International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:4, No:3, 2010

Using a Semantic Self-Organising Web Page-Ranking Mechanism for Public Administration and Education

Marios Poulos, Sozon Papavlasopoulos, and V. S. Belesiotis

***Abstract*—**In the proposed method for Web page-ranking, a novel theoretic model is introduced and tested by examples of order relationships among IP addresses. Ranking is induced using a convexity feature, which is learned according to these examples using a self-organizing procedure. We consider the problem of self-organizing learning from IP data to be represented by a semi-random convex polygon procedure, in which the vertices correspond to IP addresses. Based on recent developments in our regularization theory for convex polygons and corresponding Euclidean distance based methods for classification, we develop an algorithmic framework for learning ranking functions based on a Computational Geometric Theory. We show that our algorithm is generic, and present experimental results explaining the potential of our approach. In addition, we explain the generality of our approach by showing its possible use as a visualization tool for data obtained from diverse domains, such as Public Administration and Education.

***Keywords*—**Computational Geometry, Education, e-Governance, Semantic Web.

I. INTRODUCTION

ALTHOUGH the Web has been extended significantly of late, however, only relatively simple information-sharing mechanisms have been developed [1]. User knowledge presents unpredictable delays when retrieving Web pages from remote sites. The evident solution to improve the quality of Web services would be to increase the bandwidth, but such a choice involves increasing the financial cost. In addition, the higher bandwidth would only momentarily solve any problems since it merely encourages users to produce more and more resource-hungry applications, which work against the network [1]. A traditional method by which to deal with this situation is caching [1], [2], [3], [4]. Although caching offers several advantages like reduced network traffic and shorter response times, it has its drawbacks too, e.g., small hit rates [1], [5] and compulsory misses. To recompense for such problems, traditional caching is coupled with perfecting, 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 [1], [6] based on the recent history of requests of individual clients, which is called short-term perfecting [1], [7].

Furthermore, another mechanism, such as a Page-Rank algorithm [8], [9], [10] has proven to be very effective for ranking Web pages. However, inaccurate Page Rank results are produced because of the incomplete information about the Web structure. This problem is caused by the following phenomena:

1) The web is dynamic (temporal dimension)–The link structure develops temporally. Some links are created and modified, while others are destroyed.
2) The user is partial (spatial dimension)–For different users (or crawlers) the Web structure may be different.
3) Links are Different (local dimension)–Not all out-links are created equal. Some out-links are more significant than others.
4) The learning procedure of web page ranking can be considered to be an inaccurate mechanism [11], [12], [13], [14], [15].

Taking all this into account, in this paper we introduce a method, which combines the mechanism of caching information which comes from a dynamically self-organizing learning system.

The aim of this method is to rank the specific features of web pages using an algorithm of computational geometry and, simultaneously, to cache this information in a database. Thus, in the retrieving procedure, the user will then be able to see groups of web pages 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 [15], [16], [17]. This algorithm is tested on classification problems focusing on that of text categorization [15]. This proven [17] the speed of retrieval is increased, as is the reliability of the information.

More explicitly, we consider a simplified model consisting of a small number of IP addresses, which move semi-randomly on an equal number of vertical axes and which are unique to each IP address (see fig. 1).

M. Poulos, Archives and Library Sciences, Ionian University, Theotoki 72, 49100, Corfu, Greece, mpoulos@ionio.gr.

S. Papavlasopoulos, Archives and Library Sciences, Ionian University, Theotoki 72, 49100, Corfu, Greece, sozon@ionio.gr.

V. S. Belesiotis, Department of Informatics, University of Piraeus, 80 Karaoli & Dimitriou str., Piraeus 18534, Greece, vbel@unipi.gr.

