Text Identification via Computational Geometry

Marios Poulos[1], Vasilios S. Belesiotis[2] and Sozon Papavlasopoulos[1]
[1]Dept. of Archives and Library Sciences
Ionian University,
Palea Anaktora 49100, Corfu,
Greece
[2]Department of Informatics
of University of Pireaus
Karaoli & Dimitriou 9, Pireaus,
Greece

Abstract: - In this work, an attempt is being made to propose a mechanism for simultaneously checking the authenticity and degree of similarity between documents when the hash function of the digital signature schemes fail to do so. This work also attempts to propose a scenario of management for the control of authentication or for the detection of a degree of violation of documents, which could be adopted as a component of libraries' strategy for the protection of copyrights of documents published on the web.

Key-Words: - Text Identification, Computational Geometry, Hausdorff distance

1 Introduction

In the conventional world, the well-known International Standard Book Number (ISBN) is the most widely used identifier for books. The Book Item and Component Identifier (BICI), a draft standard for trial use, which provides unique identification of book items and of the component parts of books [1], is another example of this kind. For serials the corresponding standard identification number is the also widely known and used International Standard Serial Number (ISSN).

In the digital world things are, to a great extent, different, due to the salient properties of digital material, although standard identifiers, such as the Serial Item and Contribution Identifier (SICI), can be used to identify both print and electronic serial publications [2]. There are many other identifiers, such as the Publication Item Identifier (PII), the International Standard Textual Work Code (ISTC), and DOI [3], which is the most prevalent, being almost universal. However, none of these identifiers uses the content of the books, journals, or digital objects identified to produce the serial number.

Electronic publishing causes some concerns for publishers, since digital material is incredibly fickle. No rights-holder can be complacent about this because digital content can be easily and quickly copied, and even modified, an indefinite number of times. Digital material can be freely distributed to an enormous number of individuals with or without permission or authorization. In the recent past many cases of software [4] and music piracy have appeared. Such piracy attempts may have many adverse effects, such as financial losses, on the original owners' rights. From a user's perspective, as well as from the perspective of the library and information services, the efforts made to control access through such restrictive policies as making the material inaccessible through encryption or placing it in secure electronic containers cause complexity, frustration, disaffection, and repercussions to users [3].

Digital signature methods use public-key algorithms, but public-key algorithms are deficient for signing long documents. Digital signature protocols use a cryptographic digest, which is a one-way hash of the document. The disadvantage of this process is that it signs the hash instead of the documents' content. For example, systems such as digital signatures or DOI [3] produce completely different digital signature serial numbers for any two similar documents. This has the major disadvantage of not containing semantic features. In particular, when we examine documents that are similar to each other, the number of documents produced has no relationship to them [5].

This paper aims to solve the problems mentioned above by the use of the Hausdorff distance factor, which measures the geometrical differences between two sets of points. This will be explained in the identification procedure section of this paper.

The semantic properties of the proposed algorithm, the SCI, are proven via its application in several other areas of research which use it for text categorization and semantic keywords extraction [6], fingerprint verification [7], and person identification [8], proving that it creates a unique convex polygon for each different input.

More concretely, this method converts text into a unique set of numbers. This conversion takes place so that all the features of a document's characters may be represented in the Cartesian plane and used in the computational geometry. Furthermore, one characteristic of the onion-layers method [6] is that all the features are sorted in an ideal way and all the documents' specific features, such as spaces and punctuation marks, have a domain role in the final reduction.

The proposed method may be used as a secure method for detecting the copyright of documents, specifically in Internet publications for which copyright is a significant factor. It also addresses the needs of both the information services sector and the publishing industry.

2 Method

Our method is divided into two stages:

- Pre-processing stage: conversion of the symbolic expression (in our case an array of characters of a text) to numeric values.
- Processing stage: analysis of the proposed dimensionality reduction technique using an onion algorithm based on computational geometry for text categorization purposes.

