Jointly published by Akadémiai Kiadó, Budapest
and Kluwer Academic Publishers, Dordrecht

Scientometrics,
Vol. 54, No. 2 (2002) 291–308

Co-citations and co-sitations: A cautionary view on an analogy

CAMILLE PRIME,[1,2,3] ELISE BASSECOULARD,[1] MICHEL ZITT[1,4]

[1]*UMR EDRA, INRA-Université de Nantes, Nantes (France)*
[2]*RECODOC, Villeurbanne (France)*
[3]*Ecole des Mines, St. Etienne (France)*
[4]*Observatoire des Sciences et des Techniques (OST), Paris (France)*

Like the citation network of scientific publications, the Web is also a graph where pages are connected together by hypertext links or "sitations". In the new research field Webometrics, scholars have investigated equivalencies between citationist concepts established in bibliometrics and hyperlinks networks. This paper focuses on the possible analogy between co-citation and co-sitation to structure Web universes. It reports an experiment in the field of bibliometrics and scientific indicators. Several technical aspects that must be dealt with are reviewed. Co-sitation seems a promising way to delineate topics on the Web. However, the analogy with traditional co-citation is deeply misleading: many precautions must be taken in the interpretation of the results.

Introduction

Two types of informetric/bibliometric methods are typically used in the structural analysis and mapping of scientific networks: the linguistic way and the citation way. Among others in the linguistic family, lexical methods have proven very effective both on computational and interpretative aspects. They have both been dominant in information retrieval on the web. Their counterparts in the citation world, co-citation (*Small,* 1973; *Marshakova,* 1973) and bibliographic coupling (*Kessler,* 1963), however, are not that pervasive for a variety of reasons. Some of us have established the efficiency of improved co-citation protocols in the mapping of science, in particular in relation to "recall rates" criteria (*Zitt* and *Bassecoulard,* 1996). Co-item mapping using words and citations show several formal analogies, but also profound differences, the most important of which being the diachronic nature of citation with related phenomena (aging/immediacy), a very substantial advantage for dynamic studies of scientific fields.

0138–9130/2002/US $ 15.00

The formal kinship between the traditional "document", for example a scientific article, and a web page as identified by the specific uniform locator (URL), has encouraged a 'technological transfer' and application of information retrieval and bibliometric techniques to web analysis. Both Internet developers, and also bibliometricians have contributed to this movement. The transfers started in applications of lexicology, but have more recently extended to "citations". The natural equivalent to referencing / citation is the hyperlink. Several authors, among them *Rousseau* (1997), *Ingwersen* (1998), *Aguillo* (1999), *Boubourides* (1999), *Egghe* (2000) and *Björneborn* (2001) have investigated equivalencies between citationist concepts established in bibliometrics and those of hyperlink networks (notion of "sitation", of web impact factors, etc.). Technical problems in calculation, due to the search engines, have also been raised by *Bar-Ilan* (2001).

For their part, specialists of the web and of search engines are becoming increasingly interested in the structure of hyperlink networks in order to improve information retrieval. Such techniques range from "web structure mining" to: web content mining, relying on documents text or metadata; web usage mining based on data collected from users (log files, cookies…) etc. (*Kosala*, 2000).

"Google" (*Brin* and *Page*, 1998) is the first search engine to use links between pages in order to improve the algorithm for retrieved-web-page ranking ("PageRank"). Intuitively, PageRank is the probability that a user visits a given page by navigating at random along hyperlinks without back movements. From a bibliometric point of view, PageRank is close to both the impact factor of scientific journals (*Garfield*, 1972), (with a weighting by the number of citations emitted by citing pages – to control for differences in citing behavior between web pages), and to the influence factor (*Pinski* and *Narin*, 1976), that reckons that a citation from an influential journal may be judged more valuable than a citation from a non-influential one. A given page PageRank is all the higher if it is pointed to by high PageRank pages, using a propagation algorithm.

The other well-known example of an engine based on web-structure is the "Clever" prototype (*Kleinberg*, 1999). This engine first uses a set of documents, for example about a company X, retrieved by a lexical query on a classic engine, such as AltaVista. This set usually contains two types of documents: "authority" pages that describe the object (e.g. the company's homepage), and pages that mention the object without actually describing it, for example catalogues or portals, called "hubs". Clever proposes this distinction as a formal hypothesis on the web structure, and tries to detect the right hubs as those pointing to the right references, and vice-versa. An iterative program calculates the hub and reference functions of each page. Possible bibliometric analogies of hubs are "surveys" or "review articles".

Another method to improve ranking was proposed by *Savoy* (1996) in a study conducted on the test file CACM (3024 documents, 50 questions). The author used a propagation algorithm of scores according to hyperlinks. Three relations are accounted for: citing-cited; bibliographic coupling; and co-citation. Co-citation linkage appeared to be the most efficient on this collection. Recent experiments (*Savoy* and *Picard*, 2000) using a snapshot of around 2.3 gigabytes extracted from the Web, have tried to evaluate the usefulness of taking hyperlinks into account to improve web searching.

