

Metadata Propagation in the Web Using Co-citations

Camille Prime-Claverie and Michel Beigbeder
Laboratoire RIM/G2I
École Nationale Supérieure des Mines
158, cours Fauriel
F 42023 SAINT-ETIENNE CEDEX 2

Thierry Lafouge
Laboratoire URSIDOC
Université Claude Bernard Lyon 1
43, boulevard du 11 novembre 1918
F 69622 VILLEURBANNE CEDEX

Abstract

Given the large heterogeneity of the World Wide Web, using metadata on the search engines side seems to be a useful track for information retrieval. Though, because a manual qualification at the Web scale is not accessible, this track is little followed. We propose a semi-automatic method for propagating metadata. In a first step, homogeneous corpus are extracted. We used in our study the following properties: the authority type, the site type, the information type, and the page type. This first step is realized by a clusterization which uses a similarity measure based on the co-citation frequency between pages. Given the cluster hierarchy, the second step selects a reduced number of documents to be manually qualified and propagates the given metadata values to the other documents belonging to the same cluster. A qualitative evaluation and a preliminary study about the scalability of this method are presented.

1 Context

None of the available search engines seems to take into account the heterogeneity of the Web resources. All of them are based on a “semantic” representation of the documents, just like the traditional information retrieval systems: They stick to solely represent the *subject* and no other aspects. Though, contrary to the traditional document databases, the Web is a non controlled information repository. So the retrieved resources are heterogeneous in many points of view: their subject of course, but their type, their language, their level, their aimed audience, etc. So the users who have many needs and many expectations are not always satisfied by the results sets returned by the search engines.

We think that metadata are in the Web as in the library world a mean to address the heterogeneity problem. The HTML standard allows to embed internal metadata in the Web pages with the <META> tag. Though the use of this tag is not much spread because not well known by the au-

thors. On another hand, these metadata are often misused either by a lack of practice or objectivity by honest authors, or diverted from their intended use to get a better visibility on the Web by those who master them. That’s why most of the search engines do not take into account their content in their algorithms. In order to obtain a systematic and uniform qualification of the documents, we think that the metadata should be valued on the search engine side (external metadata). Though a manual qualification of the whole Web seems impossible because the number of documents to qualify would imply the cost to be far too large. So, only automatic or semi-automatic methods are conceivable.

2 Our qualification method

Our semi-automatic method to characterize the Web pages is composed of two steps. In the first step, homogeneous corpus are extracted. This step is fully automatic as it consists in a clusterization method which uses a similarity measure based on the co-citation frequency between pages. Given the cluster hierarchy, the second step selects a reduced number of documents to be manually qualified and propagates the given metadata values to the other documents belonging to the same cluster.

To build the co-citation graph, we only consider the citations that cross site boundaries (*external* citations), because these *inter-site* links prompt the users to leave the site they are visiting to access another one. This is a clue that the authors of the first site feel interested in the second one.

Two pages P_i and P_j are co-cited if some page P cites both of them. Let $C_{i,j}$ be the number of such pages. The greater this number, the stronger the co-citation relationship. The equivalence index, defined with $E_{i,j} = \frac{C_{i,j} \cdot C_{i,j}}{C_i \cdot C_j}$, where C_i is the number of citations received by the page P_i is usually used in scientometry [2] to measure this strength. It takes its value in the interval $[0, 1]$ and it catches the idea that two pages are close when their co-citation frequency is high compared to their respective citation frequency. As

an extreme case, when two pages are always cited together, their equivalence index is equal to 1. As we want an index that increases like a distance, let have $D_{i,j} = 1 - E_{i,j}$.

The co-citation relationship has been used with success to structure the scientific publications universe [6], and in the Web to bring together pages dealing with a mutual subject [3, 1]. Our assumption is that pages that are often co-cited share mutual metadata values.

Then a clusterization is applied to the set of cited pages given the co-citation index matrix $(D_{i,j})_{i,j}$. Only hierarchical ascendant classifications were used in our experiments. The results of these methods are dendrograms. At each level in the dendrogram, the closest classes are joined together. At the lowest level, there are as many classes as elements. At the highest level, there is only one class. Of course, the lowest level classes are homogeneous as they are singletons. At the highest level, the single class contains the whole collection and its homogeneity is that of the collection. A threshold has to be chosen that is a compromise between too small classes and too heterogeneous classes. Choosing a low threshold leads to small classes and thus to a high manual qualification cost. Choosing a high threshold leads to large classes. If the classes are too large, they become heterogeneous and our method will propagate false metadata values.