2.1 Pre-processing stage

In this stage we suppose that a selected text is an input vector $\overline{X} = (X_1, X_2, X_3 \ldots, X_n)$,

where $(X_1, X_2, X_3 \ldots, X_n)$ represents the characters of the selected text. Then, using a conversion procedure which converts a symbolic expression (in our example an array of characters of a text) to American Standard Code for Information Interchange (ASCII) characters in string arithmetic values, we obtained a numerical value vector $\overline{S} = (S_1, S_2, S_3 ... S_n)$, where these values ranged between 1-128. In our example we achieved this conversion by using the double.m function of the Matlab language. This function converts strings to double precision and equates with converting an ASCII character to its numerical representation.

For better comprehension we provide the following example via Matlab:

```
>> S = 'This is a message to test the double
"command".'
>> double(S)
ans =
 Columns 1 through 12
   84  104  105  115   32  105  115   32   97   32
109  101
  Columns 13 through 24
   115   97   103   101   32   116   111   32   116
101  115  116
  Columns 25 through 36
    32   116   104   101   32   100   111   117   98
108  101   32
  Columns 37 through 46
   34   99  111  109  109   97  110  100   34   46
```

2.2 Processing Stage

Our proposed method is based on the following proposition: The set of elements of vector $\overline{S}$ for each selected text contains a convex subset which has a specific position in relation to the original set. This position may be determined by using a combination of computational geometric algorithms known as onion-peeling algorithms [9], with an overall complexity of O(d*n log n) times, where d is the depth of the smallest convex layer and n is the number of characters in the numerical representation.

Thus, the smallest convex layer $\overline{S}_x$ of the original set of vector $\overline{S}$ carries specific information. In particular, vector $\overline{S}_x$ may be characterized as a common geometrical area of all the elements of vector $\overline{S}$. In our case, this consideration is valuable because this subset may be characterized as representing the significant semantics of the selected text.

2.3 Implementation

We consider the set of characters of a selected text to be vector S. The algorithm starts with a finite set of points, $S = S_0$, in the plane. The following iterative process is considered. Let S_1 be the set $S_0 - \partial H\ (S_0) : S_0$ minus all the points on the boundary of the hull of S_0. Similarly, define $S_{i+1} = S_i - \partial H\ (S_i)$. The process continues until the set is ≥ 3 (see Figure 1). The hulls $Hi = \partial H\ (S_i)$ are called the layers of the set, and the process of peeling away the layers is called onion peeling, for obvious reasons (Figure 1).

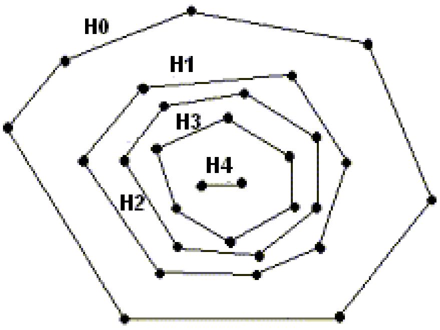


Figure 1 Onion layers of a set of points

2.4. Identification procedure

In the identification procedure we adopted the algorithm of Hausdorff distance [10]. This algorithm calculates the Euclidean positions of points between the two smallest layers regarding the Euclidean distance. In more detail, we examine a document in order to identify the authenticity percentage, and finally to determine if it is identified. Then we construct its serial number and compare it with its original serial number, which is stored in the database of an authority.

The first part of the serial number, which has to do with the number of words and the number of layers of the convex polygons, must be identical. If they are not the document cannot be identified. The next part of the serial number addresses the points of the smallest convex layer. Here we are elastic to a specific degree. We can accept a different number of points from the original, but the two sets must have a small Hausdorff distance (Atallah, 1983).