For structuring purposes, specialists of the web try to cluster web pages using the web graph structure (*Kumar* et al., 1999). Other experiments of transposition of bibliometric citation methods to the web have been tried. *Larson* (1996) used the author co-citation (*White* and *Griffith*, 1981; *White* and *McCain*, 1989) to uncover the intellectual structure of the web, by automatically detecting domains and sub-domains with no need of Yahoo-type indexes. The results were positive, with some limitations however, i.e., small sample and limits to automatization due to heterogeneity of pages. Another experiment based on articles' co-citation (*Pitkow* and *Pirolli*, 1997) aimed at representing the structure of a large web site (Georgia Institute of Technology's Graphic Visualization and Usability Center). Documents were classified according to their type: research projects, people pages, documents/contents…

In this study we focus on the power and limits of the analogy "co-citation"-"co-sitation" for the mapping of knowledge networks. Are co-citation techniques applicable to web sitations in order to delineate topics on the web? What technical problems are met? Are the results interpretable in the same way as traditional co-citations? In the particular experiment, devoted to the bibliometric field reported in the next section, these questions, their partial answers and the caveats suggested, matter more than the particular cluster structure found in the area under consideration. The discussion section tries to list a few important issues, but much remains to be done in a subject area where the literature is still relatively limited.

An experiment on a particular topic (bibliometrics)

Building the dataset

We chose a familiar field "bibliometrics and scientific indicators", as the experimental subject matter. We combined 15 queries (in both French and English) using the commercial software "Copernic" (client meta-engine) and recovered a total dataset of 7002 pages. The first difficulty lay in the identification and elimination of all invalid pages and duplicates. The elimination of duplicates is an essential stage in

this process, their presence in the dataset causing the artificial multiplication of (co-)citations. We detected two types of duplicates: “true duplicates” sharing the same URL, created by an imperfect unification process between queries used by Copernic; and “hidden duplicates” with different URLs but sharing the same content, often generated by mirror sites, whether official or not. A first cleaning operation yielded 3538 pages that we have called “citing pages”, a more elaborate cleaning proved necessary at the co-citation stage.

All hypertext links found within the pages were extracted. We made the classic distinction between intra-server links which allow movement within a website (navigation links); and inter-server links which signal interesting outside resources or materialize a partnership or a commercial advertisement. An extensive classification of links was proposed by *Ingwersen* (1988).

Characterization of hyperlinks on the citing side

The study of the ratio of external relations and of the distribution of links suggested a number of questions related to citation behavior.

We first examined the number of inter-server links as a function of total links. As shown in Figures 1 and 2, there is a striking bi-modality for highly citing pages: some pages link solely to external pages (close to the diagonal D1: $y(x) = x$), others show a profusion of internal links (close to the abscissa axis). The bi-modality fades for pages emitting less citations i.e. below ca. 60 links.

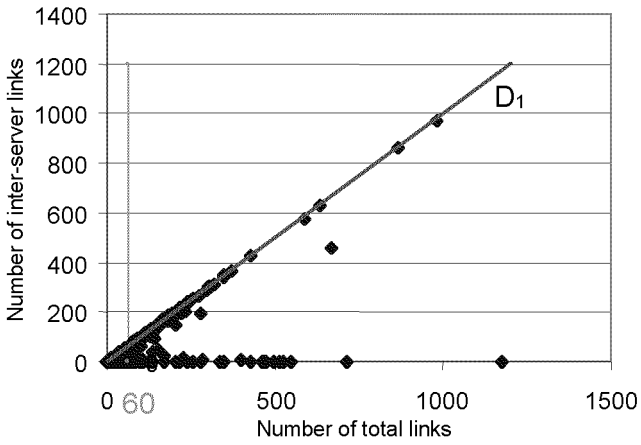

Figure 1. Number of distinct inter-server links by page

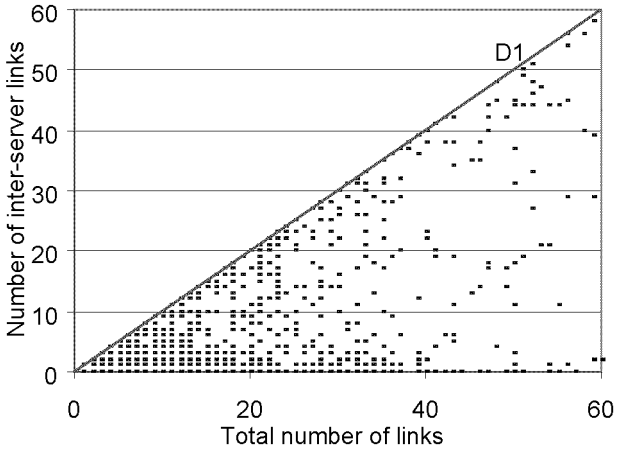

Figure 2. Zoom of Figure 1.: Number of distinct inter-server links by page

Pages emitting a large number of links can be considered as directories: they aim at organizing information and making it easily accessible. In the above dichotomy, portals point outwards, whilst indexes or summaries point at the inner resources hosted on the same server. Generally speaking, directories don't mix these two functions.

