Analysis of the Greek Web-Space

T. Mchedlidze[†], A. Symvonis[†], M. Tzagarakis[‡]

[†]National Technical University of Athens, School of Applied Mathematical & Physical Sciences,Athens, Greece.
symvonis@math.ntua.gr mchet@central.ntua.gr
[‡]Research Academic Computer Technology Institute, Patra, Greece.
tzagara@cti.gr

Abstract

In this paper, we report on our efforts to identify important Greek web sites on the web by analyzing the link structure of Greek web sites. Towards this end, five independent graph-theoretic methods have been deployed: Hub and Authority, Page Rank, Bow-Tie structure, core analysis and degree distribution. Our findings indicate that government, non-profit and educational sites are amongst the most important in the Greek web space.[1]

1. Introduction

The World Wide Web contains enormous amount of information that is growing at an exponentially rate. Taking advantage of this move, diverse types of organizations with different social roles and missions put great efforts to relocate part of their activities into the world wide web. Hence, commercial sites attempt to market products and transact business on the web, news agencies and newspapers use it to inform the their readers, governments to deliver public services and educational organizations to provide virtual learning environments. Such transfer of activities from the real to the virtual world makes the World Wide Web effectively a mirror of today's society. This is in particular true if one limits the information available on World Wide Web to the level of a particular country.

Resembling a mirror of society, analysing the information available on the World Wide Web may give insights and hints about the zeitgeist of a country i.e. the intellectual and cultural climate of an era. The structural information in particular may give valuable information with respect to this. Identifying such trends may be beneficial for a number of reasons that include: archival reasons so that the cultural heritage of a country can be preserved for future generations and development reasons giving indications of

[1] A more detailed analysis of the work presented in this paper is available from http://www.math.ntua.gr/aarg/

communities that are either on the rise or on the fall. Although a number of existing approaches attempt to mine link information on the web to identify important communities and trends [2] [4] [11], they target the entire web space thus being able to extract macroscopic information. Narrowing the context of analysis - and in particular to the local web space - may prove beneficial since new insights may be revealed [13].

In this paper we report on our efforts to identify important Greek web sites on the web by analysing link structure of Greek web sites. Towards this, five independent graph-theoretic methods have been deployed: hub and authority, page rank, Bow-Tie structure, core analysis and degree distribution. Our findings indicate that government, non-profit and educational sites are amongst the most important in the Greek web space. The paper is organized as follows: Section 2 introduces the algorithms used to mine link information on the Greek web. In section three, we give an overview of the Greek web space and in section four we present the analysis and the results of our evaluation.

2. Methods Used in the Analysis

The analysis is based on five independent graph-theoretic methods: hub and authority, page rank, Bow-Tie structure, core analysis and degree distribution.

Hubs and Authorities The *"Hub*s *an*d *Authorities*" algorithm was introduced by Kleinberg [14] in an effort to identify in the world-wide web important pages that contain information related to a specific query topic *s*. In the literature, the algorithm is also referred to as the *HIT*S algorithm (from the initials of the words in "Hypertext Induced Topic Selection") and a number of different variations have been proposed [9] [5] [15].

Page Rank Page and Brin in [17] introduced *PageRan*k which has been successfully deployed in search engine software [7]. PageRank is a method of ranking web-pages that gives an objective representation of the human notion of importance to every web-page. The intuition of the method is as follows: A web-page is supposed to be important if the pages that link to it are important as well. In contrast with a Hubs and Authorities, PageRank in its original form is not "topic-sensitive". A "topic-sensitive" version was presented in [12]. An extended comparative work of HITS and PageRank can be found in [16].

Bow-tie While the HITS and the PageRank methods study the web-graph at a microscopic level, the *bow-ti*e method examines it macroscopically. Broder and Kumar [8] observed that the web can be viewed as the bow-tie in Figure 1. According to them the web graph can be divided into several parts. The core of the web is a maximal *strongl*y *connecte*d *component(SSC*) of it’s graph. The left side of a bow-tie, named *IN*, represents the pages from which at least a path exists to some nodes in *SCC*. The right side of the bow-tie, named *OUT*, represents the pages which can be reached from nodes of *SCC*. The *TENDRIL*S contains pages that are reachable from *IN*, or that can reach *OUT*, without passages through *SCC*. *I*N can be thought to consist of new pages that link to some ”famous” ones but have not yet been discovered by *SCC*, while *OU*T can be thought to consist from some well known pages whose links point to internal pages only. As for *TENDRILS*, these pages are not yet discovered by the web, and they do not link to better-known regions.

