Limits and Feasibility of Cositation Method on the Web. An Experiment on the French Speaking Web

Camille Prime-Claverie[*], Michel Beigbeder[**] and Thierry Lafouge[***]

[*]*cclaveri@u-paris10.fr*
Laboratoire CRIS, Université Paris 10 Nanterre, 200 avenue de la république
92001 Nanterre Cedex, FRANCE

[**]*mbeig@emse.fr*
Laboratoire RIM - G2I, Ecole Nationale Supérieure des Mines de Saint-Etienne, 158, cours Fauriel,
42023 Saint-Etienne Cedex, FRANCE

[***]*lafouge@univ-lyon1.fr*
Laboratoire URSIDOC, Université Claude Bernard Lyon 1, 43, Bd du 11 novembre 1918,
69622 Villeurbanne Cedex, FRANCE

1 Introduction

There are increasing numbers tools on the Web for searching or structuring information (Brin & Page, 1998), (Gibson, Kleinberg & Raghavan, 1998), (Chakrabarti, Dom & Indyk, 1998), (Kumar and al. 1999) (Flake and al., 2002). They use the Web structure, which provides a relationship between the existing pages through hypertext links. The Web can be formally modeled as a graph G = (S,A) where S, all the nodes, represents the Web pages and A, all the arcs, represents the hypertext links between the pages. This graph is usually called a "citation graph" by analogy with the network of scientifc publications well known in scientometry.

After recalling the possible analogies between methods using the Web structure and those coming from scientometrics, this article presents experiments allowing us to study the feasibility of cositation methods on a large corpus. This study follows upon the encouraging results obtained on a small corpus by our method of propagating metadata values, which uses the page cositation method after structuring the Web (Prime-Claverie, Beigbeder & Lafouge, 2004).

2 Analysis of hypertext links in webometrics

Similarly to the world of scientific publications, the Web is a multi-author space in which documents can be linked together. This implies a structure and self-organization of the Web providing different courses of reading. Various authors (Rousseau, 1997), (Ingwersen, 1998), (Egghe, 2000) have devoted themselves to the analogies and differences between these two systems (the scientific publication and Web graph networks), which form a large part of current work in webometrics. More specifically, Egghe (2000) and Prime and al. (2002), give us some limits to this analogy.

- The major difference between a scientific article and a Web page lies in the volatility and the possibility of updating the Web page. This results in the Web graph having a
- different shape from that of the traditional citation graph. In fact, whereas two scientific articles cannot have reciprocal citation, it is perfectly possible for two pages to mutually cite each other (Barabasi, 2002). The one-way aspect of the citation graph disappears completely in the Web graph, which moreover allows a wider application of the methods for studying graphs coming from the analysis of social networks (Wasserman & Faust, 1994). The problem of updating also raises the question of the relevance of the citation; how to be sure that the content of a page cited by an author has not completely changed or even that this page has not disappeared.
- The phenomenon of graph node duplication (Web pages) occurs frequently on the Web, whereas there are none (or at least very much to be avoided) in the scientific publication network. This duplication is justified by the desire for faster access to resources. In fact, some very large and frequently consulted servers avoid obstructions by having several copies of their sites on different points of the planet. We then speak of mirror sites. However, these reproductions generate a replication of the hypertext links, which form a

Proceedings of ISSI 2005, pages 78–86.
Edited by P. Ingwersen and B. Larsen.

significant limit to the transposition of indicators based on the citation.

- In this analogy, the hypertext link materializes a citation. However, we also know that it is very often used for other purposes such as navigation within the same site. Moreover, we cannot ignore the presence of publicity links on the Web. Certain researchers question the manner of typifying the links, but such a task cannot be imposed on authors.

