International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

December 2, 2003

VOLUME 4 (2000): ISSUE 1. PAPER 1

How to hold a Virtual Library Active?

Bruno Mannina, Luc Quoniam
CRRM Centre de Recherche Rétrospective de Marseille / Université Aix-Marseille III
F-13397 Marseille Cedex 20 - France
bruno@intelligence-process.com; quoniam@univ-tln.fr

Abstract

During these past five years, the Internet phenomenon has been marked by a tremendous growth of users as well as the amount of available information. Since then, Internet has been inevitable for the information specialists and the network of networks has many advantages: a lower cost medium of communication, user-friendly, a world wide means of communication and exchange and mostly an open way of providing dynamically information to all. Because of its very different nature, searching information on the Internet becomes ever more harder. The user has to be familiar with index search, in order to find the right information. But the unbelievable growth of information sources available on the Internet makes these tools more and more inefficient. Actually, a new generation of index search has sprung off. These tools relieve the user of the repetitive commands associated with online browsing. These tools are called «Intelligent Agent». We propose in this article how can we use an intelligent agent to create and to hold a virtual library active.

Keywords

Virtual library, Robot, Intelligent Agent, Hypertext, Internet

I. Positioning of the problem

Recently, the extension of the Internet network has made the task of the webmaster difficult. Indeed the person in charge of the maintenance of the Virtual Library is confronted with the fluctuations of the Web. He must constantly be informed of the new Internet links relevant to the topic of his search. Moreover, he must have an active virtual library (**Manning, 1999**), test the old Internet links frequently to check if those are still valid.

We had to find a method enabling us to make this task automatic.

II. Definition of Automatic Virtual Library

To create an Automatic Virtual Library, you must use a program (**Chen et al., 1998**) that is able to seek automatically the new Internet links to build a

database (**Stein, 1991**). The users will be able to find what they seek (**Rostaing,1993**). The administrator (Webmaster) of the virtual Library is thus relieved of the repetitive tasks of search and validation of the links (**Paunak, 1989**) present in this one.

III. The different methods to hold a Virtual Library Active

There are 3 methods :

- Manual
- Semi-Automatic
- Automatic

III.1 Manual Method

The Webmaster receives a lot of information from the World (**Rice et al, 1989**; **Dousset et al., 1993**) For this, he uses many different sources which are:

- Congress, Seminars...
- Books, Magazines, Newspapers
- News, Forums
- Experts
- Internet (WWW)
- Mailing List
- Informal Information

After accepting the new Internet links, the Webmaster must update the Virtual Library File by editing the HTML file and adding the new link manually.

What are the advantages of operating a virtual library manually?

- Internet Links are relevant.
- There are no bad Internet links.
- Internet links come exclusively from the webmaster.

What are disadvantages of holding a virtual library manually ?

- The Webmaster must test frequently the new internet link.
- The good functioning of a Virtual Library depends on the time devoted by the webmaster.
- The list of Internet links isn't exhaustive.

What is the frequency of update ?

- First, each time that the webmaster finds a new internet link.
- Second, once a month to test the old internet links.

How much time does the webmaster need to hold this library ?

- First, five minutes to add the new internet link.
- Second, one hour a month to test all the internet links

III.2 Semi-Automatic Method

An HTML server must have a CGI program which allow users to add a new internet link and some other information about this link.

