A semantic self-organising webpage-ranking algorithm using computational geometry across different knowledge domains

Marios Poulos* and Sozon Papavlasopoulos

Archives and Library Sciences,
Ionian University,
Ioannou Theotoki 72, 49100, Corfu, Greece
E-mail: mpoulos@ionio.gr
E-mail: sozon@ionio.gr
Website: http://www.ionio.gr/~mpoulos
*Corresponding author

V.S. Belesiotis

Department of Informatics,
University of Piraeus,
80 Karaoli and Dimitriou Str.,
Piraeus 18534, Greece
E-mail: vbel@unipi.gr

Nikolaos Korfiatis

Department of Informatics,
Copenhagen Business School (CBS),
2300, Frederiksberg, Copenhagen, Denmark
E-mail: nk.inf@cbs.dk

Abstract: In this paper we introduce a method for Web page-ranking, based on computational geometry to evaluate and test by examples, order relationships among web pages belonging to different knowledge domains. The goal is, through an organising procedure, to learn from these examples a real-valued ranking function that induces ranking via a convexity feature. We consider the problem of self-organising learning from numerical data to be represented by a well-fitted convex polygon procedure, in which the vertices correspond to descriptors representing domains of web pages. Results and Statistical evaluation of procedure show that the proposed method may be characterised as accurate.

Keywords: webpage-ranking; semantic self-organising; computational geometry; information retrieval; educational applications.

Reference to this paper should be made as follows: Poulos, M., Papavlasopoulos, S., Belesiotis, V.S. and Korfiatis, N. (2009) 'A semantic self-organising webpage-ranking algorithm using computational geometry across different knowledge domains', *Int. J. Knowledge and Web Intelligence*, Vol. 1, Nos. 1/2, pp.24–47.

Biographical notes: Marios Poulos is an Assistant Professor of informatics in the Department of Archives and Library Science at the Ionian University. He received his BSc from the University of Athens, Greece (1986), and his PhD from the University of University of Piraeus (2003). In the period 2003–2008, he was member of staff (Adjuct Teatcher and Assistant Professor) in the Department of Archives and Library Science at the Ionian University. His research interests include medical informatics, semantic web and digital libraries. He is member of several technical committees and working groups on subject relates to medical informatics and information science area.

Sozon Papavlasopoulos is Lecturer of Informatics and Statistics in the Department of Archives and Library Sciences at the Ionian University. He received his BSc from the University of Thessaloniki, Greece (1978), his MSc Brunel University of London (1980) and his PhD from the Ionian University (2007). His research interests include medical informatics, semantic web-metadata, bibliometric, and pattern recognition. He has participated in several research projects funded by Greece or European Community. He also was member of several technical committees and working groups on subject relates to informatics and statistics. He is a member of several Societies.

Vasilios S. Belesiotis is a School Advisor for Informatics of the Greek Ministry of National Education and Religious Affairs for the Central Athens Greek Secondary Education. Moreover, he has been teaching for several years in the Department of Informatics of the University of Piraeus, including the subject of Didactics of Informatics. He holds a BSc Degree in Mathematics (1986), MSc (1978), and PhD Degrees in Informatics (2002). He has published over 20 papers and eight books, which have been used as textbooks both for secondary education and university courses. His research interests include education, didactics, knowledge representation and teaching.

Nikolaos Korfiatis is an Assistant Professor at the Department of Economics at the University of Copenhagen and an affiliated researcher at the Department of Informatics at Copenhagen Business School. He holds a PhD in Informatics from Copenhagen Business School (CBS) Denmark. He obtained his BSc from Athens University of Economics and Business and his MSc and Diploma in Information Systems Engineering from the Royal Institute of Technology (KTH), Stockholm in 2006. His research interests span the fields of information science and sociology with a particular emphasis to the analysis of online social networks.

1 Introduction

Undoubtedly, the web accumulates a vast amount of information resources making users heavily rely on the use of search engines to satisfy their information needs. Although the development of web retrieval mechanisms has been significantly extended to the area of understanding the meaning of content and its domain-specific properties, little attention has been given on how to include knowledge from the users themselves through

information sharing (Sidiropoulos et al., 2008). However, an important factor in that direction is the quality and the volume of the retrieved information since ambiguity in query terms might result in a large set of information being retrieved and send back to the users making the operation of a search engine a challenging task. A traditional method to deal with this situation is caching ((Katsaros and Manolopoulos, 2004, 2009; Sidiropoulos et al., 2008; Sivasubramanian et al., 2007; Vakali, 2001). Although caching offers several advantages like reduced network traffic and shorter response times, it has its drawbacks too, e.g., small hit rates (Sidiropoulos et al., 2008; Kroeger et al., 1997) and compulsory misses. To compensate for such problems, traditional caching is coupled with memory-based filtering, which aims at predicting future requests for web objects and bringing those objects into the cache in the background before a request is made for them. The most common perfecting practice is to make predictions based on the recent history of requests of individual clients, which is called short-term perfecting (Sidiropoulos et al., 2008; Nanopoulos et al., 2003).

From the side of ranking of the web objects, the Page-Rank algorithm (Page et al., 1999; Yang and King, 2006; Shivani and Agarwal, 2006) has proven to be very effective for ranking webpages. However, Page Rank order results can be often inaccurate owing to factors related with the incomplete information about the web structure that is searched and the query term that is provided. In particular, these accuracy-related problems can be attributed to the following factors:

- The web has a dynamic nature with respect to time availability of information resources (temporal dimension) – the link structure develops temporally. Some links are created and modified, whereas others are destroyed.
- The user interaction is partial and depends on the preferences that he or she has in relation with the place that he is currently residing (spatial dimension) – for different users (or crawlers), the web structure may be different. For example, a user who lives in Greece and searches for a product to purchase will be more interested to Greek electronic stores rather than international.
- The importance of the links is different (local dimension) – not all outbound-links have the same semantic strength. Some outbound-links are more significant than others.
- The learning procedure of webpage-ranking can be considered to be an inaccurate mechanism (Cohen et al., 1999; Herbrich et al., 2000; Crammer and Singer, 2002; Freund et al., 2003).

To complement these facts, the search engines have given emphasis on the development of internal taxonomies, which can contextualise the query results and terms. In practice, the main methods that are developed are focused on solutions that are based on Google directory taxonomic service.[1] The categories, which are created, suffer from the overlapping problem. In particular, we observe that the Google 'world of travel' suffers from problems similar to those of the Yahoo! directory: the categories represented 'together' do not seem to have much in common (see for instance: Attractions, Consolidators and Destinations) (Gilbert et al., 2003).

Several approaches, such as the enhanced Naïve Bayes classifier (Agrawal and Srikant, 2001), Co-Bootstrapping (Zhang and Lee, 2004a) and SVM-based methods (Zhang and Lee, 2004b), exploit the semantic overlap of corresponding categories to improve the categorisation performance. The basic idea of these approaches is that if the topics of classes A and B are known to be very similar, then there should be a large number of documents in A that also belong to B. For example, if a paper belongs to the Movie category of BBC news, it probably also belongs to the movie's category of Google directory (Shu-Hsien et al., 2008). To measure the distance between categories, we use the cosine similarity to determine the relationships and incorporate the information into a discriminative machine-learning model. Another method of web taxonomy integration uses the term vectors of neighbouring categories (Zhang and Lee, 2004b) to smooth a document's term vector with proper weights (Shu-Hsien et al., 2008). Other classification method Latent Semantic Indexing (LSI) leads to precision roughly equal to that of using a Naïve Bayesian approach; the LSI technique has a substantially higher recall and is more effective under certain conditions (Gee, 2003; Chakraborti et al., 2007).

