RANKING ALGORITHMS FOR WEB SITES
Finding Authoritative Academic Web Sites and Researchers

Dalibor Fiala, François Rousselot
INSA Strasbourg, 24 bd de la Victoire, 67084 Strasbourg, France
dalibor.fiala@insa-strasbourg.fr, francois_rousselot@insa-strasbourg.fr

Karel Jezek
Dept. of Comp. Science and Engineering, University of West Bohemia, Plzen, Czech Republic
jezek_ka@kiv.zcu.cz

Keywords: Authorities, citation analysis, prestige, ranking algorithms, Web mining.

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

1 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 are bibliographic citations in the scientific literature. This kind of analysis has become essential in the Web domain as well.

In the Web domain, citations are links among 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 ranking 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 or HITS (Chakrabarti, 2002) or simple like In-Degree which just counts in-links. Some studies have shown that all three measures are strongly positively correlated (Ding, 2002). There exist many modifications, e.g. PageRank for bibliographic citations (Sidiripoulos, 2005). Closest to our work is the research in (Thelwall, 2003), but in addition to the relations between Web sites we also studied the contents of the documents found on them. Other authors have tried to determine the importance of Web sites of Universities rather than departments as we have done. (See http://www.webometrics.info.)

2 EXPERIMENTS

In this section, we will describe our experiment with Web sites of French computer science departments. 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. First, we had to draw up a list of laboratories. To do this, we looked up in Web directories and we also submitted queries to Web search engines. From these Web pages, we manually selected 80 final sites that constituted our set of departments. The first goal was to determine the most authoritative sites as of May 2006.

2.1 Authoritative Institutions

To accelerate the process of creating the Web graph, we did not make use of a Web spider of our own, but we took advantage of a service provided by the

search engine Yahoo! We submitted to it queries in this form:

site:www.loria.fr linkdomain:www.irisa.fr

which returns the number of documents on www.loria.fr containing at least one link to documents on www.irisa.fr. For us, it is a weight of the edge from www.loria.fr to www.irisa.fr. We had to construct 6 320 queries in this way. Of course, the construction and submission of queries, storing of results, and the graph creation were automated. (The figure of the Web graph with 393 edges is available at http://home.zcu.cz/~dalfia/papers/France.svg.)

The drawbacks of relying solely upon search engines are discussed a great deal in (Thelwall, 2003). The problem consists primarily in "instability" of the results. This means that results obtained one day differ from those of another one. Another disadvantage is that the results are not transparent. We do not know which document formats are taken into account, how duplicate documents are treated, etc.

2.1.1 Results and Discussion

We applied three ranking methods to the Web graph of 80 sites of choice. 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). Then, we computed HITS authorities for the graph nodes and, finally, we generated PageRanks (HostRanks, in fact) for all of the nodes. We can see the results in tables 1 and 2. The sites are sorted by in-links (citations), i.e. by the total number of links to this site from other sites in the set (with some limitations imposed by the search engine). The first place belongs to www-futurs.inria.fr, whose positions achieved by the other methods, though, are much worse. We can suppose that the reason for this is a very strong support from a particular site. (After inspecting the Web graph, we can see that it is www.lifl.fr.) The following sites always have high ranks - www-sop.inria.fr, www.loria.fr, www.lri.fr. We can surely consider them as authoritative.

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, existence of server aliases, 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 or by the number of staff of the corresponding institution (Thelwall, 2003).

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 eight sites in Table 2, i.e. the least important sites. However, our case study (French 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 (Thelwall, 2003), where Web-based rankings correlated with official rankings.

2.2 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, we downloaded potential research papers from the sites in question. 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 about 45 thousand 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 16 000 papers in the end, i.e. over thirty thousand documents did not look like research articles.

2.2.1 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 (Seymore, 1999).

We had to construct a graph with authors (identified by surnames and initials of their first and middle names) as nodes and citations in publications as edges. The final graph (without duplicate edges and self-citations) had almost 86 000 nodes and about 477 000 edges. Strictly said, when we talk about surnames, we mean 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 French authors using the three different ranking methods. (The recognition of a French surname was done manually.) See Table 3 for details.

Table 1: Ranking of French Web sites. (1 – 40).

