Automatic Annotation Using Citation Links and Co-citation Measure: Application to the Water Information System

Lylia Abrouk, Abdelkader Gouaïch

Automatic Annotation Using Citation Links and Co-citation Measure: Application to the Water Information System

Lylia Abrouk[1,2] and Abdelkader Gouaïch[1]

[1] LIRMM, Laboratoire d'Informatique,
de Robotique et de Microlectronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5
{abrouk, gouaich}@lirmm.fr
http://www.lirmm.fr
[2] EMWIS, Euro-Mediterranean Information System
on the know-how in the Water sector
2229, route des cretes, 06560 Valbonne
l.abrouk@semide.org

Abstract. This paper describes an approach to automatically annotate documents for the Euro-Mediterranean Water Information System. This approach uses the citation links and co-citation measure in order to refine annotations extracted from an indexation method. An experiment of this approach with the CiteSeer database is presented and discussed.

1 Introduction

The Web offers technologies to share knowledge and information between organisations and users that can be distributed world-widely. In this paper we discuss the use of Web technologies for a specific professional domain that is sharing information on water management among Mediterranean countries that participate to the Euro Mediterranean Information System on the know how in the water sector (EMWIS, www.emwis.org).

EMWIS is an information and knowledge exchange system between the Euro Mediterranean partnership countries, necessary for the implementation of the Action Plan defined at the Euro Mediterranean Ministerial Conference on Local Water Management held in Turin in 1999. The objectives of EMWIS are as follows:

- Facilitate the access to information on water management;
- Develop the sharing of expertise and know-how between the partnership countries;
- Elaborate common outputs and cooperation programs on the know-how in the water field.

Using Web technologies within EMWIS to make information available is necessary but far from being sufficient. In fact, information is useful only when it can be retrieved later by users that need it. However, searching for the most

relevant information that meets user's request is still a problem especially when informations are coming from heterogeneous sources and sometimes accessible only with some rights. To solve this problem, informations, that are abstracted as *resources*, are annotated to describe both: *(i)* their context of creation: names of the authors, date of appearance and so on; *(ii)* and the semantics of their content.

The annotation of resources is very useful in order to match users' requests with resources that are available within EMWIS. However, annotating manually all the resources in a large system such as EMWIS is infeasible.

In this paper, we present an approach in order to annotate automatically a set of unannotated resources by using citation links. By contrast with classical Web approaches for automatic annotation, we use a restrained vocabulary of annotation defined in the EMWIS's global ontology.

The rest of the paper is organised as follows: Section 2 presents the context of our work and states the problem treated in this paper; Section 3 presents the backgrounds of works that have already used link analysis for different purposes such as statistical analysis, classification of resources, and meta-data propagation; Section 4 presents our approach in order to automatically annotate resources using citation links; Section 5 presents our experimentation with the Citeseer data base; and finally, Section 6 presents some perspectives and conclusions.

2 Context and Problematic

The global architecture of EMWIS defines the following entities: a National Focal Point (NFP) for each country and a single Technical Unit (TU). The NFPs are restrained teamworks that:

- create and make available a national server to access information;
- handle and manage the information system's national users.

The TU acts as a facilitator in helping each NFP to set up their information system and ensuring the coordination among all the NFPs. It is worth noting that the architecture of the EMWIS is fully distributed and Web technologies are used to share information among all EMWIS entities.

Figures 1 presents an example where a user searches some resources (documents in this case) on a specific theme. The documents are distributed among all the NFPs. To answer the user's request we face a first problem that is related to the description. In fact, the documents have to be well described by using all possible languages spoken within the EMWIS participating countries. To avoid this problem we have considered a common vocabulary to describe the resources. This is known as the EMIWIS global ontology. We have also to consider that this ontology is not static and can be updated by adding new concepts.