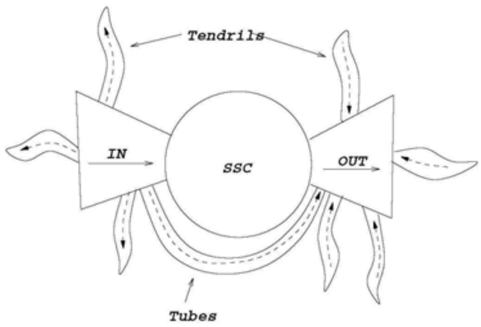

Figure 1. Bow-Tie structure of the World Wide Web

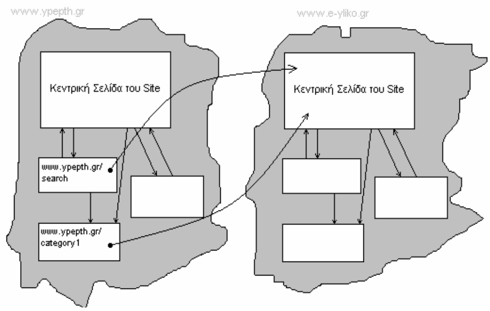

Figure 2. Types of links.

Degree distribution The macroscopic structure of the web-graph can be studied by examining the in-degree and out-degree distributions. For the WWW graph, Faloutsos et al. [10] discovered that the distribution of the in- and out-degrees of web-sites follows a power law. Namely, the number of pages with in-degree d is proportional to d^{k_I}, and the number of pages with out-degree d is proportional to d^{k_O}, for some constants k_I and k_O. Kumar et al. [15] estimated k_I to be about -2.1 and k_O to be about -2.38. The growth of the exponent values can be explained as follows: *th*e *highe*r *th*e *exponen*t *valu*e *is*, *th*e *smalle*r *portio*n *o*f *th*e *site*s *i*n *th*e *spac*e *monopolize*s a *larg*e *portio*n *o*f *th*e *link*s *include*d *i*n *space*. In their effort to explain how the web grows, Kumar et al. [15] presented the (*α*, *β*)-model. In the (*α*, *β*)-model when a new page(vertex) is created, a new link is also inserted in the graph. *α* is a probability that the just-created edge has its target at the new vertex while, *β* is a probability that the just-created edge has its origin at the new vertex. Consequently, (*1-β) and (1- α*) are the probabilities for newly created edge to have its origin and target at “old” pages. These two probabilistic parameters can be computed from the exponent values of the power law adapted to the web graph based on the equations:

$$k_I = -\frac{1}{1-\alpha} \quad k_O = -\frac{1}{1-\beta}$$

Based on the estimates of Kumar et al. [15] for k_I and k_O, it follows that α =0.52 and β =0.58. In a related work, Ishimura et al. [13] studied the local web-space of several Japanese middle-size cities. They estimated values for the in-degree and out-degree exponents were $k_I = -1.14$ and $k_O = -1.39$, respectively. The corresponding probabilistic parameters were $\alpha = 0.12$ and $\beta = 0.28$. The small values of α and β (much smaller than 0.5), according to Ishimura et al. [13], reflected the fact that a *newly generated site is less likely to link to the existing sites in the local web-space the site belongs to*.

3. The graph of the Greek web space

Experimental setup A crawler has been developed to retrieve pages from the Greek web. The crawler is written in C++ running on the Linux operating system, reaching a throughput of about 6250 pages per hour. The crawl was constrained only to web pages in the ".gr" space proceeding in a breadth first search manner. Dynamic as well as static pages were taken into consideration. Pages outside the Greek web page were discovered but the crawler did not examined any of their links. As a result, our graph contained several nodes "at the boundary" of the Greek web space, all having out degree zero. Our initial set of URLs from which the crawler started harvesting pages consisted of 73400 distinct sites in the .gr domain. From these distinct sites the crawler downloaded a total of 1045563 web pages. Initially data were collected at a URL level, in particular all web-pages with different url have been saved for any site the crawler had found. For example, taking the case on Figure 2, for www.ypepth.gr we knew all pages with the same domain name and all outgoing links of these pages. In the PageRank and Hubs and Authorities evaluation, it made good sense not to take into account the secondary pages of web-sites. [2] So it was decided to restrict the data set to domain names.