This last point leads us to present a new line of research in webometrics; the study of the citation reasons on the Web. To complete the analogy between the analysis of citations and the Web, certain scientometry specialists question an author's reasons for creating a hypertext link (Kim, 2000). Research undertaken by Thelwall (2003) and Chu (2004) distinguish between:

- Intra-site links: links connecting pages hosted on the same site
- And inter-site links: links connecting pages hosted on different sites

Links within sites connect pages created by the same authority. They serve firstly for structuring documents (links between the various parts of a document), and secondly for navigation or citation between resources on the same site. Citation between pages on the same site is similar to self-citation and does not necessarily imply a "noteworthy link" between the citing and cited pages. A link to a page hosted on another site invites the user to quit the site where he is to visit another. This kind of invitation is generally justified by the interest that the author has in the other site. The term "*sitation*" used by Rousseau since 1997 (Rousseau, 1997) (contraction of the expression "site citation") is used to describe the inter-site links.

The research of Chu (2004) shows that the sitation reasons are rather different from, and above-all less complex than, the citation reasons. A sitation is generally used for mentioning an interesting site. It is seldom used for arguing, comparing or presenting ideas. Furthermore, it is less precise. Whereas an article may be cited for the interest in a single paragraph or specific sentence, the sitation generally aims at one page at least and often at the contents of an entire site.

The results obtained by Thelwall (2003) and others (Wilkinson., Harries,, Thelwall, & Price, 2003) also show that the sitation reasons are rather different from the citation reasons. His research concerns the sitation reasons of English universities. The corpus used comprises 111 university sites and 19,438 links pointing to the home pages of these universities. For this experiment, 100 were drawn at random from among the 19,438 links. A qualitative analysis attempts to identify the reasons for these links. On completion of these observations, the author describes four types of possible link.

- General navigation links: they are for navigating towards information hosted on other sites that the author considers interesting. These links are described as general navigation because the information found on the target page does not share the topic of the source page.
- Ownership links: they make it possible to assert intellectual ownership of a document.
- Social links: they are links to partners and colleagues.
- Free links: links that, according to Thelwall, are without any particular reason for communication. Examples are links to the university where one has studied, links to a former company, etc.

This procedure obviously cannot give an exhaustive taxonomy of the sitation reasons because they are related to the corpus of the study. In this experiment, the pages of the corpus (to which the links are pointed) are of a particular type. They are home pages of English universities. In general, home pages are synopses. They give access to the site resources and seldom present basic information. This explains why no cognitive link, or at least no thematic navigation link, appears in this experiment. Without claiming to be exhaustive, we propose supplementing this classification with:

- Thematic navigation links for navigating between pages of the same topic,
- and the cognitive links (undoubtedly rare on the Web) pointing to pages evoking or arguing the ideas on the initial page.

Moreover, among these "free links", meaning those appearing without particular communication

reason (between the citing and cited pages), we could mention publicity links. These are free links within the meaning that they contribute nothing from a social or semantic point of view, but which are financially profitable.

3 Webometric methods for structuring the graph

Various relationships from the Web graph may be used for information retrieval or information structuring. The three most obvious are the sitation relationship, the relationship of link coupling and the cositation relationship (Bjorneborn & Ingwersen, 2004) (see Fig 1).