The distribution of the number of external links for each citing page is concentrated and conforms, as expected, to a hyperbolic law (Figure3). In the double-log plot of the distribution (Figure 4), the graph may be broken down into three quasi-linear parts amenable to piecewise fitting, with thresholds S1 and S2 as separators. This suggests an heterogeneous population of three groups with different functions. The ratio of the source code size to the number of external links confirms this indication:

– general or scientific portals are found in the first section (>S1); these pages with a high number of external links exhibit a correlation between the number of external links and the size of the source page;
– more specific portals are grouped in the intermediate section; these are mostly specialized in information science and/or bibliometrics. Very few articles or reference documents are found in this section;
– two types of documents are found in the lower right section (<S2): a) summaries or indexes of significant size with many internal links and few external ones; b) documents of various sizes, with a targeted content and few external and internal links. They are most probably "reference documents".

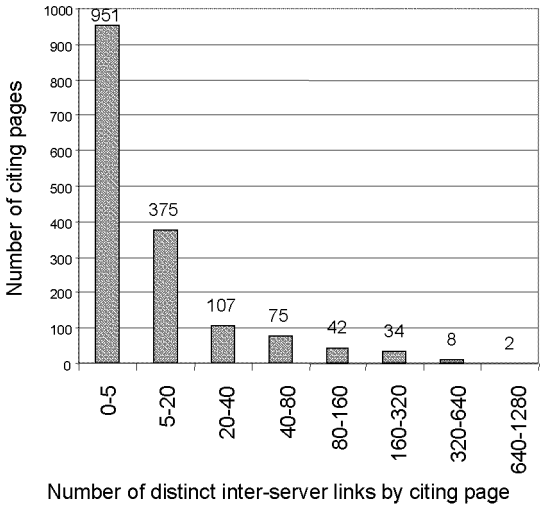

Figure 3. Distribution of citing pages according to the number of distinct inter-servers links

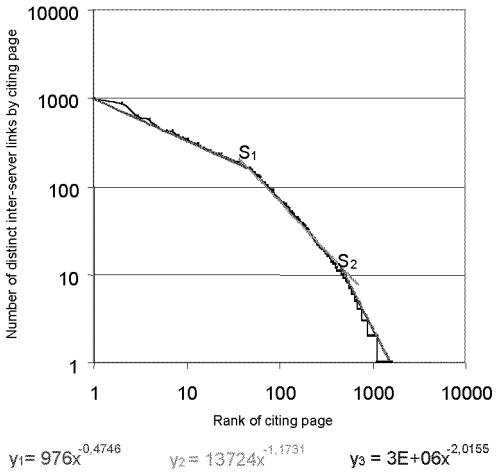

Figure 4. Number of distinct inter-server links by citing page sorted in descending order

This characterization can be used to refine the selection of citing documents generating relevant linkages for the co-citation study. A large number of often irrelevant

links are generated by the upper section documents. We can discard them to reduce noise, without much risk of losing relevant associations. This removal is consistent with the weighting of citing documents we apply in the co-citation process, lowering the role of citing documents with a large number of references.

The intermediate section gathers together information science portals that contain numerous linkages, among them some highly relevant ones. Within these portals, however, different orientations of the domain are touched upon on the same page, thereby risking the generation of loose co-citation connections. The risk of excessive noise is again reduced by weighting the co-citation linkages according to the number of emitted links.

The lower section pages have a very targeted citation behavior and generate few but expectedly relevant external links.

Characterization of hyperlinks on the cited side

In our experiment, the distribution of cited pages can be approximated by a classic power-law, with as usual for citations a smaller slope in double-log than for Zipf's law (Figure 5).

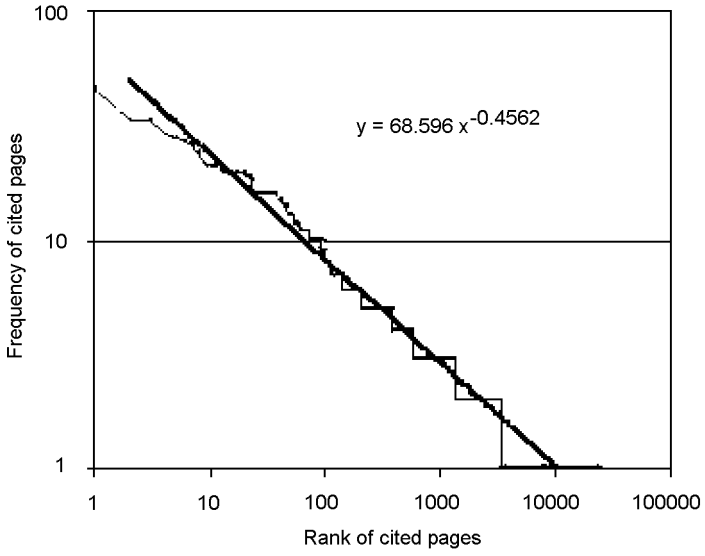

Figure 5. Distribution of cited pages ordered by descending rank

Multiple levels of "co-sitation"

Like the bibliographical reference, the hypertext link can be examined at various levels. Taken as a whole, it points towards a specific page and touches upon a particular subject. The hosting institution is revealed at the server level. Two extreme types of co-citation can be directly derived: page co-citation and server co-citation. Page co-citation evokes document co-citation rather than other forms of co-citation (authors, journals…), even if web pages are not considered homogeneous documentary units. Server co-citation can be related to journal co-citation. However these formal analogies may be partly misleading (see discussion section).