Taking all this into account, in this paper, we introduce an algorithm, which combines the mechanism of information caching that comes from a dynamically self-organising learning system. To evaluate this approach, we constructed query scenarios based on a set of seven (7) different categories of Google directory[1] thematic titles in which we used as conceptual documents. In this set, we applied a well-tested computational geometric algorithm to improve the overlapping problem of taxonomy, which was described previously.

The aim of this method is to rank the specific features of webpages using an algorithm of computational geometry and, simultaneously, to cache this information in an efficient data structure for future retrieval based on similarity of query terms. Thus, in the retrieval procedure, the user will then be able to see groups of webpages according to specific common features.

Furthermore, our method promises to reduce complexity as it is based on a well-fitted algorithm taken from our latest studies (Poulos et al., 2004, 2007, 2008). This algorithm is tested on classification problems focusing on text categorisation (Poulos et al., 2004). This approach (Poulos et al., 2008) has yielded that the speed of retrieval is increased, as well as the reliability of the information obtained. For evaluation purposes, we compared the proposed algorithm that has a 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 (Bose and Toussaint, 1995), with the well-known LSI algorithm, which has determined complexity $O(m + n)q^2$ where n is the number of the word and m is the number of terms (Cai et al., 2006). Furthermore, the reliability of this method regarding LSI method is evaluated using non-parametric statistic indicators known as Coefficients of Concordance (Kendall, 1962). To this end, this paper is structured as follows: Section 2 provides information about the method, including the description of the algorithms' steps (processing–pre-processing stage). Section 3 provides the implementation of the method and the experimental part. Section 4 provides the statistical evaluation of the proposed method in comparison with the LSI algorithm. The paper concludes with Section 5 with discussion and steps for future research.

2 Method and algorithm description

The construction of the ranking method and the algorithmic implementation involves two particular initial stages: pre-processing and processing stage where the coordinates of the convex hulls are constructed. Then, in the implementation stage, the intersection of the convex hulls is evaluated and in the categorisation and identification stage the results are extracted. These methods are analysed in the following sections.

2.1 Pre-processing stage

In this stage, we assume that a selected text is an input vector $\vec{x} = (x_1, x_2, \ldots, x_n)$, where $x_1, x_2, \ldots, x_n$ represent the characters of the selected text. Then, using a conversion procedure where a symbolic expression (in our case an array of characters of a text) is converted to ASCII characters in string arithmetic values, we obtained a numerical value vector $\vec{S} = (s_1, s_2, \ldots, s_n)$ where these values ranged between 1 and 128. In our case, this conversion was achieved 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 give 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:

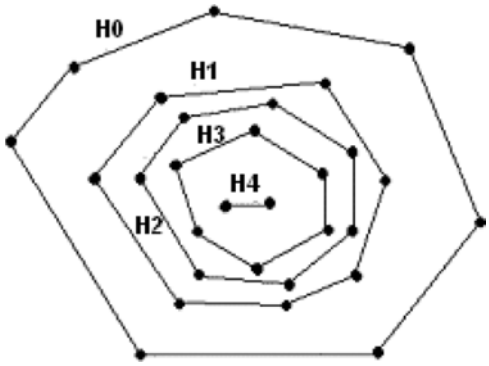
The set of elements of vector $\vec{S}$ for each selected text contains a convex subset, which has a specific position in relation to the original set (Poulos et al., 1999, 2003). This position may be determined by using a combination of computational geometry algorithms, which is known as Onion Peeling Algorithms (Bose and Toussaint, 1995) with overall complexity $O(d*n \log n)$ times, where d is the depth of the smallest convex layer and n is the number of characters in the numerical representation (in accordance with Section 2.1).

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.

2.3 Implementation

We consider the set of characters of a selected text to be vector $\vec{S}$. The algorithm (Graham, 1972) starts with a finite set of points $\vec{S} = \vec{S}_0$ in the plane. The following iterative process is considered. Let $\vec{S}_1$ be the set $\vec{S}_0 - \partial H(S_0) : S_0$ minus all the points on the boundary of the hull of $\vec{S}_0$. 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).

Figure 1 Onion layers of a set of points



2.4 Categorisation stage

The categorisation stage is divided into two concrete steps namely the formation of the onion layer for the given text and the evaluation of repeated intersections.

Stage 1: Formation of the onion layer

In our case, we considered that the smallest convex layer that has depth 3 (Figure 1) carries specific information (Poulos et al., 2003), because this position gives a geometrical interpretation of the semantics of the selected text. In other words, the smallest convex polygon (layer) depicts a particular geometrical area in which the values of these semantics range. This structure has also been used for example in statistics to define a generalisation of the concept of the median into two-dimensional data (O'Rourke et al., 1982). This feature may be characterised as unique to each selected text because the following two conditions (Poulos et al., 1999) are ensured:

- the selected area layer is non-intersected with another layer
- the particular depth of the smallest layer is variable in each case.

Thus, two variables were extracted from the proposed text categorisation method: the area of the smallest onion layer $\vec{S}_{xy}$ and the depth (i) of this layer, which is a subset of the original text set S values. In more detail, the smallest onion layer $\vec{S}_{xy}$ is a matrix that contains the set of convex coordinate values of the smallest onion layer, in other words,

$$\vec{S}_{xy} = (\vec{S}_{x_1y_1}, \vec{S}_{x_2y_2}, \ldots, \vec{S}_{x_iy_i})$$

where x represents the frequency of each converted character and y the numerical value of each character. Taking into account the specific features of the aforementioned variables, it is easy to ascertain that these may be used in the next stage.

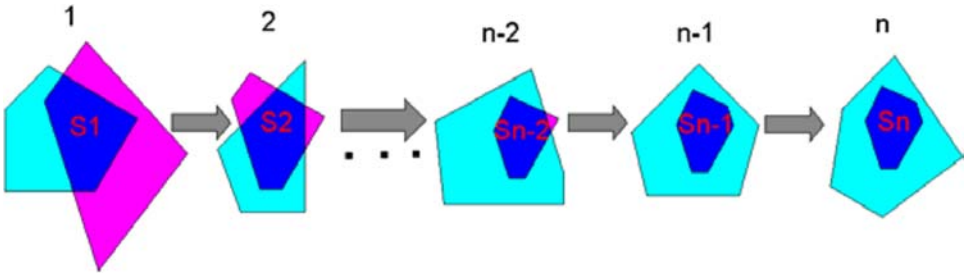
Stage 2: Repeated intersections

Let $\{x_n\}$ be the sequence of $n-1$ intersections of the $\{A_1, A_2, \ldots, A_n\}$ convex sets, where $\{x_n\}$ is

$$\{x_n\} = A_1 \cap A_2, (A_1 \cap A_2) \cap A_3, \ldots, (A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_{n-1}) \cap A_n$$

where A represents the set of the convex polygons, which have been produced in stage 1. Figure 2 provides the graphical representation of these repeated intersections. In our case, we considered that convex set $\{x_n\}$ represents the conceptual feature of each thematic area.

Figure 2 Repeated intersection until the conceptual convex $\{x_n\}$ is produced (see online version for colours)



2.5 Identification stage

An implementation of the well-known Hausdorff Distance algorithm (Normand and Bouillot, 1998) was adopted for the procedure of comparing the convex polygons' points. The creation of this algorithm was done exclusively for the computation of such problems as the computation of the mean Euclidean distance of two convex polygons in Euclidean space.