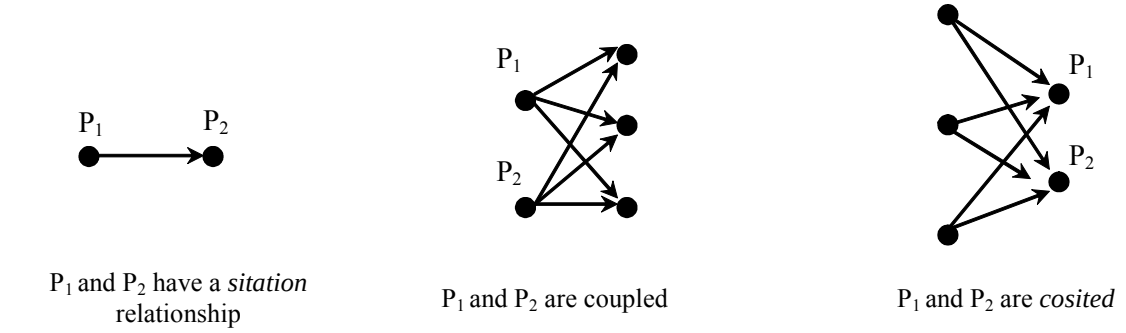

Fig 1. Possible relationships in the sitation graph

3.1 The different methods

In the preceding section, we noted that sitation reasons differ very much between hypertext links. Thus, the majority of the pages linked by a sitation (existence of a hypertext link between two pages hosted on two distinct sites) share only a small proportion of common characteristics. For example, two pages linked by a thematic or cognitive link share the same topic, which is not the case of pages linked by a social or general navigation link. This shows the limits of using only the analysis of sitations in the contexts of information search or structuring of information.

By analogy with the method of bibliographical coupling coming from bibliometrics (Kessler, 1963), two Web pages have a coupling relationship (are coupled) if they share identical hypertext links. Various empirical studies show that strongly coupled pages (sharing many identical links) are generally doublet pages (partial or total copies). Various methods , (Bharat, Broder, Dean, & Henzinger, 2000) (Prime, Bassecoulard & Zitt, 2002) used for detecting mirror sites, are based on the coupling strength (number of common output links). Generally, coupling favors the bringing together of pages with many links (directories, for example). Pages without any external links, which is often the case for contents pages, are never coupled with others.

A cositation relationship is one existing between two pages cited simultaneously by a page hosted on a site different from the cited pages. Various authors assume that two pages often cited together, as regards their respective frequency of sitation, share common properties. Here, several authors have a common reason for citing these two pages together. The good results obtained in the various transposition experiments of the cocitation method on the Web (Larson, 1996), (Piktow & Pirroli, 1997), (Prime C., Bassecoulard & Zitt, 2002) allow us to confirm this assumption. There is however a significant difference between the traditional cocitation method and the study of the cositation relationships on the Web. The traditional method aims at structuring a search domain for a restricted and generally recent period. It connects two sets: all publications belonging to the domain and the study period (the citing articles) and all the publications cited by them (the cited articles). This method proceeds in two stages: the grouping of cited articles in clusters using a similarity based on the cocitation strength of the cited articles; then the assignment of the citing articles to these clusters. (For

example, an article is assigned to a cluster with threshold "x" if it has at least "x" references to articles contained in the cluster). Thus, the matrix of cocitation of the cited articles does not contain all the cocitation relationships between the cited articles, but only those resulting from the corpus of citing articles. In the traditional method, the structuring of the cited articles is only an intermediate stage, which only has meaning from the point of view of the citing articles. Since the Web does not have diachronic nature, it is quite possible to directly structure it by examining the existing cositation relationships between the pages, with a limit however: too recent pages are not cited or only slightly, and cannot take part in the structuring.

3.2 Analogy and limits of the cositation method

In this paragraph, we call to mind the analytical limits of sitations and particularly structuring by the cositation principle.

One of the major problems concerns meaningless sitations (free links listed by Thelwall, publicity links). Whereas in traditional bibliometry, the most cited elements are structuring (contrary to words), what is the situation on the Web where publicity relationships are very much present? Do we find the same configuration as for words, where the most frequent elements are part of the noise? If studies could confirm this assumption, a Bradford selection (Bradford 1934) could be envisaged to eliminate the most cited elements (noise). Nevertheless, free links appearing at low frequency could not be removed. However, a major interest arises from the use of a cocitation based index: if there are many meaningless links on the Web, the cositation of pages formed without particular reason does not seem to have significant frequency.