Weighted Graph We call *internal links* those links which bind two pages with the same domain name, and *external links* those between two pages with different domain names. The first model we used had as vertices all different domain-names. Internal

[2] The links from secondary pages to primary ones, and vice versa, of the same domain do not generally reflect the public view, but just the view of web-page constructor.

links were treated as being of low significance. Thus, we only modelled the external links between two domains, representing them by the weight of the directed edge connecting the two domains. For example, in Figure 2 the weight of the edge from www.ypepth.gr to www.e-yliko.gr is 2. Applying the model resulted in a graph having 76506 nodes and 141791 edges.

4. Results of the evaluation

4.1 Important web-pages of our local web-space

Weightless Graph One of our objectives was to apply the HITS and the PageRank algorithms to the Greek web-graph in order to identify the most important domains of our local web-space. We used Pajec[3] due to its efficiency to compute hubs and authorities and Visone[6] due to its visualization capabilities.

Authority	Web Page
0,0164229	Europa.eu.int
0,014825	www.auth.gr
0,0144888	www.parliament.gr
0,0144344	www.in.gr
0,0144167	www.uoa.gr
0,0142828	www.ntua.gr
0,0142745	www.aegean.gr
0,0142545	www.uoi.gr
0,013984	www.geocities.com
0,0138814	users.otenet.gr
0,0137366	www.forthnet.gr
0,0134783	www.mod.gr
0,0134387	www.europa.eu.int
0,0134349	www.pi-schools.gr
0,0133442	www.teiath.gr
0,0132942	www.asfa.gr
0,0131447	www.ekt.gr
0,012955	www.eie.gr
0,0127159	www.grnet.gr
0,0125983	www.eap.gr

Table 1. The 20 best authorities obtained from the analysis of the weightless domain graph.

Hub	Web Page
0,9923293	Dir.forthnet.gr
0,0334724	www.in.gr
0,0272116	www.phantis.gr
0,0249825	www.cso.auth.gr
0,0244363	www.e-yliko.gr
0,024251	www.startpoint.gr
0,0237506	www.kosmikanea.gr
0,0236865	www.e-yliko.sch.gr
0,0215675	www.ekt.gr
0,0214549	www.translatum.gr
0,0201813	www.e-businessforum.gr
0,0200832	Users.lar.sch.gr
0,0192014	www.kalamaria.gr
0,0170332	www.de.sch.gr
0,0164346	www.noc.ntua.gr
0,0162967	www.plefsis.gr
0,0148821	www.xkatsikas.gr
0,0147811	www.uoi.gr
0,0146873	www.nad.gr
0,0141403	www.robby.gr

Table 2. The 20 best hubs obtained from the analysis of the weightless domain graph.

Web Page
www.macromedia.com
www.ast.com.gr
www.microsoft.com
www.adobe.com
www.marinet.gr
www.gozakynthos.gr
www.go-online.gr
www.celect.gr
www.statcounter.com
www.hellasnet.gr
www.google.com
www.europa.eu.int
www.next-step.gr
www.zanteweb.gr
www.ntua.gr
www.zakynthos-net.gr
www.ypepth.gr
www.adman.gr
www.wunderground.com
www.otenet.gr
www.goonline.gr
www.web-greece.gr
www.uoa.gr
www.culture.gr
www.forthnet.gr

Table 3. According to pagerank values obtained from the analysis of the weightless domain graph.

Initially, we applied the HITS algorithm to a weighted graph which resulted in having some domains in high HITS ranks due to the fact that they were heavily interlinked. Consequently, we decided to ignore the edge weights and to experiment with a weightless graph. The Hubs and Authorities result for the weightless domain graph, as computed by Pajek, are presented in Tables 1 and 2.

To get a better picture of the hubs and authorities results, we used the layered layout[3] provided by Visone. The results of the visualization appear in Figures 3 and 4.