In this work, the convex pairs of values (x, y) represent both the smallest onion layer that came from the original audio file and the smallest onion layer that came from a second audio file. From the computation of the distance between these two sets of points it emerges if, and to what degree, the second audio file is identical to the original one.

Specifically, the illustration of this algorithm is analysed here:

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 Euclidean distance.

In this case, if $h(A, B) = 0$ all the points of the smallest polygon (Figure 2) of the two audio files that are compared are the same, then the two audio files are identical. Otherwise, the degree of similarity between the two audio files is decided by the value of Hausdorff distance (Normand and Bouillot, 1998). A typical indicator of high similarity is decided empirically to be any value equal to or less than 0.1, $h(A, B) \leq 0.1$ and in all other cases, in which the Hausdorff distance is bigger than 0.1, the audio files are considered different.

3 Method implementation: experimental part

In our case, for the sake of simplicity, we create a collection of titles that were extracted by Google Directory (Google, 2009). In particular, we aggregated seven different scientific titles (for example see Table 1) of the corresponding universal resource identifier – URI (Knowledge Management, Method Biophysics, Genetics, Methods and Techniques, Archaeology Methodology, Computational Geometry, Mathematical Physics) however many of these such as Mathematical Physics and Biophysics have common scientific areas creating an overlapping conceptual sectors. In the proposed method, we isolated all the titles of these categories and we isolated 28 titles of Biophysics category, 17 titles of the Genetics category, 32 titles of Mathematical Physics category, 14 titles of Computational Geometry category, eight titles of Methods and Techniques category, 16 titles of Archaeology Methodology category and ten titles of Knowledge Management category. Totally, we aggregated 125 conceptual titles. In the next step, each category is submitted in a number of repeated interactions as to the number of the titles per category such as described in Section 2.4 (see Table 2).

3.1 A scenario of searching procedure

To test the existing method, we test the extracted polygons with a combination of words that are belonging to two overlapping areas. For the implementation of this purpose, we used three pairs of words of which we extract corresponding convex polygons and we tested with extracted polygons using the Hausdorff distance (see Section 2.5) and then we obtain $3 \times 7 = 21$ sections (see Table 3). As can be seen, the evaluation of this scenario yields satisfactory results.

Table 1 Biophysics "http://directory.google.com/Top/Science/Biology/Biophysics/" Retrieved 5-4-09

Domain	*Title*
http://www.biophysics.org/	The Biophysical Society
http://www.bmrb.wisc.edu/	BioMagResBank
http://www.ebsa.org/	European Biophysical Societies' Association
http://www.smb.org/	Society for Mathematical Biology
http://physiology.med.cornell.edu/WWWVL/PhysioWeb.html	Physiology and Biophysics
http://www.iupab.org/	International Union for Pure and Applied Biophysics
http://en.wikipedia.org/wiki/Biophysics	Biophysics
http://www.ks.uiuc.edu/Research/biocore/	BioCoRE
http://www.uwo.ca/biophysics/	Medical Biophysics
http://www.bioelectromagnetics.org/	Bioelectromagnetics Society
http://nerve.bsd.uchicago.edu/	Electrophysiology and the Molecular Basis of Excitability
http://www.britishbiophysics.org.uk/	British Biophysical Society
http://xray.bmc.uu.se/sbnet/	Structural Biology Network (SBNet)
http://www.life.uiuc.edu/crofts/bioph354/index.html	Biological Energy Conversion
http://web-mcb.agr.ehime-u.ac.jp/english/biophys/default.htm	Electrophysiology of Plants
http://nanotech.about.com/science/nanotech/library/weekly/aa062500a.htm	Quantum Biology
http://www2.imperial.ac.uk/ssherw/physflow/pfn/	Physiological Flow Network
http://biophysics.centenary.edu/Biophysics%20course/index.htm	Biophysics and Bioimaging
http://www.lsbu.ac.uk/water/nucleic.html	Nucleic Acid Hydration
http://www.isebi.org/	International Society for Electrical Bio-Impedance
http://www.life.uiuc.edu/biophysics/ibs/	UIUC Chapter of Illinois Biophysics Society
http://home.hccnet.nl/ja.marquart/	SPR Web Pages
http://www.geocities.com/ResearchTriangle/Node/5345/	Thermosynthesis
http://centrofermi-nmr.phys.uniroma1.it/	International School on Magnetic Resonance and Brain Function
http://www.bio-epidermis.org	Information Biophysics Epidermis
http://www.physicsplanet.com/articles/biophysics	Biophysics
http://web.mac.com/thorkona/Water_in_Trees/Water_In_Trees.html	How Does Water go Up in a Tree?
http://biomag.wikidot.com/	Biomag: Biomagnetic Wiki and Forum
http://www.physiology-physics.blogspot.com/	Physiology: Physics Woven Fine

Table 2 Procedure of repeated intersection of url title obtained by Google categories (see online version for colours)

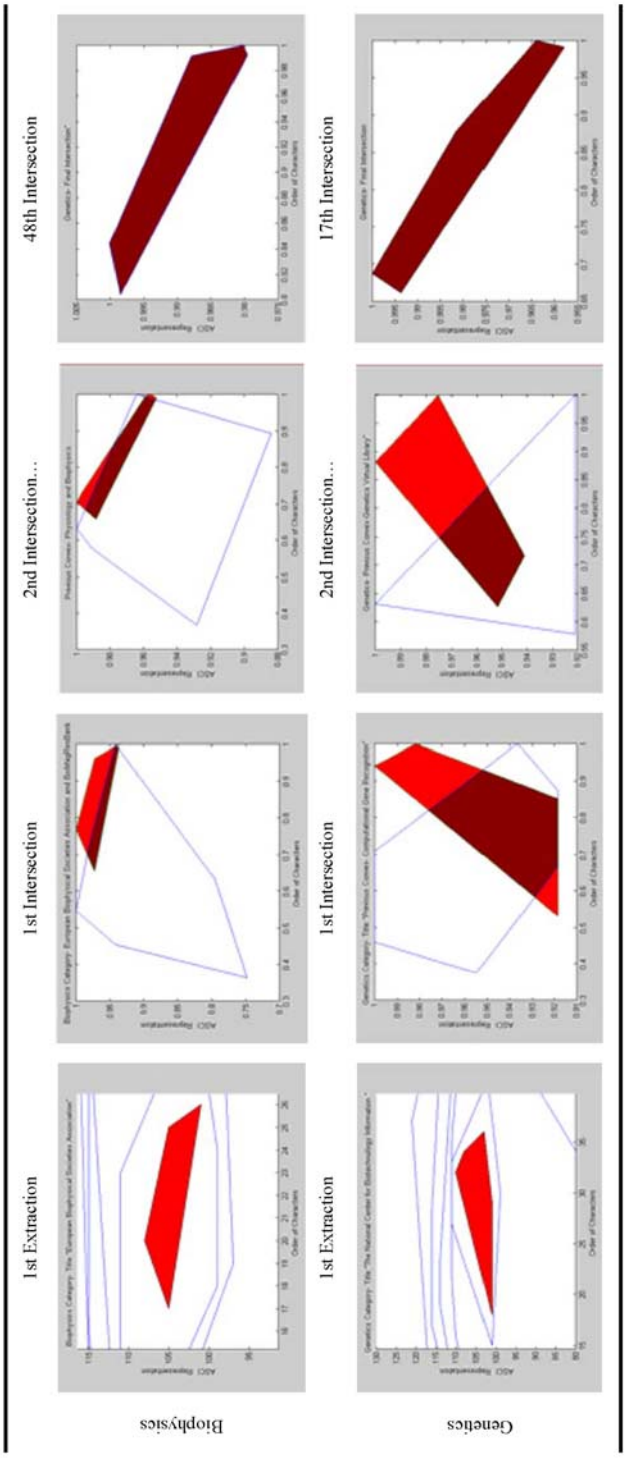


Table 2 Procedure of repeated intersection of url title obtained by Google categories (see online version for colours) (continued)