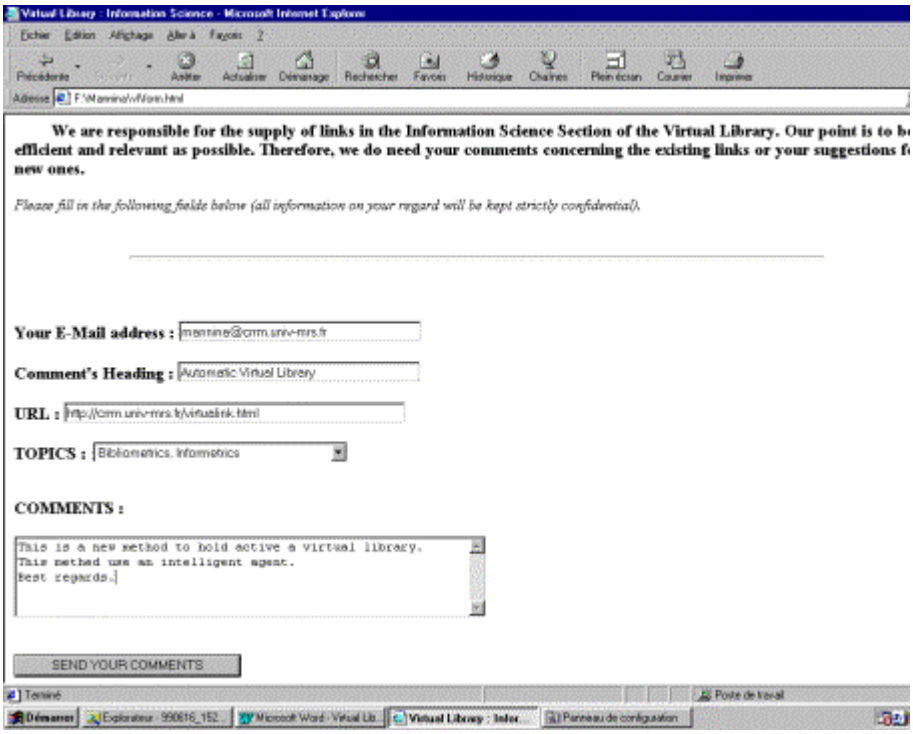

Figure 1 : CGI Form to add a new link.

The user must fill out a form with:

- His E-mail address
- The Title of the new link
- The URL of the new link
- The Description of the new link

When the form is accepted, the new internet link is added to the virtual library file. At the same time, the webmaster receives an Email which informs him that a new link has been added into the virtual library.

It's only at this point that the Webmaster is going to verify the new link added and if the new link isn't relevant, then the Webmaster will delete it.

What are the advantages of holding a virtual library semi-automatically?

- Internet links are relevant.
- The new link doesn't come exclusively from the Webmaster.

What are the disadvantages of holding a virtual library semi-automatically?

- The Webmaster must test the new and old internet links.
- The list of internet links isn't exhaustive.
- A lot of new internet links aren't relevant, the webmaster must delete them.

What is the frequency of update?

- First, each time that the webmaster receives an Email of the CGI program.
- Second, Once a month to test the old internet links.

How much time does the webmaster need to hold this library?

- First, a few minutes every time that he receives an Email of the CGI program.
- Second, one hour every month to test all the internet links.

III.3 Automatic Method

The Webmaster must configure the software (**Mannina, Dou & Florence, 1997**) only once to create the new Virtual Library.

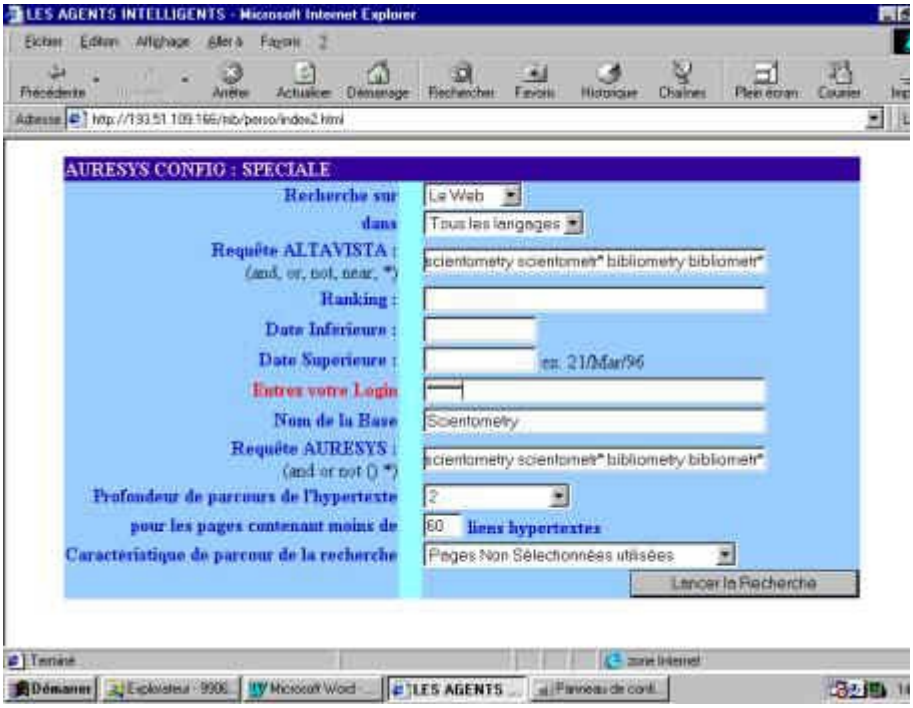

Figure 2 : Form to configure the intelligent agent.