If a class is fully homogeneous, *i.e.* if the metadata values are the same for every metadata field for every element, then it's enough to qualify any element of this class and then to propagate the metadata values to the other elements. In a less favourable case where a class is not fully homogeneous, the goal is to choose the most representative element of the class, *i.e.* the element that generates the least number of errors when its metadata values are propagated to the other elements of the class. This element can be chosen according to the co-citation subgraph induced on the class and valued with the values $D_{i,j}$. We chose the most *central* element and assumed that it is the most representative one. This element is the one that maximizes the centrality function $C(P_i) = \frac{n-1}{\sum_{j=1}^n d(P_i, P_j)}$ where $d(P_i, P_j)$ is the geodesic distance (the length of a shortest path) between P_i and P_j .

3 Evaluation

To evaluate the first step of our method, we have conducted an experiment on a corpus dealing with astronomy [4]. In this experiment, the pages were manually qualified with respect to four metadata fields: 1. authority type (*institution, association, person, company*); 2. site type (*home server, search site, resource site, Web service*); 3. information type (*self-descriptive* or not); 4. page type (*home page, portal, index, content page, form*). It should be noted

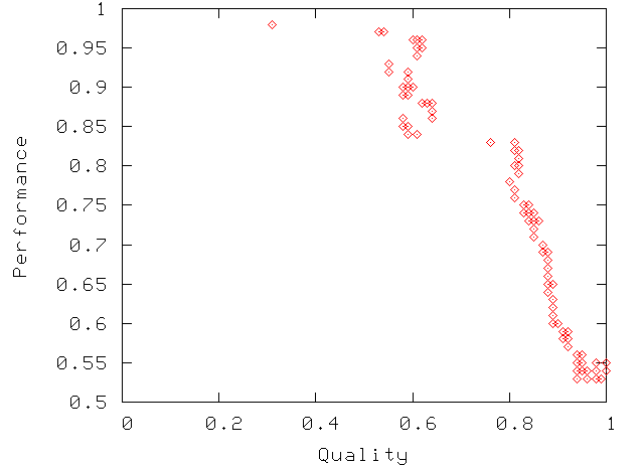


Figure 1. Performance vs. quality with the average link classification method

that sometimes some pages could not be qualified for some metadata fields, not because the metadata values used were not complete but because some information were not available. For instance, some resource sites don't present the authority in charge of the site.

The homogeneity degree was evaluated at the last but highest level – just before all the classes are grouped into one – by measuring the class entropies. We obtained the following results: 77% of the classes were homogeneous for the site type field, 85% for the authority type field, 91% for the information type field. Though for the page type field, only 55% of the classes were homogeneous.

These results show that the external citation links provide information about the document typology. But a high co-citation similarity does not follow up in similarity in the physical characteristics of the pages. This is because there are no standardized rules to split the documents in nodes to obtain a hypertextual document. So co-citing homogeneous documents in their whole does not results in co-citing pages similar in their physical aspects. Thus, in the sequel, the page type field is not taken into account for the evaluation of our propagation method.

In order to evaluate our propagation module, we defined two indices [5] to measure the propagation quality and the method cost, the latter in terms of the quantity of work to be manually done. The quality index Qual is the ratio between the number of rightly propagated metadata values and the number of propagated metadata values. It measures the *precision* of the propagation method and highlights the cohesion within the classes. The performance index Perf is the ratio between the number of propagated metadata values and the number of qualified metadata values (either manu-

ally or automatically qualified). This index is related to the relative cost of the whole qualification. Its value would be 1 if all the metadata values were obtained through propagation, and tends toward 0 as more and more are obtained by manual qualification. The figure 1 displays the points of coordinates (Qual, Perf) obtained at the different dendrogram levels. This dendrogram is the result of the average link classification method. The most favourable point – the closest to the point (1, 1) – is about (80%, 80%). Forwards, a slight increase in quality ends up in a large decrease in performance.