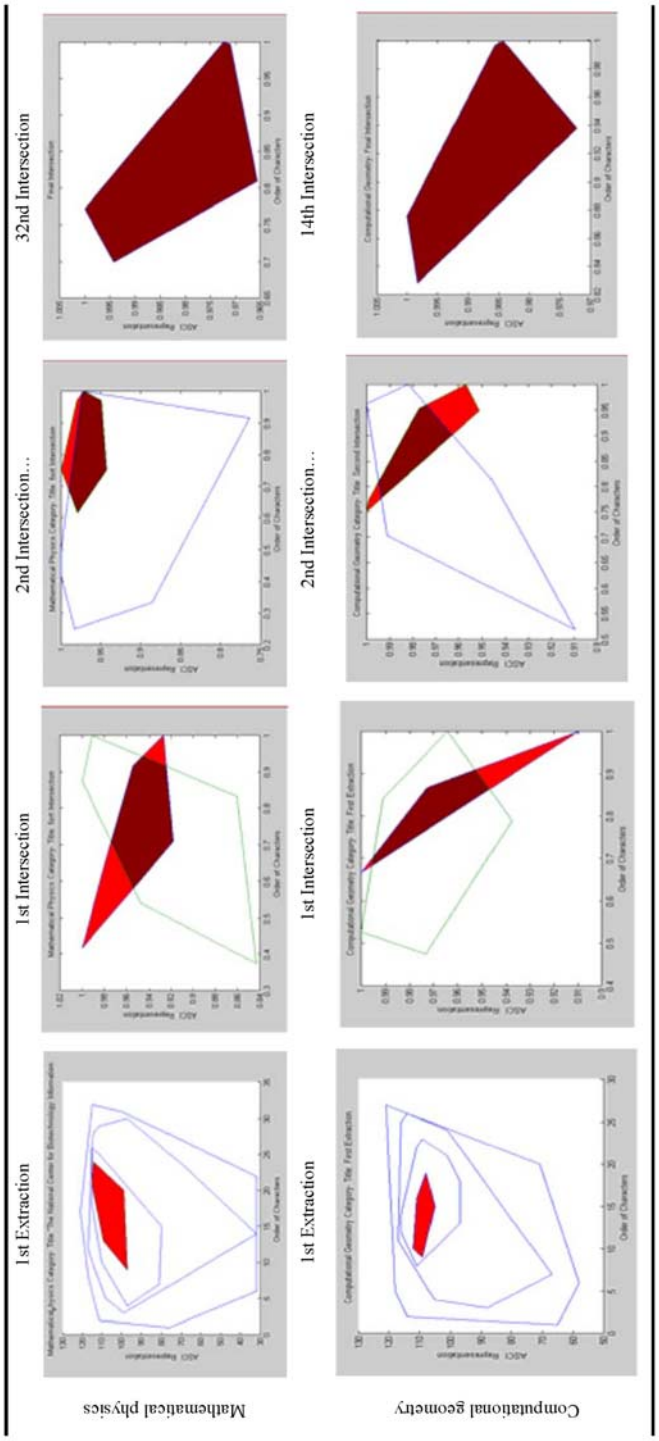


Table 2 Procedure of repeated intersection of url title obtained by Google categories (see online version for colours) (continued)

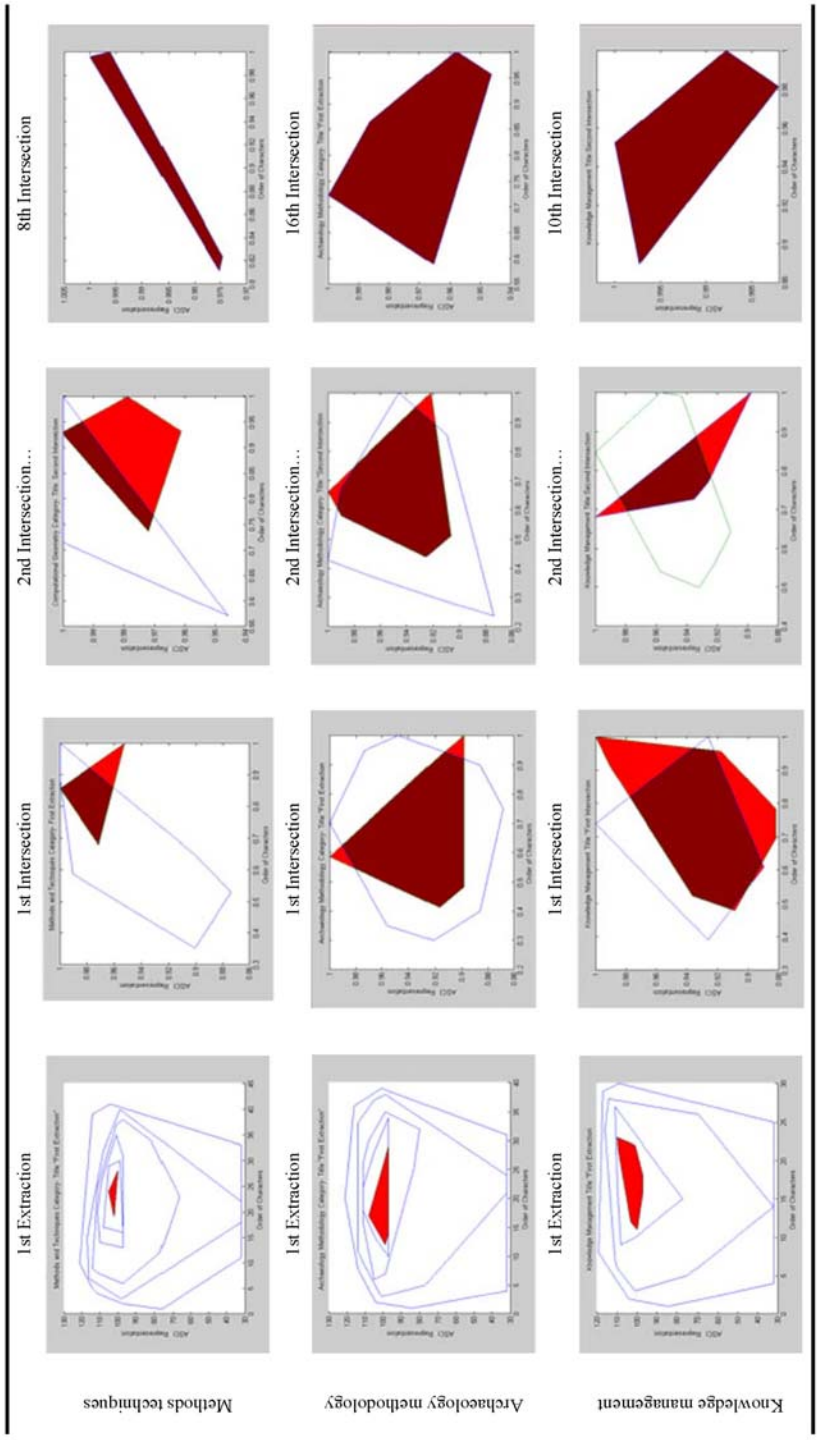


Table 3 A scenario searching using pair of words (see online version for colours)

