International Journal of Innovative Computing, Information and Control
Volume **3**, Number **3**, June **2007**

ICIC International ©2007 ISSN 1349-4198
pp. 525–537

MEASURING SOURCE CODE SIMILARITY USING REFERENCE VECTORS

Asako Ohno

Graduate School of Cultural Studies and Human Science
Kobe University
Tsurukabuto 1-2-1, Nada, Kobe, Japan
uiko@ccs-srv.cla.kobe-u.ac.jp

Hajime Murao

Faculty of Cross-Cultural Studies
Kobe University
Tsurukabuto 1-2-1, Nada, Kobe, Japan
murao@i.cla.kobe-u.ac.jp

Received September 2006; revised December 2006

Abstract. *This paper discusses a method for measuring the similarity between program source codes. Unlike many existing methods, our method compares the source codes indirectly using reference vectors instead of using the original source codes. Elements of the vectors are textural features of token-cooccurrence matrices generated from the source codes. Since similarity is simply defined as a distance between two reference vectors, it can be calculated in a very short time. Another advantage is that we do not need to retain a huge storage area for source codes when applying this method to a source code retrieval system. We implemented the proposed method on a simple source code retrieval system and made experiments with Java program source codes submitted as assignments for a programming class. Results confirmed that our method can effectively find reasonable similarity between pairs of source codes in a short time.*

Keywords: Source code the similarity, Refactoring, Software maintenance

1. **Introduction.**

A similarity measuring method for program source codes is beneficial for managing source code repositories, refactoring large scale software systems, maintaining open source softwares, or finding plagiarisms. A number of methods for the similarity measurements have been reported. Many of them can calculate the similarity in a short time but their output contains a considerable amount of false positives, while others achieve high precision but require complex computation for the similarity calculation. Methods using original source codes for the similarity comparison might require a huge storage capacity to store millions of source codes when implemented on applications such as a retrieval system for a source code repository. Regarding copyright protection, such methods are not safe without applying precautionary measures against data leakage. Therefore, there is a strong demand for a development of a similarity measuring method which is versatile, easy to introduce, can measure the similarity in a short time, and does not require its users original source codes to calculate the similarity.