DUPE: The Depository of Universal Plagiarism Examples

Robert Zeidman
Software Analysis & Forensic Engineering Corporation
15565 Swiss Creek Lane
Cupertino, CA 95014 USA
1 (408) 741-5809

bob@SAFE-corp.biz

ABSTRACT

The field of software plagiarism detection may be traced back to 1987 when professors Faidhi and Robinson published their article entitled “An empirical approach for detecting program similarity and plagiarism within a university programming environment” in *Computer Education*. Since that time, professors at academic institutions have created interesting and complex software plagiarism detection algorithms and included them in plagiarism detection programs including Plague, YAP, MOSS, and JPlag.

While these algorithms led to some interesting ways of analyzing source code, none of the algorithms, or the programs that embodied them, was accurate enough to be used in intellectual property litigation. The programs were often research projects of limited duration, and the detection algorithms sacrificed accuracy for speed. The algorithms all discarded source code comments, something that was not interesting to academics, but which often provided clues to copying upon which experts relied in court. In addition, no standards were ever created to objectively determine that copying had indeed occurred, much less been detected, or to independently compare the conclusions of the various detection programs.

This changed in 2003 when I created the CodeMatch program that very quickly became used in software IP litigation. I created a test bench of purposely plagiarized code that could be used to independently and objectively compare the results produced by different plagiarism detection programs. Some in the academic community claimed that my tests were biased toward the algorithms used by CodeMatch, which explained why CodeMatch faired so well compared to the other programs. However, these same critics, despite my requests, never produced their own set of standard tests.

Although I believe that the standard tests I have used are not biased, it occurred to me that there could be a better way to eliminate even unintentional bias. The solution would be to take the source code for certain open source programs and announce a new open source project that would involve purposely plagiarizing the code. Programmers from around the world would be invited, perhaps in a competition, to change the source code while retaining the functionality. The original programs and the plagiarized versions submitted from others would be stored in a database known as the Depository of Universal Plagiarism Examples or DUPE. Plagiarism detection programs would then be run on DUPE and comparisons could be made. Also, important statistics about plagiarized code could be determined, as well as patterns identified in order to improve the plagiarism detection programs.

SAFE Corporation has begun looking into creating this database. However, we realize that we would like to work with partners in academia. First, we believe that computer science academics have done groundbreaking research that can be utilized. Second, practicing lawyers and academics in law schools have a better grasp of the evolving laws dealing with intellectual property. Third, working with academia will provide a better chance of reducing any inadvertent bias on our part.

We believe that there are several key issues that need to be resolved in creating DUPE. These are:

1. Choosing appropriate open source projects.
2. Creating a minimum definition of software plagiarism.
3. Creating the database.
4. Determining policies including who can access it, how it will be used, and who will maintain it.
5. Determining how to run the tests, how to generate the results, and how to distribute the results.
6. Understanding legal issues including privacy issues, copyright issues, and licensing issues.

IMF 2009 will be an ideal venue to present ideas about DUPE, solicit feedback from the knowledgeable participants, and find others who are willing to partner with SAFE Corporation to implement DUPE.

AUTHOR

Robert Zeidman is a Senior Member of the IEEE and president of SAFE Corporation, the leading provider of tools for comparing and measuring software intellectual property. Among his publications are technical papers on hardware and software design methods as well as three textbooks -- *Designing with FPGAs and CPLDs*, *Verilog Designer's Library*, and *Introduction to Verilog*. He has taught courses at engineering conferences throughout the world. Robert holds six patents and earned a master's degree in electrical engineering at Stanford University and bachelor's degrees in physics and electrical engineering at Cornell University.