Finding Authoritative Researchers on Academic Web Sites

Dalibor Fiala, Karel Jezek, and Francois Rousselot

***Abstract*—**In this paper, we present a methodology for finding authoritative researchers by analyzing academic Web sites. We show a case study in which we concentrate on a set of Czech computer science departments' Web sites. We analyze the relations between them via hyperlinks and find the most important ones using several common ranking algorithms. We then examine the contents of the research papers present on these sites and determine the most authoritative Czech authors.

***Keywords*—**Authorities, citation analysis, prestige, ranking algorithms, Web mining.

I. Introduction

NOTIONS of importance, significance, authority, prestige, quality and other synonyms play a major role in social networks of all types. They denote an object that has a large impact on the other objects in the community. Perhaps the best example is bibliographic citations in the scientific literature. Counting citations of research publications is a relatively objective manner to determine quality or useless research known since a long time ago. With the fast growth of the World Wide Web in the past ten years, this kind of analysis has become essential in this domain as well.

In the Web domain, citations are links between Web pages or Web sites (when we talk about site level). Therefore, current Web search engines make use of various link-based quality ranking algorithms whose rankings they combine with the keyword search results to offer the user not only topic-relevant but also high quality Web pages. These algorithms may be recursive, such as PageRank [1], [2], [11] or HITS [3], [7], [9] or simple like In-Degree which just counts in-links. Some studies [5], [6] have recently shown that the rankings produced by the three algorithms are highly positively correlated. Recursive methods have a strong probabilistic background [4]. Closest to our work is the research in [13], [14] but in addition to the relations between Web sites we also studied the contents of the documents found on them.

Manuscript received September 22, 2006. This work was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009 within the National Program for Research II.

D. Fiala is with the Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Univerzitni 22, 30614 Plzen, Czech Republic and with the Design Engineering Laboratory, Graduate School of Technology of Strasbourg (INSA Strasbourg), 24 boulevard de la Victoire, 67084 Strasbourg, Cedex France (phone: +420-377-63-24-01, +33-388-14-47-53; fax: +420-377-63-24-02, +33-388-14-47-99; e-mail: dalfia@kiv.zcu.cz, dalibor.fiala@insa-strasbourg.fr).

K. Jezek is with the Department of Computer Science and Engineering, University of West Bohemia in Pilsen, Univerzitni 22, 30614 Plzen, Czech Republic (e-mail: jezek_ka@ kiv.zcu.cz).

F. Rousselot is with the Design Engineering Laboratory, Graduate School of Technology of Strasbourg (INSA Strasbourg), 24 boulevard de la Victoire, 67084 Strasbourg Cedex, France (e-mail: francois.rousselot@insa-strasbourg.fr).

II. Experimental Framework

Our first objective was to determine authoritative institutions among Czech computer science University departments. We have chosen this area because we know it well and we could expect that there would be enough data on the Web to analyze. At the same time, we supposed the data volume to be easily manageable. Even though we limited our experiments by topic and scope, the methodology we used was sufficiently general to be able of applying to a completely different scientific field.

A. Constraints

We have selected seventeen computer science Web sites from a Web directory of Czech academic institutions. Our selection had several constraints. First, we wanted to take account of their geographic location so as to include various regions of the Czech Republic. Second, each department had to have its home page on its own server. That means, we did not consider home pages being on a URL's path such as www.someuniversity.cz/somedepartment but only those like www.department.university.cz. Therefore, we had to eliminate departments whose home pages were located in their University domain, which was sometimes the case.

The reason for this is the fact that stand-alone servers can be manipulated more easily by a machine. A Web spider recognizes quickly whether or not a link on a department's Web page is internal (within department). And third, we wanted the departments to correspond in the University hierarchy approximately to the level of our home department. This is somewhat tricky because not all of the Universities have the same structure of schools consisting of departments. For this reason, some institutions in our list are schools rather than departments.

