The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0033-0337.htm

Towards text copyright detection using metadata in web applications

Marios Poulos
Information Technology Laboratory, Department of Archives and Library Sciences, Ionian University, Corfu, Greece

Nikolaos Korfiatis
Institute of Informatics and Mathematics, Goethe University Frankfurt, Frankfurt, Germany, and

George Bokos
Information Technology Laboratory, Department of Archives and Library Sciences, Ionian University, Corfu, Greece

Received August 2011
Accepted August 2011

Abstract

Purpose – This paper aims to present the semantic content identifier (SCI), a permanent identifier, computed through a linear-time onion-peeling algorithm that enables the extraction of semantic features from a text, and the integration of this information within the permanent identifier.

Design/methodology/approach – The authors employ SCI to propose a mechanism for simultaneously checking the authenticity and degrees of similarity between different information objects, and present an empirical investigation of the method. A management scenario for the control of the authentication process and the detection of the degree of violation of documents is proposed.

Findings – Such a mechanism could be adopted as a component of libraries' strategy for the protection of the copyrights for documents published on the web.

Practical implications – The use of the proposed numeric code can be utilised efficiently as a constituent part of the digital object identifier (DOI) system, making its computation more efficient and meaningful.

Originality/value – The identifier proposed in the paper can result in a more efficient index for identifying and retrieving objects in a digital library, as well as online repositories and commercial applications that can handle information retrieval requests more effectively.

Keywords Text identification, Information retrieval, Semantics, Persistent identifiers, Data handling, Copyright, Research work

Paper type Research paper

1. Introduction

Information is the keyword in the hyper-information society in which we live now. Consumption and use of information is highly related to its retrieval and the way a piece of information can be identified and characterised by an information retrieval system (Baeza-Yates and Ribeiro-Neto, 1999). Libraries and other memory institutions have developed several ongoing initiatives to develop and implement persistent identifiers, with the most acknowledged being the Digital Object Identifier, or DOI[1] (Paskin, 2002), which is an ongoing initiative established by the International DOI Foundation.

Program: electronic library and information systems
Vol. 45 No. 4, 2011
pp. 439-451
© Emerald Group Publishing Limited
0033-0337
DOI 10.1108/00330331111182111

Identifiers uniquely identify entities of several kinds, making it possible for them to be referenced accurately and effectively. An identification system is employed in order to establish the identity of a given object out of a closed pool of N objects (one-to-N matching). The goal of this study is to argue for the adoption of a semantic identifier (we call it the semantic object identifier or SCI), which will combine the identification approach with the semantic features of the information object associated with the information resource, by exploiting the use of an authentication system.

In contrast to an identifier, an authentication system involves confirming or denying the identity claimed by an object. In the offline world, there are many identifiers in use today; they vary in their attributes and objectives. An example of an identifier from the conventional world is the well-known International Standard Book Number (ISBN), which is the most widely used identifier for books. The Book Item and Component Identifier (BICI), which is a draft standard for trial use that provides unique identification of book items and of the component parts of books (Hilse and Kothe, 2006), is another example. For serial publications, the corresponding standard for identification and authentication is the widely used International Standard Serial Number (ISSN)[2].

In the digital world things are, to a great extent, different, because of the salient features of the digital material. However, standard identifiers, such as the Serial Item and Contribution Identifier (SICI), can be used to identify both conventionally printed and electronic publications[3]. Several different identifiers exist, like the Publication Item Identifier (PII), the International Standard Textual Work Code (ISTC), and DOI, which as mentioned previously is the most prevalent, since it is almost universal. However, one of the issues that we have found among these is that none of them makes use of the content of the books, journals, or digital objects identified to produce the relevant identification code (Bellini *et al.*, 2008).

Electronic publishing raises some serious concerns, since digital material is incredibly fickle. No rights-holder can be complacent about this, because digital content can be quickly and easily copied, and possibly modified, reproducing massive amounts of new instances. Digital material can be distributed to an enormous number of individuals, even in the absence of permission or authorisation. In the recent past, many cases of software and music piracy have been reported (Hill, 2007). Such piracy attempts can have many adverse effects, such as financial loss 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 restrictive policies, such as making the material inaccessible through encryption or placing it in secure electronic containers, can cause undesirable repercussions like complexity, frustration, or disaffection. (Tenopir, 2003; Fuhr *et al.*, 2007).