- Within Web sites, all pages are generally cited at least once. This allows one to navigate fully within the site and read all the pages. However, distinctly fewer pages receive external citations. In fact, each site only has a few entry points. Entry points are pages by means of which arrival on the site occurs coherently. These pages are found at various levels in the sites. At the top, of course, are the home pages, but also at lower levels are pages representing the beginning of a logical document. The results of the Chu study (Chu, 2004) show that a sitation is not very accurate and that it often targets the contents of an entire site. Thus, it is often the home pages that are cited. However, the sitation of pages other than entry points is quite possible, but rare. This is why the analysis of sitations is mainly concerned with the links received through the entry points of each site. Sitation based methods (meaning on external links) applied to the Web, particularly take into account the entry points and not the totality of the Web pages.
- Another limit concerns the transposition of the cocitation principle onto the Web. In traditional bibliometry, the distribution of out-degrees (emitted citations) follows Gaussian laws. The number of bibliographical references per article varies between 20 and 50 . On the Web, the distribution of out-degrees follows hyperbolic laws (Broder and al., 2000). The probability that a page only emits a single sitation is relatively strong. Thus, it is quite possible that an entry point, even if frequently cited, is never cocited. In other words, it is possible that a page is always cited by pages emitting only one sitation.

Moreover, various technical problems need to be overcome to use these different methods on a sitation graph. The analysis of sitations is based on inter-site relationships. However, it is quite difficult to automatically identify sites. One solution is to examine the URL . Remember that the standard form of a URL is written as follows:

```
protocol://engine-name.domain-name/file-
address;parameters?request#argument
```

Two possible approximations would imply that a site corresponds to:

* The pages hosted under a domain name (for example `emse.fr`)
* Or the pages hosted under the concatenation of engine name and domain name (for example `www.emse.fr` or `rim.emse.fr`).

These two approximations generate errors. Firstly, several sites can coexist under the same domain name. The majority of access providers offer to host their customers' sites under their domain

name. Secondly, big sites can be hosted on several engines.

- Another difficulty is obtaining the actual Web graph. The graph is revealed by browsing through the Web from link to link. This involves managing:
    * Errors in the use of HTML syntax and in writing URLs
    * Alias URLs, meaning the different links possible for naming the same source file
    * Duplicated sites (or mirror sites)
    * Pages inaccessible to the robots, which are described in the file at the root of the sites (robots.txt)

Finally, there is not just one Web graph but several possible graphs, depending on the methods chosen for revealing the graph.

4 Experiments on the French-speaking Web: feasibility of the cositation methods

To solve the difficulties of information retrieval due to the heterogeneous nature of the Web, we have proposed a semi-automatic procedure for characterizing sites and Web pages (Prime-Claverie, Beigbeder & Lafouge, 2004). This procedure is based on an analysis of the links, and makes particular use of the cositation principle. The encouraging results allow us to envisage experimenting on a larger scale and think about integrating search engines into our method. But first of all, we think it is necessary to evaluate some of the previously described limits of the analysis of the sitations, on a big corpus representative of the Web. More specifically, we want to study the two limits resulting in significant reductions in corpus. These are firstly the fact that only the entry points are mainly cited; secondly the fact that the probability of an entry point being never cocited is considerable.

In this section, we present an exploratory study allowing us to estimate the proportion of cosited pages on the Web, and which can be characterized by cositation based methods. This is the first step towards a study of the shape of the Web cositation graph.

4.1 Presentation of the collection

For this exploratory study, we have used a corpus named Wfr4, consisting of 5,057,642 pages. These pages were collected on the Web in December 2000 using the robot[1] CLIPS-Index developed by Mathias Géry and Dominique Vaufreydaz, members of the CLIPS[2] laboratory of Grenoble university (France). This robot can collect up to three million pages a day going through the Web. It performs the following operations: choice of URL to be collected from within a local list; retrieval and storage of the corresponding page; analysis of its source code and extraction of the URL cited; addition of the URL cited to the list of pages to be collected. In the Wfr4 collection, all the pages belong to French-speaking geographical domains (table 1), which does not mean that all the documents are in French. Indeed, many sites offer versions of their documents in several languages.