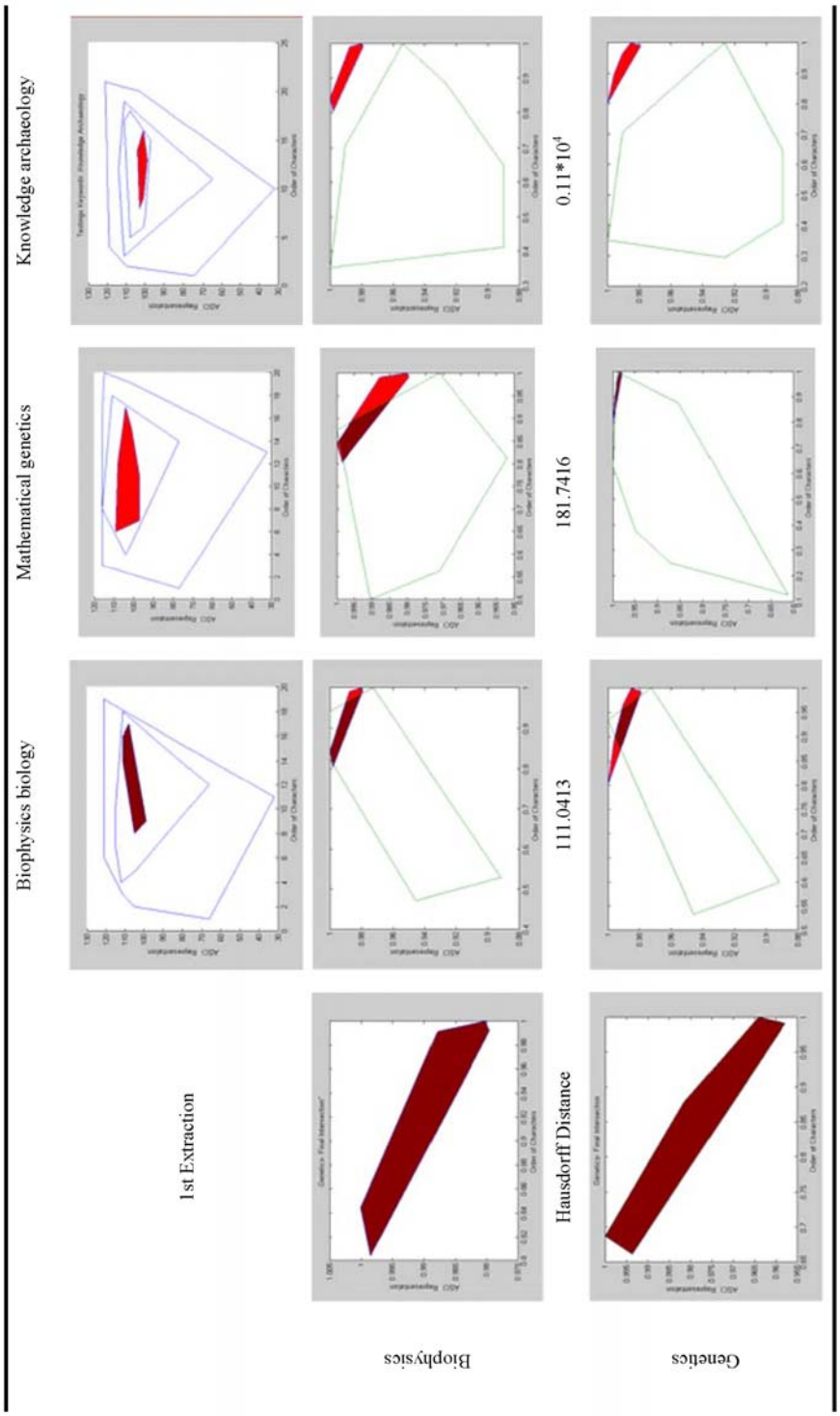


Table 3 A scenario searching using pair of words (see online version for colours) (continued)

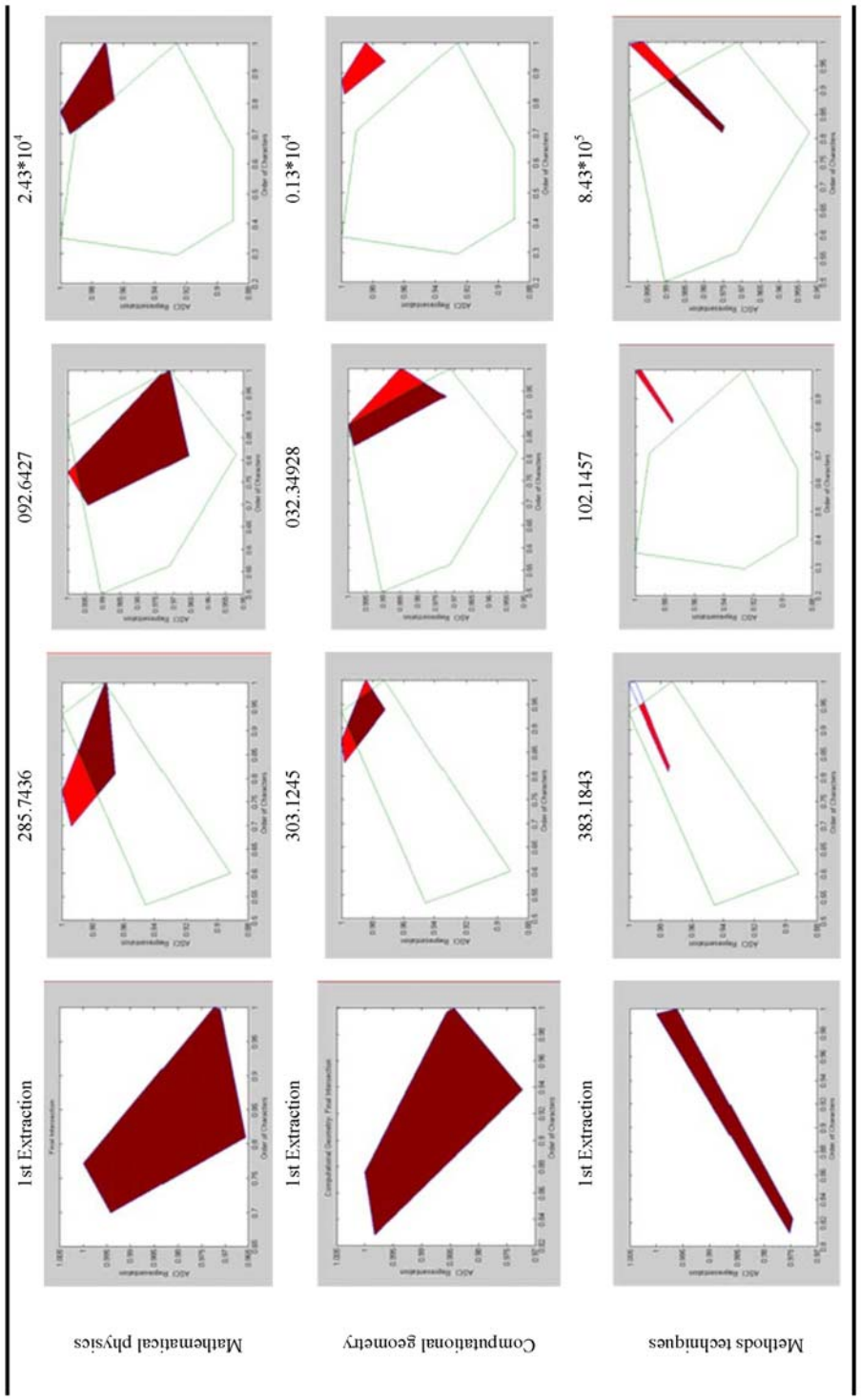
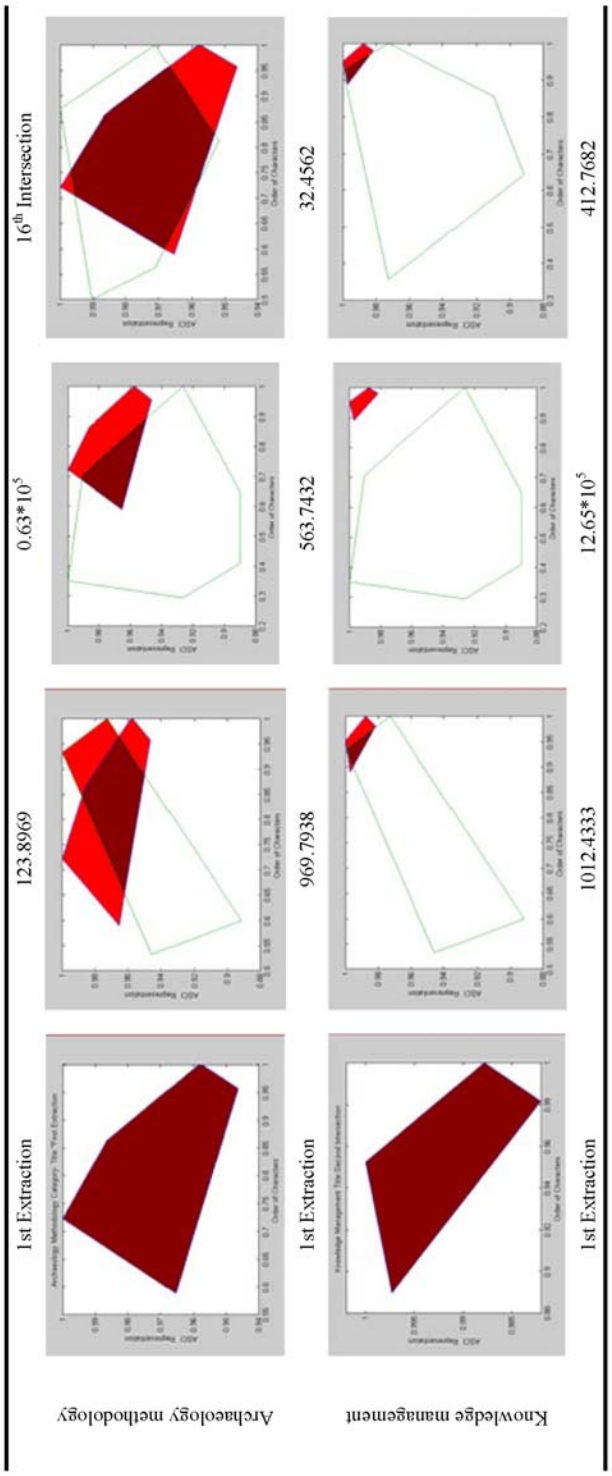