Digital signature methods use public-key algorithms, but public-key algorithms are insufficient for signing long documents (Bawa *et al.*, 2009). 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. As a result, systems such as digital signatures or DOI produce completely different digital signature for two similar documents. This is a major disadvantage, since the resulting identifier does not contain any of the semantic features of the document. In particular, if we employ any current digital signature method on similar documents, their resulting signatures will be totally related.

This paper aims to solve the aforementioned problems by the use of the Hausdorff distance factor, which measures the geometrical differences between two sets of points. This method will be explained in detail in the section describing the identification.

The semantic properties of the proposed SCI algorithm, are proven via its application in several other research areas for text categorisation and semantic keywords extraction (Poulos *et al.*, 2006), image verification (Poulos *et al.*, 2008), fingerprint verification (Poulos *et al.*, 2003), and person identification (Poulos *et al.*, 2002), 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 transformed through computational geometry. Furthermore, one characteristic of the onion-layers method 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 (Poulos *et al.*, 2002).

For demonstration reasons we considered the DOI structure, in order to obtain semantic features. DOI has two components, known as the prefix and the suffix, which are separated by a forward slash.

Our proposal, as an example of the possible applications of SCI, is to use the SCI number in DOI, in a way similar to the one ISBN, PII, SICI, and other identifiers of this kind are used. In this way, DOI acquires semantic features, while retaining and reinforcing its uniqueness. DOI also has the objective of permanently locating and identifying digital objects irrespective of any potential change in their location and address within the worldwide web.

To this end this paper is organised as follows. The section that follows describes the method used to extract the semantic elements of the text and the related processing required for the implementation stage discussed in the subsequent section along with an experimental demonstration on a set of information objects with their semantic properties altered. An example application scenario in the field of digital libraries is discussed by making use of the proposed identifier. The paper concludes in the final section with suggestions for future research.

2. Method description

The method is based on two stages where the information object (represented in input vectors) is fed into a geometric mapping object where the construction of convex hulls is made by extracting the semantic futures of the text and mapping them in a plane.

In particular, our method is divided into the following two stages:

(1) *Pre-processing stage* – Conversion of the symbolic expression (in our case an array of characters of a text) to numeric values.

(2) *Processing stage* – Analysis of the proposed dimensionality reduction technique using an onion algorithm based on computational geometry for text categorisation purposes.

We proceed with a detailed description of the stages below.

2.1 Pre-processing stage

In this stage we suppose that a selected text forms an input vector $\vec{X} = (X_1, X_2, X_3, \ldots, X_n)$, 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 obtain a numerical value vector $\vec{S} = (S_1, S_2, S_3, \ldots, S_n)$, with values ranging from 1 to 128. In our example, we achieve 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 $\vec{S}$ for each selected text contains a convex subset that has a specific position in relation to the original set (Poulos *et al.*, 2003). This position may be determined through the combination of a series of computational geometric algorithms known as onion-peeling algorithms, 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 $\vec{S}_x$ of the original set of vector $\vec{S}$ carries specific information. In particular, vector $\vec{S}_x$ may be characterised as a common geometrical area of all the elements of vector $\vec{S}$. In our case, this consideration is valuable because this subset may be characterised as representing the significant semantics of the selected text.

3. Implementation

Having provided the algorithmic description for the method, we continue with the implementation part of this method. In order to do that, we start with a vector S as the set of characters of the selected text. 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 $H_i = \partial H(S_i)$ are called the layers of the set, and the process of peeling away the layers is called onion peeling, for obvious reasons (Figure 1).

3.1 Explanation of the numerical code

The above algorithm provides the components constituting the proposed numerical code. This number is unique for each text, or each concatenation of words. The