B. Procedure

In December 2005, we let our Web spider crawl all of the seventeen servers. The spider stored information about hyperlinks between Web pages on the servers to a database and built a corpus of downloaded documents for further analysis (see Section 4). We repeated the same procedure two more times in a-few-days intervals and the results we obtained

remained almost unchanged. We show those from the last experiment in Table I.

We have to mention briefly a few Web crawling related issues which may have impact on the parameters we examined. We were interested only in links via the HTTP protocol and pointing to documents in certain formats. For instance, we did not consider video or audio documents, which are natural, but we also left out documents with extensions doc, rtf, txt, and ppt, which is more arguable. (However, taking account of these formats in one of the experiments caused only one change in the middle part of the chart in Table I.) To prevent the spider from getting stuck in Web traps, we set the maximum depth of nesting in the Web graph to eight, which is empirically a good estimate for yielding reasonable results. (Documents in greater depths are usually duplicates with different names – URLs.)

C. Results

Our spider collected over 250 000 documents (in specific formats) and created a roughly 7 GB corpus. We found about 3.3 million links to those documents within the set of servers. We removed duplicate links and self-links (intra-site links). Duplicate links have the same source and target URL; self-links have a source and a target within the same server. After removal, there were 1 850 links left. The sites in Table I are ordered descendingly by the number of in-links (citations).

We can notice in Table I that the hosts are grouped into three clusters. At the top, there are three Web sites that are clearly ahead of the others. At the bottom, there are sites that have no or very few in-links. In between, there is the largest block of average departments. We show the number of the documents of our interest found on the individual servers as well. Of course, the number of in-links often depends on the number of documents on the target site. Their numbers vary greatly due to different sizes of hosting institutions (see also Section 2.A), preference of various document formats and document generation (dynamic Web pages), etc. One way of tackling this problem is to normalize the number of citations somehow. For instance, it is possible to divide the number of citations by the number of documents on a particular site (the ratio in the last column of Table I) or by the number of staff of the corresponding institution [14]. In this context, it is interesting to note the very low total ratio. This means that in a closed set of Czech computer science institutions, the departments cite one another very rarely, which is somewhat astonishing.

D. Issues

There are some facts that may severely influence the ordering by in-links. One of them is the existence of server aliases. For instance, www.siteA.cz and www.siteB.zcu.cz is one machine with the same content. Thus, citations to both should be counted together. There may be a large number of aliases and ignoring them could lead to wrong results. It is not possible to replace host names with IP addresses either since more virtual servers can share one IP address.

