extent avoid the shortcomings of full-text retrieval. The component classification system is shown in Fig. 1.

As shown in Fig. 1, firstly while inputting components to the library, separate each component's description into corresponding facet's value according to faceted classification scheme; then extract keywords form each facet's value to establish component vector space; lastly cluster components to classify components by the use of full-text retrieval. The method of combining faceted classification with full-text retrieval has two advantages: on the one hand, it can make facet's value convert traditional controlled terms into text, so as to reduce the subjectivity of manually establishing and maintaining term space; on the other hand, through faceted classification the text content is more concentrative, which is beneficial to calculate similarity between component texts, and then promote the effect of component clustering. Here, from the view of facet's integrity and independence, five facets are defined for component faceted classification scheme.

1) Function: Component function description and its application area.

2) Operation: Object: Objects of input components.

3) Application Environment: Software and hard Ware environment.

4) Representation Method: Component pattern and representation.

5) Performance: component reliability evaluation. On the basis of this component representation, components can be clustered. Before clustering, the similarity between components needs obtaining. Here, we use a similarity calculation method based on grade strategy.

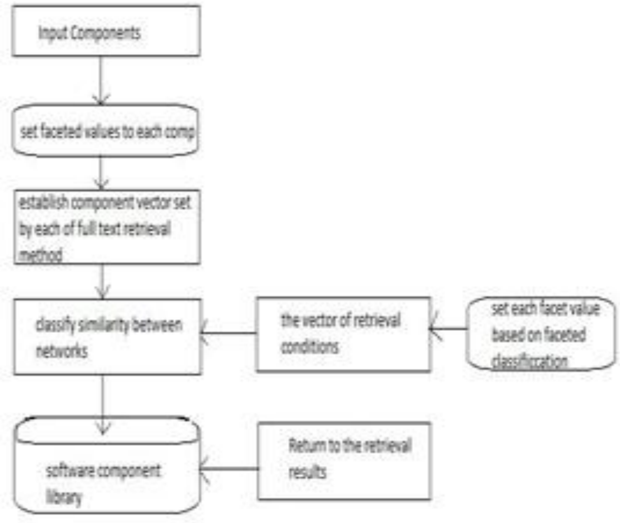
Input Components
set faceted values to each comp
establish component vector set by each of full text retrieval method
classify similarity between networks
the vector of retrieval conditions
set each facet value based on faceted classification
software component library
Return to the retrieval results

Figure 1.Component Classification System

III. COMPONENT SIMILARITY CALCULATION METHOD BASED ON GRADE STRATEGY

A. Similarity Calculation

The component clustering is usually based on Latent Semantic Analysis (LSA) [7] or Vector Space Model (VSM); and essentially, LSA is the extension of VSM [8]. The mode of component representation is the same for the two models. That is to say, a component set A= $(A_{1,}A_{2,}\ldots\ldots A_n)$ is represented by a mxn words-text matrix X= $[a_{ij}]$ (l<=i<=m 1<=j<=n). Here, m stands for the number of keywords in component set; n stands for the number of component s; a_{ij} stands for the weight of keyword in component set, which is greater or equal to 0. Then, the similarity between components is calculated by the distance between column vectors of matrix A. Usually, Cosine coefficient method is used, which is to calculate the cosine value between column vectors in matrix A. That is shown as formula 1.

$$\mathrm{Cos}(t_{ij)}=(p_iC_n)(\,p_iC_n)^T \;/\; \|\,p_iC_n\|_2\|(\,p_iC_n)^T\|_2$$

Here, p_i stands for i column of n-order unit matrix. C_n stands for column vectors of matrix A.

B. Component Similarity Calculation Based on Grade Strategy

Before released, every component will be-furnished with relevant description. According to these descriptions, one component representation method is needed to represent components. Then, these representation texts can mark corresponding components. Here, a method of combining faceted classification with full-text retrieval is adopted. And by means or calculating similarity between components, component clustering will be achieved and get the goal of classifying components.