In this experiment we focus on page co-citation, which is expected to provide a fine-grain structure of a topic. We have concentrated on inter-server links only. Only 1594 out of the 3538 pages of our dataset possessed at least one external link. As discussed above, we had to eliminate 27 "general" portals (determined empirically according to the distribution).

Final cleaning of citing pages: detection of duplicates through bibliographic coupling

A second cleaning operation involved a bibliographic coupling calculation (*Kessler*, 1963). The probability that two different pages exhibit exactly the same list of references is very low, especially for long lists. We made the assumption that an identity of references was a strong indication of duplication. We calculated a bibliographic coupling index on all pairs of citing documents and 525 pages were found to exhibit an Ochiai index of 1, indicating that they shared all their references with another page, either on the same server or on another one. Bibliographic coupling enables users to identify perfect mirrors or less official "clones". The algorithm is fairly heavy (cartesian product, however on sparse co-occurrences matrices) and generates large files, but can be accelerated *ex ante* by selecting pairs of pages with the same number of references.

This duplication of references at the page level may concern both large lists and several duplicates: for example, two pages were found to be identical with 90 references, and five pages with 87 references. Other duplicates shared less than 20 references. The duplicates belong to several types. First, there are indexes that copy all hypertext links from other pages. In our dataset, most of these mirror indexes are created by an engine at CRRM-Marseille (*Mannina*, 1997), aimed at monitoring the web content on a subject and extracting hyperlinks. We also found indexes from Yahoo, with several versions for different markets/languages. From the contents/presentation point of view, these pages may not be *verbatim* duplicates, but their hyperlinks list is identical,

and, in a co-citation context, they can be considered as practical duplicates that must be unified. Secondly, many pages on the same host share the same external linkages. These hyperlinks generally express a commercial (publicity) link or an industrial partnership. This is the case for the www server ainet.com (American infometrics) which offers on each page to download Netscape and Internet Explorer products.

The presence of clones is likely to jeopardize both citation and co-citation studies (see discussion section). Without cleaning, some clusters can even appear as pure mirror illusions. The issue is still more serious, because more difficult to detect, for quasi-clones that do not share all references, but whole sections of references obtained by partial duplication. These pages escape straightforward detection by bibliographic coupling. Of course other classic informetric approaches, for example comparison of terminology, may be used to detect these duplicates.

Clustering cited pages

After removal of most duplicates, clustering of cited pages was carried out using the average group linkage procedure "proc cluster" of the SAS software package . The similarity index used is weighted by the number of citations emitted by the citing pages, which can be seen as reflecting a fractional count of the citations (*Zitt* and *Bassecoulard*, 1996). Finally, to allow a manual checking of clusters' contents, we selected pairs of URLs (external links) with at least 5 co-occurences. The remaining 230 URLs were grouped into 27 clusters, and 4 "singletons".

An initial examination uncovered eight suspect clusters:

– four were artefacts generated by large sites divided into several hosts for technical reasons (for instance alerting.isinet.com or isinet.com). Such links, internal to a site but pointing towards another machine, were incorrectly identified as external by our algorithm;

– the other four clusters, created by publicity links between commercial sites, were outside the field of interest. In this particular study, these irrelevant clusters were isolated and easily detected and did not jeopardize further analyses. However these kinds of spurious link, due in fact to the multiple functions of a web site, are a major issue for co-citation applications.

Relevant clusters (131 pages) are shown on Table 1. Seven of them (46 pages) deal with central topics in the field bibliometrics and scientific indicators. The others cover closely related domains such as information science, competitive intelligence, scientific edition and electronic publishing, libraries and data-sources etc.