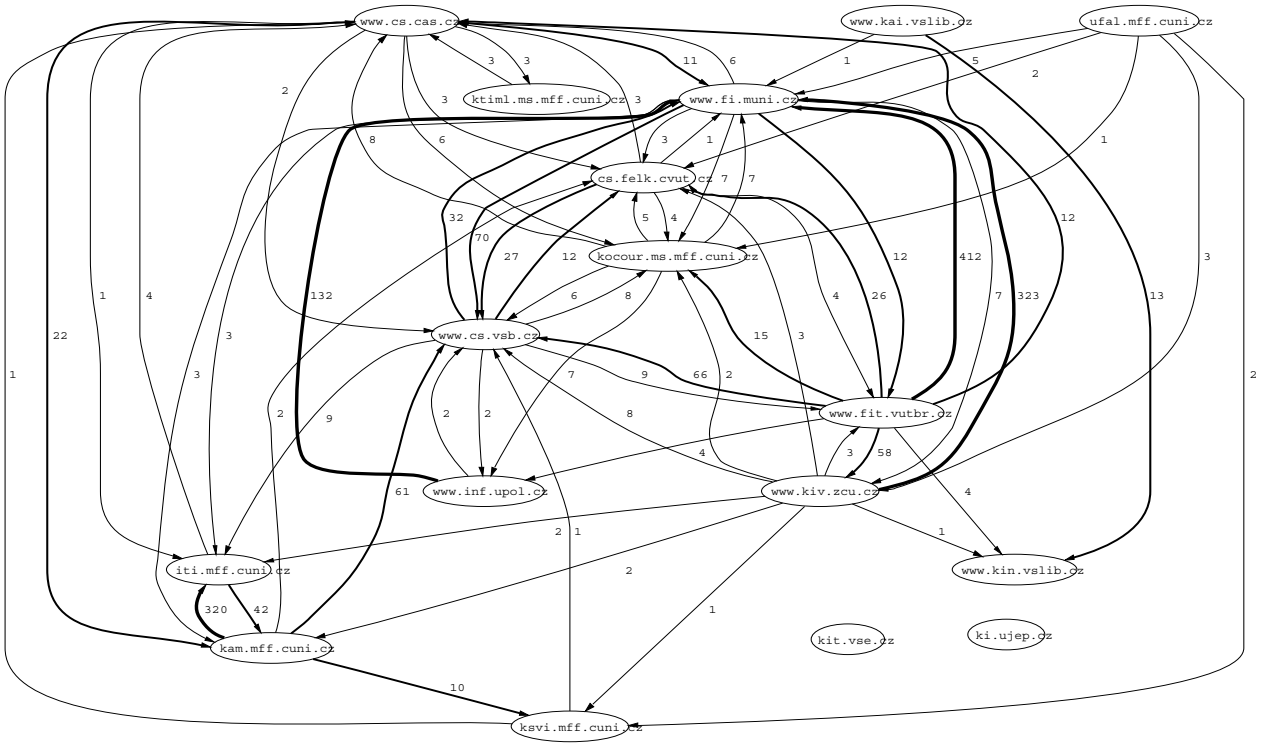


Fig. 1 Citation graph of Web sites

Another problem is dynamically generated Web pages (see the Web site with a significantly higher number of documents). In such a case, two and more URLs (and two or more possible references) represent one document and citations should be counted only once then. This is very annoying, especially regarding the low inter-connectivity of the Web sites. Last but not least, there is a problem with document formats. If a server hosts documents in a format we ignore (e.g. rtf) to a greater extent than the other servers, it can automatically lose citations. All these issues must be taken into account when declaring the most authoritative institutions.

TABLE I
WEB SITES ANALYZED

Server	# Docs	# In-Links	Ratio
www.fi.muni.cz	15 438	924	0.0599
iti.mff.cuni.cz	632	335	0.5301
www.cs.vsb.cz	18 325	243	0.0133
kam.mff.cuni.cz	10 952	69	0.0063
www.kiv.zcu.cz	12 309	68	0.0055
cs.felk.cvut.cz	16 422	56	0.0034
kocour.ms.mff.cuni.cz	11 860	43	0.0036
www.cs.cas.cz	3 226	37	0.0115
www.fit.vutbr.cz	148 682	28	0.0002
www.kin.vslib.cz	46	18	0.3913
www.inf.upol.cz	1 230	13	0.0106
ksvi.mff.cuni.cz	472	13	0.0275
ktiml.ms.mff.cuni.cz	847	3	0.0035
ki.ujep.cz	240	0	0
kit.vse.cz	273	0	0
ufal.mff.cuni.cz	8 316	0	0
www.kai.vslib.cz	2 423	0	0
Total	251 693	1 850	0.0074

III. AUTHORITATIVE INSTITUTIONS

The relations between the examined servers from Table I are depicted in Fig. 1. The citation network is a directed graph with edge weights set to in-link numbers. To enhance visual perception we use three types of edges – normal width lines (less than ten citations), medium width lines, and thick lines (more than 99 citations). By simply looking at the network, we can immediately identify two major candidates for the most important hosts – www.fi.muni.cz and www.cs.vsb.cz. To verify it, we took advantage of the methods from Section 1. First, we computed in-degrees of the nodes in the citation graph without respect to edge weights (i.e. each edge has a weight of one). Note that the in-links in Table I are actually in-degrees respecting edge weights. Then, we computed HITS authorities for the graph nodes and, finally, we generated PageRanks (HostRanks, in fact) for all of the nodes. Table II summarizes the rankings produced by all three algorithms.