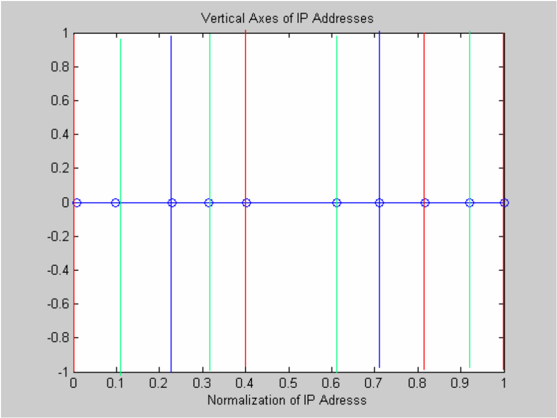

Fig. 1 The construction of semantic vertical axis

The criterion of this movement is based on the self-organizing convexity of the groups which contain IP addresses with at least one (1) common conceptual feature (see fig. 2).

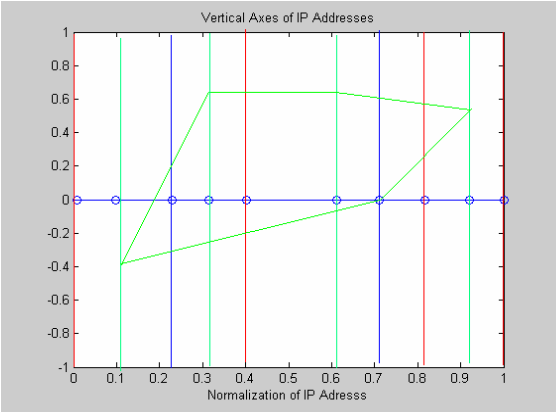

Fig. 2 A visual interpretation of the proposed algorithm

This semantic grouping is achieved via a well fitted algorithm of computational geometry. Finally, in the experimental part, this model is tested on a small number of real IP addresses.

II. Method

A. Algorithm Description

The proposed method is divided into the following stages:

1) The IP addresses are transformed into numeric form and the dots are removed, for example, google.com has an IP address 72.14.207.99 and this is transformed into 721420799. In this way, a vector S=[a1, a2,…, an] is constructed where [a1, a2,…, an] are the sites in the proposed numeric form, which in our case is n=10.
2) We normalize the vector Sn=S/max(S)
3) We put the elements of Sn on the axis of X (see fig. 1). Furthermore, at these points n vertical lines are set on axis X.

In our case, a different coloured line represents each category. The first category is represented by red, the second by green and the two (2) overlapping categories by blue.

In the next steps the Graham algorithm [18] is applied in a semi random proposed algorithm.

4) Element(s) k=1,..n of Sn, which have the most common features (in our case belonging to two categories) are set on the Cartesian plane with coordinates (Sk, 0). In the case where there is more than one element with the same maximum number of features then, pivot point P0 with the highest value max(Sn) is selected (see fig. 3).

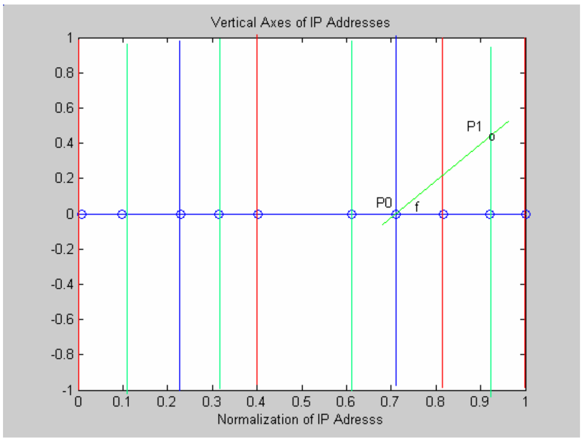

Fig. 3 An explanation of the first principal steps of the semi-random algorithm

5) The horizontal ray with a particular colour (red or green), emanating from the pivot (P_0), begins to rotate counter-clockwise at a random angle f, where $\hat{f} < 90^0$, and stops at the first point P1, which is determined by the section between this ray and the first vertical line of the same color. These points, the first and the second (P0, P1), then form the new axis of rotation. Using point P1 as the pivot, the third point of the convex polygon is determined as the first point of contact on random rotation of the new axis (see fig. 4).

B. Algebraic explanation of the area of a convex polygon

Suppose the x and y coordinates of a point pi are denoted by (xi, yi) and there are n total points in the plane. If we considered that the algebraic area (or the signed area) of the oriented triangle p_1, p_2, p_3 is $\frac{1}{2}\Delta(p_1, p_2, p_3)$, then $\Delta(p_1, p_2, p_3)$ determinant $\begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{pmatrix}$.

