J.T. Yao, V.V. Raghvan, G.Y. Wang (Eds.): WSS'04, pp. 163-170, 2004.

FIM-MetaIndexer: a Meta-Search Engine Purpose-Built for the French Civil Service and Statistical Classification of the Interrogated Search Engines.

Katarzyna Wegrzyn-Wolska
ESIGETEL, Ecole Supérieure d'Informatique et Génie de Télécommunication
77-200 Avon-Fontainebleau, France
katarzyna.wegrzyn@esigetel.fr

Abstract

Searching for specific information on the Web is becoming more and more difficult. To facilitate this task, it is necessary to use specialized tools. Our objective is to build a searching system for the French Civil Service. We describe here our work done within the framework of French gouvernmental contract by FIM[1] during the PhD thesis in Ecole des Mines de Paris (ESNMP). We survey the problems related to Web information retrieval and particularly meta-search techniques. We describe the implementation of a system to search for administrative documents. The goal is to retrieve the documents corresponding to a question simultaneously submitted to several Civil Service Web servers. We present our study, choice of methodology and the implementation of our meta-system. We describe some evaluation results of the experiments, performed to evaluate the relevance of the answers received from the search engines.

1. Introduction

This paper describes a Meta-Search system developed for searching the documents produced by the French Civil Service[2]. The main purpose of this system is to obtain the information directly from its source location, from the government website.

First, we describe the general problems of information retrieval on the Web, then we describe the FIM-MetaIndexer system in terms of architecture, information flow and configuration possibilities. We describe also some of our methods of analysis and classification of the different Search-Engines implemented on the government sites. The conclusion summarizes our results and provides suggestions for future improvements.

2. Searching for information on the Web

2.1. General problems of information retrieval on the Web

The Web is an enormous source of useful information. Surfing the Web can be a great pleasure and adventure, but searching for specific information can often be very difficult.

Most of the information retrieval problems are due to [3][7][12] the size of the Web, the quantity and quality of documents [2], reliability of links [5], index database updating, indexing of dynamic pages, the heterogeneity of the document formats and resources, problems related to multiple languages, etc.

The most popular searching methods are: subject catalogues, Search Engines, Meta-Search tools, on-line database agents, and other tools like robots (spiders and crawlers) and monitoring agents [8][4].

The Web-based information retrieval and searching techniques are necessary, the new systems like a WSS[3] and WIRSS[4] [15][16] are steel under development, and the new concepts are introduced.

2.2. Meta-Search Engines

The Meta-Search Engines [6][9][10] are tools, which transmit an individual client's query simultaneously to several different Search Engines (Figure 1).

Meta-Search Engines do not have their own database of Web pages (corpus of Web documents) but can only access the documents from the result pages given by search engines which have been queried.

[1] Fonds Interministériel de Modernisation de l'Administration Française

[2] FIM-MetaIndexer system developed for the *Fonds Interministeriel de Modernisation de l'Administration Française.*

[3] WSS Web-based Support Systems

[4] WIRSS Web-based Information Retrieval Support Systems

There are two kinds of Meta-Search Engines implementations either on the server or on the client.

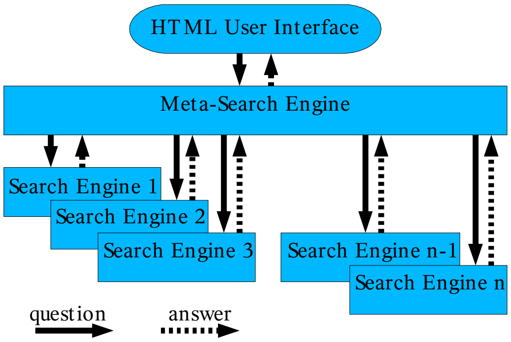


Figure 1. A general Meta-Search structure

2.2.1. Advantages and disadvantages of Meta-Search Engines. Some of the most significant advantages [11] are: simultaneous Search Engines interrogation, uniform request form, homogeneous response presentation, and the possibility of including additional functions such as: link verification, double page elimination, answer relevance evaluation etc.

The most significant disadvantages are: the growth of response time, the risk of elimination of some relevant answers, copyright problems, etc.

2.2.2. General Meta-Search Engines. The general Meta-Search Engines are neither dedicated nor adapted to information retrieval from the Civil Service Web sites. They consult general Google-like Search Engines, so it is not possible to find a particular Civil Service's documents. These documents are rarely indexed, if at all. It is difficult to retrieve the dynamic pages by autonomous retrieval agents.

2.3. Answer relevance checking

2.3.1. Checking relevance methods. There are two major types of relevance checking methods [14]; textual and topological. The topological method encompasses the textual, whilst taking into account hyperlinks.