From the hubs-visualization (Figure 4) we observe that the difference between hub values is too large, while the authorities-visualization (Figure 3) reveals a large group of nodes displayed at the same layer. Trying to identify the reason for this fact, we noticed that there are 10114 nodes (out of the 76441 in the entire graph) that have identical authority value equal to 0,0081123. All these nodes have just one incoming link from *dir.forthnet.gr*. Knowing this, it is not surprising that the domain *dir.forthnet.g*r is the best hub. The PageRank analysis of the weightless graph has been done using Visone. The results are visualized in Figure 5 and the top 25 domains according to the PageRank evaluation are presented in Table 3. Comparing the results of PageRank and HITS evaluation, we distinguish two types of popular domains: Domains that are very popular (high pagerank) and are authorities for their contained information; these domains are listed in Table 4.[4] Domains that are very popular portals and search engines; these domains are listed in Table 5. [5]

www.europa.eu.int
www.ntua.gr
www.uoa.gr
www.forthnet.gr

Table 4. Most popular pages of Greek web.

www.in.gr
www.ekt.gr
www.uoi.gr

Table 5. Most popular portals and search engnes of Greek web.

Component	Vertex Number
SCC	2898
IN	15
OUT	73403
TUBES	0
TENDRILS	72
OTHERS	53

Table 6. Bow-Tie Analysis. Components sizes.

The results of Table 4 and Table 5 are partially consistent with top 100 pages list given by ALEXA [1]. At ALEXA's web-page the top 100 sites of the Greek web space are ranked according to traffc data of Alexa Toolbar users. The domains that are highly ranked both by

[3] A layered layout emphasizes the importance of the nodes of a graph according to the weight assigned to the nodes. The layout places the nodes to different layers according to these weights. In our case we used the hub and authority weights.
[4] The order of every site in this list is based on the sum of its authority and pagerank order numbers.
[5] The order of every site in this list is based on the sum of its authority and hub order numbers.

our experiment and Alexa are : *www.uoa.gr*, *www.ntua.gr*, *www.forthnet.g*r and *www.in.gr*.

4.2 Macroscopic Structure

Bow-Tie Analysis The bow-tie analysis of the weightless graph was facilitated by Pajek. By an analysis similar to that by Broder and Kumar [8], we found out that the *SCC(strongl*y *connecte*d *component*) and the IN components of our graph were rather small. Table 6 presents the sizes of all components.

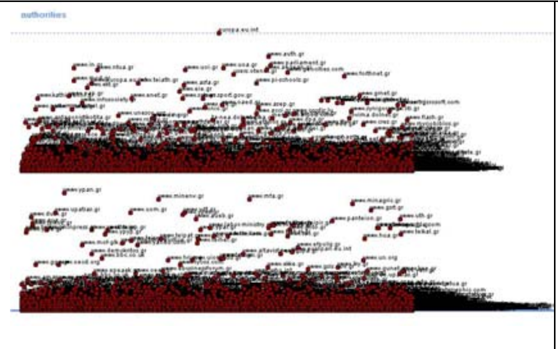

Figure 3. Authority Weights. Layered Layout.

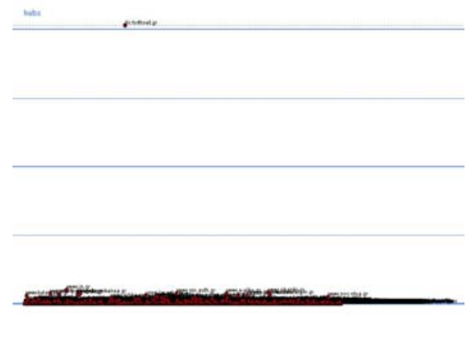

Figure 4. Hub Weights. Layered Layout.

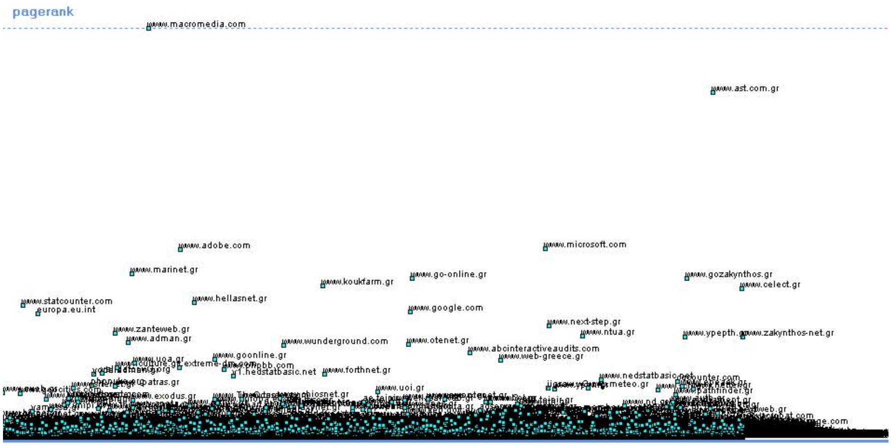

Figure 5. Pagerank. Layered Layout.