The sign of the algebraic area is positive if and only if the orientation of the triangle is counter-clockwise. Consider any n-vertex simple polygon P whose vertices around its boundary in some orientation (clockwise or counter-clockwise) are $p_1, p_2, ..., p_n$. Let O denote the origin in the plane of P. Then the algebraic area of the convex polygon P is $\frac{1}{2}\Delta\sum_{i=1}^{n} O,p_i, p_{i+1})$, where we assume $p_{n+1}= p_i$.

Here again, the sign is positive if and only if the given orientation is counter-clockwise around the boundary of P. Thus, any simple polygon P with more than three vertices (i.e. n > 3) has at least one diagonal. A diagonal is a line segment that connects two non-adjacent vertices of P and has no intersection with the exterior of P.

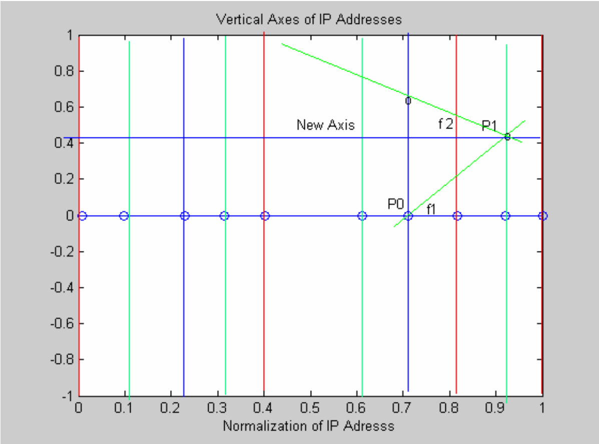

Fig. 4 An explanation of the two principal steps of the semi-random algorithm

III. The Implementation of the Method

In our case, for simplification, we consider a system that investigates 10 sites. These sites are grouped into two different categories, Medicine and Physics, with two (2) of the ten (10) belonging to both categories. In this way we construct a simple geometric model in which the overlapping topics are included (see table 1).

In applying the algorithm described in section II.B we follow the following steps:

Step 1.The IP addresses are transformed into numeric form and the dots are removed (see table 1)

Step 2. We normalize the above numeric form (see table 1)

Step 3. We put the elements of the numeric form on the axis of X (see fig. 1)

Step 4. We apply the Graham algorithm such as has been described in section II.B (method) and we obtain two semi-random convex polygons (see fig. 5).

Table 1
Applying the Algorithm

URL	IP	Normalization	Thematic Area	Colour Line's
www.bmj.com	171.66.121.222	0.8175	Medicine	Red
www.goldcopd.com	209.97.218.160	1	Medicine	Red
www.docguide.com	84.51.225.238	0.4025	Medicine	Red
www.medmatrix.org	12.35.245.111	0.0082	Medicine	Red
www.astro.org	4.79.229.34	0.2282	Medicine & Physics	Blue
www.iupesm.org	149.28.118.28	0.7110	Medicine & Physics	Blue
exambot.com	66.197.32.134	0.3153	Physics	Green
www.physlink.com	206.82.194.44	0.0985	Physics	Green
physicsworld.com	193.128.223.163	0.6107	Physics	Green
www.iop.org	193.131.119.137	0.9198	Physics	Green

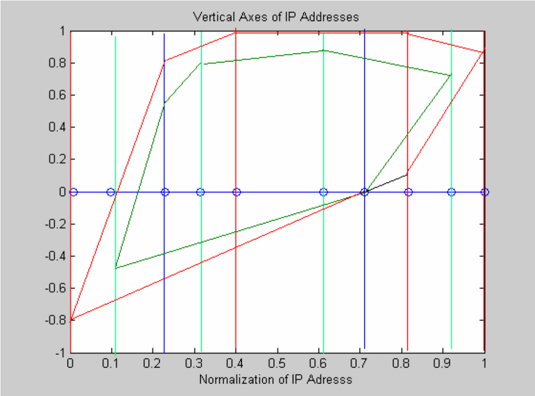

Fig. 5 Construction of the two groups of web-pages. The journals of medicine and physics content are marked with red and green colour accordingly, while the journals belonging to both categories are marked with blue colour.