Country	France	Belgium	Luxembourg
Number of sites	29,441	8,851	1,152

Table 1: The most represented extensions in the collection

4.2 Characteristics of the collection graph

The first step in this study consists of building the graph of the relationships between the pages in the collection. This consists of 5,057,642 distinct URLs, each one pointing to a data file corresponding to a Web page. To reduce the size of the corpus (number of graph peaks), we have decided not to consider Web pages created dynamically. Thus, URLs containing requests are truncated at the question mark[3]. The term that we use to describe these truncated URLs is "netpath".:

[1] `http://www-mrim.imag.fr/membres/mathias.gery/CLIPS-Index/`

[2] `http://www-clips.imag.fr/`

[3] Remember that the standard form of a URL address is written in the following way: protocol://engine.domain-name/file-address;paramétres?request#argument

For example, the following URL :

`http://www.enssib.fr/article.php?id=170&cat=La+recherche&id_cat=170`
designates a page ;
`http://www.enssib.fr/article.php` is its netpath ;
`http://www.enssib.fr/` is its site.

To reveal the graph, meaning the relationships between the netpaths in this collection, the following steps are performed:

- Extraction of the URLs of the citing pages and the URLs of the cited pages
- Standardization of the URLs; for example by writing the engine names in lower-case letters or deleting the port number when it is the default port, etc.
- If necessary, cutting the URLs at the question mark (obtaining the netpath)
- Construction of tables for associating a number to each netpath and each site. In this experiment, a site is defined as the concatenation of the engine name and domain name.

The characteristics of the Wfr4 collection are:

- 5,057,642 Web pages
- 3,823,589 netpaths (URLs truncated at the question mark)
- 43,462 Web sites (concatenation of the engine name and domain name)

4.2.1 Distribution of the number of pages per site

Firstly, we were interested in the size of the sites as regards number of pages and number of netpaths. Figure 2 shows the distribution of the number of pages and number of netpaths per site. We see that these distributions have the same shape and that they conform to the laws of information (hyperbolic laws). However, we were surprised at the quantity by sites containing a very low number of pages: approximately 20,700 sites contain only one page or one netpath (which is 47.7% of the sites in the collection).

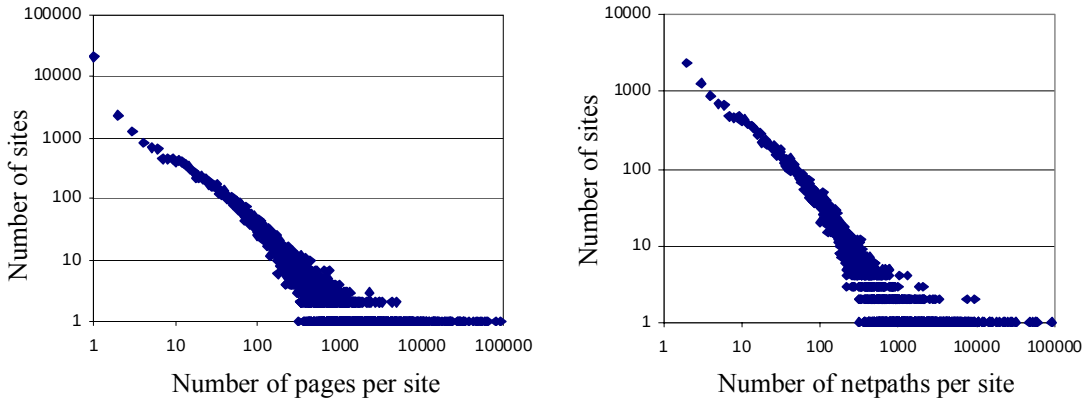

Fig. 2: Distribution of the number of pages and netpaths

4.2.1 Distribution of the number of entry points per site

In this experiment, all netpaths receiving at least one external citation are regarded as entry points. Figure 3 gives the distribution of the number of entry points per site. This distribution also has a hyperbolic shape. Figure 4 shows that there is no correlation between the number of netpaths per site and the number of entry points per site (coefficient of correlation equal to 0.46). Navigation in the data file shows that generally, when a site has several entry points, only a few have a very high citation frequency.

