COMPGS04/M024: Tools and Environments

Coursework 2 (Group Coursework)

Authors:
Jihyun Han
Elliott Omosheye
Sachin Pande
Luke Richardson
Yasaman Sepanj

March 26, 2014

1 Introduction

1.1 Sub-section

2 Normalized Compression Distance

Vitányi, Paul MB and Balbach, Frank J and Cilibrasi, Rudi L and Li, Ming. Information theory and statistical learning: Normalized Information Distance. Springer. 2009.

2.1 Background

Normalized compression distance is a method of computing the similarity between two documents of any kind whether this be two text files or two music files. It measures the difficulty of being able to turn one document to the other. This method of determining similarities between two files is based upon the Normalized Information Distance (NID) between them.

2.2 Normalized Information Distance

The information distance between two strings $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined to be the length of the shortest program, $\boldsymbol{p}$, for a universal computer to transform $\boldsymbol{x}$ into $\boldsymbol{y}$ and vice versa. The length of this program can be defined with the use of Kolmogorov complexity as follows:

$$|\boldsymbol{p}| = \mathbf{max}\{\boldsymbol{K}(\boldsymbol{x}|\boldsymbol{y}), \boldsymbol{K}(\boldsymbol{y}|\boldsymbol{x})\} \tag{1}$$

This distance is absolute and therefore needs to be normalized in order to provide the similarity in relative to both the files. Such an normalization provides the NID.

$$\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbf{max}\{\boldsymbol{K}(\boldsymbol{x}|\boldsymbol{y}), \boldsymbol{K}(\boldsymbol{y}|\boldsymbol{x})\}}{\mathbf{max}\{\boldsymbol{K}(\boldsymbol{x}), \boldsymbol{K}(\boldsymbol{y})\}} \tag{2}$$

2.3 Normalized Compression Distance

The NID is however impractical as it uncomputable. Therefore requiring a computable algorithm is requires, such a algorithm is Normalized Compression Distance (NCD). Transforming the equation through substitution of the uncomputable function $\boldsymbol{K}$ with a real-world compression function $\boldsymbol{Z}$ provides the NCD, defines by

$$\boldsymbol{e}_{\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{Z}(\boldsymbol{xy}) - \mathbf{min}\{\boldsymbol{Z}(\boldsymbol{x}), \boldsymbol{Z}(\boldsymbol{y})\}}{\mathbf{max}\{\boldsymbol{Z}(\boldsymbol{x}), \boldsymbol{Z}(\boldsymbol{y})\}} \tag{3}$$

where $\boldsymbol{Z}(\boldsymbol{x})$ is the length of the compression of string $\boldsymbol{x}$ using the compression function $\boldsymbol{Z}$.

2.4 Application

For the case of this project we are interested in similarities between Java source files, for which a variant of the NID can be used to measure the similarity. This variant named sum distance is defined by

$$\boldsymbol{e}_{\mathbf{sum}}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{K}(\boldsymbol{x}|\boldsymbol{y}) + \boldsymbol{K}(\boldsymbol{y}|\boldsymbol{x})}{\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y})} \tag{4}$$

The two files are tokenized then compressed with a customized compressor to approximate the sum distance.

2.5 Conclusion

NCD is able to provide a good score for the similarity between two files. Whilst the NCD algorithm and the variant sum distance are relatively easy to implement, the latter requires the need for the implementation of a custom compressor which is out of scope for this project. Furthermore these algorithms are unable to provide details about the nature of the similarities between the two Java source files.

3 Plagiarism Detection

3.1 Introduction

Plagiarism detection in code is the method and process used for locating instances of cloned code within a set of documents. The tool we have chosen to demonstrate and explore plagiarism detection is JPlag.