Core The subgraph G' of graph G is called a *k*-core, if every node of G' is connected with at least k nodes in G' and G' is a maximal. A core of the domain graph is thought to identify domains with strong ties of some form. So, we tried to identify the ”heart” of the Greek web space, as it is represented by a core.

Applying the core search algorithm, implemented by Batagelj[3], and focussing on the in-degree, we found that the maximal core of the studied graph is 11-core with 66 nodes. It is evident that the domains in the core are related to Greek government, non-profit and educational institutions. The domains of the core are presented in Table 7.

Power-Law As mentioned in Section 2, the in-and out-degrees of vertices of the World Wide Web graph are power-law distributed. This fact was proved also for the local-web space of Japanese cities. In this section, we describe the results we have obtained by a similar analysis. For every number $i \in N$ (let us call parameter *i* the *“in-degree”*) we computed the number of nodes (we call this parameter *“number of domains”*) with in-degree i that are contained in our graph. We do the same for the out-degrees.

Adapting the linear model to *ln(in-degree)*, *ln(number of domains),* where the first is an independent and the second is a dependent parameter, we ’ve got that the best adapted lines were (see Figures 6 and 7):
For in-degree: *ln(number of domains)* = 4. 00 – 1.91 *ln(in-degree)*, while
for out-degree: *ln(number of domains)* = 2.34 – 0.903 *ln(out-degree)*.

URL	URL	URL	URL
www.ypepth.gr	www.europa.eu.int	www.grnet.gr	www.gsrt.gr
www.gunet.gr	www.auth.gr	www.minagric.gr	www.primeminister.gr
www.culture.gr	www.sport.gov.gr	www.yen.gr	www.ministryofjustice.gr
www.ypyp.gr	www.minpress.gr	www.labor-ministry.gr	www.ydt.gr
www.mof-glk.gr	www.minenv.gr	www.ypes.gr	www.mathra.gr
www.mfa.gr	www.mod.gr	www.ypan.gr	www.ypai.gr
www.yme.gr	www.teilar.gr	www.teimes.gr	www.teipat.gr
www.teipir.gr	www.teiser.gr	www.teihal.gr	www.parliament.gr
www.et.gr	europa.eu.int	www.adobe.com	www.aegean.gr
www.cordis.lu	www.asfa.gr	www.teithe.gr	www.teikoz.gr
www.teilam.gr	www.teiath.gr	www.teiep.gr	www.teikal.gr
www.uoa.gr	www.un.org	www.europarl.eu.int	www.oecd.org
www.infosociety.gr	www.gnto.gr	www.uoi.gr	www.uom.gr
www.ntua.gr	www.aueb.gr	www.unipi.gr	www.upatras.gr
www.government.gr	www.efpolis.gr	www.neagenia.gr	www.ypetho.gr
www.uth.gr	www.aua.gr	www.teikav.edu.gr	www.panteion.gr
www.hua.gr	www.duth.gr		

Table 7. The 66 domains of the 11-Core (based on the in-degree of the nodes).

We note that the power law explains the in-degree distribution by 90%, while the out-degree distribution is explained by 72%. The 72% of the out-degree may be explained by the very big percentage of pages with out-degree 1. Thus, the exponent values of power-law for our graph are $k_I = -1.91$ and $k_O = -0.9$.

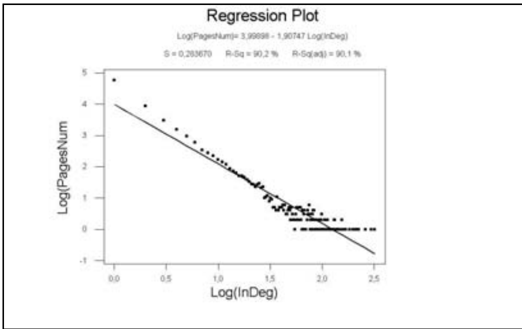

Figure 6. Power-Law for in-degree.