	Site	InD	HITS	PR
1	www-futurs.inria.fr	45	41	53
2	www-sop.inria.fr	1	1	9
3	www.loria.fr	1	5	3
4	www.lri.fr	6	6	10
5	www-rocq.inria.fr	13	12	28
6	www.irisa.fr	4	3	18
7	www.lifl.fr	5	7	4
8	www.lix.polytechnique.fr	20	17	26
9	dpt-info.u-strasbg.fr	39	53	43
10	www.inrialpes.fr	6	8	2
11	www.irit.fr	9	4	8
12	www.liafa.jussieu.fr	13	15	39
13	www.lirmm.fr	1	11	1
14	www.labri.fr	13	13	30
15	www-leibniz.imag.fr	10	14	13
16	liris.cnrs.fr	13	16	11
17	www.prism.uvsq.fr	13	25	5
18	www.di.ens.fr	34	26	44
19	www.lip6.fr	20	21	40
20	www.laas.fr	6	2	27
21	dep-info.u-psud.fr	61	58	69
22	www-lil.univ-littoral.fr	25	34	35
23	www-verimag.imag.fr	25	37	16
24	www.i3s.unice.fr	25	31	7
25	eurise.univ-st-etienne.fr	25	23	32
26	www-lsr.imag.fr	34	26	37
27	www.info.unicaen.fr	13	10	14
28	www-timc.imag.fr	12	9	17
29	www-sic.univ-poitiers.fr	45	46	50
30	cedric.cnam.fr	25	22	38
31	www.dil.univ-mrs.fr	39	54	25
32	www-lmc.imag.fr	25	29	34
33	www.info.univ-angers.fr	34	44	24
34	lifc.univ-fcomte.fr	20	32	21
35	eric.univ-lyon2.fr	10	19	6
36	www-id.imag.fr	25	33	15
37	www-lipn.univ-paris13.fr	13	24	29
38	dept-info.labri.fr	25	18	36
39	www.isima.fr	39	43	48
40	sis.univ-tln.fr	20	28	12

Table 2: Ranking of French Web sites. (41 – 80).

	Site	InD	HITS	PR
41	www-clips.imag.fr	25	30	22
42	www.lisi.ensma.fr	39	40	33
43	www-info.iutv.univ-paris13.fr	61	69	72
44	www.lif.univ-mrs.fr	34	36	31
45	www.cril.univ-artois.fr	39	35	41
46	www.li.univ-tours.fr	34	42	45
47	citi.insa-lyon.fr	45	45	54
48	deptinfo.unice.fr	39	38	46
49	msi.unilim.fr	52	55	64
50	www.iut-info.univ-lille1.fr	61	62	65
51	www.lia.univ-avignon.fr	20	20	23
52	lil.univ-littoral.fr	52	48	57
53	lisi.insa-lyon.fr	45	39	47
54	www.isc.cnrs.fr	45	71	19
55	www.if.insa-lyon.fr	61	72	52
56	sirac.inrialpes.fr	61	62	62
57	phalanstere.univ-mlv.fr	45	65	20
58	www.lalic.paris4.sorbonne.fr	45	47	61
59	www.icp.inpg.fr	52	51	49
60	www-valoria.univ-ubs.fr	52	57	51
61	lihs.univ-tlse1.fr	52	48	60
62	www.epita.fr	52	67	42
63	llaic3.u-clermont1.fr	52	51	56
64	lsiit.u-strasbg.fr	52	48	57
65	liuppa.univ-pau.fr	52	56	66
66	wwwhds.utc.fr	61	66	55
67	www.depinfo.uhp-nancy.fr	61	68	59
68	lrlweb.univ-bpclermont.fr	61	62	62
69	www-lium.univ-lemans.fr	61	70	67
70	www.dptinfo.ens-cachan.fr	61	58	68
71	www.ai.univ-paris8.fr	61	58	69
72	www.lita.univ-metz.fr	61	58	69
73	dept-info.univ-brest.fr	73	73	73
74	lina.atlanstic.net	73	73	73
75	lis.snv.jussieu.fr	73	73	73
76	psiserver.insa-rouen.fr	73	73	73
77	www.listic.univ-savoie.fr	73	73	73
78	www-info.enst-bretagne.fr	73	73	73
79	www.info.iut.u-bordeaux1.fr	73	73	79
80	www.info.iut-tlse3.fr	73	73	79

2.2.2 Results and Discussion

The rankings produced by In-Degree and HITS are very similar (the top five researchers are exactly the same) whereas that by PageRank is rather different. The authors in In-Degree and HITS are more or less the same (only in various positions), but PageRank introduces some new names. However, there are two authors ("Halbwachs N" and "Berry G") occurring in top five of each ranking. We can certainly call these researchers authorities.