3.2 JPlag

JPlag is a system that finds similarities, software plagiarisms, among multiple source code files. It is robust against many attempts to disguise the copied code because it is aware of programming language syntax and program structure. Thus, it is very hard to deceive: 90% of plagiarisms are detected and the rest raise suspicion. [1] Moreover, JPlag is able to process a hundred programs with several hundred lines of code in seconds, making it a very robust and scalable tool. It has been successfully used for detecting plagiarism among students' Java programs but support for other languages such as C and C++ are also available.

JPlag's comparison algorithm is based on the "Greedy String Tiling" [2], which takes in a set of program source code as input and outputs a similarity score between pairs of programs and their corresponding similarity regions. To do so, it must initially convert the program source code in to token strings, this is the only language dependent process involve din JPlag's plagiarism detection. Tokens should be chosen in a manner, which captures the essence of the program's structure rather than the surface aspects, as a program's structure is harder for plagiarists to modify. JPlag does this phase whilst ignoring whitespaces, comments and names of identifiers. Figure 1 shows an example Java code and its corresponding tokens generated by JPlag. A list of possible JPlag tokens can be found in this technical report [3].

Once the token strings have been generated, JPlag beings the comparison phase by compar-

	Java source code	**Generated tokens**
1	`public class Count {`	BEGIN_CLASS
2	`  public static void main(String[] args)`	VAR_DEF, BEGIN_METHOD
3	`            throws java.io.IOException {`	
4	`    int count = 0;`	VAR_DEF, ASSIGN
5		
6	`    while (System.in.read() != -1)`	APPLY, BEGIN_WHILE
7	`      count++;`	ASSIGN, END_WHILE
8	`    System.out.println(count+" chars.");`	APPLY
9	`  }`	END_METHOD
10	`}`	END_CLASS

Figure 1: A sample Java Source code with the generated token strings

ing the two generated token strings. When comparing two strings A and B, JPlag's objective is to discover a maximal set of contiguous substrings. Each substring must occur in both A and B and must be as long as possible. These substrings (matches) must be unique in that each substring should not cover tokens already covered by other substrings. To avoid false matches, a minimum match length M is enforced.

3.3 Greedy Tiling Algorithm

The Greedy String Tiling algorithm itself has two steps shown in Figure 2. Firstly, all substrings common to both token strings are found with lengths equal to or greater than the minimum M. Secondly, all the tokens found in the substrings are marked so that they are no longer picked up by subsequent iterations of the first step. By marking its tokens, a match becomes a "tile". Thus, the similarity score between is calculated based on the fraction of tokens that were found as matches:

$$sim(A, B) = \frac{2 * coverage(tiles)}{|A| + |B|} \tag{5}$$

```
0   Greedy-String-Tiling(String A, String B) {
1       tiles = {};
2       do {
3           maxmatch = M;
4           matches = {};
5           Forall unmarked tokens A_a in A {
6               Forall unmarked tokens B_b in B {
7                   j = 0;
8                   while (A_{a+j} == B_{b+j} &&
9                           unmarked(A_{a+j}) && unmarked(B_{b+j}))
10                      j++;
11                  if (j == maxmatch)
12                      matches = matches ⊕ match(a, b, j);
13                  else if (j > maxmatch) {
14                      matches = {match(a, b, j)};
15                      maxmatch = j;
16                  }
17              }
18          }
19          Forall match(a, b, maxmatch) ∈ matches {
20              For j = 0 ... (maxmatch − 1) {
21                  mark(A_{a+j});
22                  mark(B_{b+j});
23              }
24              tiles = tiles ∪ match(a, b, maxmatch);
25          }
26      } while (maxmatch > M);
27      return tiles;
28  }
```

Figure 2: Greedy String Tiling algorithm

4 Winnowing: Local Algorithms for Document Fingerprinting

Schleimer, Saul and Wilkerson, Daniel S and Aiken, Alex. Winnowing: local algorithms for document fingerprinting. 2003.