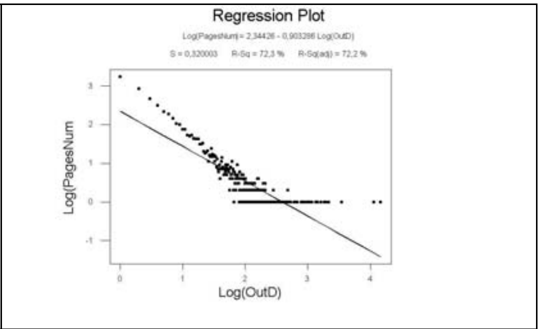

Figure 7. Power-Law for out-degree.

Recall that the exponent values of Kumar et al. [15] for the WWW are: $k_{I_{WWW}} = -2.1$, $k_{O_{WWW}} = -2.38$ while the values of Ishimura et al. [13] for local web-space Japanese cities are $k_{I_{Japan}} = -1.14$ and $k_{O_{Japan}} = -1.39$. Refraining from drawing any conclusions for the out-degrees due to the low-percentage of data explanation by the model, we conclude that *in the Greece web space there are a lot of domains that are pointed to by a large number of other domains.* This conclusion can be made by contrasting our $k_I = -1.91$ with $k_{I_{Japan}} = -1.14$ and noting that their difference is greater than the difference of $k_I = -1.91$ and $k_{I_{WWW}} = -2.1$. In simple words, we can conclude that the growth of the Greek web-space is more likely to resemble the growth of global web than the growth of web of one town. Interpreting the exponent value of in-degree in terms of (α, β)-model, we get that α = 0.48, therefore a new created link has approximately the same probabilities to link to a new web-page of our local space or to link to an already existing web-page.

References

1. http://www.alexa.com/site/ds/
2. Amitay, E., Carmel, D., Herscovici, M., Lempel, R., and Soffer, A. (2004), *Trend Detection Through Temporal Link Analysis*, Journal Of The American Society For Information Science And Technology, 55(XX),pp. 1-12.
3. Batagelj, V., Mrvar, A., Mutzel, M. (2003), *Graph Drawing Software*, pp. 77-103,

Springer, Berlin.
4. Berberich, K., Bedathur, S., Vazirgiannis, M. and Weikum, G. (2006), *BuzzRank. . . and the trend is your friend*, In Proceedings of the 15th international conference on World Wide Web, pp. 937 -938.
5. Bharat, K., Henzinger, M. (1998), *Improved algorithms for topic distillation in a hyperlinked environment*, In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,ACM Press, New York, NY, pp. 104-111.
6. Brandes, U., Wagner, D. (2002), *VisOne-Software for Visual Social Network Analysis*, Web Site: http://visone.info/
7. Brin, S., Page, L. (1998), *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Computer Networks and ISDN Systems, 30, 1-7, pp. 107-117.
8. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Weiner, J. (2000), *Graph Structure in the web*, In Proceedings of the 9th International World Wide Web Conference.
9. Chakrabarti, S., Dom, B., Raghavan, P., Rajogopalan, S., Gibson, D., Kleinberg, J. (1998), *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text*, In Proceedings of the 7th World-Wide Web conference, Amsterdam.
10. Faloutsos, M., Faloutsos, P., Faloutsos, C. (1999), *On Power-Law Relationship of the Internet Topology*, In Proceedings of SIGCOMM'99, pp. 251-261.
11. Gibson, D., Kleinberg, J. and Raghavan, P. (1998), *Inferring Web communities from link topology*, In Proceedings of the ninth ACM conference on Hypertext and hypermedia, pp. 225-234, June 20-24, Pittsburgh, Pennsylvania, United States.
12. Haveliwala, T. (2003) *Topic-Sensetive PageRank: A Context-Sensitive Ranking Algorithm for Web Search*, In IEEE Transactions on Knowledge and Data Engineering.
13. Ishimura, Y., Shin, K., Kamibayashi, N. (2005), *Graph-theoretic nature of local web-spaces*. In Proceedings of the 16th International Workshop on database and Expert Systems Applications.
14. Jon M. Kleinberg. (1999), *Authoritative Sources in a Hyperlinked Environment*, Journal of the ACM (JACM), v.46 n.5, pp. 604-632.
15. Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A. (1999), *Extracting Large Scale Knowledge bases from the Web*, In Proceedings of the 25th VLDB Conference, Edinburgh, Scotland.
16. Langville, A., Meyer, C.D. (2005), A *Survey of Eigenvector Methods for Web Information Retrieval*, SIAM Rev.47, 1, pp. 135-161.
17. Page, L., Brin, S., Motwani, R., Winograd, T. (1998), *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report, Stanford University.