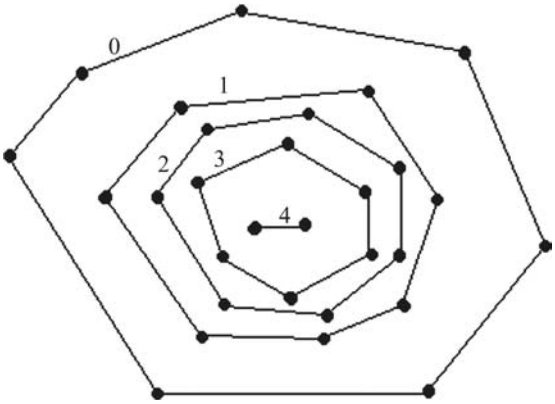


Figure 1.
Onion layers for a set of points

structure of this numerical code is simple and exceptionally functional. It consists of four parts, as shown in Figure 2. The first part, number of words, is the number of words that constitutes the text in question (see the red area). The next three parts are results from the run of the algorithm in the Matlab. The second part of the proposed numerical code is the number of layers of the convex polygons (see the green area). The x digital number of smallest onion layer (the flesh-coloured area) and y digital number of smallest onion layer (the azure area) follow. These two areas constitute the points of the convex polygon in the Cartesian plane. Through assembling all these parts the proposed number (SCI), which is unique for each text, is constructed.

3.2 Identification procedure

In the identification procedure we adopted the Hausdorff distance algorithm adapted to a geometric transform for the set of points shown in Figure 1 (Atallah, 1983). This algorithm calculates the Euclidean positions of points between the two smallest layers with respect to the Euclidean distance. In more detail, we examine a document in order to identify the authenticity percentage, and finally to determine if it can be identified. Then we construct its numerical code and compare it with the original numerical code, which is stored in the database of the authority.

The first part of the numerical code, which is relevant to the number of words and the number of layers of the convex polygons, must be identical. If this is not the case, the document cannot be identified. The next part of the numerical code addresses the points of the smallest convex layer. The algorithm is less sensitive to this part of the identifier, since it is acceptable for a certain number of points to be different, as long as

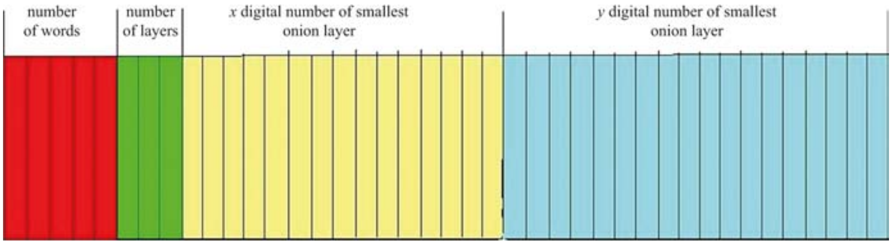

number of words
number of layers
x digital number of smallest onion layer
y digital number of smallest onion layer

Figure 2.
The proposed numerical code

the two sets have a small Hausdorff distance (Huttenlocher *et al.*, 1993; Dubuisson and Jain, 1994).

Given two sets of points, $A = \{a_1, a_2, \ldots, a_m\}$ and $B = \{b_1, b_2, \ldots, b_n\}$, the Hausdorff distance is defined as $H(A,B) = \max[h(A,B), h(B,A)]$, where:

$$h(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|,$$

and $\|a - b\|$ is any metric between the points $a(x_1,y_1)$ and $b(x_2,y_2)$, such as the Euclidian distance.

For the experimental investigation, we consider the two compared smallest layers to be the same if the Hausdorff distance is equal to zero. If the distance exceeds an empirically obtained threshold, the layers are considered to be different. Otherwise, they are considered to be similar.

3.3 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 one (with differences in the underlined and bold words), while the fourth is completely different. These texts are presented below:

> Text 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.
>
> Text 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.
>
> Text 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.
>
> Text 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.

After executing the aforementioned algorithm for text 1, as described in the Method section (see Figure 3), the latest onion polygon (colour green, see Figure 3) can be isolated.