We can see indeed that all three measures are strongly positively correlated. The hosts www.cs.vsb.cz and www.fi.muni.cz are in the top three servers whichever ranking method we applied; cs.felk.cvut.cz is highly ranked by In-Degree and HITS whereas www.cs.cas.cz is favoured by PageRank only. Number two by citations, iti.mff.cuni.cz, is handicapped by its strong support from more or less just one server as we may see in Fig. 1. Naturally, the nodes (sites) with a zero in-degree end up at the bottom of each chart. Perhaps, we could prefer those with some out-links at least to those with a zero out-degree. These nodes with no in-links and out-links are entirely isolated and do not participate in the community.

TABLE II
ALGORITHMS AND RANKINGS

Server	In-Degree	HITS	PageRank
www.fi.muni.cz	1 – 2	3	3
iti.mff.cuni.cz	6	5	6
www.cs.vsb.cz	1 – 2	2	1
kam.mff.cuni.cz	7 – 8	8	7
www.kiv.zcu.cz	9 – 12	9	10
cs.felk.cvut.cz	3	1	4
kocour.ms.mff.cuni.cz	4 – 5	4	5
www.cs.cas.cz	4 – 5	6	2
www.fit.vutbr.cz	7 – 8	7	8
www.kin.vslib.cz	9 – 12	11	13
www.inf.upol.cz	9 – 12	10	9
ksvi.mff.cuni.cz	9 – 12	12	11
ktiml.ms.mff.cuni.cz	13	13	12
ki.ujep.cz	14 – 17	14 - 17	14 – 17
kit.vse.cz	14 – 17	14 – 17	14 – 17
ufal.mff.cuni.cz	14 – 17	14 – 17	14 – 17
www.kai.vslib.cz	14 – 17	14 – 17	14 – 17

The phase of finding significant institutions enables us to reduce the set of Web sites that we are going to analyze in the next stage. For example, we might discard the last four sites in Table I, i.e. the least important sites. However, our case study (Czech academic computer science Web sites) has a sufficiently small data set so that no reduction is necessary. Measuring the quality of academic institutions with webometric tools is justified in [14], where Web-based rankings correlated with official rankings.

IV. AUTHORITATIVE RESEARCHERS

In addition to studying links in a collection of computer science Web sites, we were also interested in the documents themselves found on these Web sites. Thus, besides files containing hyperlinks (mainly HTML documents), we downloaded potential research papers as well. In practice, that meant collecting PDF and PostScript files because most research publications publicly accessible on the Web are in these two formats. First, we had to preprocess our download corpus. We unpacked archives and converted observed files to plain text via external utilities. So, at the beginning, we had a 12 thousand set of potential research papers. We discarded duplicates and examined the remaining documents. We used a

simple rule to categorize the documents. In case they included some kind of references section they were considered as papers. In this way, we obtained some 3 600 papers in the end, i.e. over eight thousand documents did not look like research articles.

A. Information Extraction

The next task is to extract information from the papers needed for citation analysis, i.e. names of authors, titles of papers, etc. We employ the same methodology with use of Hidden Markov Models (HMM) as in [10], [12]. The difference is that we work with complete papers, not just with preprocessed headers and references. Moreover, the resulting text files analyzed by HMMs may often have been incorrectly converted to text before. Existence of diacritics in the Czech spelling also worsens the extraction. We did not measure the extraction accuracy due to lack of training objects but, for the above reasons, we suppose it to be significantly lower than those 90 - 93% reported in [12].