He must fill out the form, including:

- Keywords describing the virtual library
- The name of the new Virtual Library
- Depth of research

After filling out the form, the webmaster must run the software. The software (Intelligent Agent) will generate the new Virtual Library (**Dou et al., 1999**)

Once the research is completed, the webmaster must verify the new virtual library. He has to check that the research is well done.

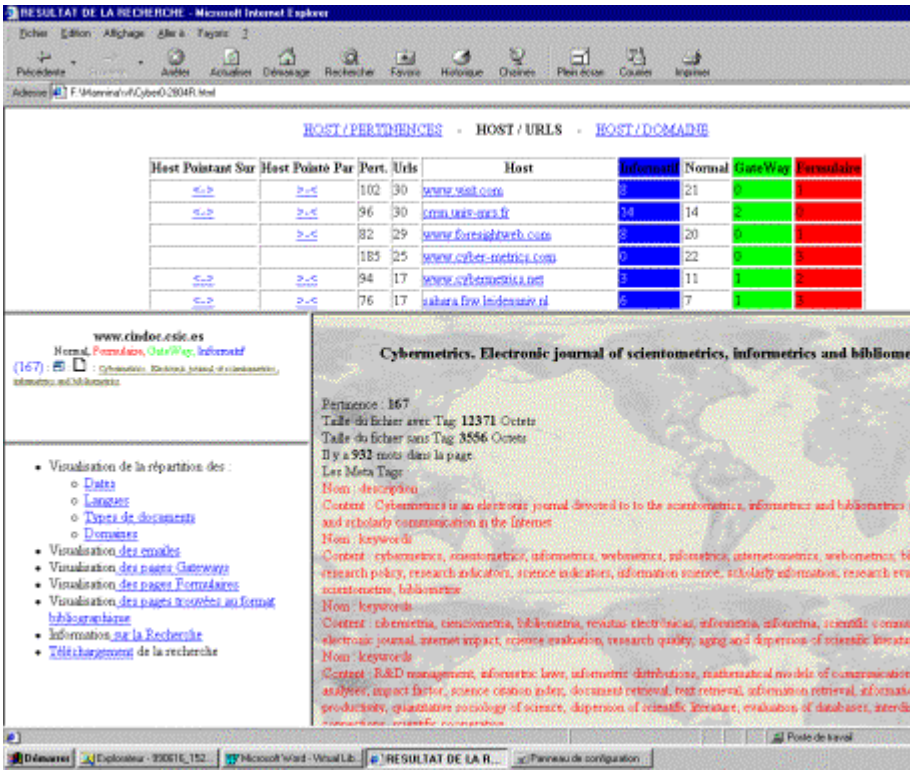

Figure 3 : Virtual Library Automatic

What are the advantages of holding a virtual library automatically?

- Internet links are relevant
- Saves time
- Update is quick
- Update can be also automatic
- The exhaustively is nearly total
- Consultation of the Virtual Library is intuitive
- It gives a lot of information (**Smith, 1997**) (emails of expert, gateway files, form files, server citations (**Rousseau, 1997**; **Small, 1973**), etc.).