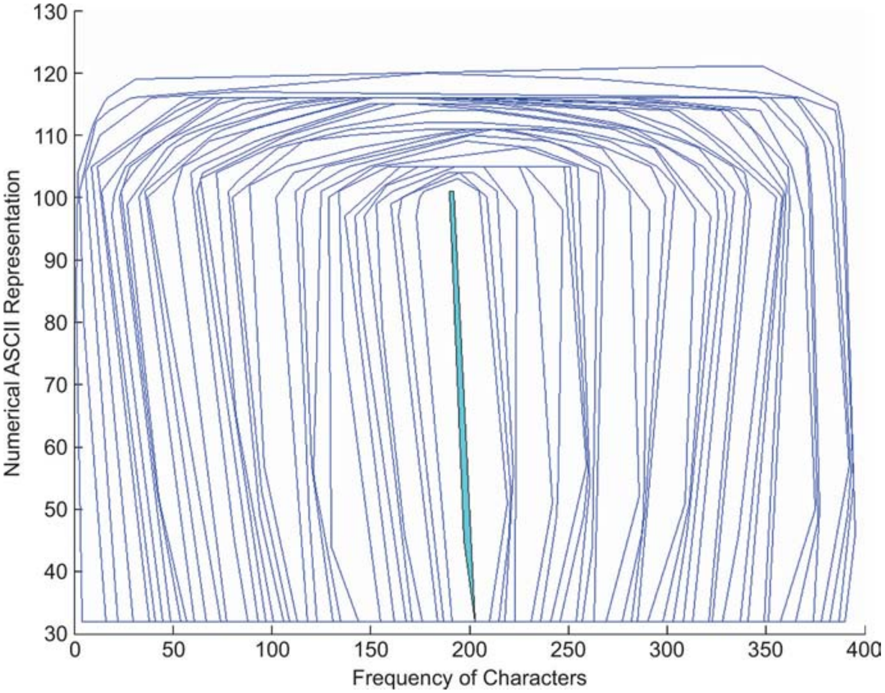
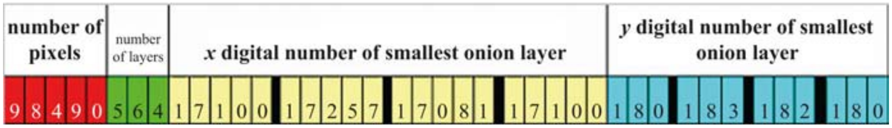


Figure 3.
Implementation of the onion polygon for Text 1

3.4 SCI incorporation in DOI

As shown in Figure 2 and explained in the Introduction, SCI can be employed as the DOI suffix string. SCI incorporation in DOI is extremely useful, as it adds semantic features. For example, DOI and SCI can form the following number for Text 1 of our experiment: 10.xxxxx/SCI70351921901972031921011014532101. In the aforementioned algorithmic implementation the numerical code is constructed via the coordinates of the green latest onion and is adapted in the particular proposed form (see Figure 4).

Finally, in Figure 5 the results of the forward Hausdorff distance (FHD) values are presented. According to these results, the following empirical rule can be formulated. Tthe proposed procedure utilises the numerical code for deciding whether the text is authentic, similar, or different from the reference text. In particular, if the number of words in the numerical code 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.



number of pixels	number of layers	*x* digital number of smallest onion layer	*y* digital number of smallest onion layer
9 8 4 9 0	5 6 4	1 7 1 0 0 ■ 1 7 2 5 7 ■ 1 7 0 8 1 ■ 1 7 1 0 0	1 8 0 1 8 3 ■ 1 8 2 ■ 1 8 0