To implement technically EMWIS objectives, we have considered the following goals for our work:

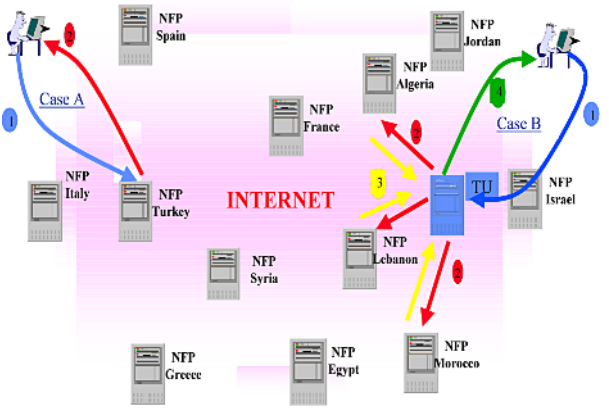


Fig. 1. EMWIS Architecture

1. resource annotation: in this part, we are focused on how to annotate automatically resources that have not been annotated by the experts of the domain;
2. global ontology enhancement: in this part, we are focused on how to add new concepts and relationships within the global ontology and how to update automatically the existing annotation of resources.

This paper targets only the first part of the work and presents means in order to annotate automatically a large set of resources using the citation links that structurally exist among resources. In fact, within a large system like EMWIS it is not feasible to assume that the content of all the resources can be described manually by experts. Our goal is then to provide a mean to assist experts and content managers to annotate the resources by suggesting automatically some annotations after analysing the citation links of already existing resources.

3 Backgrounds

Before presenting the state of the art, we provide some general definitions that will be used in the rest of the paper:

1. A document is the material that supports the encoding of information. The document can be either a hard copy, a web page, or any other medium that makes information persistent.
2. A resource is a generic concept that we use in order to talk about documents when these documents are needed to be used. There are several relationships between resources such as: citation, access link (for instance hyperlinks).
3. A citation occurs between documents: in this case, the document that *cites* another document, indicates that it is 'talking about' some parts of the *cited* document.

4. An access link, or hyperlink, indicates that the *target* document can be accessed directly from the *source* document.

This section presents works that have already used the relationships among the resources in order to:

- Extract statistical information;
- Classify the documents according to their importance;
- Propagate annotations and meta-data among documents.

3.1 Bibliometry

The Bibliometry is a statistical analysis of scientific publications [11]. It provides some qualitative and quantitative mesures about the activity of producers (scientists, laboratories and so on) and broadcasters (journals, editors and so on) of scientific documents.

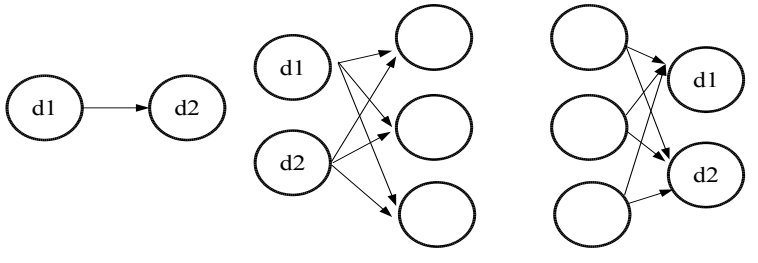
The bibliometry field considers the citations among the documents: *citation analysis*. The citation analysis is about establishing relations between the authors and documents and defining other more complex relations such as the co-reference and co-citation. These relationships are described in more details in the next paragraph.

3.2 Citation Analysis

Scientific documents can be modelled as an directed graph $G = (N, A)$ where the nodes represents articles and the arrows citation relationship.

Figure 2 illustrates some relationships among documents:

- Citation relationship: when a document d_1 references a document d_2 for instance. Generally the citation analysis determines the impact of one author on a given field by determining the amount of time that this author is cited by others.



(A) d1 cites d2 (B) Bibliographic coupling of d1 and d2 (C) co-citation of d1 and d2

Fig. 2. Relations among documents

- [8] has introduced the *bibliographic coupling* relationship. Documents are considered as bibliographically coupled when they share one or more bibliographic references. However, the bibliographic coupling is now displaced by co-citation clustering.
- The co-citation relationship represents documents that are cited by the same documents. The co-citation method [5], that has been used in bibliometry since 1973, aims to create relationships between the documents that are in the same domain field or *theme*. The hypothesis is that documents which are cited jointly share the same theme.