Table 1. List of relevant clusters

Cluster	Number of elements	Subject
CL43	15	Bibliometric Research Groups
CL46	10	Bibliometric and infometric Societies
CL60	8	Infometrics
CL115	5	Bibliometric Conferences
CL54	4	Bibliometric resources
CL100	2	Science evaluation
CL103	2	Bibliometric publications
CL114	29	Journals of information science, libraries and publishing
CL42	14	Search Engines
CL61	8	Competitive intelligence
CL41	7	Databases and information services
CL53	7	E-journals and E-resources
CL106	6	Scientific journals
CL40	4	Library associations and booksellers online
CL66	4	Internet and information science societies
CL120	2	National Research Council Canada
CL124	2	Competitive intelligence (French societies)
CL65	2	Resource centers
CL96	2	University of New Jersey

Table 2. Contents of Cluster 43: research centers

URL	Research Center
http://crrm.univ-mrs.fr/sfba/home.html	Société Francaise de Bibliométrie Appliquée, SFBA
http://crrm.univ-mrs.fr/sfba/sfba.html	Société Française de Bibliométrie Appliquée, SFBA – Presentation
http://coombs.anu.edu.au/Depts/RSSS/REPP/repp.htm	Research Evaluation and Policy Project
http://www.sri.com/policy/econpract/stpp.html	Science and Technology Policy Program at SRI International
http://www.elsevier.nl	Elsevier Science Publisher
http://www.csic.es	Consejo Superior de Investigaciones Científicas, CSIC
http://ai.iit.nrc.ca/II_public/WebBird/	BIRD: Bibliometric Retrieval of Documents – BIRD is a bibliometric "query by example" search engine.
http://crrm.univ-mrs.fr	CRRM, Univerity of Marseille
http://crrm.univ-mrs.fr/commercial/software/software.html	Centre de Recherche Rétrospective de Marseille, CRRM – software
http://meritbbs.rulimburg.nl	Maastricht Economic Research Institute on Innovation and Technology, MERIT
http://sahara.fsw.LeidenUniv.nl/cwts/cwtshome.html	Centre for Science and Technology Studies (CWTS) Leiden University – The Netherlands
http://crrm.univ-mrs.fr	CRRM, Univerity of Marseille
http://www.isi.fhg.de	Fraunhofer Institute for Systems and Innovation Research (ISI)
http://www.umu.se/soc/inforsk/inforsk2.htm	Inforsk (The Information Research Group)
http://www.unites.uqam.ca/cirst/	Centre Interuniversitaire de Recherche sur la Science et la Technologie, CIRST
http://www.isinet.com/prodserv/rsg/rsghp.html	ISI – Products and Services

The clustering allows one to identify most of the players in the field and tends to group them by "type". For example, cluster CL43, detailed in Table 2, gathers bibliometric research centers, cluster CL115 specialized conferences etc.

We have found only one cluster with a true document content (CL103, bibliometric publications). In all the other cases, the clustered web pages are institutional in nature. Whether co-sitation tends to group homogeneous nodes of the web network remains to be tested. In our experiment, more highly cited, institutional pages were more likely to be kept for the clustering stage. This may be due to the threshold setting (of 5 co-occurences) being too high to retrieve document nodes.

Discussion

The analogy of classic co-citation to web co-citation should be considered with care, because of a number of fundamental differences. We shall not expand on general problems of web structure and access (volatility, invisibility of parts of web, engine's low recall rates and linkage coding in some site-building softwares etc.), already covered extensively in the literature. We shall, rather, focus on the technical question of duplication, and on fundamental limits associated with the status of objects and linkages.

Mirrors, clones and illusions: a major hindrance for web citation and co-citation studies

The noise due to large portals can be reduced without much difficulty by appropriate selection and weighting. Portals likely to generate noisy associations may then be discarded. The issue of topical portals should be discussed for each case study – depending on their degree of specialization. Clear conclusions of this study on these points are that there are (a) constraints of selection on citing sources, based on bibliometric distributions or other means (b) constraints of weighted measures for co-citation indexes, consistent with the elimination of massive citing sources.

Greater trouble comes from the multiplication of mirror sites/pages, that generate a replication of references and introduce an abnormal level of redundancy in the web citation processes. Unification issues are central in many applications of bibliometrics, but the problem is particularly intricate for web sources.

The "official" mirror sites can be easily detected and unified. Among other methods, bibliographic coupling is a means to detect hyperlink list similarity between pages. Another difficulty comes from the general mimetic behavior in referencing. In the scientific community, the tendency to reproduce the referencing practices of colleagues may be seen as one of the sources of the Matthew effect and related accumulated advantages studied by founders of bibliometrics (*Price,* 1976). The duplication of lists of references through search engines, the internal structures of sites, or the copying of hyperlinks, is a more trivial and mechanistic version of a citation amplificator but it creates a key issue for interpretation of web "sitations". Hindrance is perhaps less critical for "co-sitations" if "web illusion" clusters that emerge remain isolated.

Analysis at the page level is appealing for a relatively fine-grain approach. However, this level amplifies the linkages associated with publicity and partnership that would certainly be considered as irrelevant in a bibliometric approach.

The status of objects: qualification of documents

In the traditional context of co-citations the status of citing and cited objects is clear. The major problems encountered in traditional co-citation are the recurrent issue of citation interpretation(that has given birth to a huge literature) and the specific technical problems of clustering optimization or structuring techniques associated with the bibliometric properties of the field (for example the recall rate problem).