We stored the information to a database for a comfortable subsequent querying. We found out that there were about 34 000 distinct author surnames counting together authors in paper headers and references. Strictly said, words identified as surnames. Of course, many of these words were not surnames (they were incorrectly classified) or they were foreign surnames which we did not wish to consider. From the citation graph with "surnames" as graph nodes we determined the most authoritative Czech authors using the three different ranking methods. (The recognition of a Czech surname was done manually.) See Table III for details.

TABLE III
TEN MOST AUTHORITATIVE CZECH CS RESEARCHERS

Rank	In-Degree	HITS	PageRank
1	Hajic	Kucera	Pokorny
2	Kucera	Matousek	Hajic
3	Nesetril	Hajic	Jancar
4	Jancar	Jancar	Matousek
5	Matousek	Nesetril	Brim
6	Panevova	Pala	Kucera
7	Sgall	Smrz	Kratochvil
8	Pala	Sgall	Pultr
9	Kratochvil	Kratochvil	Tronicek
10	Smrz	Panevova	Pala

Let us underline several facts. First, we did not disambiguate the names. Thus, a couple of authors may actually be represented by one name. Even adding first names does not resolve this problem. One solution would be to cluster authors according to their co-authors or publication topics as it is done in [8]. Authors report that this method works well with European (English) names but it achieves accuracy of only 60 – 70% with Chinese names. Second, duplicate citations are handled only in the sense that we remove duplicate documents before analysis. We do not examine whether two or more papers having perhaps only small differences are one publication in reality. Their references to another paper are counted separately.

Third, Czech names often contain diacritics. In international publications written in English, though, diacritics are left out sometimes. The spelling is not unified. Furthermore, conversion to plain text from PDF and PostScript files does not work well and produces more variants of one name. For instance, we found seven commonly used variations of the name "Hajic" (without diacritics here) in our database. In other words, names with no diacritics in their original spelling have a better chance to have their citations counted correctly. All the surnames in Table II are written without diacritics, but we tried to include their frequent versions in citations. The two-way name ambiguity (one author may be known under more names and one name may represent a couple of authors) is to be reflected in future improvements. For all these reasons, the actual citation numbers are less interesting than the ranking itself. Let us not forget that the ranking is a result of those 3 600 papers we got. The question is how it would change if more papers were analyzed.

B. Discussion

Again, we removed duplicate edges and self-citations from the citation graph of authors. The only author occurring among the top three researchers for each method is "Hajic". Other highly ranked names include "Kucera" or "Matousek", but these names are quite common and represent several scientists as we may easily convince ourselves by submitting them to a Web search engine.

Looking mostly just at the first page of results returned by the search engine we can make a guess about the probable affiliations of the authors. For example, for "Hajic" we got ufal.mff.cuni.cz, for "Kucera" we obtained www.fi.muni.cz and kam.mff.cuni.cz, and for "Matousek" we got kam.mff.cuni.cz and www.fit.vutbr.cz. When comparing the sites of these authoritative researchers to those in Table II, we may observe that ufal.mff.cuni.cz, kam.mff.cuni.cz, and www.fit.vutbr.cz have no high positions there. Only www.fi.muni.cz is ranked high. Therefore, it is unclear what impact highly cited authors have on the importance of their institutions' Web sites. We shall have a closer look at this problem in the next section.

V. FURTHER ANALYSIS

In Table IV, we show all the researchers from Table III along with their probable affiliations (i.e. Web sites of their hosting institutions) as we obtained them from a Web search engine. Of course, one researcher may be affiliated with several institutions. This table enables us to produce yet another ranking for Web sites which we will call a ranking by authors. A site is assigned one point for each occurrence in Table IV. Thus, we do not take into account the rank of occurrence. The sum of points obtained for occurrences is the key for ordering by authors (see Table V). The most successful site in this respect is kam.mff.cuni.cz which receives five points. The interpretation of this ranking is that

the more highly cited authors site hosts, the higher rank this site has. So it is a Web site ranking based on paper citations between researchers.