3.3 Propagating Meta-data Using Links

Marchiori [12,13] has used link analysis to propagate meta-data among documents (Web pages). His idea is that when a document d_1 owns some meta-data (or keywords) (a_v) (which indicates that the keyword a has a weighting equal to v) and there is a document d_2 with a hyperlink to d_1, then the keywords of d_1 are propagated to d_2 but with a loss factor, f, such that the keywords are equal to $(a_{v\times f})$. The same mechanism is then applied to all pages that are linked to d_2. This time, the resulting keywords weighting will be $(a_{v\times f\times f})$. Consequently, the keywords of the initial page d_1 are propagated to all accessible and indirectly accessible pages with a loss factor until reaching a defined threshold.

Prime [17] has also used links in order to propagate meta-data among documents. The core idea of this work is to add nonthematic meta-data to thematic meta-data that have been added by search engines. As Marchiori, Prime considers that when a link exists between two documents then these documents share the same thematic. However, Prime does not propagate meta-data using directly the Web graph but by using a subset called *co-citation graph*. The first step of this methodology is to determine the similarity between Web documents using a similarity index: two pages are close according to their citation frequency and co-citation frequency. The second step gathers closest pages in clusters.

3.4 Link Analysis for the Web

The classification of web pages is a known example that uses link analysis to find most important pages. The most known algorithms are: *Page Rank* [2,3,14] and *HITS*[9,7].

The *Page Rank* algorithm is used by Google[1] to classify web pages. The principle of this approach is to consider that a page is more important if there are several pages that point on it. This measure assumes three hypothesis [1]:

1. the popularity of a page depends on the popularity of the pages that point on it;
2. the links of a page do not have the same importance;
3. the popularity of a page does not depend on the users' requests.

[1] www.google.com

The *Hits* algorithm uses a search engine to identify in a set of web pages *authorities* and *hubs*. Hubs and authorities exhibit what could be called a *mutually reinforcing relationship*. A good hub is a page that points to many good authorities; a good authority is a page that is pointed to by many good hubs. An iterative algorithm HITS (Hypertext Induced Topic search) calculates the values of hubs and authorities for each page. The first step consists of posting a query; Hits assembles a root set s of pages, returned by the search engine on that query; it expends then this set to a larger base set t by adding in any pages that point to, or are pointed by, any page on s. The second step is to associate with each page a *hub-weight* and an *authority-weight*. The update operations are performed for all the pages, and the process is repeat until convergence that is proven to be reached.

We can also mention other uses of links such as for: *(i)* geographical categorisation of Web pages such as in work of [4]. *(ii)* discovery of similar Web documents, for instance [15] calculates using links the level of similarity between Web documents; *(iii)* discovery of communities in the Web such as in work of [7,10,19].

4 Automatic Annotation of EMWIS Documents

Before describing our approach, the types and the organisation of EMWIS documents are presented. EMWIS documents can be one of the following types: news, event document, legal documents, technical document, slide presentation document, Web document. The events are seminars, workshops, conference, courses that are organised by EMWIS.

For an event there is a Web document that includes a description and links to other documents related to this event. In the rest of the paper, when there is no ambiguity the term 'event' is used directly to talk about the 'event document'. Each event cites other documents, such as the Web page of the NFP that organises the event, a document that describes the topic of the event, and a set of presentations and publications. Most of the EMWIS documents are not annotated and this task is impossible to perform manually.

Section 2 has presented two main questions related to: *(i)* the uniform description of documents to avoid translating annotations in each language; *(ii)* the annotation of all documents using terms defined in the global ontology.

To answer the first question, we have defined a global ontology of the EMWIS community. This ontology is a set of concepts structured as a tree. The links among the concepts are semantic relationships (synonymy, aggregation, composition) or inheritance. To each concept we associate a set of terms in each language. Figure 3 describes a small part of the EMWIS ontology.