2.3.2. Problems. Relevance checking of the answers given by the Search Engines involves a number of difficulties. The most serious problems are: different and nonhomogenized searching methods, heterogeneous relevance evaluation algorithms and different classification methods implemented on the Search Engines queried.

The Meta-Search method involves other problems: real time interrogation and checking slows down search speed (it is necessary to find a compromise between the response quality and the speed of the search system); the unknown total document corpus, which means that it is not possible to use standard retrieval performance checking methods. All the standard methods need fixed limits to the total document corpus. The Web corpus is unlimited and cannot be accessed by the Meta-Search Engine.

3. FIM-MetaIndexer

FIM-MetaIndexer is a Meta-Search system developed to search for documents produced by the French Civil Service[5]. The main purpose of this system is to retrieve the information directly from its source location, the government web site thus giving us the possibility to obtain immediately all of the existing and available information on the web site, as well as irretrievable documents in the search results of a standard Search Engine. These are documents not yet indexed because of the necessary delay time for the indexing task, or the documents named 'Invisible Web", which are not indexed at all by the standard Google-like Search Engine. These documents are dynamically generated as an answer to the query.

3.1. Methods choice and servers choice for queries

3.1.1. Method choice. Our Meta-Search Engine (FIM-MetaIndexer [13]) was installed on a distant accessible from Internet server. The server uses 50 independent agents to interrogate Search Engines simultaneously.

The technology used is based on the HTTP protocol. We used the GET, and the POST methods for page retrieval, and the HEAD method for page existence verification. The system was developed using the Perl programming language.

3.1.2. Choice of Search Engines. Our system interrogates three kinds of Search Engines: standard Google-like Search Engines[6]; more then 30 specialized French Civil Service Search Engines[7] (like the "French Senate" site Search Engine) and the Civil Service portals[8].

[5] FIM-MetaIndexer system developed for the *Fonds Interministeriel de Modernisation de l'Administration Française.*

[6] e.g. : Google (http://www.google.com), AltaVista (http://www.altavista.com), Yahoo.fr (http://www.yahoo.fr)

[7] e.g. : *Assemblée Nationale* (http://www.Assemblee-nationale.fr), *Sénat* (http://www.senat.fr), *Ministère de la Justice* (http://www.justice.gouv.fr), *Ministère de l'Education Nationale* (http://www.education.gouv.fr)

[8] *Adminet-Journal Officiel* (http://www.admi.net/jo), *Service-Publique* (http://www.service-publique.fr)

4. FIM-MetaIndexer architecture

FIM-MetaIndexer is based on modular architecture. There are three major modules (Figure 2):

- **MAD** (*Module d'Analyse des Données*) data input analyzer;
- **MIM** (*Module d'Interrogation des Moteurs*') Search Engine interrogation module;
- **MAR** (*Module d'Analyse des Réponses récupérées*) answer analyzer module.

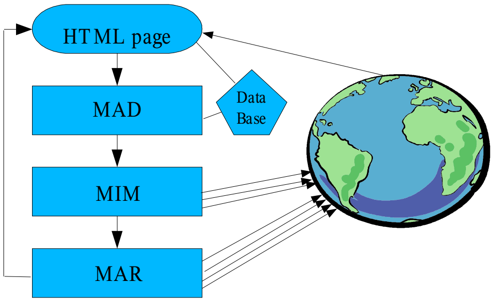


Figure 2. FIM-MetaIndexer architecture

4.1. MAD: Data Input Analyzer

The data input analyzer is the module in charge of comprehension and analysis of client queries and all selected parameters (for example : choice of engine for the interrogation, maximal research time, number of results to be recovered, mode for searching and checking the answers). These parameters are chosen by the user using the HTML based input interface. All necessary data for the interrogation are used to build the general request form which is transmitted to the Search Engine interrogation module (MIM).

4.2. MIM: Search Engine Interrogation Module

This is an agent-based interrogation module which works with selected Search Engines. This module is in charge of three important tasks: translation of client queries and the whole set of parameters into a form which is recognizable by the Search Engines, selection and interrogation of Search Engines, answer recovery and checking that the documents exist.

4.2.1. Query building. There are many different types of Search Engines. They use different search methods, they use and accept only their own specific request format. So, before interrogating the Search Engine, it is necessary to transform the general format of the query prepared by the MAD module into a format which is comprehensible to, and accepted by, each Search Engine. The request building algorithm is presented in Figure 3.