TABLE IV
AUTHORITATIVE RESEARCHERS AND THEIR AFFILIATIONS

Author	Affiliation
Brim	www.fi.muni.cz
Hajic	ufal.mff.cuni.cz
Jancar	www.cs.vsb.cz
Kratochvil	kam.mff.cuni.cz
Kucera	www.fi.muni.cz, kam.mff.cuni.cz
Matousek	kam.mff.cuni.cz, www.fit.vutbr.cz
Nesetril	kam.mff.cuni.cz
Pala	www.fi.muni.cz
Panevova	ufal.mff.cuni.cz
Pokorny	kocour.ms.mff.cuni.cz, cs.felk.cvut.cz
Pultr	kam.mff.cuni.cz
Sgall	ufal.mff.cuni.cz, www.cs.cas.cz, iti.mff.cuni.cz
Smrz	ufal.mff.cuni.cz, www.fit.vutbr.cz
Tronicek	cs.felk.cvut.cz

TABLE V
WEB SITES AND THEIR RANKING BY AUTHORS

Site	Points	Rank
cs.felk.cvut.cz	2	4
iti.mff.cuni.cz	1	6
kam.mff.cuni.cz	5	1
ki.ujep.cz	0	10
kit.vse.cz	0	10
kocour.ms.mff.cuni.cz	1	6
ksvi.mff.cuni.cz	0	10
ktiml.ms.mff.cuni.cz	0	10
ufal.mff.cuni.cz	4	2
www.cs.cas.cz	1	6
www.cs.vsb.cz	1	6
www.fi.muni.cz	3	3
www.fit.vutbr.cz	2	4
www.inf.upol.cz	0	10
www.kai.vslib.cz	0	10
www.kin.vslib.cz	0	10
www.kiv.zcu.cz	0	10

At this stage, we have five different rankings: by in-links, in-degree (each edge has a weight of one), HITS (authority), PageRank, and authors. Table VI summarizes these rankings introduced gradually in Tables I, II and V. Naturally we were interested in the correlations between these orderings. The Spearman correlation coefficients for each pair of rankings are presented in Table VII. They are all significant at the 0.02 level. The very high positive correlation between the first four rankings was expected as it had already been reported before [5]. However, there is still a relatively high correlation between the Authors ranking and the others – more than 0.6. This implies that we can answer the question from Section IV.B by saying that highly cited authors do have a positive impact on the importance of their departments' Web sites.

TABLE VI
RANKINGS SUMMARY

Site	In-Links	In-Degree	HITS	PageRank	Authors
cs.felk.cvut.cz	6	3	1	4	4
iti.mff.cuni.cz	2	6	5	6	6
kam.mff.cuni.cz	4	7	8	7	1
ki.ujep.cz	14	14	14	14	10
kit.vse.cz	14	14	14	14	10
kocour.ms.mff.cuni.cz	7	4	4	5	6
ksvi.mff.cuni.cz	11	9	12	11	10
ktiml.ms.mff.cuni.cz	13	13	13	12	10
ufal.mff.cuni.cz	14	14	14	14	2
www.cs.cas.cz	8	4	6	2	6
www.cs.vsb.cz	3	1	2	1	6
www.fi.muni.cz	1	1	3	3	3
www.fit.vutbr.cz	9	7	7	8	4
www.inf.upol.cz	11	9	10	9	10
www.kai.vslib.cz	14	14	14	14	10
www.kin.vslib.cz	10	9	11	13	10
www.kiv.zcu.cz	5	9	9	10	10

TABLE VII
CORRELATION BETWEEN RANKINGS

	In-Links	In-Degree	HITS	PageRank	Authors
Citations	X	0.89	0.89	0.86	0.63
In-Degree	X	X	0.96	0.96	0.65
HITS	X	X	X	0.95	0.64
PageRank	X	X	X	X	0.63
Authors	X	X	X	X	X