Table 3 A scenario searching using pair of words (see online version for colours) (continued)



3.2 Latent Semantic Indexing

In the fist stage, we created Table 4 to sort out all the conceptual contents of the documents. In particular, the local weight (L_{ij}) of each term is represented by each row in Table 3 and every document is represented by each column. Thus, an individual entry in Table 3, a_{ij}, represents the weights of the terms i in document j, where, in our case, $i = 1, \ldots, 57$ and $j = 1, \ldots, 7$. More specifically, the term – document matrix – can be expressed as a vector consisting of the weights of each term and mapped in a vector space (Tokunaga, 1999):

$$A = \begin{pmatrix} L_{11} & \cdots & L_{1j} \\ \vdots & \ddots & \vdots \\ L_{i1} & \cdots & L_{ij} \end{pmatrix} \tag{1}$$

where L_{ij} is the local weighting. This approach defines the weights w_{ij} in each document. Let P_{ij} denote the occurrence probability of t_i (terms frequency) and d_j (titles). We ascribe significance to a term's occurrence, on the grounds that it represents a document's topics more than other factors do. Therefore, we base w_{ij} on the term occurrence probability P_{ij} in each document, and we define a local weighting L_{ij} as follows:

$$L_{ij} = P_{ij} \times \log(1 + P_{ij}). \tag{2}$$

Table 4 Hausdorff distance calculation

Texts	*Biophysics biology*	*Mathematical genetics*	*Knowledge archaeology*
Biophysics	111.0413	181.7416	0.11×10^4
Genetics	285.7436	092.6427	2.43×10^4
Mathematical physics	303.1245	032.34928	0.13×10^4
Computational geometry	383.1843	102.1457	8.43×10^5
Methods techniques	123.8969	0.63×10^5	9.13×10^4
Archaeology methodology	969.7938	563.7432	32.4562
Knowledge management	1012.4333	12.65×10^5	412.7682

4 Statistical evaluation

In this section, a statistical evaluation between the most used information retrieval method "Latent Semantic Indexing (LSI)" and the proposed method is attempted. In particular, we tested the seven (7) semantic areas of titles with the pair of selected words. For the comparison of two techniques, we used Rank Correlation among several variables of the Chi-square control. In the following sections, we develop the procedure followed to yield the normalised matrixes for each method and, thereinafter, apply them to the selected statistical method.

We constructed two equivalent tables of each method. In particular, the Hausdorff distances of the proposal method (using computational geometry) are presented in Table 4 and in normalised transformation in Table 5 (3 × 7 of dimensionality).

Table 5 Normalisation of Hausdorff distance values

Texts	*Biophysics biology*	*Mathematical genetics*	*Knowledge archaeology*
Biophysics	0.0836	0.1368	0.0086
Genetics	0.2151	0.0697	0.1903
Mathematical physics	0.2282	0.0244	0.0102
Computational geometry	0.2885	0.0769	0.6664
Methods techniques	0.0933	0.0498	0.7149
Archaeology methodology	0.7301	0.4244	0.0244
Knowledge management	0.7622	1.0000	0.3108

The LSI method uses an algebraic model of document retrieval, using a matrix technique known as singular value decomposition. This technique is achieved by a normalised matrix (Table 6) of dimensionality 3 × 7 to conceptually compare the LSI method with our proposed method. For the construction of the values of this table, we show that if two LSI probabilities of the words *A* and *B* are independent then the joint probability is

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B). \tag{3}$$

Table 6 The normalise values of LSI output probability

Texts	*Biophysics biology*	*Mathematical genetics*	*Knowledge archaeology*
Biophysics	0.0534 × 0.1982 = 0.0106	0.0172 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000
Genetics	0.0172 × 0.0000 = 0.0000	0.2689 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000
Mathematical physics	0.0000 × 0.0000 = 0.0000	0.1163 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000
Computational geometry	0.0172 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000
Methods techniques	0.0172 × 0.000 = 0.0000	0.0000 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000
Archaeology methodology	0.0000 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000	0.0000 × 0.2062 = 0.0000
Knowledge management	0.0000 × 0.0000 = 0.0000	0.0000 × 0.0000 = 0.0000	0.0724 × 0.0000 = 0.0000

Then, we selected the LSI probabilities of the selected words, and the output of equation (3) consists of the elements of Table 7.