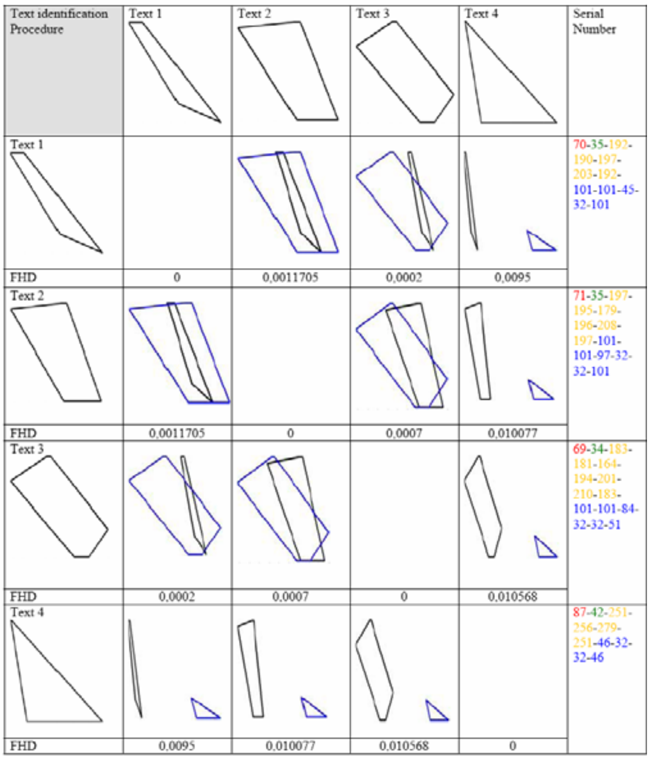
Given two sets of points, A={a1,a2,…,am} and B={b1,b2,…,bn}, the Hausdorff distance is defined as H(A,B)=max(h(A,B), h(B,A)), where: $h(A,B) = \max_{a\in A} \min_{b\in B} \|a-b\|$, ||α - b|| is any metric between the points a(x1,y1) and b(x2,y2), such as the Euclidian distance. In the experimental decision, we say that two compared smallest layers are the same if the Hausdorff distance is equal to zero, or are similar or different when these are a lesser empirical number, which is determined in the experimental part.

3. Experimental Identification procedure

In this stage we used four different texts. The second and third texts are similar to each other and to the first text, while the fourth is different from the first text. These texts are presented below:

1. The Philippines trade deficit widened to 542 mln dlrs in the eight months to end-August from 159 mln dlrs in the same 1986 period, the National Statistics Office said. It said exports in the eight-month period rose to 3.58 billion dlrs from 3.18 billion in 1986, while imports rose to 4.12 billion dlrs from 3.34 billion a year earlier. The country s trade deficit totalled 202 mln dlrs in 1986.
2. The Philippines trade deficit widened to 542 mln dlrs in the eight months to end-August from 159 mln dlrs in the same 1986 period, the National Statistics Office said. It said that exports in the eight-month period rose to 3.58 billion dlrs from 3.18 billion in 1986, while imports rose to 4.12 billion dlrs from 3.34 billion a year earlier. The country s trade deficit totalled 202 mln dlrs in 1986.
3. The Taiwan trade deficit widened to 542 mln dlrs in the eight months to end-August from 159 mln dlrs in the same 1986 period, the National Statistics Office said. The exports in the eight-month period rose to 3.58 billion dlrs from 3.18 billion in 1986, while imports rose to 4.12 billion dlrs from 3.34 billion two years earlier. The country s trade deficit totalled 202 mln dlrs in 1986.
4. The Bank of France sold 1.6 billion francs of 8.50 pct March 1987/99 Caisse de Refinancement Hypothecaire (CRH) state-guaranteed tap stock at an auction, the Bank said. Demand totalled 6.82 billion francs and prices bid ranged from 93.50 to 96.60 pct. The minimum accepted price was 95.50 pct with a 9.13 pct yield, while the average price was 95.69. At the last auction on February 19, two billion francs of CRH tap stock was sold at a minimum price of 91.50 pct and yield of 9.73 pct.

These texts were submitted in the algorithmic procedure, which is described in the Method section. Table 1 presents the results and the Forward Hausdorff Distance (FHD) values.