4.1 Introduction

If you wanted to compare two whole files to see if the are a clone then the obvious method would be to hash the files and compare the hash values. This algorithm applies this to finding partial clones. In short it works by removing irrelevant features from the text, splitting the text into parts of length k called k-grams and then hashing each k-gram. A small subset of these hashes is the derived and this is called the called the fingerprints of the document. The idea is that if two documents share one or more fingerprints then they likely share the same text.

The problem comes in finding the best way to decide which hash to use as one of the fingerprints of the file. This paper purports to give an efficient algorithm to select the fingerprints and also guarantees that at least part of any sufficiently long match is detected.

The paper compares their solution, called winnowing, to other algorithms. It uses the density of the algorithm and it defines the density of a fingerprinting algorithm to be the expected fraction of hashes that are selected from all the hash values computer when given a random input. It defines a local algorithm, which captures certain properties of a fingerprinting algorithm. It must define a window of size w to be w consecutive hashes of k-grams in a document and selects at least one fingerprint from each window. The algorithm is considered local iff the choice of fingerprint of each window only depends on the hashes in that window.

4.2 Algorithm

2 thresholds are chosen by the user, noise threshold k is the lower bound (no matching strings shorter than k), and the guarantee threshold t (match all substrings at least as long).
Noise threshold k is used to divide the text into k-grams. Bigger k reduces noise, but also reduces sensitivity to the reordering of contents.

Window size $w = t–k + 1$

For each window select the minimum hash value to be the fingerprint of the document. If there is a tie use the rightmost minimal value.
The winnowing algorithm is first compared to a $0 \bmod p$ algorithm, which is where the hash is selected if it is a divisor of p. It is then compared to other local algorithms, to see if there is another algorithm that performs better than winnowing. They prove that the winnowing algorithm has a lower bound on the density as good as the optimum local fingerprinting algorithm and therefore there does not exist a local fingerprinting algorithm with lower density.

4.3 Conclusion

Finally the paper shows the real world results with web data, and with plagiarism detection in the MOSS system. They find a problem due to the fact that there are large passages of repetitive low-entropy strings. Here winnowing encounters many identical hash values and therefore many ties for the minimum hash in the window. This causes poor behaviour; to solve this they define Robust Winnowing, which is not a local algorithm. Robust winnowing breaks ties by selecting the same hash as the previous window if possible. Otherwise it reverts back to normal winnowing behaviour. This is used to reduce the density on low-entropy strings. Moss uses this algorithm for its plagiarism detection. They discuss the implementation of Moss and how it stores the fingerprint data, how the comparisons are done against all the entries in the database, as well as how it presents its results.

5 Latent Semantic Indexing

Manning, Raghavan and Schütze. Introduction to Information Retrieval: 18 Matrix decompositions & latent semantic indexing. Cambridge University Press. 2008.

5.1 Overview of technique

Latent Semantic Indexing (LSI) makes use of a term-document matrix, this is simply an MxN matrix, where columns represent the number of documents in the collection and rows represent the terms. Calculation can then be carried out over this matrix to compute the LSI.

In the case of our tool, there would be two columns, for the pair of documents. The terms would be tokens, produced by some simple whitespace parser or potentially more complex grammar.

5.1.1 Low-rank approximation and computing similarity

LSI first requires an approximation of the term-document matrix into lower-dimensional space using singular value decomposition, an extension of symmetric diagonal decomposition. Symmetric diagonal decomposition is a technique in which a square matrix can be factored into the product of matrices derived from its eigenvectors.

This approximation is required as even medium sized term-document matrices can contain tens of thousands of terms, so this is required for LSI to be scalable.

The mathematics of low-rank approximation is too complex to cover in a short summary, but the end result is a reduction of the size of the term-document matrix, while still maintaining differences and similarities between documents close to that of the original matrix.

Once the low-rank approximation has been computed, the similarity between two documents can be calculated using cosine similarity.