Table 7 LSI calculations

	Knowledge management		*Methods and techniques*		*Archaeology methodology*		*Computational geometry*		*Mathematical physics*		*Biophysics*		*Genetics*	
Terms	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}
Information	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0179	0	0.0000
Net	1	0.0172	1	0.0179	0	0.0000	0	0.0000	0	0.0000	2	0.0350	0	0.0000
Knowledge	4	0.0724	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Board	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
wikipedia	2	0.0350	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0179	0	0.0000
Managment	4	0.0724	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Forum	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0179	0	0.0000
Research	1	0.0172	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Programme	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Archive	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Lab	0	0.0000	3	0.0534	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Protocol	0	0.0000	4	0.0724	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Book	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Protein	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Molecular	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000	1	0.0172	1	0.0172
Biology	0	0.0000	1	0.0172	0	0.0000	1	0.0172	0	0.0000	3	0.0534	1	0.0172
Biomedical	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Reasearch	1	0.0172	1	0.0172	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Archaeology	0	0.0000	0	0.0000	10	0.2062	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Social	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000

Table 7 LSI calculations (continued)

	Knowledge management		*Methods and techniques*		*Archaeology methodology*		*Computational geometry*		*Mathematical physics*		*Biophysics*		*Genetics*	
Terms	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}
Photogrammetry	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Dental	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Microwear	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Mailing	0	0.0000	0	0.0000	2	0.0350	0	0.0000	0	0.0000	0	0.0000	0	0.0000
List	0	0.0000	0	0.0000	2	0.0350	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Alloy	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Experimental	0	0.0000	0	0.0000	2	0.0350	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Method	0	0.0000	1	0.0172	1	0.0172	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Computational	0	0.0000	0	0.0000	0	0.0000	8	0.1600	0	0.0000	0	0.0000	1	0.0172
Geometry	0	0.0000	0	0.0000	0	0.0000	9	0.1828	5	0.0919	0	0.0000	0	0.0000
Algorithm	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000
Literature	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000
Voronoi	0	0.0000	0	0.0000	0	0.0000	2	0.0350	0	0.0000	0	0.0000	0	0.0000
Diagram	0	0.0000	0	0.0000	0	0.0000	1	0.0179	0	0.0000	0	0.0000	0	0.0000
Toussaint	0	0.0000	0	0.0000	0	0.0000	2	0.0350	0	0.0000	0	0.0000	0	0.0000
Database	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000	1	0.0172
Polygon	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000	0	0.0000
Maths	0	0.0000	0	0.0000	0	0.0000	0	0.0000	6	0.1163	1	0.0172	0	0.0000
Physics	0	0.0000	0	0.0000	0	0.0000	0	0.0000	10	0.1982	1	0.0172	0	0.0000
Solutions	0	0.0000	0	0.0000	0	0.0000	0	0.0000	3	0.0534	0	0.0000	0	0.0000

Table 7 LSI calculations (continued)

	Knowledge management		*Methods and techniques*		*Archaeology methodology*		*Computational geometry*		*Mathematical physics*		*Biophysics*		*Genetics*	
Terms	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}	*T/O*	L_{ij}
Equation	0	0.0000	0	0.0000	0	0.0000	0	0.0000	3	0.0534	0	0.0000	0	0.0000
Algebra	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Function	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Model	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Analysis	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Topology	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Applied	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000	0	0.0000
Biophysical	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	10	0.1982	0	0.0000
Physiology	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000
Bioelectromagnetic	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	2	0.0350	0	0.0000
Electrophysiology	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000
Thermosynthesis	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172	0	0.0000
Biotechnology	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172
DNA	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172
Code	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172
Genome	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	13	0.2689
Polygenic	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172
Chromosome	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172
Virtual	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1	0.0172

4.1 Concordance: rank correlation among several variables

The concept of correlation between two variables can be expanded to consider association among more than two, as shown for multiple correlations (Kendall, 1962). Such association is readily measured non-parametrically by a statistic known as Coefficients of Concordance. In our case, this correlation, W, gave the perfect criterion of agreement among three (3) selected sets of data. In other words, this test investigates if the data of columns are in agreement. The aforementioned correlation W was employed by the following equation:

$$W = \frac{\sum R_i^2 - \frac{\left(\sum R_i\right)^2}{n}}{\frac{M^2(n^3 - n)}{12}} \tag{4}$$

where M is the number of variables being correlated and n is the number of data per variable, R_i is the sum of the values of each column for $i = 1, \ldots, n$.

We can ask whether a calculated sample, W, is significant, i.e., whether it represents an association different from zero in the population that was sampled. A simple way to find out the relationship between the Kendall Coefficients of Concordance, W, and the Friedman Chi-square x_r^2 involves using the following formula:

$$x_r^2 = M(n-1)W. \tag{5}$$

Therefore, we can convert the calculated W to its equivalent x_r^2 and then employ the critical values of x_r^2.

4.2 Implementation of concordance: rank correlation control

In the next stage, the data of Tables 5 and 6 can be tested using the top-down correlation technique (see equation (4)). The null hypothesis (H_0) of this test is that there is an agreement regarding the investigated data of the columns of Table 4. In our case, according to equation (2), $M = 3$ was the number of investigated documents, $n = 7$ was the conceptual values after reduction. The results of this control are present as follows:

Computational Geometry Results

Then, $W = 0.029$ and thus $x_r^2 = 3(7-1) \times 0.029 = 0.5227$ (see equation (3)). Comparing it with the Chi-square distribution with a degrees of freedom $i = 3 - 1 = 2$ and $(x_r^2)_{0.05,7} = 0.829$, we had non-rejection of the null hypothesis (H_0) of the agreement regarding the conceptuality of the compared documents ($0.99 < P < 0.95$).

Latent Semantic Indexing (LSI) Results

Then, $W = 0.0034$ and thus $x_r^2 = 3(7-1) \times 0.0034 = 0.0612$ (see equation (3)). Comparing it with the Chi-square distribution with a degrees of freedom $i = 3 - 1 = 2$ and $(x_r^2)_{0.05,7} = 0.829$, we had non-rejection of the null hypothesis (H_0) of the agreement regarding the conceptuality of the compared documents ($0.99 < P < 0.95$).

5 Discussion: further research

In this paper, an algorithm is described that is based on a suitably transformed algorithm of computational geometry (Poulos et al., 2004), and which aims to introduce a new webpage-ranking approach.

According to the problems, a webpage connection mechanism has been developed with a property as criterion. This property is the semantic values, which can be acquired from the suitable application of the webpages on the Cartesian level after repeated intersections.

This connection mechanism was realised via an algorithm, which was reformed from computational geometry. With this proposal, we can assume that the problems of the overlapping page rank taxonomy can be improved in two levels.

In the first level, the statistical results showed that the computational geometric solution gave more stable significant probability of the taxonomic ability of the system. Then, the LSI method in order 0.522 >> 0.0612. Furthermore, as justified in introduction part the proposed method may be characterised as more fast because the complexity of proposed method is significantly less than LSI method or $O(d^*n \log n) << O(m+n)q^2$.

The utility of this mechanism can be found in the dynamic self-organising feature of the algorithm, which automatically ranks the webpages into groups. Furthermore, the web retrieval procedure is simplified because the user may select the relevant group of pages from a more limited choice of homogeneous groups by entering specific keywords, which are represented by unique convexities.

A particular future application of the discussed method and algorithmic implementation could be used on the Educational/computer-based training domain and, in particular, the evaluation of essays by students using some key characteristics as grading indicators. For example, an expert could provide a set of significant keywords/references required by an essay to be evaluated as pass and then the algorithmic method described could provide an indication as to which essays meet the minimum criteria or not. In this application domain, another interesting application of the proposed method is on the recommendation of new pages for a particular user based on the degree of similarity obtained by running the algorithm on a predefined set of the user's browsing history. For example, in the case of a set of books, an application scenario could involve the recommendation of new books and learning resources on a user for that particular domain, since cases where domain membership is ambiguous the algorithm could compute the distances with the user's history as a reference point.