The situation is strikingly different for web objects. Let us focus on the status of documents. The qualification of "web documents" and sources is far from being achieved, despite current attempts at normalization (*Dublin Core Project*, 1999). The web offers a variety of information of all kinds at any level of generality. Moreover, the formalism of a URL (path server/…/pages) does not necessarily reflect a hierarchical structure of information. This creates a much more intricate situation than in classic document retrieval, where the types of documents are known and standardized. A few technical problems in this respect are: the lack of correspondence between logical and technical pages; the structure of sites, not always hierarchical; the frequency of catalogues; and the above-mentioned replication or quasi-replication of sites/pages.

Given the variety of types and levels of URL, there is not one but many types of web co-citation study: web co-citation of servers, of institutional pages (such as labs), of subject-defined pages, of classic document pages, etc. The relation to the formal object (right truncation assumed to identify the server; intermediate truncations; whole URL

for the final pages) may not be as straightforward as in traditional co-citation (document, author, journal). In this exploratory study we limited ourselves to a formal definition (final pages, i.e., whole URLs).

Interpretation of hyperlinks is naturally heavily dependent on the type and level (site, pages…) of objects. The way hyperlinks are generated is also fundamental. Even for purely scientific items posted on the web, *Wouters* (2001) stresses the role of editors (webmasters), besides the authors, in the management of hyperlinks. Though the formal analogy of web page co-citation is rather document co-citation, the author/institutional dimension is very present at the URL level, and as a result in the clusters obtained.

Acknowledging the heterogeneity of sources and targets is of course fundamental for a sound adaptation of co-citation for structuring web networks. Some difficulties, due to poor qualification, are alleviated in targeted studies based on relatively short *ex ante* list of players (for instance institutions), or when ad hoc identification and unification of corresponding URLs can be afforded. But when the purpose is the mapping of information sources in a large field, with little prior knowledge, the technical obstacles can be serious. The page level, as studied here, seems the most promising for fine-grain structuring, but other levels may be interesting in some cases, provided a sufficient homogeneity is obtained. For instance, if an institutional view is sought, relevant pages can be selected either by threshold settings (institutional pages are more cited than document pages) or by appropriate truncation (when institutional web sites have a hierarchical structure, institutional descriptions are more likely to be found near the root).

Last but not least, many URL pages are often a-chronic. Whether they do not matter for some type of pages, or because of the current lack of normalization, time references are often missing or available dates can refer to site management aspects rather than the creation/update of the underlying documents. This a-chronicity may be related to the absence of archival functions and/or to the volatility of the web. The loss of the temporal dimension, that will perhaps be overcome with the expected norms on meta-data, is of major importance for the interpretation of citationist analogies.

The nature of hyperlinks and the possible loss of diachrony: two major differences

The founders of the co-citationist approach at ISI and Philadelphia University developed an interpretative framework where co-cited cores are viewed as “intellectual bases” dynamically designed by the current citing literature, and the corresponding citing sets as “research fronts” (*Small*, *Griffith*). Two aspects of co-citation are undoubtedly appealing in the dynamic description of scientific advances, first the ability

to disclose the combinatory nature of the advancement of research, and secondly the temporal dimension common to citation techniques. This gives to co-citation strong capabilities in the historic sketching of intellectual sequences and conceptual associations, many examples can be found in the former ISI's Atlas of Science. We have found many examples in our own studies (e.g., the "evapo-transpiration" case in *Zitt* and *Bassecoulard*, 1998). When applying co-citation to the web objects, the two pillars, the combinatory vision and diachrony, are threatened to different degrees.

Firstly, the actual processes of combination in science are depicted by co-citation, and in a different manner by "a-chronic" co-word (or co-classification) techniques. Each technique can be applied to map the web structure. The transposition is more direct for co-word, since co-citation, ideally adapted to a quasi-normalized way of communication (the scientific article), must face a much more difficult situation when this quasi-norm for documents is missing.

Secondly, with web a-chronic documents, the time dimension is lost. The diachronic and asymmetric aspect of citation is very powerful and has given rise to a tremendous amount of works on the dynamic aspects of citation and aging (e.g., *Glänzel*, 1994). *Egghe* (2000) argued that the absence of diachrony in hyperlinks condemned the citationist analogy. The citing/cited asymmetry itself may also be challenged by the frequent practice of reciprocal linkages.

In the traditional applications, the diachronic structure of citation extends to co-citation. For example we combined dynamic characterization both on the cited and citing side to qualify clusters (*Zitt* and *Bassecoulard*, 1994). The diachronic capabilities of (co-)citation feed arguments in the co-words vs. co-citation dispute (*Leydesdorff*, 1997; *Braam* et al., 1991). If co-citation were just a particular instance of a structuring/mapping technique, using the token "reference" instead of "words" or "classification codes", the loss of diachrony would not be so damageable. In fact, the very originality of co-citation, in the wide range of "co-item" techniques, fades with the loss of the temporal depth. The loss of diachrony also deprives the comparison co-citation/ bibliographic coupling from an important dimension. In traditional application of citation analysis for structuring/mapping of science, the criteria for choosing co-citation or bibliographic coupling are largely linked to the time dimension: capability of disclosing intellectual structure and chronology for co-citation, better immediacy for coupling. These criteria largely vanish in web structuring applications. It remains a fact that co-citation is typically used with a prior bradfordian selection of cited items (through citation scores or co-occurrence levels), while coupling operates on proximity between citing pages, calculated on all references. Coupling tends to generate less "silence", at the expense of noise level, less robustness and usually larger

storage/computational requirements. This difference persists in web oriented applications. Co-citation remains an efficient method to disclose connections between the "brightest" objects.