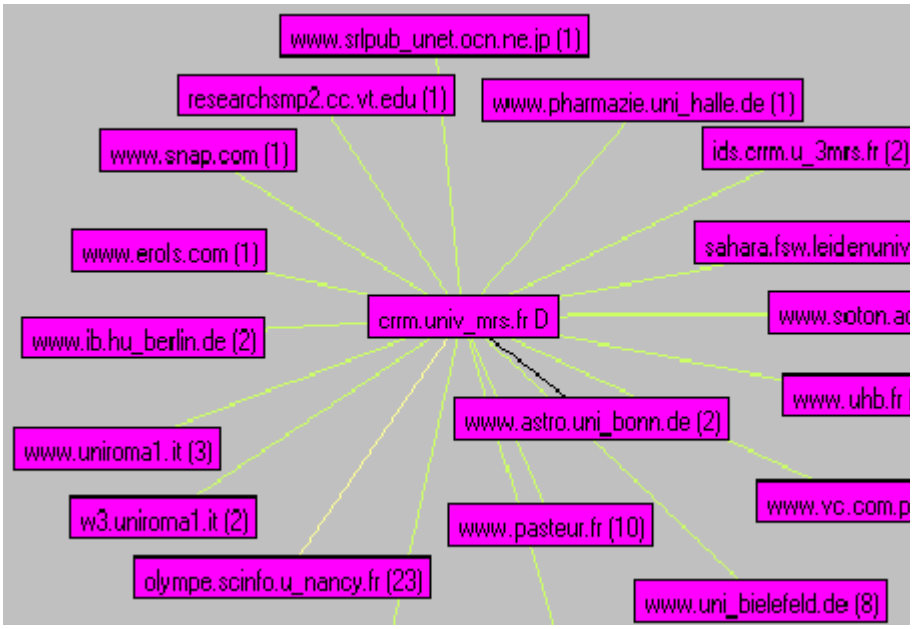

Figure 4 : Server citation : e.g : CRRM is the destination.

What are the disadvantages of holding a virtual library automatically?

- Virtual Library is an instantaneous picture of the web
- A relevant virtual library depend on the initial request

What is the frequency of update?

- Every month a new search is done. (A new search is done every month)

How much time does the webmaster need to hold this library?

- Five minutes the first time and after the software reruns automatically.

IV. Problems with the Automatic Virtual Library

To better understand how to run an automatic virtual library, we shall give examples. We made several searches, with the scientometry and the bibliometry as topics. The request was:

scientometri* or bibliometri* or scientometry or bibliometry

The number of results returned by Altavista for this request (in Advanced mode) is:

3518 answers.

Altavista makes it impossible to look for more then the first 200 results in normal mode so we have used the advanced mode to obtain the maximum number of readable answers, 1010.

The Intelligent Agent having been used for automatic harvest of the documents is Auresys, software developed by laboratory CRRM. This software was selected for several reasons such as the possibility to modify several parameters of search (depth of search, limitation in the size of page HTML selected, navigation,..). Auresys uses only one engine of search: Altavista. Useful for our demonstration.

Note : All searches have be made within one week, in order to have the least possible number of differences possible in the number of results returned by Altavista.

V. Example of Automatic Virtual Library

This is an example of automatic virtual library made by Auresys. This software used the results found by Altavista to build, the new virtual library. Auresys obtained the links (**Mannina, Dou & Florence, 1997**) one by one from Altavista, and checked to see if they answered the initial request.

The table below shows the values resulting of this search.

Depth of search	***1***
Urls Selected	512
Urls Not selected	192

Urls Not selected (size > 40 KB of text)	173
Urls in Errors	133
Urls Visited	1010
Urls in errors of protocol (except protocol HTTP)	0
Time to search	13h 13'

After having analyzed the contents of the virtual library, it appears that some hosts do not have all their HTML files (those answering the request), present in the new library. Altavista limits the number of readable results to 1010 answers. It was necessary, to access the other results, in order to find a new parameter in the Auresys software enabling us to resolve this problem. Auresys proposes to its users to modify the search parameter, which is the depth of the search.

VI. Virtual Library and Depth of search

Before showing you a new search, we will explain an in-depth search.

Navigation, known as in-depth, consists in using hypertext links present in an HTML file to follow the search. This method makes it possible to find files which still have not been indexed. However, the number of documents found by this method is fewer because the search engines generally index a whole Internet host.

The benefits can be explained by the fact that the search engine spends several months before updating its database. Some new HTML files appear on some host which have been already indexed.

The figure below shows a traditional hierarchy of a host. The depth 1 corresponds generally to the "Home Page". In theory, an in-depth search of a Web server, would make it possible for the user to navigate through all the HTML files of the server.

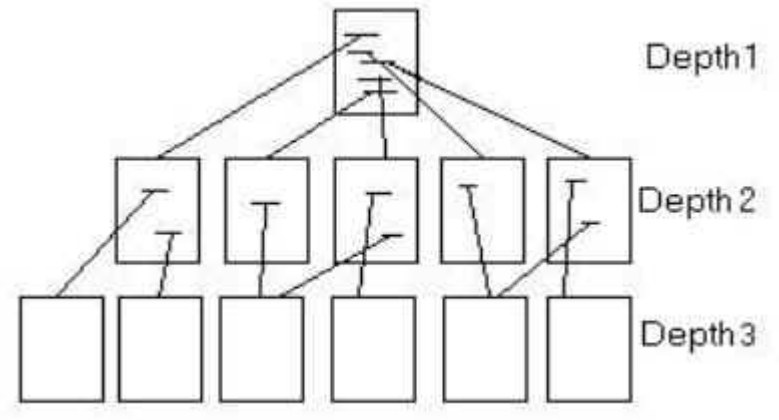

Fig 1: Structure of HTML server.

