

# Use of Kolmogorov distance identification of web page authorship, topic and domain

David Parry

*School of Computer and Information Sciences, Auckland University of Technology, Auckland, New Zealand*

*Dave.parry@aut.ac.nz*

## Abstract

*Recently there has been an upsurge in interest in the use of information entropy measures for identification of similarities and differences between strings. Strings include text document languages, computer programs and biological sequences. This work deals with the use of this technique for author identification in online postings and the identification of WebPages that are related to each other. This approach appears to offer benefits in analysis of web documents without the need for domain specific parsing or document modeling.*

## 1. Introduction

Kolmogorov distance measurement involves the use of information entropy calculations to measure the distance between sequences of characters. Information retrieval is a potentially fruitful area of use of this technique, and it has been used for language and authorship identification [1], plagiarism detection in computer programs [2] and biological sequences, such as DNA and amino acids [3].

Authorship, genre identification and measures of relatedness remain an important issue for the verification and identification of electronic documents. Related document searches have been identified as an important tool for users as information retrieval systems [4]. Computers have been used for a long time to try and verify the identity of authors in the humanities [5] and in the field of software forensics [6]. Various techniques have been used in the past including Bayesian Inference[7], neural networks[8] and more sophisticated methods using support vector machines [9]. However, such approaches tend to be extremely language and context specific although often very effective.

Briefly, this approach is based around the concept of the relative information entropy of a document. The concept of the relative information of a document is closely related to that of Shannon [10]. One way of expressing this concept is to view a document as a

message that is being encoded over a communication channel. A perfect encoding and compression scheme would produce the minimum length of message. In general, a document that can undergo a high degree of shortening by means of a compression algorithm has a low information entropy – that is there is a large degree of redundancy, whereas one that changes little in size has a high degree of information entropy, with little redundant information. A good compression algorithm should never increase the size of the “compressed” document. As the authors of [11] point out, a good zipping algorithm can be considered as a sort of entropy meter.

The Lempel-Ziv algorithm reduces the size of a file by replacing repeating strings with codes that represent the length and content of these strings [12], and has been shown to be a very effective scheme. To work efficiently, the Lempel-Ziv algorithm “learns” effective substitutions as it examines the document sequentially to find repeating sequences that can be replaced in order to reduce the file size. This algorithm is the basis of the popular and rapid zip software in its various incarnations including Gzip, Pkzip and WinZip. Importantly this method relies on a sequential examination of the document to be encoded, so concatenation with other documents can have dramatic effects on the efficiency of zipping, as rules for encoding created at the start of the process are found to be useless at the end. By adding, a document of unknown characteristics to one of known properties (for example language, author, genre etc.) then is suggested that the combined relative entropy is smallest when the two documents are most similar.

The work of [1] demonstrated that it was possible to identify the language used in a document by comparison with known documents. This method is therefore complementary to other methods that concentrate on the understanding of the document, much as handwriting or voice analysis widens the possibilities of author identification, even if the content is not distinctive [13].

The rest of this paper describes one implementation of these types of algorithm (section 2), along with a number of experiments (Section 3). Section 4 discusses the results and section 5 describes other approaches and draws conclusions about this approach.

## 1. Algorithms

The Kolmogorov distance is based on the method of [14], and earlier work such as [15] which deals with the identification of minimum pattern length similarities. By using compression algorithms the following formula for the distance between two objects may be computed. Assuming  $C(A|B)$  is the compressed size of A using the compression dictionary used in compressing B, and vice versa for  $C(B|A)$  and  $C(A)$ ,  $C(B)$  represent the compressed length of A and B using their own compression dictionaries. The distance between A and B,  $D(A,B)$  is given by:

$$D(A, B) = \frac{C(A|B) + C(B|A)}{C(A) + C(B)}$$

This formula is explicitly derived in [16]. Various methods of compression have been used for this, for this work a method was used that did not need explicit access to the compression dictionary, so that standard zip programs could be used. Concatenating files and then compressing them allows the compression algorithm to develop its dictionary on the first file and then apply it to the second. The algorithm used is given by:

*Obtain the two files – file<sub>1</sub> and file<sub>2</sub>*

*Concatenate them in two ways, file<sub>1</sub>+ file<sub>2</sub> = (file<sub>12</sub>) and file<sub>2</sub>+ file<sub>1</sub>=(file<sub>21</sub>)*

*Calculate the compressed length of:*

*file<sub>1</sub> as zip<sub>1</sub>*

*file<sub>2</sub> as zip<sub>2</sub>*

*file<sub>12</sub> as zip<sub>12</sub>*

*file<sub>21</sub> as zip<sub>21</sub>*

The distance (D) is then given by:

$$D(\text{file}_1, \text{file}_2) = \frac{(\text{zip}_{12} - \text{zip}_1) + (\text{zip}_{21} - \text{zip}_2)}{\text{zip}_1 + \text{zip}_2}$$

This approach depends on the compression algorithm being lossless. Previous work had demonstrated that if the file<sub>1</sub> is the same as file<sub>2</sub> the distance is minimal.