While the traditional method to calculate similarity between components is that: firstly obtain similarity between components under each facet, and then get the total similarity between components by the way of calculating the average of similarity. In this method, we can find that every facet is treated equally. But users often have different attention point on each facet. For example, for most cases, users pay more attention to component function, while don't care about component application environment and representation method, et al. Thus, if we still use the traditional method, it will lead to lower the accuracy of similarity between components; furthermore affect the effect of component clustering. Therefore, from the view point of user and according to user requirements and attention, grade strategy is introduced to calculate the similarity between components, namely similarity calculation method based on grade strategy.

The basic thought about this method is that on the basis of component representation of combing faceted classification with full-text retrieval, each facet is given different grade,. in other words, each facet is given different grade weight; and then the total similarity between components is calculated by weighing synthesis account. The similarity calculation method is shown as formula 2:

$$\mathrm{Sim}(A_i,A_j)=\mathrm{Sigma}(P)\ a_p\,\mathrm{Sim}_p(A_i,A_j)$$

Here, (1<=P<=n), A_i,A_j — two components; a_p — the grade weight of the pth facet, and a_1,+a_2+…..+a_n,=1 0<a_p<1; Simp(A_i,A_j) — the similarity between components under the pth facet. n — the number of facets. Moreover, $\mathrm{Sim}_p(A_i,A_j)$ is calculated by the use of cosine coefficient method.

IV. COMPONENT CLUSTERING

After the similarity between components is obtained, components can be clustered. The improved algorithm of component clustering by introducing grade strategy is described as below:

Step1: Use the method of combining faceted classification and full-text retrieval together to represent components, and then each component corresponds to a text. Thus, a component text set is established as input data.

Step2: By the use of Chinese words segmentation technique and the method of extracting keywords, the component text set is represented as a words-text matrix.

Step3: Calculate each keyword's weight in each component text, and convert primary words-text matrix to a numerical matrix.

Step4: If component clustering is based on VSM, jump to step5; if component clustering is based on LSA, the primary numerical matrix should be substituted by lower-dimension matrix by SVD [9](10].

Step5: Calculate $\mathrm{Sim}_p(A_i,A_j)$.

Step6: Calculate the total similarity between components by the use of the component similarity method based on the grade strategy.

Step7: Cluster components and output the component clustering result. This algorithm flow chart is given as figure 2

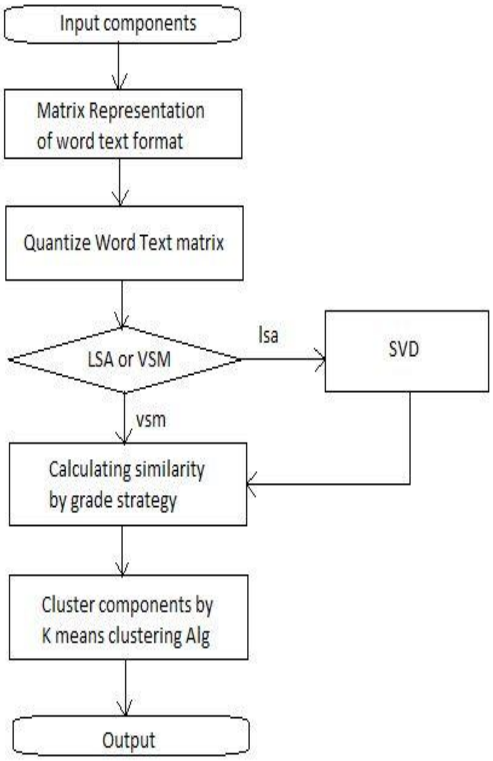
Input components
Matrix Representation of word text format
Quantize Word Text matrix
LSA or VSM
lsa
SVD
vsm
Calculating similarity by grade strategy
Cluster components by K means clustering Alg
Output

Figure 2. Algorithm Flow Chart

V. RESULTS AND ANALYSIS

In order to verify the effectiveness of the grade strategy, component clustering based on VSM and LSA with grade strategy, and traditional component clustering based on VSM and LSA are compared between each other by means of experiments. In our experiments, k-means clustering algorithm is adopted to cluster components, and TF-IDF [II] method is used to calculate weight of keywords of each component text, which is most widely used currently.