Figure 4.
An example serial number extraction of Text 1

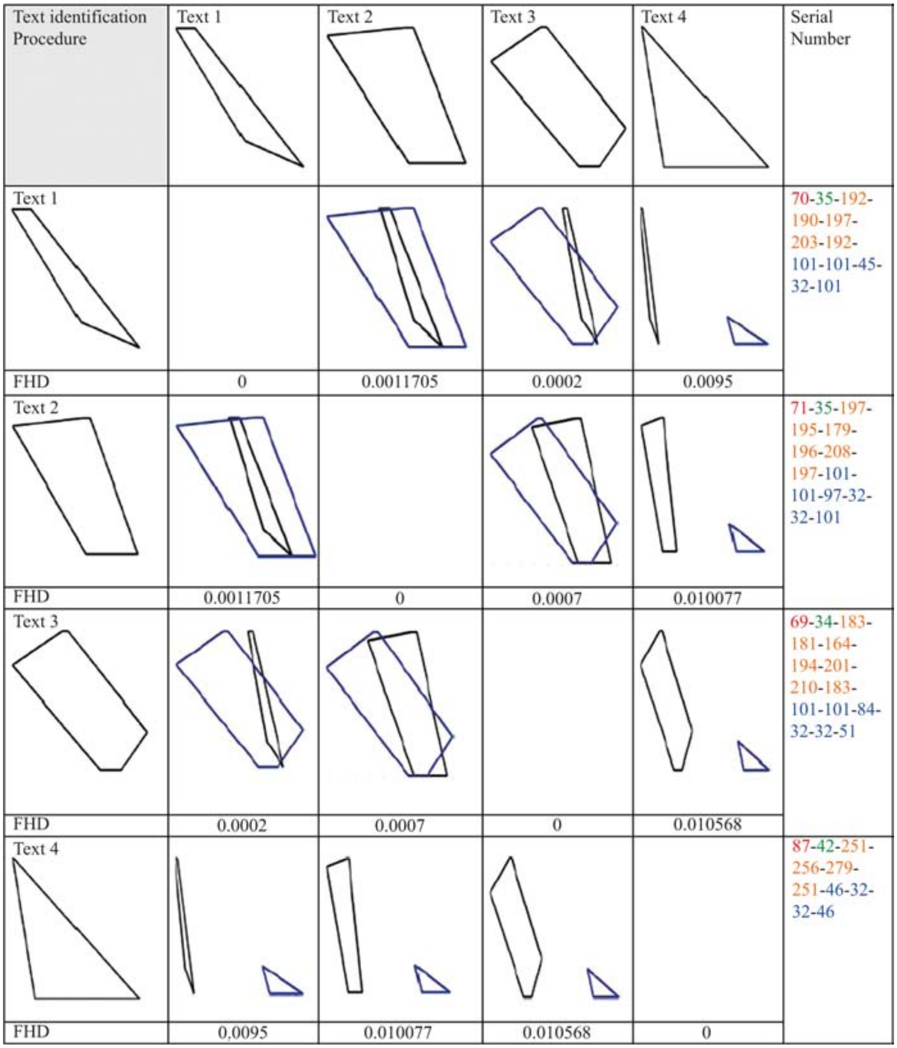


Figure 5.
The identification procedure between four texts

4. Algorithm's example application scenario

As mentioned in the introduction, the proposed numerical code (SCI) could be applied, as an application example, in the construction of the unique, Digital Object Identifier (DOI) for a text.

DOI is a system for identifying digital objects in a digital environment in a unique way. A DOI is an alphanumeric string, e.g. "10.1000/182", which can be assigned as an identifier to any digital entity, such as a text document, for use in digital networks.

However, a DOI Resource Metadata Declaration (RMD) is a message designed specifically for metadata exchange between registration agencies (RAs). The format may also be used for input of service metadata, but it is not intended as a replacement of other domain or service specific schemes. An RMD has the form of an XML document that conforms to an XML schema (xsd). All its elements and allowed values are mapped into the index Data Dictionary (iDD). In our case, this mechanism could be used to adapt the SCI model in the RDM form (see Figure 4). In greater detail, our extension is positioned in the current DOI metadata model as a mediator between the DOI resources and the RMD flows (see Figure 6).

The DOI provides current information, including the internet address where the entity can be found. The DOI is paired with the object's electronic address, or URL, in an updateable central directory, and is published in place of the URL in order to allow the content to move in another location without the need to change the link itself.

DOI's basic form has the following syntax: 10.xxxx/sssss. This sequence contains two sections: the prefix 10.xxxx, which involves the identification of the Publishing Authority, the "10." indicating that this string is a DOI alphanumeric string, and the suffix, which provides the capability for any meaningless or meaningful identification string to be added. The publishing authority usually adds the strings that identify the object in the suffix.

As an example application of the proposed numerical code for texts, the SCI may play the role of the suffix of the text's DOI. In particular, a valid DOI is 10.xxxx/sssss, in which "xxxx" is the code number of the registrant-publisher, e.g. 1039, and "sssss" is the unique alphanumeric string of the specific object, e.g. b505574c. The SCI can take the place of the "sssss". For instance, for Text 1 from the aforementioned example, which has SCI equal to 70-35-192-190-197-203-192-101-101-45-32-101, we could calculate the DOI as:

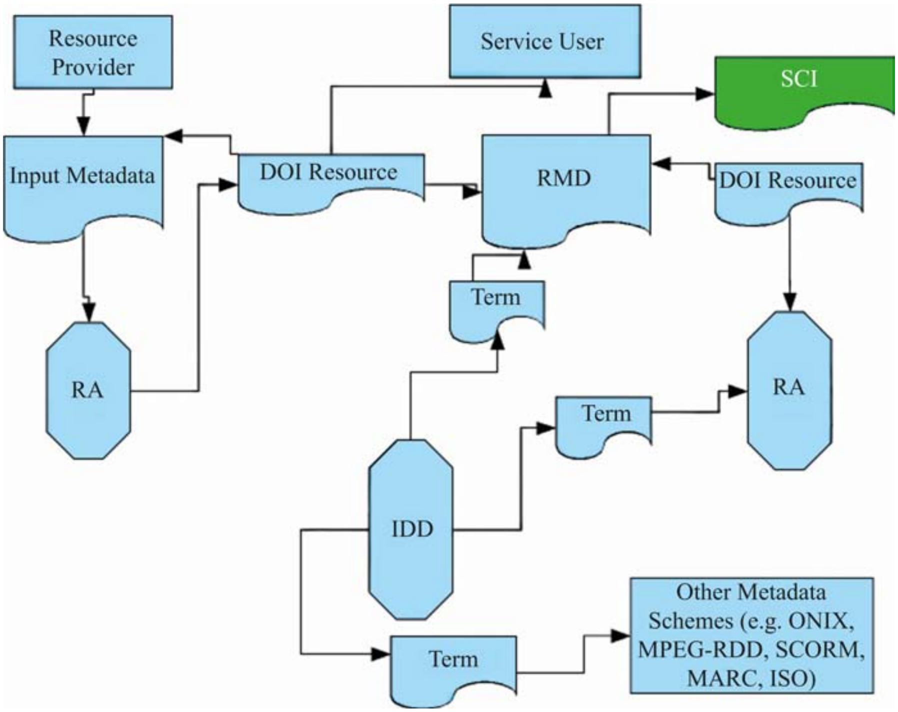


Figure 6.
The existing interface of DOI resource metadata declaration (RMD) using the SCI model of the standard DOI resource metadata flow

10.xxxx/SCI-70-35-192-190-197-203-192-101-101-45-32-101. With the existing model of RMD descriptions in the DOI model, an RMD representation of a text resource in the DOI model, enclosing the Hausdorff distance, named as SCI Criterion can be represented as:

```
Resource
   Identifier [“URI”]
      Subtype DOI
   SCIIdentifier [“70-35-192-190-197-203-192-101-101-45-32-101”]
      Subtype DOI
         SCI Criterion [“0 ≤ n ≥ 1”] (which is the Hausdorrf criterion)
         Subtype DOI
      Name[“1”]
         Subtype[“title”]
      Category [jpeg]
            Subtype [MimeType]
```

The resulting DOI suffix is: 10.xxxx/SCI-70-35-192-190-197-203-192-101-101-45-32-101.

Considering the text description limitations that DOI has, we consider the integration of the SCI extension as an extra field to the existing DOI mechanism an important improvement of the standard design RMD descriptors. With respect to the existing interfaces we consider the application of the proposed extension as a transformation of the simple DOI text retrieval mechanism to a more sophisticated tool, as for instance digital rights management. In such a case the proposed numerical code as part of the DOI suffix may play a meaningful role, since apart from being the part of a unique identifier, it can also be used to check the authenticity of the text it is assigned to.

4.1 Library retrieval case study

The proposed method could be applied in three virtual scenarios of text retrieval using realistic presuppositions. In particular, we considered a number of DOI database SCI model texts, which are registered in a database named Multimedia Ionian Library, as depicted in Figure 7.

- *Scenario of control identification of a text* – In this case we considered a text, for example 1.1.jpg, published with a DOI number

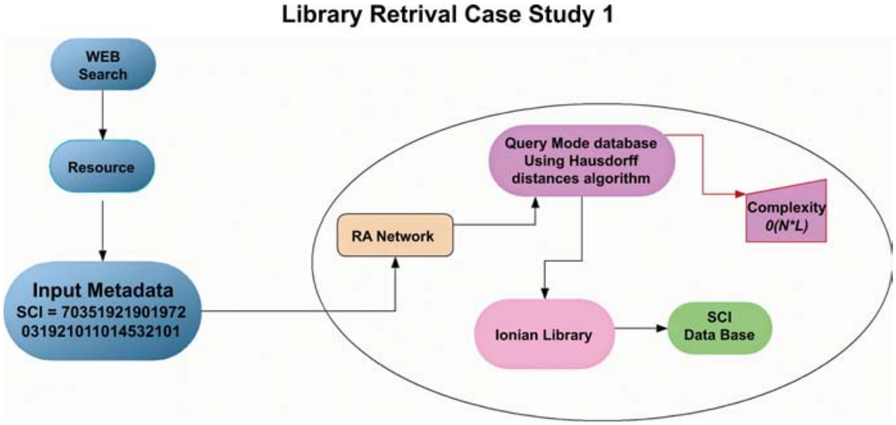

Library Retrival Case Study 1
WEB Search
Resource
Input Metadata
SCI = 70351921901972 031921011014532101
RA Network
Query Mode database Using Hausdorff distances algorithm
Complexity 0(N*L)
Ionian Library
SCI Data Base