VI. CONCLUSIONS AND FUTURE WORK

Notions of popularity or authority, commonly used in social networks such as scientific publications, have also been adopted for the World Wide Web in recent years. The most popular ranking techniques are link-based methods like In-Degree, PageRank, and HITS. We present a methodology and a case study of finding authoritative researchers on the Web. We applied the ranking algorithms to a small set of Czech academic computer science Web sites and determined the most authoritative ones. (We also tried to examine Slovak computer science departments, but the data set was too small.) This step normally enables reducing the volume of data to be analyzed since we could continue finding researchers on the more important sites only. Further, we analyzed the research papers publicly available on the sites and we determined the most significant researchers by applying several ranking techniques to the citation graph. The results we achieved are not quite reliable due to the constraints and problems mentioned above, but we believe that our methodology is practical as we have shown in our experiments.

Further, we generated yet another ranking for institutions based on citations in papers. This meant assigning affiliations

to each researcher in Table III. We were interested in the difference between the top ranked sites determined via analysis of Web links (Table II) on one hand and those based on paper citations on the other hand (Table V). We have discovered that there is a relatively high correlation between the link-based (Web) and citation-based (papers) ranking. This result will have to be verified with larger data. (Currently, we are working on French academic sites.) In our future research we would like to concentrate on the issue of combining Web and paper authorities. The methodology we have developed is general, which will enable us to focus on other areas of the Web as well.

ACKNOWLEDGMENT

We would like to thank the authors of many freely available tools we used during our research – SQLite, XPDF, GhostScript, QuickGraph, Unzip, Gzip, BsdTar and others.

REFERENCES

[1] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," in *Proc. 7th World Wide Web Conference*, Brisbane, Australia, 1998, pp. 107–117.

[2] S. Chakrabarti, *Mining the Web: Analysis of Hypertext and Semi Structured Data*. San Francisco, CA: Morgan Kaufmann Publishers, 2003, pp. 209–218.

[3] S. Chakrabarti, B. E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Spectral Filtering for Resource Discovery," in *Proc. ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, Melbourne, Australia, pp. 13-21, 1998.

[4] M. Diligenti, M. Gori, and M. Maggini, "A Unified Probabilistic Framework for Web Page Scoring Systems," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, no. 1, 2004, pp. 4–16.

[5] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a Unified Framework for Link Analysis," in *Proc. 25th ACM SIGIR Conf. Research and Development in Information Retrieval*, Tampere, Finland, 2002, pp. 353–354.

[6] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a Unified Framework for Link Analysis," Lawrence Berkeley National Laboratory, University of California, Berkeley, CA, Technical Report 49372, Nov. 2001.

[7] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring Web Communities from Link Topology," in *Proc. 9th ACM Conference on Hypertext and Hypermedia*, Pittsburgh, PA, 1998, pp. 225–234.

[8] H. Han, H. Zha, and C. L. Giles, "Name Disambiguation in Author Citations Using a K-way Spectral Clustering Method," in *Proc. 5th ACM/IEEE-CS Int. Conf. Digital Libraries*, Denver, CO, 2005, pp. 334-343.

[9] J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM*, vol. 46, no. 5, 1999, pp. 604-632.

[10] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, "Automating the Construction of Internet Portals with Machine Learning," *Information Retrieval Journal*, vol. 3, no. 2, 2000, pp. 127-163.

[11] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Computer Science Department, Stanford University, CA, Technical Report 1999-66, Nov. 1999.

[12] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction," in *Proc. AAAI'99 Workshop Machine Learning for Information Extraction*, Orlando, FL, 1999, pp. 37–42.

[13] M. Thelwall, "Extracting Macroscopic Information from Web Links," *Journal of the American Society for Information Science and Technology*, vol. 52, no. 13, 2001, pp.1157-1168.

[14] M. Thelwall, "The Relationship between the WIFs or Inlinks of Computer Science Departments in UK and Their RAE Ratings or Research Productivities in 2001," *Scientometrics*, vol. 57, no. 2, 2003, pp. 239-255.