Finally, in the current and future research plan, we aim to test this proposed method against a larger number of webpages and data sets obtained for classification purposes. One additional objective that we have is to implement this algorithm in an ontological approach taking also semantic-based similarity indexes as an input.

Acknowledgements

The authors thank the referees for their very helpful comments and suggestions.

References

Agrawal, R. and Srikant, R. (2001) *Method and System for Merging Hierarchies*, US Patent No. 6,687,705, Issued 3 February, 2004 (filed January 8).

Bose, P. and Toussaint, G. (1995) 'No quadrangulation is extremely odd', *Proceedings in 6th International Symposium on Algorithms and Computation (formerly SIGAL International Symposium on Algorithms)*, 4–6 December, Cairns, Australia, pp.340–358.

Cai, D., He, X. and Han, J. (2006) 'Tensor space model for document analysis', *Proceedings of the 29th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Seattle, Washington, USA, August 06–11, 2006). SIGIR '06*, ACM, New York, NY, pp.625, 626, DOI= http://doi.acm.org/10.1145/1148170.1148287

Chakraborti, S., Mukras, R., Lothian, R., Wiratunga, N., Watt, S. and Harper, D. (2007) 'Supervised latent semantic indexing using adaptive sprinkling', *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, 6–12 January, Hyderabad, India, pp.1582–1587.

Cohen,W.W., Schapire, R.E. and Singer, Y. (1999) 'Learning to order things', *Journal of Artificial Intelligence Research*, Vol. 10, pp.243–270.

Crammer, K. and Singer, Y. (2002) 'Ranking with ranking', *Proceedings of the Advances in Neural Information Processing Systems*, 23–26 July, Edmonton, Alberta, Canada, pp.133–142.

Freund, Y., Iyer, R., Schapire, R.E. and Singer, Y. (2003) 'An efficient boosting algorithm for combining preferences', *Journal of Machine Learning Research*, Vol. 4, pp.933–969.

Gee Kevin, R. (2003) 'Using latent semantic indexing to filter spam', *SAC'03: Proceedings of ACM Symposium on App. Comp.*, 9–12 March, ACM Press, Melbourne, Florida, pp.460–464.

Gilbert, A., Gordon, M., Paprzycki, M. and Wright, J. (2003) 'The world of travel: a comparative analysis of classification methods', *Annales UMCS Informatica*, pp.393–398.

Graham, R.L. (1972) 'An efficient algorithm for determining the convex hull of a finite planar set', *Inform, Process, Lett.*, pp.132–133.

Herbrich, R., Graepel, T. and Obermayer, K. (2000) 'Large margin rank boundaries for ordinal regression', *Advances in Large Margin Classifiers*, pp.115–132.

Katsaros, D. and Manolopoulos, Y. (2004) 'Caching in web memory hierarchies', *Proceedings of the ACM Symposium on Applied Computing (SAC)*, 14–17 March, Nicosia, pp.1109–1113.

Katsaros, D. and Manolopoulos, Y. (2009) 'Prediction in wireless networks by Markov chains', *IEEE Wireless Communications Magazine*, Vol. 16, No. 2, pp.56–64.

Kendall, M.G. (1962) *Rank Correlation Methods*, 3rd ed., Charles Griffin, London, p.199.

Kroeger, T., Long, D.E. and Mogul, J. (1997) 'Exploring the bounds of web latency reduction from caching and perfecting', *Proceedings of the USENIX Symposium on Internet Technologies and Services (USITS)*, 8–11 December, Monterey, pp.2–2.

Nanopoulos, A., Katsaros, D. and Manolopoulos, Y. (2003) 'A data mining algorithm for generalized web prefetching', *IEEE Trans. Knowl. Data Eng.*, Vol. 15, No. 5, pp.1155–1169.

Normand, G. and Bouillot, M. (1998) 'Hausdorff distance between convex polygons', *Computational Geometry Web Project*, Retrieved 8 January, 2008, from http://www.cgrl.cs.mcgill.ca/godfried/teaching/cg-projects/98/normand/main.html

O'Rourke, J., Chien, J., Olson, C. and Naddor, T. (1982) 'A new linear algorithm for intersecting convex polygons', *Computer Graphics and Image Processing*, Vol. 19, No. 4, pp.384–391.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1999) *The Pagerank Citation Ranking: Bringing Order to the Web*, Tech. Rep. Stanford Digital Library Technologies Project, Paper SIDL-WP-1999-0120 (version of 11/11/1999).

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 (ICECS99)*, Institute of Electrical and Electronics Engineers, Pafos, Cyprus, Vol. 2, pp.1005–1012.

Poulos, M., Papavlasopoulos, S., Chrissicopoulos, V. and Magkos, E. (2003) 'Fingerprint verification based on image processing segmentation using an onion algorithm of computational geometry', *Sixth Int. Conf. on Mathematics Methods in Scattering Theory and Biomedical Technology (BIOTECH'6) Tsepelovo-Ioannina*, Word Scientific, Singapore.

Poulos, M., Bokos, G., Kanellopoulos, N., Papavalsopoulos, S. and Avlonitis, N. (2007) 'Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes', *Journal of Graph Algorithms and Applications*, Vol. 11, No. 1, pp.277–307.

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, No. 5, pp.708–718.

Poulos, M., Papavlasopoulos, S. and Chrissicopoulos, V. (2004) 'A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm', *Journal of Digital Information (JoDI)*, Vol. 6, No. 1.

Shivani, S. and Agarwal, R. (2006) 'Ranking on graph data', *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, pp.25–32.

Shu-Hsien, L., Chen, C-M. and Wu, C-H. (2008) 'Mining customer knowledge for product line and brand extension in retailing', *Expert Systems with Applications*, Vol. 34, No. 3, pp.1763–1776.

Sidiropoulos, A., Pallis, G., Katsaros, D, Stamos, K., Vakali, A. and Manolopoulos, Y. (2008) 'Prefetching in content distribution networks via web communities identification and outsourcing', *World Wide Web*, Vol. 11, No. 1, pp.39–70.

Sivasubramanian, S., Pierre, G., van Steen, M. and Alonso, G. (2007) 'Analysis of caching and replication strategies for web applications', *IEEE Internet Computing*, Vol. 11, No 1, pp.560–566.

Tokunaga, T. (1999) *Computation and Language Volume 5: Information Retrieval and Natural Language Processing*, University of Tokyo Press, Tokyo Japan, pp.234.

Vakali, A. (2001) 'Proxy cache replacement algorithms: a history-based approach', *World Wide Web J.*, Vol. 4, No.4, pp.277–298.

Yang, H. and King, I. (2006) 'Predictive random graph ranking on the web', *Proceedings of International Joint Conference on Neural Networks*, 16–21 July, Vancouver, Canada, pp.1825–1832.

Zhang, D. and Lee, W.S. (2004a) 'Web taxonomy integration through co-bootstrapping', *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 25–29 July, Sheffield, United Kingdom, pp.410–417 [doi>10.1145/1008992.1009062].

Zhang, D. and Lee, W.S. (2004b) 'Web taxonomy integration using support vector machines', *Proceedings of the 13th international conference on World Wide Web*, 17–20 May, New York, NY, USA, pp.472–481 [doi>10.1145/988672.988736].

Notes

[1]http://directory.google.com/Top/Health/Medicine/Basic_Sciences/, retrieved (5-4-09).

[2]http://directoty.google.com