## 3. Methods

Three experiments were performed to validate the algorithm used. One used author identification, the second used WebPages from different domains, and the third used different topics within a particular web corpus.

### 3.1 Experiment One

One particularly rich source of testing data is achieved newsgroup and list server postings that often contain particularly relevant information in a concise format. Newsgroup postings provide a rich corpus of material to

study classification schemes – for example the use of readability or other scores[17] to characterize discussion.

Postings from an online teaching system – Business on line [18], were used. A total of 160 initial messages were used. The Kolmogorov distance (KD) was calculated between this message and 10 other messages, only one of which by the same author as the first. The message combination with the shortest KD was then noted, and the results are shown in Table 1.

**Table 1: Kolmogorov Distance for Messages**

Status	Percent Shortest KD	Percent in sample
Author1<>Author2	51.88%	90%
Author1=Author2	48.13%	10%

Using Chi-Squared, this result is significant at the  $p < 0.001$  level (SPSS 11)  $\chi^2 = (1, N=160) = 258, p < 0.001$ . The proportion of messages with common authors having the smallest distance is a great deal higher than expected by chance.

### 3.2 Experiment Two

4,389 Web Pages were downloaded using a web spider from 6 root sites. A similar comparison was done for the website domain-based group, with each of 80 pages compared with one from the same domain, and nine from others. The results are shown in Table 2. Again, using Chi-Squared, this result is significant at the  $p < 0.001$  level (SPSS 11)  $\chi^2 = (1, N=80) = 451, p < 0.001$ . The proportion of websites from common domains having the smallest distance is a great deal higher than expected by chance.

**Table 2: Kolmogorov Distance for Domains**

Status	Percent lowest KD	Percent in sample
Different Domain	18.75%	90%
Same Domain	81.25%	10%

### 3.3 Experiment Three

This experiment used the British Medical Journal (BMJ) Website that includes a large number of pages grouped by topic. The process began by selecting those topics that had at least 5 valid pages available for download. For each of these valid domains (n=133), 5 initial pages were chosen. One page from the same domain, and nine different pages from other domains were then selected, in a similar manner to that described above. Again, the files were selected to be of similar length, and the pages zipped together, using the Kolmogorov distance by zipping algorithm. Self-

comparison i.e. where file1=file2 was not permitted. The results are shown in Table 3.

**Table 3: Kolmogorov distance BMJ topics**

Source	Number of occurrences with shortest distance	Percent in sample
Different topic domain	17.89%	90%
Same topic domain	82.11%	10%

Using CHI-Squared implemented in SPSS version 11 the results show that the minimal distance is significantly more likely to occur using files from the same domain, rather than ones of similar length from other domains.  $\chi^2=(1,N=665)=3839,p<0.001$

#### 4. Discussion and Future Work

The Kolmogorov distance measure approach demonstrates effective identification of related documents. This relatedness may be intrinsic to the text, as in the case of content, authorship or language, or related to the structure of the webpage, that is the arrangement of tags or formatting information.

Drawbacks to the practical implementation of this method centre around two main areas, combinatorial explosion and confounding similarity.

As stated this method requires each file to be compared with each other file, thus the number of calculations needed to find the distance between n documents is given by n!

Current work is concentrating on the clustering of documents using this approach. One approach has been to find documents that are close in terms of KD, to use these as cluster centroids, and measure the distance of new examples from these. This approach, by identifying the centroid of a cluster in terms of a limited number of documents would remove the issue of n! comparisons. Work by [3], has emphasized the importance of clustering.