IV. Experimental Part

In this paper, we attempted to produce a global bibliometric index for a published article. Thus, we constructed a global ideal article with (5) five high score indicators for each citation. For this implementation, we used the fuzzy logic theory in order to classify in (5) five degrees the bibliometric indicators. Finally, we extracted a bibliometric indicator for a simple article by correlating this with the ideal article. This paper may be considered as the first attempt to construct the proposed IGBI indicator and thus the research presented here should be considered as research in progress. Our major concern is to evaluate the IGBI index against well known comparisons such as journal status [11] using an extensive citation graph. Furthermore, the extension of a semantic vocabulary such as COAP and the construction of a citation index to provide real time IGBI analysis is also an important step.

V. Applying the Methodology in an Evaluation Test

In this section we will explain how our algorithm can be used in different domains as a visualization tool. We focus on the educational domain, and present how our method can be applied in data obtained during the evaluation process as an expert visualization tool producing diagrams showing the way students answered the questions of a certain test. More specifically, a methodology for the aforementioned purpose is the following:

A) Create an evaluation test

1) We create a test, the students are going to work with, which consists of questions of all types.
2) For every question of the test the thematic axis the student must cover with his answer are predetermined.

B) Evaluation of the students' answers.

For every question of the test and for each student evaluate:

1) How many axis the student covered.
2) The way the thematic axis is covered.

The evaluation can be translated in marks.

Thus, we have recorded for every question of the test and for every contestant how many thematic axis are covered and in which way each one.

A code numbers is assigned for each response. For example a code number can be aaa.bbb.ccc.ddd, which is broken down as follows:

aaa : examination center code, e.g. 193
bbb : contestant code, e.g. 092
ccc: the test code , e.g. 177
ddd: question number, e.g. 005

The following example describes the use of our methodology. At the examination center 193, 30 students are examined in a test with the code number 177, which had 15 questions. We will focus on question 9, where we expected the students to cover three thematic axis, and on four students with code numbers 092, 093, 094 and 095.

Three curved polygons are rendered, by applying our algorithm, to each of the three categories of expected answer axis. The resulting figure describes the success of the test that the students took and we believe that offers a lot to the evaluation process.

From all the above we believe that the methodology can be used as a visualization tool, in several domains, such as:

A) Wide public sector, as.

1) In Public Administration examination centers
2) Document indexing, regarding the axis the document is related.
3) Damage registering (type of damage) in public sectors.
4) Classification of citizens' applications to Public Service centers, Ministries, Organizations.
5) Indexing Offers to Competitions

The following table (table 2) shows possible results of the test for question 9:

Table II
Results of the Test for Question 9

Code	Number of covered thematic axis
193.092.177.009	1
193.092.177.009	2
193.092.177.009	3
193.093.177.009	1
193.093.177.009	3
193.094.177.009	2
193.094.177.009	3
193.095.177.009	2

B) Education and Teaching.

1) At schools as a processing method of test results by the teacher.
2) At the education procedure, for visualizing the students' answers with respect to relevant educational techniques, like a hale of ideas.

VI. Discussion –Further Research

In this paper an algorithm is described that is based on a suitably transformed algorithm of computational geometry [15], and which aims to introduce a new web page-ranking approach.

For this purpose, we used only ten (10) web pages with at least two (2) specific features in order to simplify the method and make it more understandable. In this method we show that it is possible for a number of web page groups to be created which have shared convexity.

The utility of this mechanism can be found in the dynamic self-organizing feature of the algorithm, which automatically ranks the web pages 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.

We describe how our algorithm can be employed as a visualization tool for data obtained from several different domains, such as Public Administration and Education domain in the students' evaluation process.

Finally, in our future research plan we aim to test this proposed method against a larger number of web pages, our further aim being to implement this algorithm in an ontological approach.

References

[1] A. Sidiropoulos, et al., "Prefetching in Content Distribution Networks," *Communities Identification and Outsourcing World Wide Web*, vol. 11, pp. 39-70, 2008.