5.2 Performance

Dumais 1993 and Dumais 1995 conducted experiments on TREC documents and tasks. They managed to achieve results with precision at or above the TREC median, achieving a top score on 20% of topics. Therefore we can conclude that the accuracy of LIS is, in general, good.

The performance of LSI is noted to be poor in terms of computational complexity. It is noted that "there have been no successful experiments with over one million documents". Although this is something of concern for some, for us this would be a tolerable limitation, considering we are only dealing with two input files.

LSI is noted to "work best in applications where there is little overlap between queries and documents". For calculating document similarity, ideally we would have more linear performance and be similarly good at working with considerable overlap.

LSI is a vector space model and as such treats documents as a 'bag of words'. This could either prove to improve similarity matching or degrade it, as source code can be sufficiently reordered to produce a program with distinct behaviour, but LSI will still produce results that show a close match. Conversely if someone was to rearrange code to attempt to disguise plagiarism, LSI matching would be unaffected.

Finally there are two other limitations of LSI related to its vector space representation, its inability to cope with synonymy and polysemy. Synonymy refers to two distinct words having the same meaning and polysemy one word being having different meanings dependent on use. This would be an issue for programming language code, where certain words, such as access modifiers and Java keywords are used frequently throughout text.

5.3 Conclusion

Overall LSI seems to be a good similarity matching algorithm. However, for our particular use case, matching Java source code, the weaknesses due to it being a vector space model are likely to put it at a disadvantage to other models which take into account word ordering.

6 Paper 5

6.1 Sub-section

7 Tool requirements

7.1 Overview

The following list of requirements are formed from the defined set of tool requirements in the coursework brief and based on our algorithm choice, JPLAG plagiarism detection.

7.2 Requirements

1. The tool will compute a similarity measurement for two input Java source code files using an implementation of the JPLAG plagiarism detection algorithm.
2. The tool must work via command line and should not have a graphical interface.
3. Input taken should be the names of two files as command lines arguments.
4. The output should be a similarity measure to STDOUT as a percentage.
5. The tool will output additional results from running the plagiarism detection algorithm
6. Input Java source files will be parsed using a Java parser library, ANTLR.
7. The tool will be written in Java (version 6) to meet the requirements for the parser library.
8. The tool will not invoke the JPLAG binary or any other similarity testing tool.
9. The output of similarity measurement by plagiarism detection algorithm will complete in reasonable time for expected inputs. Expected inputs are generally Java source code files and specifically under testing TOH.java, variants of TOH.java and arbitrary dissimilar test inputs.
10. The tool will have an automated testing mechanism, using the JUnit testing framework to carry out pairwise comparison.
11. The tool should work with all platforms the JVM works on that support ANTLR. The tool will be tested and run on Linux and Windows platforms.

8 UML Diagrams of architecture

8.1 Sub-section

9 Tool implementation

9.1 Sub-section

10 Testing

The aim of testing in this project is to gain an understanding of the chosen algorithm's effectiveness in identifying file similarity. This has been achieved through the writing of a simple testing script in Java a number of test file inputs.

The test script simply

11 Results of pairwise comparison

11.1 Sub-section

12 Evaluation of results and project review

12.1 Sub-section

13 Responsibilites

Sachin Pande - Group Leader

- Responsible for setting up the project
- Implemented the parse listener
- Overridden the default ANTLR listener method in order to create a customized token strings
- Researched Normal Compressions Distance
- Implementation of testing code

Yasaman Sepanj

- Implemented the Greedy String Tile algorithm in Java
- Researched Plagiarism Detection

Elliott Omosheye

- Modified the Java grammar
- Implemented methods that analyses the result and returns the similarity percentage
- Researched Winnowing

Jihyun Han

- Written the report
- Testing methodology and writing of tests
- Researched Clone Detection

Luke Richardson

- Implementation of testing code
- Testing methodology and writing of tests
- Researched Latent Semantic Indexing
- Written the report