4 Scalability of our method

Our method relies on external co-citations. So only pages that receive at least one external citation are of concern. On a web site, every page receive at least one citation, otherwise this page would not be accessible. Though only some of them receive *external* citations because the number of entry points on a site is rather low. The entry points are either the site home page, or some lower level pages that start logical documents.

Among these pages, only those that share citations with other ones can be classified with our method and then can receive their metadata values through propagation. In traditionnal bibliometry, the outgoing degree distribution follows a Gaussian law: the mean number of outgoing links varying between 20 and 50 from corpus to corpus. On the Web, the outgoing degree distribution follows a hyperbolic law and the probability for a page to have only one outgoing link is high. Thus it is quite possible that a given entry point even with numerous citations has no co-citation. So, some statistics on the Web are to be evaluated to indicate whether or not our method is usable. We used a collection – that we call WFR4 – consisting of 5 057 642 pages¹. In our experiments, the dynamic pages were not considered for themselves but collectively. So the URL were not considered in their completeness, as defined in the RFC 2396, but we only used the part `<host>:<port>/<path>` that we will call “netpath”. This leads to 3 823 589 netpaths in 43 462 sites. Among them, only 1 004 152 netpaths (26.2%) have external citations with other ones: 831 009 (21.7%) emit outgoing external citations and 250 558 (6,5%) receive incoming citations and thus are entry points (cf. fig. 2). Among the former ones, 384 655 emit only one external citation; then 831 009 – 384 655 = 446 354 have many outgoing external links and are to be considered to build the co-citation graph.

The subgraph composed of the set of 446 354 vertices

¹These pages were collected on the Web in December 2000 by Mathias Géry and Dominique Vaufreydaz, members of the CLIPS laboratory of the Grenoble university (France), using the robot CLIPS-Index they developed <http://www-mrim.imag.fr/membres/mathias.gery/CLIPS-Index/>.

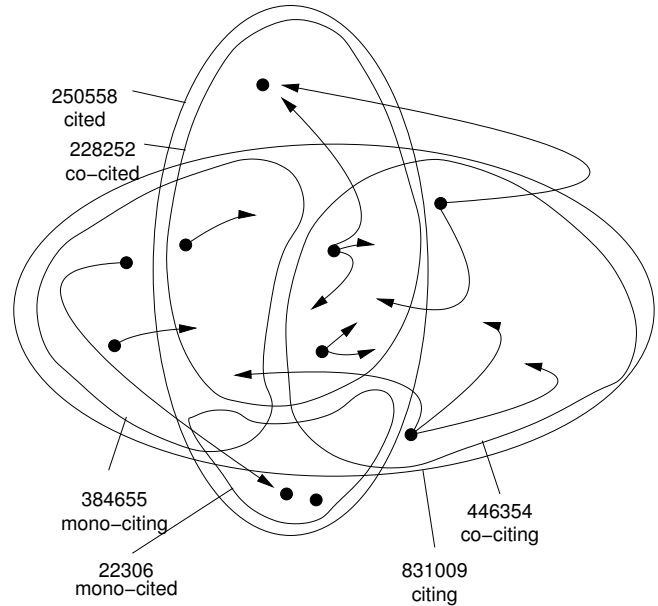


Figure 2. Cardinals of the citing, co-citing, co-cited and cited netpath sets

that emit at least two external citations and the vertices cited by the elements of the former set is composed of:

- 635 535 vertices,
- 2 940 048 arcs,
- 446 354 vertices with an outgoing degree greater or equal to 2; these netpaths belong to 12 229 sites,
- 228 252 vertices receiving citations (and thus they are co-cited); these netpaths belong to 40 239 sites.

From these statistics, we can conclude that among the 3 823 589 netpaths contained in the corpus, 228 252 (5,96%) are co-cited and can participate to the semi-automatic qualification method. This is a quite low ratio. But if we consider the site level rather than the netpath level, 40 239 sites over the 43 462 sites of the corpus (*i.e.* 92,6%) have at least one co-cited entry point.