Figure 4 describes the major steps for the annotation and the global ontology enhancement processes:

1. For the document d a first annotation is generated using an indexation method. The result is a set of concepts belonging or not to the global ontology. Let E_{og} be the set of concepts that belong to the global ontology and

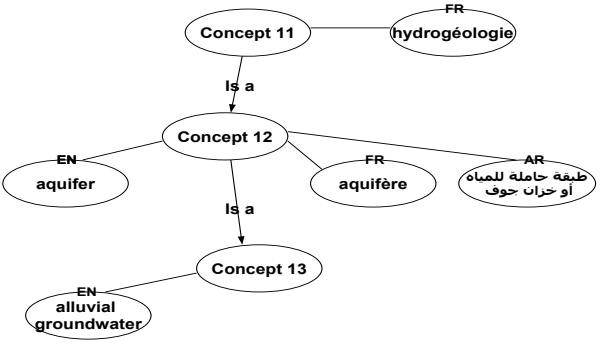


Fig. 3. EMWIS ontology

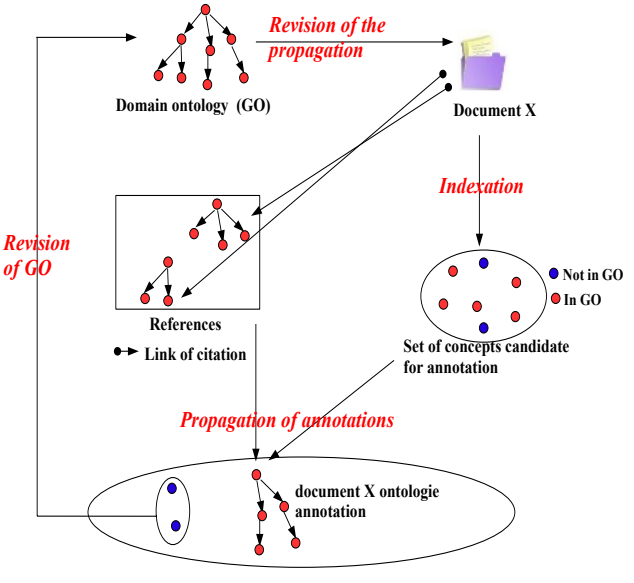


Fig. 4. Global solution

E_{ong} the set of concepts that do not belong to the global ontology. The result of the annotation of a document is then $E_{og} \bigcup E_{ong}$. It is worth noting that the annotation generated by the indexation is not precise enough to describe to content since it contains a lot of terms and noises.

2. On the basis of the assumption that all the documents are annotated by using only concepts of the global ontology, the second step refines the first annotations by using the annotations coming from the citations of d. This is performed by adding or removing concepts from the set E_{og}. This step is known as *the propagation of the annotations*.

3. The third step which is the enhancement of the global ontology updates the global ontology by concepts defined in E_{ong}.
4. The update of the global ontology might imply the revision of the propagation process (step 2).

This article is focused only on the propagation of the annotations. So, having the structure of EMWIS documents, we suggest to use the citation links ,similarly too [13] and [17], to select meaningful annotations. To implement this solution, one has to answer the following questions: *(i)* what citations should be taken into account? In fact, not all the citations in a document are meaningful to determine the theme of the document; *(ii)* How to annotate the document? *(iii)* and finally, how to merge annotations that come from the selected documents.

The answer of these questions is provided by the following steps:

- structuring the documents using the co-citation analysis;
- selecting a subset of cited documents;
- importing and selecting the annotations which are coming from the selected documents.

