F-035

Detecting Natural Similarities in Scientific Documents – Author versus Content

Tomohiro Uno† Thomas Zeugmann†

1. Introduction

In recent years, many data mining methods and clustering algorithms have been devised, additionally, those methods are applied to a variety of fields.

Nowadays, there exists a huge quantity of scientific documents in the world. Since a huge quantity of scientific documents exists, when we search the documents, sometimes we meet documents which are written by authors who have the same name. But if the name is common, it is unknown whether or not these authors are the same person. Therefore, it seems to be useful for us to identify the author of a document based on its contents.

To identify the author of a document, the author should be characterized by his habit of writing, the theme of the document, and so on. The objective of this research is to attempt to detect natural similarities from the given documents only.

In order to detect natural similarities, in this research, we apply the concept of the so-called Normalized Compression Distance (abbr. NCD) to documents to detect similarities. Incidentally, the concept is available as an open-source software tool, i.e., the "CompLearn" program and library [2].

2. The Similarity Distance

The present idea for the NCD was proposed by P. Vitányi and his co-workers and it is based on the Kolmogorov complexity theory (cf. [4]). We omit the underlying theory and refer the reader to [6] for a detailed exposition.

The NCD is, roughly speaking, that two objects are deemed close if we can significantly "compress" one given the information of the other, and vice versa.

2.1 Kolmogorov Complexity

Technically, the Kolmogorov complexity of x given y is the length of the shortest binary program, for the reference universal prefix Turing machine, that on input y outputs x; it is denoted by $K(x|y)$. For the precise definitions, the theory and application, see [4]. The Kolmogorov complexity of x is the length of the shortest binary program that on input the empty string λ outputs x; it is denoted as $K(x) = K(x|\lambda)$. Essentially, the Kolmogorov complexity of a file is the length of the ultimate compressed version of the file. However, there is no way to compute $K(x)$, that is, the Kolmogorov complexity is uncomputable.

2.2 Normalized Compression Distance

In [1], the information distance $E(x, y)$ was introduced.

Definition 1. The normalized version of information distance $E(x, y)$, namely the *normalized information distance*, or NID, is defined as

$$\mathrm{NID}(x, y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}. \qquad (1)$$

The NID is, however, impractical since it cannot be computed. Then, we approximate the NID by using a real compressor C.

Definition 2. The compressor C (such as bzip2, gzip) based approximation of the normalized information distance (1), is called the *normalized compression distance* or NCD:

$$\mathrm{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}. \qquad (2)$$

The NCD is the main concept applied in this paper. It is the real-world version of the ideal notion of normalized information distance (1). By [3], the NCD is a non-negative number $0 \leq r \leq 1 + \varepsilon$ representing how different the two files are. Smaller numbers represent more similar files. The ϵ in the upper bound is due to imperfections in our compression techniques, but for most standard compression algorithms one is unlikely to see an ε above 0.1.

3. CompLearn

CompLearn is a suite of simple-to-use utilities that we can use to apply compression techniques to the process of discovering and learning patterns.

3.1 Quick Start

A short example of using CompLearn is shown in Figure 1.

By using the **ncd** command we compute the NCD, by the **maketree** command we generate a best fitting binary tree from a given distance matrix. For a precise explanation of these commands, see [2].

†Graduate School of Information Science and Technology, Hokkaido University.

Step 1	**$ ncd -b -d directory directory**
Step 2	**$ maketree distmatrix.clb**
Step 3	**$ neato -Tps treefile.dot > tree.ps**

Figure 1: Example of using CompLearn

3.2 Standardized Benefit Score

The algorithm which generates the tree also computes the indicator which measures the quality of the tree representation of the overall order relations between the distances in the matrix. The indicator is called standardized benefit score or $S(T)$ value in the sequel. The $S(T)$ value ranges from 0 (worst) to 1 (best). For details of the idea concerning the $S(T)$ value, we refer the reader to [3].

4. Analysis

4.1 Data

There are 29 data for this research. The data which we use are scientific documents down-loaded from SpringerLink [5]. The common author name of these documents is "Wei Wang" and there is also a single co-author. However, these Wei Wangs are not always the same person. Documents are written in English. The oldest document is published in 2002, the newest one is from 2009.

4.2 Computing the NCD

We input the data to CompLearn to compute the NCD. The compressor used to compute the NCD matrix was bzip2. The result obtained from applying the data to the NCD is shown in Figure 2.

In Figure 2, some pairs of data which seem to be the same author are adjoined to one another. Incidentally, the $S(T)$ value for Figure 2 is 0.942821.

5. Conclusion

Through this research, we obtained favorable results. As our results show, it was often possible to identify the same person "Wei Wang" as author from the contents of the relevant paper by using the NCD. Note that the documents which are classified into the same group should have been written by same the author with high probability.

On the other hand, not all classifications obtained are correct. There may be various reasons explaining the misclassifications, e.g., it is conceivable that the co-author did most of the writing. In such cases, the sought similarity may be hard or impossible to detect. So, future research should also explore methods that can analyze the documents from a different point of view.

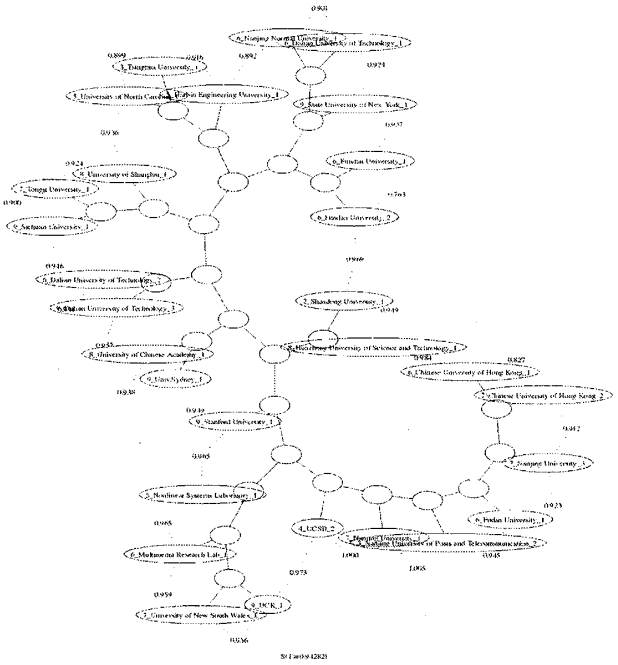

Figure 2: Analyzing documents written by Wei Wang by using the NCD and bzip2 as compressor

References

[1] C. H. Bennet, P. Gács, M. Li, P. M. B. Vitányi, and W. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4):1407–1423, 1998.

[2] R. Cilibrasi, A. L. Cruz, S. de Rooij, and M. Keijzer. Comptearn toolkit. http://www.complearn.org/.

[3] R. Cilibrasi and P. Vitanyi. Similarity of objects and the meaning of words. In *Theory and Applications of Models of Computation*, volume 3959 of *Lecture Notes in Computer Science*, pages 21–45. Springer Berlin / Heidelberg, 2006.

[4] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 3rd edition, 2008.

[5] SpringerLink. http://www.springerlink.com/.

[6] P. M. B. Vitányi, F. J. Balbach, R. L. Cilibrasi, and M. Li. Normalized information distance. In *Information Theory and Statistical Learning*, pages 45–82. Springer, New York, 2008.