Rip (1988) argued that co-citation was able to depict the "legitimatory repertoire" more than to disclose the conceptual structure of a field. This criticism, too severe for document co-citation studies conducted in good conditions, may be rephrased for co-cited networks of web-objects. With a bradfordian selection, what may emerge through co-citation when applied to a scientific area – as in our study – is a kind of "laboratory life", a display of bright objects whatever they may be: concepts, advertisements, authors/institutions, products…

Conclusion

The type of semantic interpretation from web co-citation clusters depends primarily on the status of classified objects. Sophisticated analyses, validated in classic co-citation, can only survive to a minor degree in the web context. In the experiment we have reported, a cautionary interpretation, in terms of topical/institutional proximity, seems legitimate, both on the cited side (co-citation proximity) and the citing side (assignment to co-citation clusters; or direct bibliographic coupling). Institutional aspects are seen to play a prominent role.

Clustering techniques, among other methods, are meant to reflect macro-structures, which are in principle more robust than individual features. The existence and identity of a cluster hardly depends on the presence or absence of a particular item in the cluster, whether on the citing or the cited side. This fact is especially valuable when other contingent factors deeply affect the impact, visibility or persistence of individual items – this is the case for volatile web objects.

This does not prevent web co-citation or coupling keeping track of temporal change of topics, provided appropriate methods of data analysis are used. However, they lose diachrony, one of their distinct advantages over a-chronic methods. But keeping the cited/citing asymmetry, in spite of frequent reciprocal postings, and remaining anchored in a logic of source rather than of contents, web co-citation offers an efficient alternative to describe web structures, with informetric properties different to those offered by lexical techniques. We can also extrapolate from co-citation studies the hypothesis that "bradfordian" co-citation clusters, with a prior citation score or co-occurrence thresholding, will be more appropriate than bibliographic coupling to make major structures apparent, at the expense of weak signals.

To summarize, the transposition of co-citation or related techniques (coupling) to the mapping of web topics seems promising but only with "down-sized" ambitions. It encompasses a wide range of possible applications, depending on the type of citing source, the type of cited item and the homogeneity of sources/items. The qualification of items and linkages, for a given objective, is central. The new standards on meta-data should be helpful in the future. A serious issue, however, is the prevalence of duplications or quasi-duplications in citing lists. Among other citation-based methodologies that have met with success in new search engines, web co-citation or coupling can yield helpful auxiliary tools for information retrieval and mapping. The choice between the two techniques is largely a matter of a signal/noise trade-off.

Finally, it should be remembered that "sitation" and citation are different matters entirely. Beyond formal analogies, structuring a landscape of communication through web page networks in science, and structuring a landscape of scientific outputs on calibrated databases, are very distinct in nature. Taking advantage of the partial analogy sitation-citation to assess an institutions' performance in the traditional meaning would be deeply misleading. The picture can be as different as a city's nightscape of neon signs and its roadmap. The related indicators, Web visibility and academic performance, are completely different species. Progress in the qualification of pages and generalization of electronic publication are likely to bring the two worlds closer and closer, however a high degree of caution is necessary as long as the status and coverage of web sources and pages remains unregulated.

*

The authors wish to thank Thierry *Lafouge*, at RECODOC, for helpful discussions and Loïc *Vinet*, at UMR EDRA, for his collaboration.

References

AGUILLO, I. (1999), Statistical Indicators on the Internet: The European Science-Technology-Industry System in the World-Wide Web, at: http://diotima.math.upatras.gr/weborg/aguillo2

BOUDOURIDES, M., B. SIGRIST, P. ALEVIZOS (1999), *Webometrics and Self-Organization of the European Information Society*, at: http://hyperion.math.upatras.gr/webometrics

BRAAM, R. R., H. F. MOED, A. F. J. VAN RAAN (1991), Mapping of science by combined co-citation and co-word analysis, II. Dynamical aspects, *Journal of the American Society of Information Science*, 42:252–266.

BRIN, S., L. PAGE (1998), The anatomy of a large-scale hypertextual Web search engine, *Proceedings of the 7th International World Wide Web Conference, 1998*.

BAR-ILAN, J. (2001), How much information the search engines disclose on the links to a Web Page ? A case study of the "Cybermetrics" home page, *Proceedings of the 8th International Conference on Scientometrics and Infometrics , ISSI 2001, Sydney, Australia, July 16-20, 2001,* pp. 63–73.