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 (Han, 2005). 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.

Table 3: Authoritative French CS researchers.

	In-Degree	HITS	PageRank
1	Halbwachs N	Halbwachs N	Cahon S
2	Caspi P	Caspi P	Berry G
3	Sifakis J	Sifakis J	Filiol E
4	Berry G	Berry G	Halbwachs N
5	Benveniste A	Benveniste A	Zhang Z
6	Abiteboul S	Nicollin X	Benveniste A
7	Maler O	Cousot R	Lavallée S
8	Nicollin X	Raymond P	Dombre E
9	Cousot P	Cousot P	Boudet S
10	Cousot R	Abiteboul S	Dégoulange E
11	Raymond P	Maler O	Gourdon A
12	Bouajjani A	Asarin E	Abiteboul S
13	Asarin E	Comon H	Charpin P
14	Comon H	Bouajjani A	Carlet C
15	Zhang Z	Coupaye T	Cohen G
16	Berstel J	Berstel J	Troccaz J
17	Meyer B	David B	Abdalla M
18	Florescu D	Arnold A	Payan Y
19	Baccelli F	Pilaud D	Cousot R
20	Leroy X	Bruneton E	David R
21	Bruneton E	Maraninchi F	Cousot P
22	Flajolet P	Meyer B	Caspi P
23	Arnold A	Leroy X	Sifakis J
24	Graf S	Bensalem S	Deransart P
25	Cohen J	Graf S	Maler O
26	Coupaye T	Tripakis S	Bouajjani A
27	Pilaud D	Lakhnech Y	Dubois D
28	Lakhnech Y	Bozga M	Caron P
29	David R	Gautier T	Pierrot F
30	Faugeras O	Liu J	Raymond P

Third, deciding whether or not a researcher is French is inherently subjective. Our decision was based on searching with several general and specialized search engines. Ideally, we found the researcher's home page hosted by a French Web site or affiliation to a French institution given in an article. Of course, by French authors we also mean those who had lived and worked in France for a long time. We are aware that this feature is particularly fuzzy. The 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 16 000 papers we got. The question is how it would change if more papers were analyzed.

3 CONCLUSIONS

We present a methodology and a case study of finding authoritative researchers on the Web. We applied several ranking algorithms to a set of French academic computer science Web sites and determined the most authoritative ones.

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. The methodology we have developed is general, which will enable us to focus on other areas of the Web as well.

This work was supported in part by the Ministry of Education of the Czech Republic under Grant 2C06009.

REFERENCES

Chakrabarti, S. (2002). *Mining the Web: Analysis of Hypertext and Semi Structured Data*. Morgan Kaufmann.

Ding, C., He, X., Husbands, P., Zha, H., and Simon, H. (2002). PageRank, HITS and a Unified Framework for Link Analysis. Proc. 25th ACM SIGIR Conference Research and Development in Information Retrieval, 353–354, Tampere, Finland.

Han, H., Zha, H., and Giles, C. (2005). Name Disambiguation in Author Citations Using a K-way Spectral Clustering Method. *Proc. JCDL'05*, Denver, CO.

Seymore, K., McCallum, A., and Rosenfeld, R.. (1999). Learning Hidden Markov Model Structure for Information Extraction, *Proc. AAAI'99 Workshop Machine Learning for Information Extraction*, 37–42.

Sidiropoulos, A., Manolopoulos, Y. (2005). A Citation-Based System to Assist Prize Awarding. SIGMOD Record, Vol. 34, No. 4, 54-60.

Thelwall, M. (2003). 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, 239-255.