Our experimental component data are from Shanghai Component Library [12], which includes four topics: buttons, menus, tree and text processing. According to user general custom (often concentration on component function) and through many experiments, we get fairly reasonable grade weights for five facets, which are respectively 0.59, 0.13, 0.09. 0.07 and 0.12. Then, components can be clustered. Here, three criterions clustering precision, clustering recall ratio and F-Score coefficient are used to verily the validity of the grade strategy by means of comparing clustering effect.

1) Clustering precision: Reflect the degree of merging similar of same type and into dissimilar components into a same cluster, and reflects the distinguishing ability for different clusters. The clustering precision is higher, and then the components. in each cluster are more similar. The computing method is as formula 3:

$$P = Preision(i,j)= N_{ij} / N_{ij}$$

2) Clustering recall ratio: Reflect the degree of merging components of same type into a same cluster. The clustering recall ratio is higher, and then the similar components are more concentrative, and the less components spited into her clusters are. The computing method is as formula-4:

$$R= Re\ call(i, j) = N_{ij} / N_{ij}$$

3) F-Score coefficient: It is a more comprehensive evaluation for component clustering, which is integration of clustering precision and recall ratio. The computing method is as formula 5:

$$F – Score\ (i,j)= 2*P*R / P+R$$

Based on the three criterions, the figure of contrast effect for the four components clustering methods is shown as can be seen from the figure, clustering precision, clustering recall ratio and F-Score coefficient are increased by introducing the grade strategy. And we can obtain two important conclusions:

- The effect of component clustering based on LSA is better than based on VSM;
- The effect of component clustering with grade strategy is better than without grade.

We can found that among the four clustering methods, component cloistering based on LSA with grade strategy is the best. In a word, to some extent, the effect of component clustering with the grade strategy is improved, either based on VSM or LSA, so that components are classified more reasonably and user requirements are better satisfied.

IV. CONCLUSION

Bsed on the Faceted classification and Full Text retrieval methods the Grade strategy was proposed. By the use of this method, the similarity between components is calculated more accurately, the clustering effect is improved better either based on LSA os VSM, and the components are classified more reasonable. Furthermore, it may raise the efficiency and accuracy of component retrieval and promote software reuse. However, in this paper, we obtain the grade weight of each facet by means of experiments for many times. So how to get the grade weight more reasonable needs to be further researched.

REFERENCES

i. *Yang Fuging, Mci Hong. Design and Implementation of Component Based Software, Qinghua University Press, 2008.1.*

ii. *Frakes W, Gandel PI, Representing Reusable Software Information and Software Technology,1 990,32(1 0):654-664.*

iii. *Pan Y, Zhao J F, Xic B. The research and development of the technologies in component library. Computer Science.2003, 5:90-93.*

iv. *Prieto-Diaz R Implementing facetel classification for Software reuse. Cormnunications of the ACM, 1991, 34(5)'88-97.*

v. *Nilsson NJ. Introduction to Machine learning [EB/0L]. http://robotics.stanford edu/people/nilsson/mlbook.html, 1996.*

vi. *Liu Daxin, Zhao Lei, Wang Zhuo. A component classincation retrieve algorithm based on faceted classification and clustering analysis[J] Computer Application. 2004, 24: 89-90.*

vii. *Dumais S T. Using Latent Semantic Analysis to Improve Information Retrieval al[C]. In CHI'88 Proceedings.1988:281-285.*

viii. *Wang Chunhong, Zhang Mei. Analysis end study on the Latent semantic Analysis Index Retrieval Model [J], Compater Application, 2007, 27(5):1283-1285.*

ix. *FURNAS GW. Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure[A], Proceedings of SIGIR'88[C].1988.36-40.*

x. *Zhou Shuigeng, Guan Jihong, Ilu Yanfa. Latent Semantic Analysis and its application in Chinese text processing [J], Mini-,Micro System, 2001,22(2):239-2,13.*

xi. *Joachims T. A probabilitic analysis of the rocchio algorithm with TFJDF for text categorization. Proceedings of the 14th International Conference on Machine Lcaming(ICML-97),1 997: 143-151.*

xii. *Shanghai Component Library. [DB/OL], http://www.sstc.org.cn/.*