Clustering Dynamic Class Coupling Data to Measure Class Reusability Pattern

Anshu Parashar and Jitender Kumar Chhabra
Department of Computer Engineering,
National Institute of Technology, Kurukshetra,
Kurukshetra 136 119, India
conf.ppr@gmail.com

Abstract. Identification of reusable components during the process of software development is an essential activity. Data mining techniques can be applied for identifying set of software components having dependence amongst each other. In this paper an attempt has been made to identify the group of classes having dependence amongst each other existing in the same repository. We explore document clustering technique based on tf-idf weighing to cluster classes from vast collection of class coupling data for particular java project/program. For this purpose firstly dynamic analysis of java application is done using UML diagrams to collect class import coupling data. Then in second step, this coupling data of each class is treated as a document and represented using VSM (using TF and IDF). Then finally in the third step basic K-mean clustering technique is applied to find clusters of classes. Further each cluster is ranked for its goodness.

Keywords: Coupling, Data Mining, Software Reusability.

1 Introduction

Software Reuse is defined as the process of building or assembling software applications from previously developed software [16] The success of reusability is highly dependent on proper identification of whether a particular component is really reusable or not. Class coupling plays a vital role in measuring the reusability and selecting classes for reuse in combination because the highly coupled classes are required to be reused as a group [17] for ensuring the proper functioning of the application [8]. So for reuse, issues like maintaining class code repository, deciding what group of classes should be incorporated into repository and identifying exact set of classes to reuse, need to be addressed. By using clustering, one can find frequently used classes in the same cluster and can know their coupling with other classes in a particular application.

1.1 Data Clustering and Reusability

Data mining techniques can be used to analyze software engineering data to better understand the software and assist software engineering tasks. Clustering can be used
for document clustering in the information retrieval task. The Vector Space Model (VSM) is the basic model for document clustering. In this model, each document, \( d_j \), can be represented as a term-frequency vector in the term-space:

\[
d_{jT} = (t_{f_{1j}}, t_{f_{2j}}, \ldots, t_{f_{vj}}) \quad j = 1, 2, \ldots, D
\]

where \( t_{f_{ij}} \) is the frequency of the \( i \)th term in document \( d_j \), \( V \) is the total number of the selected vocabulary, and \( D \) is the total number of documents in the collection.[18]. One can weight each term based on its Inverse Document Frequency (IDF) [18,3]. After having VSM representation, K-mean algorithm can be applied to cluster the documents [15]. Clustering technique can be applied to cluster the Classes/components that may often be reused in combinations [19]. Due to the popularity of open source concept large amount of source code of classes is available on internet as software repositories. For this reason, it is desirable to have clustering mechanism that form cluster of classes based on their association or coupling patterns. In this paper, we explore document clustering technique based on tf-idf weighing [3] to cluster classes from vast collection of class coupling data for particular java project/program. For this purpose firstly dynamic analysis of java application is done using UML diagrams to collect class import coupling data. Then in second step, these collected coupling data of each class are treated as document and represented as VSM(using TF and IDF). Then finally in the third step basic K-mean clustering technique is applied to find cluster of classes. Further each cluster is ranked for their goodness based on some user specified threshold and clusters which are not satisfying threshold are discarded as bad cluster.

The rest of the paper is organized as follows. Section 2 discusses the related works. Section 3 describes the proposed methodology. Section 4 presents conclusion and future scope.

2 Related Works

For object-oriented development paradigm, class coupling has been used as an important parameter effecting reusability[7]. Arisholm et al[4] have provided a method for identifying import coupled classes with each class at design time using UML diagrams. Few algorithms[1,13] like CLARANS [14], BIRCH [21], UPGMA [11] have been proposed for clustering large data sets. K-Mean’s and its family of algorithms have also been extensively used in document clustering [12]. Fung et al [9] proposed to use the notion of frequent itemsets, document clustering and they used an approach proposed by Agrawal et al [2]. There are some distance measures available in literature like Absolute distance, Euclidean distance and cosine distance [22,23,24,6]. Alzghool et al [3] also proposed a technique based on clustering the training topics according to their tf-idf (term frequency-inverse document frequency) properties. For Evaluation of cluster quality authors in [20,15] proposed cluster ranking to quickly single out the most significant clusters based on goodness and quality of cluster.