4.1 Building the Co-citation Graph

When an author cites another document, this is done to indicate that the cited document contains some information that relevant to the context of the citation. However, we can also find citations that contribute to a small part of the document and do not necessarily determine the general theme of the whole document. Consequently, we have to consider only citations that contribute to determine the thematic of the source document. The co-citation method has been proven to be a good measure to determine the similarity on theme among documents. In fact, when documents are often cited together by different documents, we can assume that they target the same subject. We use the similarity index as described by [16] as follows:

$$SI_{(i,j)} = \frac{C^2_{(i,j)}}{C_{(i)} * C_{(j)}}$$

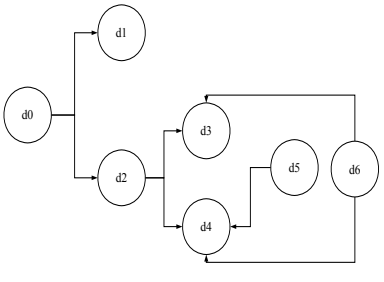


Fig. 5. An example of a citation graph between documents

- $C_{(i,j)}$ is the co-citation frequency or the number of time that i and j are cited together;
- $C_{(i)}$ is the citation frequency or the number of time that the page i is cited;
- $C_{(j)}$ is the citation frequency or the number of time that the page j is cited.

A distance function $d(i,j)$ is then defined as $d(i,j) = 1SI_{(i,j)}$. Using this distance function the co-citation matrix and graph are built as shown by the example presented in Figure 5.

The co-citation matrix of the example presented in Figure 5 is:

$$\begin{pmatrix} 0.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 0.00 & 0.50 & 1.00 & 1.00 & 1.00 & 1.00 \\ 1.00 & 0.50 & 0.00 & 1.00 & 0.83 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 0.00 & 0.33 & 1.00 & 1.00 \\ 1.00 & 1.00 & 0.83 & 0.33 & 0.00 & 1.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 0.00 \end{pmatrix}$$

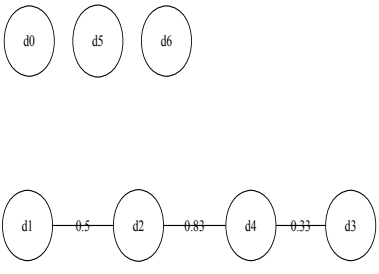


Fig. 6. The co-citation graph of the example presented in Figure 5

This matrix defines the co-citation graph that is presented by Figure 6: documents are linked with a weighted link that is equal to the co-citation distance; values equal to 1 are ignored.

The next step is to determine some clusters by defining a threshold distance f. The maximum distance following the paths within a cluster cannot exceed this threshold. We use classical clustering methods in order to have clusters with maximum documents and were the maximum distance between the documents following paths do not exceed the threshold. For instance, when $f = 0.5$ then we build two clusters as presented in Figure 7. These clusters are interpreted as themes were the documents are aggregated on.

4.2 Selecting the Meaningful Citations

Figure 8 presents a case when a new document is being included to the system. d_7 cites some exiting documents : $\{d1, d3, d4\}$. We assume then a document can

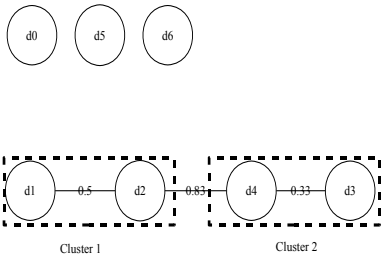


Fig. 7. Clusters with a threshold set to $f = 0.5$

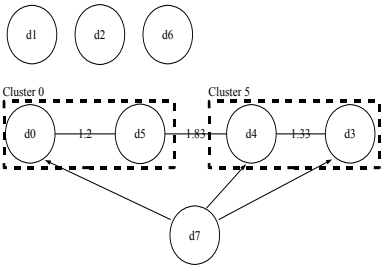


Fig. 8. Adding a new document d_7 to the system

target more that one theme. So, one has to provide a mean in order to select which citations are considered for the import. We suggest to define an order relationship between the clusters relatively to a document d:

$$cl_1 \leq_d cl_2 \equiv (\#\{cl_1 \cap citations(d)\}, \#cl_1) \leq (\#\{cl_2 \cap citations(d)\}, \#cl_2)$$

In this order relationship the first criterion considers the numbers of citations that belong to the cluster; the second order criterion considers the importance of the cluster.