Text identification Procedure	Text 1	Text 2	Text 3	Text 4	Serial Number
Text 1					70-35-192-190-197-203-192-101-101-45-32-101
FHD	0	0,0011705	0,0002	0,0095	
Text 2					71-35-197-195-179-196-208-197-101-101-97-32-32-101
FHD	0,0011705	0	0,0007	0,010077	
Text 3					69-34-183-181-164-194-201-210-183-101-101-84-32-32-51
FHD	0,0002	0,0007	0	0,010568	
Text 4					87-42-251-256-279-251-46-32-32-46
FHD	0,0095	0,010077	0,010568	0	

Table 1 The identification procedure between (4) four texts

According to these results we adopted the following empirical rule:

The proposed procedure uses the extracted serial number for deciding whether the text is authentic, similar, or different from the reference text. In particular, if the number of words in the serial number is not identical, the examined text is not authentic. If it is, we then have identification and the FHD is 0. In the case of two texts with different authors and with FHD ≥ 0.001, one of them may be the result of plagiarism. Generally, SCI can identify the extent of similarity between texts irrespective of the reasons that may be responsible for the alteration or dissimilarities between them.

4 Conclusion

In this work we presented a new identification technique based on the onion-peeling algorithm with a complexity of size O(d*n log n) times in order to create a serial number for a text. For the implementation of this purpose we used the onion algorithm technique of computational geometry, and for the identification procedure we used the Hausdorff distance algorithm. Results showed that the proposed method may be used as an accurate method for identifying same, similar, or different conceptual texts. This unique identification method of texts can solve many problems that the information society faces, such as plagiarism, problems related to copyright, and tracking problems, with the combination of SCI and DOI. The possibility provided by the proposed algorithm to present the FHD helps to identify the copy percentage of copyright-protected content, as well as to safeguard the intellectual rights of authors and other digital rights owners. The advantages of the exported serial number are obvious, and we aim to highlight them and discuss the probable combination with DOI. Finally, this method may be used by the information services sector and the publishing industry for standard serial-number definition identification, as a copyright management system, or both.

References:

[1] Salton, Gerard and Hans-Jochen Schneider, ed. Research and Development in Information Retrieval: Proceedings, Berlin, May 18-20, 1982. NewYork; Springer- Verlag, 1983.

[2] ANSI/NISO 239.5611996 (Version 2) Serial Item and Contribution Identifier(SICI) Published by NISO Press 4733 Bethesda Avenue, Suite 300 Bethesda, MD 20814 ISBN: l-880124-28-9 USA 1999

[3] Norman Paskin, "DOI: Current Status and Outlook May 1999. D-Lib Magazine, May 1999.

[4] Thomson Gale, BSA Unveils Model Business Practices For Internet Auction Sites. Software Industry Report (Newsletter) Vol 32 24(3) Millin Publishing, Inc. 2000

[5] Davidson, L.A., and Douglas, K. (1998). Promise and problems for scholarly publishing. *The Journal of Electronic Publishing,* Vol. 4, Is. 2, ISSN 1080-2711,

[6] Poulos M., Papavlasopoulos S. and Chrissikopoulos V. (2004). A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm. *Journal of Digital Information,* Vol 6, Is. 1,

[7] M. Poulos, S. Papavlasopoulos and V. Chrissicopoulos (2004). An application of the onion-peeling algorithm for fingerprint verification purposes. Journal of Information & Otimization Sciences. V. 26. 3. pp. 665-681.

[8]Poulos M., Rangoussi M., Chrissicopoulos V., and Evangelou A (1999). Parametric person identification from the EEG using computational geometry. Proceedings of the Sixth International Conference on Electronics, Circuits and Systems (ICECS'99), Pafos, Cyprus, September, pp. 1005-1012.

[9]Bose, P. and Toussaint, G. (1995). No quadrangulation is extremely odd. In 6th International Symposium on Algorithms and Computation (formerly SIGAL International Symposium on Algorithms), pp. 340-358.

[10]Atallah M. J. (1983). A linear time algorithm for the Hausdorff distance between convex polygons. *Information Processing Letters*, Vol. 17, pp. 207-209.