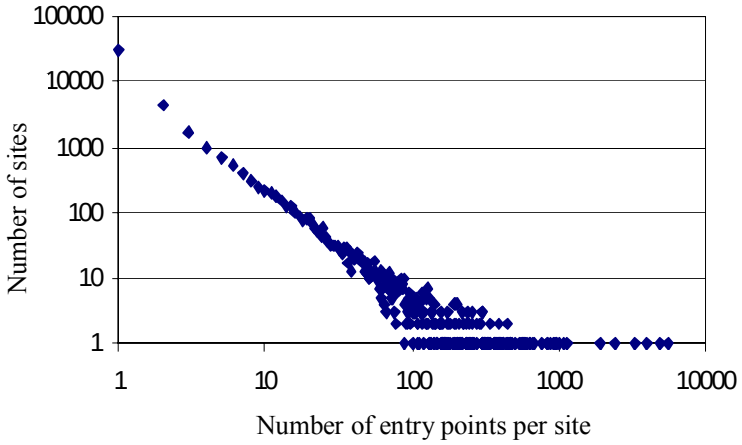

Fig. 3: Distribution of the number of entry points per site

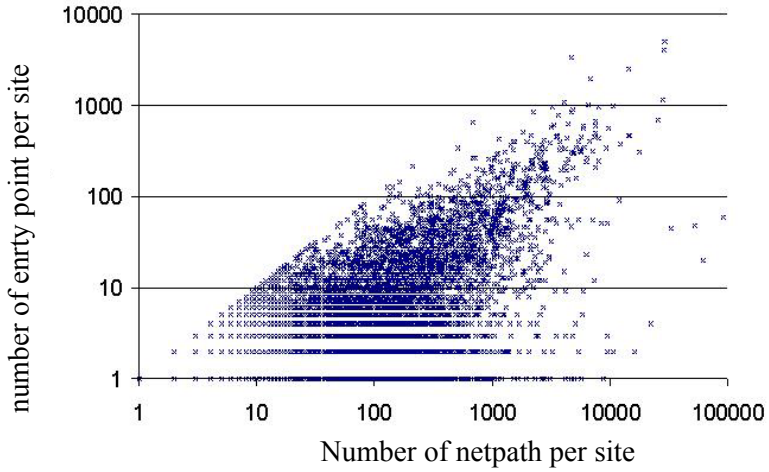

Fig. 4: Relation between number of entry point and netpaths

4.3 Characteristics of the sitation graph

Various authors (Kumar & al., 1999) (Albert, Jeong, & Barabasi, 1999) (Broder et al., 2000) have been interested in the distributions of the citations received (in-degree) and emitted (out-degree) by the Web pages.

The originality of our work has been to study the graph of external citations, meaning the graph of inter-site citations or the sitation graph. Our sitation graph is an oriented graph (not valued), which contains:

- 1,004,152 peaks (netpaths)
- 3,324,703 arcs (external relationships between the netpaths)
- 831,009 citing peaks, (meaning peaks where the out-degree is greater than 0), the citing peaks belong to 18,834 sites
- 250,558 cited peaks (entry points), meaning peaks where the in-degree is greater than 0; the cited peaks belong to 43,312 sites. The Wfr4 corpus was formed by the robot's route, which followed the links. This means that all the sites in this corpus were reached by the robot and consequently should have an entry point. The presence of 150 sites appearing without entry point is perhaps explained by the phenomenon of domain name aliases.