Figure 7.
An application scenario of text retrieval using a DOI descriptor

10.xxxx/SC-70-35-192-190-197-203-192-101-101-45-32-101 in a web location. Then the identification of the source using the SCI identifier could be satisfied using the Hausdorff distance algorithm, setting as input criterion data $n = 0$.

- *Scenario for a collection of similar texts* – In this case, we extended this procedure for a selection of texts that obtained a particular degree of Hausdorff distance similarity $10 - 3 \leq n \leq 10 - 2$.
- *Scenario for a collection of different texts* – In this case, we extended this procedure for a selection of texts that obtained a particular degree of Hausdorff distance similarity $10 - 1 \leq \mathrm{n} \leq 10 - 2$.

5. Conclusions and future work

This method presented could be used as an efficient approach for detecting the copyright of documents, specifically in internet publications for which copyright is a significant factor (Poss *et al.*, 2004). It also addresses the needs of both the information services sector and the publishing industry.

As a possible future application (see Figure 8) the numerical code can be embedded into the suffix of the DOI system to enable the text retrieval capabilities through OpenURL queries. This solution was adopted because the DOI co-operates perfectly with such metadata information services as the CROSSREF and the protocol OPENURL. Specifically, it is known that an OpenURL consists of a base URL, which addresses the user's institutional link-server, and a query-string, which contains the data of this entry, typically in the form of key-value pairs. For example, an OpenURL Query that utilises the SCI Identifier can be as follows:

http://resolver.example.ionio.gr/cgi?content = text&SCI = 70351921901972031921011014532101&title = 1&format = text

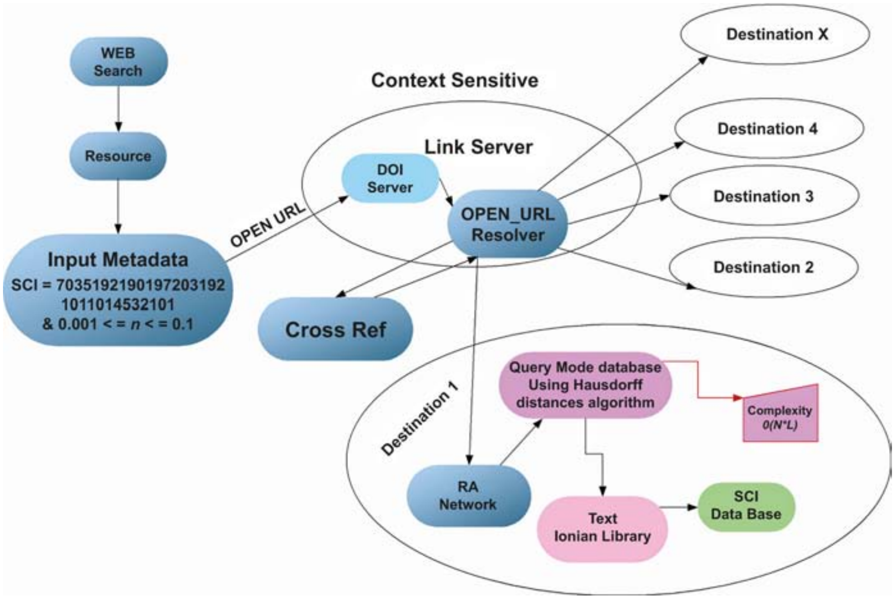


Figure 8.
An application scenario of text retrieval using OpenURL technology

In this work we have 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 numerical code for a text. For the implementation of this purpose we used the onion algorithm technique of computational geometry, and for the identification procedure we employed the Hausdorff distance algorithm. Empirical investigation showed that the proposed method may be used as an accurate method for identifying same, similar, or different conceptual texts.

This unique identification method for texts with the combination of SCI and DOI can be a solution to many problems that the information society faces, such as plagiarism, copyright related issues, and tracking (Stein *et al.*, 2007). It can also be useful in many aspects of the education process, such as in lesson planning and student evaluation.