5 Complexity and computing time

Building the co-citation graph needs several steps. Given the HTML data, the first one consists in parsing the HTML pages to extract the cited URL. In the second step, numerical identifiers are associated to the netpaths of these URL. In the third step, the citation graph and the external citation graph matrices are built. Lastly, the co-citation graph is obtained through a matricial product.

Most of the computing time of the first step (93%) was dedicated to input/output to read the data files and to processor computing time to parse the HTML content. The remainder of the time was used to combine the relative URL found in the HTML files with the base URL of the parsed page to build the absolute URL of the cited pages. About 22 hours were needed to parse the 15 gigabytes². The complexity of this step is linear in the HTML data size. Notice that we could have only processed absolute URL as the relative URL never are external ones; though the gain would have been small.

The second step was implemented with GDBM, the GNU associative table manager. With this tool a dictionary of the URL is built and each URL is associated to its numerical identifier. Similarly, the netpaths and the site names are processed. Managing the 5 057 642 pages, the 3 823 589 netpaths and the 43 462 sites of the collection WFR4 was done within 13 hours. This step is overlinear in the data size, and its complexity is that of the dictionary algorithm. Very larger dictionaries are manageable. Notice that the computing time could have been reduced by two if only the netpaths and the site names would have been processed because in the sequel the URL are no more useful.

In the third step, the previously built dictionaries are used to convert the netpath graph into the (sparse) citation matrix, consuming 7.5 hours of computing time. As in the previous step, the complexity is that of the dictionary algorithm.

The last step consists in computing the co-citation graph by a matrix product. An *ad hoc* algorithm had to be developed to manage the large matrices that are involved. The citation matrix is very sparse and using a general matrix representation would limit its size to about one thousand vertices with some hundred megabytes of central memory, which is quite inadequate. With m citing vertices and n cited vertices, the time complexity is $O(m \cdot n^2)$. With our *ad hoc* implementation, computing the product was achieved within one hour at the site level ($m = 12\,229$ and $n = 40\,239$). Extrapolating this measure leads to a computing time of 1175 hours at the netpath level ($m = 446\,354$ and $n = 228\,252$). This rather long time (49 days) seems accessible because very more powerful computers could be used. Moreover this computation concerns a part of the Web which is rather stable and don't need to be updated frequently, so a computing time of some days is acceptable.

Once the co-citation graph is obtained, a clusterization has to be done. The algorithm we used is not scalable due to its complexity. As there are many works about the clusterization algorithms and their scalability, a dedicated study concerning which scalable algorithm to use in our method should be conducted.

²The quoted times were measured on a computer with a 1 GHz Pentium and 512 megabytes of memory.

6 Conclusion

We presented a semi-automatic qualification method of the Web resources. If only a little ratio (6%) of the pages could be qualified with our method, these pages cover 92% of the sites. So another method is needed to complete this one at the site level if the goal is to qualify a more significant part of the whole Web and not only the co-cited site entry points. We showed that our method would be applicable in the large, though the question of the clusterization algorithm remains to be studied.

The possibility to qualify the Web resources with reliable metadata at low cost opens a path to new search engines that would use these metadata. These search engines could answer the users by taking into account some aspects of the information needs such as the type, the level, etc. Our approach is somewhere between the current search engines which try to automatically index the whole Web and the directories which classify and index only at the site level by manual methods.

References

- [1] R. Larson. 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*, 1996. <http://sherlock.berkeley.edu/asis96/asis96.html>.
- [2] B. Michelet. *Analyse des associations*. PhD thesis, Université de Paris 7, 1988.
- [3] C. Prime, E. Bassecoulard, and M. Zitt. Co-citations and co-sitations: A cautionary view on an analogy. *Scientometrics*, 54(2):291–308, 2002.
- [4] C. Prime-Claverie, M. Beigbeder, and Th. Lafouge. Clusterisation du web en vue d'extraction de corpus homogènes. In Florence Sédes, editor, *actes de INFORSID 2002, 20e congrès informatique des organisations et des systèmes d'information et de décision*, pages 229–242, juin 2002.
- [5] C. Prime-Claverie, M. Beigbeder, and Th. Lafouge. Transposition of the co-citation method with a view to classifying web pages. *Journal of the American Society for Information Science and Technology*, 55(14):1282–1289, 2004.
- [6] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.