BJÖRNEBORN. L., P. INGWERSEN (2001), Perspectives of Webometrics, *Scientometrics*, 50 (1) : 65–82.

DUBLIN CORE PROJECT (1999), *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, at: http://dublincore.org/documents/dces/

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

GARFIELD, E. (1972), Citation analysis as a tool in journal evaluation, *Science*, 178 : 471–479.

GLÄNZEL, W., U. SCHOEPFLIN (1994), A stochastic model for the aging of scientific literature, *Scientometrics*, 30 (1) : 49–64.

INGWERSEN, P. (1998), The calculation of web impact factors, *Journal of Documentation*, 54 : 236–243.

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

KLEINBERG, J. (1999), Authoritative sources in a hyperlinked environment, *Journal of the ACM*, 46 : 604–632.

KOSALA, R., H. BLOCKEEL (2000), Web mining research: A survey, *SIGKDD Explorations*, 2 : 1–15.

KUMAR, R., P. RAGHAVAN, S. RAJAGOPALAN, A. TOMKINS (1999), Trawling the Web for emerging cyber-communities, *In the Proceedings of the Eighth World Wide Web Conference*.

LARSON, R. (1996), Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of the cyberspace, In: *Proceedings of the Annual Meeting of the American Society of Information Science*, (Baltimore, Md., Oct. 19-24, 1996).

LEYDESDORFF, L. (1997), Why words and co-words cannot map the development of the sciences, *Journal of the American Society for Information Science*, 48 (5) : 418–427.

MANNINA, B., L. QUONIAM, *Cybermetrics*, 4(1) paper 1. [OnLine] at: http://www.cindoc.csic.es/cybermetrics/articles/v4i1p1.html

MARSHAKOVA, I. V. (1973), Document coupling system based on references taken from Science Citation Index (in Russian), *Nauchno – Tekhnicheskaya Informatsiya*, Ser. 2 (6) : 3.

PINSKI, G., F. NARIN (1976), Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, *Information Processing and Management*, 12 : 297–312.

PITKOW, J., P. PIROLLI (1997), Life, death and lawfulness on the electronic frontier, In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System (CHI'97),* ACM New York, pp. 118–125.

PRICE, D. J. DE S. (1976), A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science*, 27 : 292-306.

RIP, A. (1988), Mapping of science: Possibilities and limitations, In: *Handbook of Quantitative Studies of Science and Technology*, A. F. J. VAN RAAN (Ed.), Elsevier Science Publishers, Amsterdam, 253–273.

ROUSSEAU, R. (1997), Sitations: an exploratory study, *Cybermetrics*, 1 : 1. at: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

SAVOY, J. (1996), Citation schemes in Hypertext information retrieval, In: AGOSTI, M., SMEATON, A. (Eds), *Information Retrieval and Hypertext*, Kluwer, pp. 99–120.

SAVOY, J., J. PICARD (2000), Recherche documentaire sur le web: Les hyperliens sont-ils vraiment utiles ? *Actes JADT2000*, 27–34.

SMALL, H. G. (1973), Co-citation in the scientific literature, *Journal of the American Society for Information Science*, 24 : 265–269.

SMALL, H. G., GRIFFITH, B. C. (1974), The structure of scientific literature 1&2, *Science Studies*, 17–40 and 265–269.

SMALL, H. G., E. SWEENEY (1985), Clustering the Science Citation Index using co-citation. I. A comparison of methods, *Scientometrics*, 7 : 391–409.

SMALL, H. G., E. SWEENEY, E. GREENLEE (1985), Clustering the Science Citation Index using co-citation. II. Mapping science, *Scientometrics*, 8:321–340.

WHITE, H. D., B. C. GRIFFITH (1981), Author co-citation: A literature measure of intellectual structure, *Journal of the American Society for Information Science*, 32:163–172.

WHITE, H. D., K. W. MCCAIN (1989), Bibliometrics. In: *Annual Review of Information Science and Technology*, 24:119–186. Elsevier, Amsterdam.

WOUTERS, P., R. DE VRIES (2001), *Formally Citing the Web*, at: http://home.pscw.uva.nl/lleydesdorff/avril/April2001/wouters.htm

ZITT, M., E. BASSECOULARD (1994), Development of a method for detection and trend analysis of research fronts built by lexical of cocitation analysis, *Scientometrics*, 30:333–351.

ZITT, M., E. BASSECOULARD (1996), Reassessment of co-citation methods for science indicators : Effect of methods improving recall rates, *Scientometrics*, 37:223–244.

ZITT, M., E. BASSECOULARD (1998), Méthodes de structuration pour l'analyse stratégique des univers scientifiques: les techniques de citation, *Veille Stratégique, Scientifique et Technologique, Actes VSST'98*, Toulouse, France, 19-23 oct.1998, pp. 31–41.

Received October 12, 2001.

Address for correspondence:
CAMILLE PRIME
Ecole des Mines, 158 Cours Fauriel
F–42023 St. Etienne, Cedex 2, France
E-mail: prime@emse.fr