Here the values resulting from different searches:

Depth Of search	***2***	***3***	***4***
Urls Selected	645	729	783

Urls Not selected	2684	9480	28835
Urls Not selected (size > 40 KB of text)	264	445	1178
Urls in Errors	436	1958	6733
Urls Visited	4029	12612	37529
Urls in errors of protocol (except protocol HTTP)	73	205	585
Time to search	21h 20'	44h 40'	149h 34'

Certainly we find here the different urls of server forget, but according to the value of the table we see that for a profit of 20% of Urls selected, we have a time of search which increase exponentially. It was thus necessary to find a new additional parameter in order to decrease the search time.

With the in-depth navigation Auresys allows to indicate if we want that the link present in a "bad" HTML file (HTML file which does not answer to the initial request) be used. This is the selection of hypertext links parameter.

VII. Virtual Library and Selection of the hypertext links parameter

The following table presents various values resulting from several searches

Depth Of search	***2***	***3***	***4***
Urls Selected	641	692	701
Urls Not selected	1961	2287	2361
Urls Not selected (size > 40 KB of text)	248	267	282
Urls in Errors	358	462	470
Urls Visited	3208	3708	3814
Urls in errors of protocol (except protocol HTTP)	58	52	61
Time to search	20h 12'	20h 54'	20h 53'

The profit of urls selected is less important but it's obvious that the saving of time is enormous and constant.

VIII. Conclusion

It is obvious that the time savings offered by the automatic virtual library is considerable. Moreover the administrator is relieved from searching the various Internet links. This method also has the big advantage of being rather exhaustive. Indeed, to manually make a virtual library with as many Internet links is almost impossible.

The only constraint with which the administrator is confronted is that he must start building the virtual library occasionally (monthly).

IX. Bibliography

Manning, G. (1999). **The WWW Virtual Library. Subject Catalog, May 1999** . <http://vlib.org/AlphaVL.html>. May, 1999.

Rousseau, R. (1997). **Sitations: an exploratory study**. **Cybermetrics**, 1 (1). Paper 1. <http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.htm>

Mannina, B.; Dou, H. & Florence, B. (1997). The internet and regional innovation : Auresys 2.0, The Provence/Côte d'Azur regional robot. IDT'97, Le salon de l'information électronique. Paris, 3-5 Juin 1997.

Smith, M. A. (1997). Measuring the social structure of the Usenet. In P. Kollock & M. A. Smith (Eds.). Communities in cyberspace. Berkeley, CA: University of California Press.

Paunak, H. (1989). Hypermedia Topologies and user navigation. Proceedings of HyperTexte'89 Conference, 43-50, Pittsburgh, Pa, Nov 1989.

Stein, R. M. (1991) Browsing through terabytes. **Byte**, State of Art, May.

Chen, H; Chung, Y. M.; Ramsey, M. & Yang, C. (1998). A smart itsy bitsy spider for the web. **Journal of the American Society for Information Science**, 49(7):604-618, 1998.

Dou, C.; Leitzelman, M.; Mannina, B. & Giraud, E. (1999). **Building gateway by intelligent agent**. **ISDM**, 3:25-34. <http://crrm.univ-mrs.fr/isdm/journal/1999/issue3.pdf>. February 1999

Small, H.G. (1973). 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.

Rostaing, H. (1993). Veille technologique et bibliométrique: concepts, outils, applications. Thèse: Aix-Marseille III, Janv, 353 p, 1993.

Rice, R.; Borgman, C. L. & Hart, P. J. (1989). Journal to Journal Citation data : issue of validity and reliability. **Scientometrics**, 15(3-4):257-282.

Dousset, B.; Dkaki, T.; Longevialle, C. (1993). Qualité de l'information et analyse des données. **Revue française de bibliométrie**,12:198-204

Notes:

- This paper is an elaborated version of a communication presented during Cybermetrics'99 Seminar
- The CRRM WebSite has changed : **http://crrm.u-3mrs.fr**
- Bruno Mannina created **Intelligence Process**, with CRRM students. <http://www.intelligence-process.com>
- **Luc Quoniam** is working now at IUT SRC -St Raphaël France <http://src.univ-tln.fr/~quoniam/>
- A search agent is now free and on-line: **SearchProcess** <http://www.searchprocess.com>

Received 9/July/1999
Accepted 6/April/2000
Revised 5/May/2000