The probability value provided by the proposed algorithm presenting the FHD helps in identifying the matching percentage between any content and copyright-protected content, as well as safeguarding the intellectual rights of authors and other digital rights owners. The advantages of the exported number are obvious, and we aim to highlight them while discussing the combination with DOI. Finally, this method may be used by the information services sector and the publishing industry for standard number definition identification, as a copyright management system, or both.

Future research, therefore, should be focused on further investigating the properties of the method, through experiments with large collections of documents. This way it will be possible to extensively evaluate the validity of the algorithm against a larger sample base and examine the utilisation of DOI as a metadata platform-co-operation model with other information networks such as neural networks. Our long-term goal is the adoption of the proposed model as a strategic component for organisations and libraries for the identification and control of the authenticity of electronically published documents on the web as well as the linkage with a users' information seeking behaviour (McKechnie *et al.*, 2006).

Notes

1. See www.doi.org
2. See www.issn.org
3. The SICI standard is developed by National Information Standards organisation of the United States (NISO); see www.niso.org

References

Atallah, M.J. (1983), "Dynamic computational geometry", *Proceedings of the 24th Annual Symposium on Foundations of Computer Science*, pp. 92-9.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999), *Modern Information Retrieval*, Addison-Wesley, Harlow.

Bawa, M., Bayardo, R.J., Agrawal, R. and Vaidya, J. (2009), "Privacy-preserving indexing of documents on the network", *The VLDB Journal*, Vol. 18, pp. 837-56.

Bellini, E., Cirinnà, C., Lunghi, M., Damiani, E. and Fugazza, C. (2008), "Persistent identifiers distributed system for cultural heritage digital objects", *Proceedings of iPRES 2008*, p. 242.

Dubuisson, M.P. and Jain, A.K. (1994), "A modified Hausdorff distance for object matching", *Proceedings of the International Conference on Pattern Recognition*, IEEE, Piscataway, NJ, p. 566.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovas, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C. and Solvber, I. (2007), "Evaluation of digital libraries", *International Journal on Digital Libraries*, Vol. 8, pp. 21-38.

Hill, C.W.L. (2007), "Digital piracy: causes, consequences, and strategic responses", *Asia Pacific Journal of Management*, Vol. 24, pp. 9-25.

Hilse, H.W. and Kothe, J. (2006), "Implementing persistent identifiers", European Commission on Preservation and Access, Amsterdam, available at: http://xml.coverpages.org/ECPA-PersistentIdentifiers.pdf

Huttenlocher, D.P., Klanderman, G.A. and Rucklidge, W.A. (1993), "Comparing images using the Hausdorff distance", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 850-63.

McKechnie, L., Julien, H., Pecoskie, J.L. and Dixon, C.M. (2006), "The presentation of the information user in reports of information behaviour research", *Information Research*, Vol. 12, p. 7.

Paskin, N. (2002), "Digital object identifiers", *Information Services and Use*, Vol. 22, pp. 97-112.

Poss, R., Bauer, T.W. and Heckman, J.D. (2004), "Copyright, ownership, and truth in data in the electronic age", *The Journal of Bone and Joint Surgery*, Vol. 86, p. 669.

Poulos, M., Bokos, G. and Vaioulis, F. (2008), "Towards the semantic extraction of digital signatures for librarian image-identification purposes", *Journal of the American Society for Information Science and Technology*, Vol. 59, pp. 708-18.

Poulos, M., Magkos, E., Chrissikopoulos, V. and Alexandris, N. (2003), "Secure fingerprint verification based on image processing segmentation using computational geometry algorithms", *Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition, and Applications, Rhodes*.

Poulos, M., Papavlasopoulos, S. and Chrissikopoulos, V. (2006), "A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm", *Journal of Digital Information*, p. 6.

Poulos, M., Rangoussi, M., Alexandris, N. and Evangelou, A. (2002), "Person identification from the EEG using nonlinear signal classification", *Methods of information in Medicine*, Vol. 41, pp. 64-75.

Stein, B., Koppel, M. and Stamatatos, E. (2007), "Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07", *ACM SIGIR Forum*, Vol. 41, pp. 68-71.

Tenopir, C. (2003), *Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies*, Council on Library and Information Resources, Washington, DC.

Corresponding author
Nikolaos Korfiatis can be contacted at: Korfiatis@em.uni-frankfurt.de

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**