Confounding similarity, represents the case where documents have a great deal of similarity that is unrelated to their content – for example in the case of documents converted to HTML by popular editors with supplied templates or conversion programs. This does not seem to be an issue in the case of the BMJ topic corpus, but may become important in other cases. If necessary text extraction and separation from formatting tags could be used. The ultimate length of documents that can be effectively processed in this way should be investigated, it seems reasonable to suppose that extremely long

documents of very short documents would not be suitable because of the likelihood of common of repeating motifs in the former case and the absence of repeating motifs in the latter. Other compression techniques, including those where the compression dictionary is stored separately, should be investigated.

It is important to note that this approach is generally complimentary to existing ones and has not been compared with other methods – such as comparisons using textual information, This method is attractive in areas where there is difficulty in performing domain specific parsing or there is no knowledge relating to document structure.

In terms of open-source implementation, this approach could easily be added as a plug-in to browser technology, allowing individual users to compare new documents to those cached already, or by allowing users to collaboratively compare documents with a central or dispersed repository. The decreasing cost of storage implies that document cache comparison will become increasingly important, and simple, general comparison tools will be important in this regard.

#### 5. Conclusion

Comparing electronic documents using the Kolmogorov technique is easily implemented and is not constrained by any proprietary technology. This approach seems particularly useful for short, unstructured documents such as newsgroup postings and emails. Web logs (Blogs) are also becoming more popular and this approach could be used for comparison and validation of these. Use of this technique, in addition to current methods may allow improved characterization of electronic communication and searching of electronic databases. For search engine technology, such approaches may allow improved ranking of results. Particular applications include relatedness and clustering applications, email filtering, fraud and plagiarism detection and genre identification. Further research in this area may increase the value of this approach.

#### 6. References

- [1] D. Benedetto, E. Caglioti, and V. Loreto, "Language Trees and Zipping," *Physical Review Letters*, vol. 88, pp. 048702-1 to 048702-4, 2002.
- [2] X. Chen, B. Francia, M. Li, B. McKinnon, and A. Seker, "Shared information and program plagiarism detection," *Information Theory, IEEE Transactions on*, vol. 50, pp. 1545-1551, 2004.
- [3] R. Cilibrasi and P. M. B. Vitanyi, "Clustering by compression," *Information Theory, IEEE Transactions on*, vol. 51, pp. 1523-1545, 2005.

- [4] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: a study of user queries on the Web," *SIGIR Forum*, vol. 32, pp. 5-17, 1998.
- [5] S. Y. Sedelow, "The Computer in the Humanities and Fine Arts," *ACM Computing Surveys (CSUR)*, vol. 2, pp. 89-110, 1970.
- [6] P. W. Oman and C. R. Cook, "Programming style authorship analysis," pp. 320--326, 1989.
- [7] Mosteller F. and Wallace D., *Applied Bayesian and Classical Inference: the case of the Federalist Papers*: Addison-Wesley, 1964.
- [8] S. T. Singhe, F.J., "Neural networks and disputed authorship: new challenges " in *Artificial Neural Networks, 1995., Fourth International Conference on*, 1995, pp. 24-28.
- [9] O. d. Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," *ACM SIGMOD Record*, vol. 30, pp. 55-64, 2001.
- [10] Shannon., "Mathematical Theory of Communication," in *Bell Systems Technical Journal*, 1948.
- [11] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, "Data compression and learning in time sequences analysis," *Physica D: Nonlinear Phenomena*, vol. 180, pp. 92-107, 2003.
- [12] J. L. Ziv, A., "A universal algorithm for sequential data compression," *Information Theory, IEEE Transactions on*, vol. 23, pp. 337-343, 1977.
- [13] S. N. Srihari and S. Lee, "Automatic handwriting recognition and writer matching on anthrax-related handwritten mail," in *Eighth International Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 280-284.
- [14] A. Kolmogorov, "Logical basis for information theory and probability theory," *Information Theory, IEEE Transactions on*, vol. 14, pp. 662-664, 1968.
- [15] A. Kolmogorov, "Three Approaches to the quantitative definition of Information," *Problems of Information Transmission*, vol. 1, pp. 1-17, 1965.
- [16] M. Li, X. Chen, X. Li, B. Ma, and P. Vitenyi, "The similarity metric," presented at SODA - Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Baltimore, Maryland., 2003.
- [17] P. Sallis and D. Kasabova, "Computer-Mediated Communication, Experiments with e-mail readability," *Information Sciences*, pp. 43-53, 2000.
- [18] A. Sallis, G. Carran, and J. Bygrave, "The Development of a Collaborative Learning Environment: Supporting the Traditional Classroom," presented at WWW9, Netherlands, 2000.