[2] D. Katsaros, Y. Manolopoulos, "Caching in Web memory hierarchies," In: *Proceedings of the ACM Symposium on Applied Computing (SAC)*, Nicosia, 14–17 March, 2004.

[3] S. Sivasubramanian, G. Pierre, M van Steen, G. Alonso, "Analysis of caching and replication strategies for web applications," *IEEE Internet Computing*, vol. 11, no. 1, pp. 60–66, 2007.

[4] A. Vakali, "Proxy cache replacement algorithms: a history-based approach," *World Wide Web J*, vol. 4, no. 4, pp. 277–298, 2001.

[5] T. Kroeger, E. D. Long, J. Mogul, "Exploring the bounds of Web latency reduction from caching and perfecting", In: *Proceedings of the USENIX Symposium on Internet Technologies and Services (USITS)*, Monterey, 8–11 December, 1997.

[6] D. Katsaros, Y. Manolopoulos, "Prediction in wireless networks by Markov chains," *IEEE Wireless Communications magazine*, 1977.

[7] A. Nanopoulos, D. Katsaros, Y. Manolopoulos, "A data mining algorithm for generalized Web prefetching," *IEEE Transactions on Knowedgeand Data Engineering,* vol. 15, no. 5, pp. 1155–1169, 2003.

[8] L. Page, S. Brin, R. Motwani and T. Winograd, *The pagerank citation ranking: Bringing order to the web.* Stanford Digital Library Technologies Project, Tech. Rep. Paper SIDL-WP-1999-0120 (version of 11/11/1999).

[9] H. Yang, I. King, "Predictive Random Graph Ranking on the Web," *2006 International Joint Conference on Neural Networks*, Vancouver Canada, July 16-21, 2006.

[10] A. Shivani, "Ranking on Graph Data," In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[11] W. W. Cohen, E. R. Schapire, & Y. Singer, "Learning to order things," Journal of Artificial Intelligence Research, vol. 10, pp. 243–270, 1999.

[12] R. Herbrich, T. Graepel, & K. Obermayer, *Large margin rank boundaries for ordinal regression, Advances in Large Margin Classifiers.* Liu Press, 1997, pp. 115–132.

[13] K. Crammer, Y. Singer, "Ranking with ranking," *In Proceedings of the Advances in Neural Information Processing Systems*, 2002.

[14] Y. Freund, R. Iyer, E. R Schapire, Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[15] M. Poulos, S. Papavlasopoulos and V. Chrissicopoulos, "A Text Categorization Technique based on a Numerical Conversion of a Symbolic Expression and an Onion Layers Algorithm," *Journal of Digital Information (JoDI)*, v. 6. 1, 2004.

[16] M. Poulos et al., "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, 2007.

[17] M. Poulos, G. Bokos, F. Vaioulis, "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, 2008.

[18] L. R. Graham, *An efficient algorithm for determining the convex hull of a finite planar set.* Inform, Proc. Letters, pp. 132-133, 1972.

M. 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 Piraeus (2003). In the period from 2003 to 2008 he was a member of the staff (adjunct lecturer 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 a member of several technical committees and working groups on subjects related to the medical informatics and information science areas.

S. 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 M.Sc 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.

Dr Papavlasopoulos 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 the Greek Mathematical Society and of the Greek Computer Society.

V. S. Belesiotis holds a B.Sc. in Mathematics from the Univessity of Patras (Greece, 1986), studies (M.Sc. level) in Operational Research and Informatics from the University of Athens (Greece, 1978) and a Ph.D. in Informatics from the University of Piraeus (Greece, 2002).

He is a Secondary Education School Advisor for Informatics in Central Athens at the Greek Ministry of National Education and Religious Affairs. Moreover, he has been teaching for several years in the Department of Informatics in the University of Piraeus, including the subject of Didactics in Informatics.

Dr. Belesiotis has published over 20 papers and 8 books, which have been used as textbooks both for secondary education and university courses. His research interests include didactics, education in informatics, and knowledge representation. Furthermore, he has been a member of several technical committees and working groups on subjects related to informatics and education.