We notice that, of the 3.82 million netpaths in the Wfr4 collection, approximately only 1 million (26.26% of the netpaths) maintain a relationship with other sites in the corpus (receive or emit

sitations). Moreover, among these 3.82 million netpaths, only 250,558 netpaths receive external citations and are thus presented as entry points. This figure allows us to estimate 6.5% for the proportion of pages (netpaths) entry points on the Web, meaning the maximum proportion of pages use in an analysis of the sitations procedure. Furthermore, we note the almost exclusive character of the citing or cited function of the pages. In fact, only 7.7% of the peaks (77,415) of the sitation graph are citing and cited at the same time.

Figure 5 shows the distributions of in- and out-degrees for the netpaths. Although we only consider citations external to the sites (sitations), the adjustments obtained approach the results presented by Broder et al. (Broder et al., 2000) and Albert et al. (Albert et al., 1999).

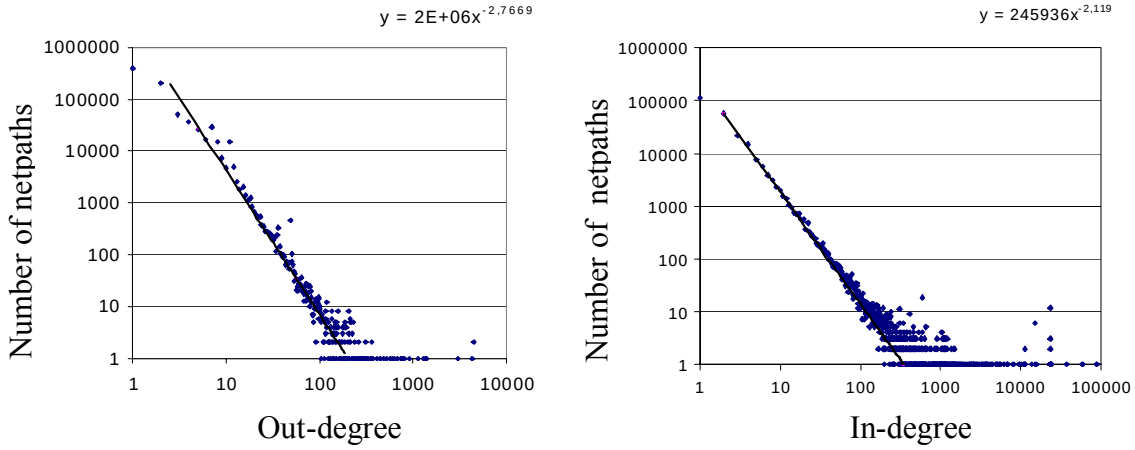

Fig. 5: Distribution of in- and out-degrees in the corpus

A necessary condition for a page to generate cositations is that it must emit at least two sitations. Thus, all netpaths in our corpus not meeting this condition are eliminated. In the Wfr4 collection, 384,655 netpaths only emit one external link. In the sitation graphs, 384,655 sitation relationships (arcs) are therefore deleted, involving the disappearance of 384,655 citing peaks (or 46% of the citing peaks) and 22,306 cited peaks, meaning 8.9% of the entry points. The new sitation graph thus obtained contains:

- 635,535 peaks (netpaths)
- 2,940,048 arcs (relationships between the peaks)
- 446,354 citing peaks, where the out-degree is equal to or greater than 2; these peaks belong to 12,229 sites
- 228,252 cited and cocited peaks belonging to 40,239 sites

These indications enable us to conclude that:

- Of the 3,823,589 netpaths in the collection, 228,252 netpaths are cocited (or 5.96% of the netpaths). In this collection, almost 6% of the pages are cocited entry points classifiable by the cocitation method.
- Of the 43,462 Web sites in the collection, 40,239 sites, or approximately 92%, have entry points that can be qualified by our method of propagating metadata values (Prime-Claverie, Beigbeder & Lafouge, 2004). We find this last figure very encouraging.