When considering the previous example we have:

$$cluster_1 \leq_{d_7} cluster_2 \; as: \quad (1, 2) \leq (2, 2)$$

By using this order relationship an ordered list of clusters for the incoming document is created. We add another parameter that is the maximum number of themes allowed for a document: *max_theme*. The document that are selected for the annotation import are those that are cited by the article and that belong to the highest *max_theme*-clusters of the ordered list.

For instance, if we consider that $max_theme = 1$ for the simple example; then we select only documents that belong to $cluster_2$ and that are cited by c_7, which means $\{d4, d3\}$.

4.3 Selecting and Importing the Annotations

The last step has produced a set of articles for the import. However, one has to make a choice on: *(i)* what annotation to select; *(ii)* and what to import in an annotation knowing that every annotation is a tree of concepts defined within the EMWIS global ontology.

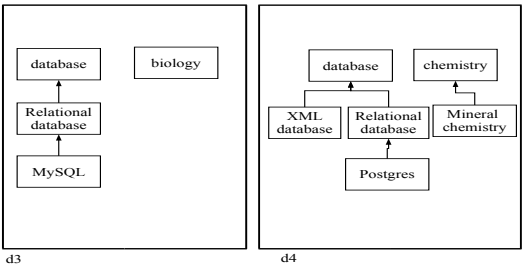


Fig. 9. An example of the annotation of documents d_3 and d_4

Figure 9 presents an example of annotation of documents d_3 and d_4. The naive solution would be to import the whole annotations of the document d_7. However, some annotation concepts are either not relevant with the content of d_7 or too specific for d_7. To solve these problems we use the result of the indexation process of the document. In fact, the indexation will generate the set of terms that appear frequently in the document. Consequently, only the intersection between in set of terms produced by the indexation and the concepts of the annotations is considered. This allows to remove concepts that are not found in the document and to select the right level in the annotation tree. For instance, if the term *'relational database'* appears several times within the document d_7, it will be produced by the indexation process. The intersection of this term with the annotations of d_3 and d_4 will remove the *'chimestry'* and 'biology' concepts as d_7 does not use these terms. Concepts that are too specific to d_3 and d_4 such as the type of database system used (MySQL, Postgres) will also be remove since the intersection stops the depth of the tree to *'relational database'*. So, the final annotation of d_7 will be the tree starting by the concept *'database'* and until the concept *'relational database'*. It is worth noting that in this simple description of the example we have simplified the process. In fact, we have used some means (such as synonyms and so on) in order to map indexation results , which are general terms, to the concepts of the global ontology.

5 Realisations and Experiments

To experiment with our approach we have considered the CiteSeer[2] collection as a test database. CiteSeer [18],[6] is a digital library for scientific literature. CiteSeer localises scientific publications on the Web and extracts some information

[2] http://citeseer.ist.psu.edu/

such as citations, title, authors and so on. This collection has been selected for two reasons: *(i)* the important number of documents; *(ii)* the fact that is contains scientific documents that use several citations. We have built a database that contains more than 550 000 documents.

However, CiteSeer description of documents cannot be used directly. In fact, CiteSeer uses a general vocabulary to describe the content of a document. But, we are interested only to description of documents using a controlled vocabulary or an ontology. We have used the ACM controlled vocabulary as an ontology to annotate CiteSeer documents during the experiment.

The presented approach has been implemented and the automatic annotation of unannotated articles has been performed. The experimentations show that the indexation keywords have been considerably refined when considering the citations of the document. Furthermore, for a concept, x,that has been selected for the annotation the fact that all its parents will be included for the annotation this adds information that can be useful during the search. In fact, if the user request is not directly related to the concept x but about his father concepts, then the document will be selected as potentially interesting for the user.

For the CiteSeer database we are remarked that the co-citation graph naturally express the clusters and themes so there is not need for the f parameter. In fact, this parameter has been expressed by [16] to split clusters on themes but of a specific domain all the documents are transitively cited together express a cluster for a specific theme. We have also remarked that setting *max_theme* higher to 3 does not affect the results on annotations. This can be explained by the fact that scientific and technical papers targets specific themes and do not uses more than 3 themes.