5 Conclusion

In this article we presented possible parallels between the methods of structuring the Web and the networks of scientific publications. The application limits of the cositation method on the Web are above all dependent on the theoretical and practical limits of the analysis of sitations (small proportion of pages receiving external citations, meaningless sitations, impossibility of obtaining a graph of the Web itself, etc.).

The results of the exploratory experiment on the Wfr4 collection are positive. They allow us to conclude that a majority of sites are cosited on the Web (approximately 92% of the sites) and that these can be classified by the cositation method. Thus, applied to the Web, the cositation method is

presented as a method for structuring the entry points on the sites (6% of the Web pages) but not all the pages. Note that there are advantages in structuring the entry points alone. In fact, these are generally accessible through stable URLs and are durable. Moreover, even if the information contained on the entry points is updated, their typological characteristics are seldom called into question.

6 References

Albert, R., Jeong, H. & Barabasi, A.-L. (1999). Diameter of the World Wide Web. *Nature*, 401, 130-131.

Barabasi A. (2002). Linked : The New Science of Networks. Perseus Publishing, Cambridge.

Bharat, K., Broder, A., Dean, J. & Henzinger, M. (2000). A comparison of techniques to find mirrored hosts on the www. *IEEE Data Engineering Bulletin*, 23 (4), 21-26.

Bjorneborn, L. & Ingwersen, P. (2004). Toward a Basic Framework for Webometrics. *Journal of the American Society for Information Science and Techology*, 55 (14), 1216-1227.

Bradford, S. (1934). Sources of information on specific subjects. *Engineering*, 137, 85-86.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *WWW7/Computer Networks and ISDN Systems*, 30(7), 107-117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the Web. *WWW9/Computer Networks*, 33(1-6), 309-320.

Chakrabarti, S., Dom, B. & Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *Proceedings of the ACM SIGMOD International Conference on Management of data*. New York: ACM Press.

Chu, H. (2004). Taxonomy of inlinked web entities : What does it imply for webometric research ? Library and Information Science Research : An International Journal, (In press.)

Egghe, L. (2000). New informetric aspects of the Internet : some reflections, many problems. *Journal of Information Science*, 26(5), 329-335.

Flake, G. W., Lawrence, S., Giles, C. & Coetzee, F. (2002). Self-organization of the Web and identification of communities. *IEEE Computer*, 35(3) :66-71.

Gibson, D., Kleinberg, J. & Raghavan, P. (1998). Inferring web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, pp. 225-234. New York: ACM Press.

Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2), 236-243.

Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10-25.

Kim, H. (2000). Motivations for hyperlinking in scholarly articles : a qualitative study. *Journal of the American Society for Information Science,* 51(10) :887-899.

Kumar, R., Raghavan, P., Rajagopalan, S. & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *WWW8/Computer Networks*, 31(11-16), 1481-1493.

Larson, R. (1996). Bibliometrics of the World Wide Web : An exploratory analysis of the intellectual structure of Cyberspace. *In Proceedings of the Annual Meeting of the American Society of Information Science, Baltimore.*

Pitkow, J. & Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97, pages 118-125.

Prime, C., Bassecoulard, E. & Zitt, M. (2002). Cocitations and co-sitations : a cautionary view on an analogy. *Scientometrics*, 54(2) :291-308.

Prime-Claverie C., Beigbeder M.& Lafouge T. (2004).. Transposition of the cocitation method with a view to classifying web pages. *Journal of the American Society for Information Science and Techology*, 55 (14), 1282-1289.

Rousseau, R. (1997). Sitations : an exploratory study. *In Cybermetrics*, volume 1.

Thelwall, M. (2003). What is this link doing here ? beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research*, 8(3).

Wasserman, S. & Faust, K. (1994). Social network analysis : methods and applications. Cambridge university Press.

Wilkinson, D., Harries, G, Thelwall, M. & Price L. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication . *Journal of Information Science*, 29 (1), 49-56.