However, we have not defined until now an objective evaluation method to prove the efficiency of our approach. In fact, all the evaluations are subjectives and tries to compare the automatic annotation with the annotations of a human expert. As a perspective, we have to provide an objective method as an evaluation of this approach. This problem can be faced in almost all similar works on the same field.

6 Conclusion

This paper has described an approach in order to automatically annotate documents used by a specific community, namely EMWIS users. Annotating manually all documents in a distributed and large information system is a hard task and the classical indexation methods generate too fuzzy and imprecise keywords. We have exploited the citation relationships an information about the context of the document in order to refine its annotation and to add general the concepts defined within the ontology of the domain.

The work that has been presented is this paper differs from other works that target general and open communities such as the Web. In fact, we address here a specific quite close community, which facilitate the elaboration of an ontology of the domain. This makes the annotations independent from the multiple

languages spoken within the community and helps also for the searching the appropriate documents that meet users' requests by structuring the search from the specific to the general concepts of the annotations.

The experiments with the CiteSeer database has shown the feasibility of the approach and have allowed the automatic annotation of scientific articles. However, we still need an objective measure to evaluate our approach independently from human experts.

References

1. F. Aguiar. *Modélisation d'un systme de recherche d'information pour les systmes hypertextes. Application la recherche d'information sur le World Wide Web.* PhD thesis, Ecole suprieure des Mines de Saint-Etienne, 2002.
2. A. Arasu, J. Novak, A. Tomkins, and J. Tomlin. Pagerank computation and the structure of the web: Experiments and algorithms, 2001.
3. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Australia, 1998.
4. O. Buyukkokten, J. Cho, H. Garcia-Molina, L. Gravano, and N. SHivakumar. Exploiting geographical location information of web pages. In *WebDB (Informal Proceedings)*, 1999.
5. E. Garfield. Co-citation analysis of the scientific literature: Henry small on mapping the collective mind of science. *Essays of an Information Scientist: Of Nobel Class, Women in Science, Citation Classics and Other Essays*, 15(19), 1993.
6. S. Ghita, N. Henze, and W. Nejdl. Task specific semantic views: Extracting and integrating contextual metadata from the web. In *In Submitted for publication, L3S Technical Report*, 2005.
7. D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, 1998.
8. M. Kessler. Bibliographic coupling between scientific papers. In *American Documentation*, pages 10–25, 1963.
9. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Journal of the ACM*, pages 139–146, 1999.
10. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Computer Networks*, Amsterdam, Netherlands, 1999.
11. P. Lauri. The bibliometrics a trend indicator. In *International Journal Information Sciences for Decision Making*, page 2836, 1997.
12. M. Marchiori. The quest for correct information of the web: hyper search engines. In *The Sixth International WWW Conference*, Santa Clara, USA, April 1997.
13. M. Marchiori. The limits of web metadata, and beyond. In *Proceedings of the Seventh International World Wide Web Conference*, pages 1–9, Australia, 1998.
14. R. Mihalcea, P. Tarau, and E. Figa. Pagerank on semantic networks, with application toword sense disambiguation. In *Proceedings of the 20th international conference on computational linguistics (COLING2004)*, Geneva, Switzerland, 2004.
15. D. Phelan and N. Kushmerick. A descendant-based link analysis algorithm for web search. 2002.

16. C. Prime-Claverie. *Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web.* PhD thesis, Ecole suprieure des Mines de Saint-Etienne, 2004.
17. C. Prime-Claverie, M. Beigbeder, and T. Lafouge. Propagation de métadonnes par l'analyse des liens. In *Journés Francophones de la Toile - JFT2003*, France, juillet 2003.
18. J. Stribling, I. G. Councill, J. Li, M. F. Kaashoek, D. R. Karger, R. Morris, and S. Shenker. Overcite: A cooperative digital research library. In *International Workshop on Peer-to-Peer Systems*, 2005.
19. V. Vandaele, P. Francq, and A. Delchambre. Analyse d'hyperliens en vue d'une meilleure description des profils. In *Proceedings of JADT 2004, 7es Journes internationales d'Analyse statistique de Donnes Textuelles*, 2004.