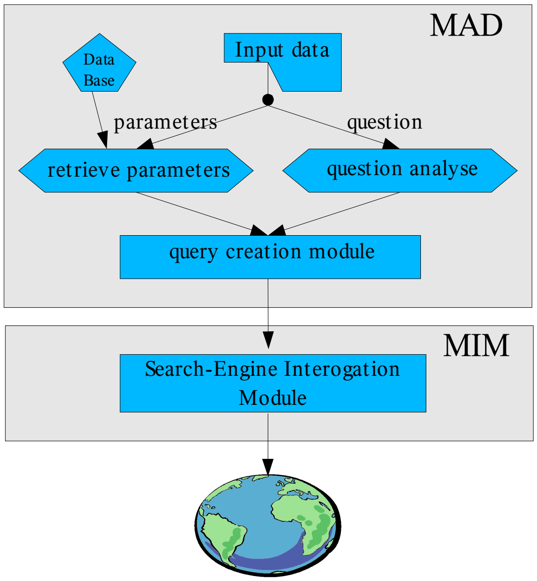
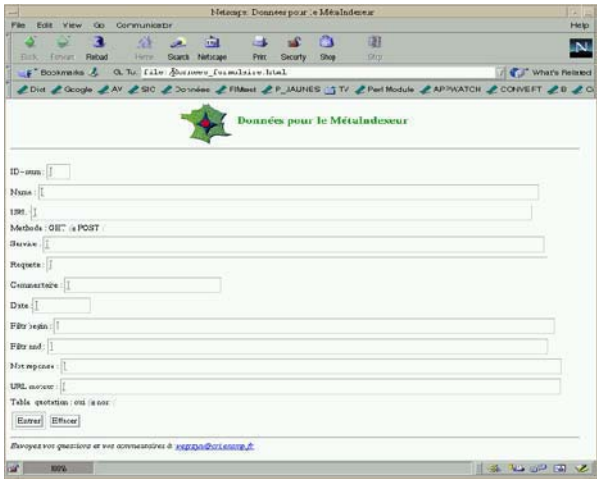


Figure 3. Request construction algorithm

In order to translate from one format to another FIM-MetaIndexer needs to know the configurations data concerning every Search Engine. This data is collected during the integration process of each Search Engine and saved in FIM-MetaIndexer data base.


Netscape: Données pour le MétaIndexeur
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Go To: file:/Donnees_formulaire.html
What's Related
Dict Google AV SIC Données FIMeet P_JAUNES TV Perl Module APPWATCH CONVERT

Données pour le MétaIndexeur

ID-num :

Name :

URL :

Methode : GET ○ POST ○

Service :

Requete :

Commentaire :

Date :

Filtr begin :

Filtr end :

Nbr reponses :

URL moteur :

Table quotation : oui ○ non ○

Entrer Effacer

Envoyez vos questions et vos commentaires à [illegible]

Figure 4. Input data form for new Search

Modification of the Search Engine description or insertion of a new Search Engine to the list is possible in two ways : directly in the data bases, or by filling in the special insertion HTML form (Figure 4).

4.2.2. Query translation. Query translation is carried out with special attention to : Boolean expression, accentuated letters, optional parameters and seeking the exact phrase.

The Boolean operation can be expressed in varied forms [1], so it is necessary to modify the expression given by the user according to Boolean functions and Boolean operators appropriate to each Search Engine. For unacceptable functions it is necessary to change Boolean expression according to the possibilities offered by a selected Search Engine. For example, if the Search Engine does not accept an alternative function, each component of the alternative function is sent separately to collect the results.

The treatment of the accentuated letters is also full of pitfalls. It is necessary to take note of and distinguish two kinds of treatment, one for the interrogation and the other for the analysis of the document answers.

In the request different options may be used which are proposed by FIM-MetaIndexer (searching mode, number of results, number of results on one page). This will only work with a Search Engine which uses the same, or very similar parameters, for search configuration.

4.2.3. Recovery and answer checking. FIM-MetaIndexer offers some configuration possibilities for searching methods: with or without checking the document-answers existence. The default option without verification is faster and requires minimal network occupation.

4.3. MAR: Answer analyzer module

This module checks the relevance of the answers according to the selected mode (verification on/off) and presents the results sorted by score. FIM-MetaIndexer sorts the results according to the relevance and prevalence of the same documents. FIM-MetaIndexer eliminates repeated answers. The more often the document is cited, the more relevant it is considered to be. This is referred to as weight of relevance and is increased by a specific coefficient. This coefficient is calculated dynamically for each multiple answer, according to the number of repetitions.

Relevance checking with the "verification on" mode is concerned with existence verification and calculation of the number of repetitions of every significant word used in the question.

In "verification off" mode FIM-MetaIndexer cannot itself evaluate or measure the relevance of the document-answers. It will retain the relevance score supplied by each questioned engine. This method of document sorting is illustrated in Figure 5.

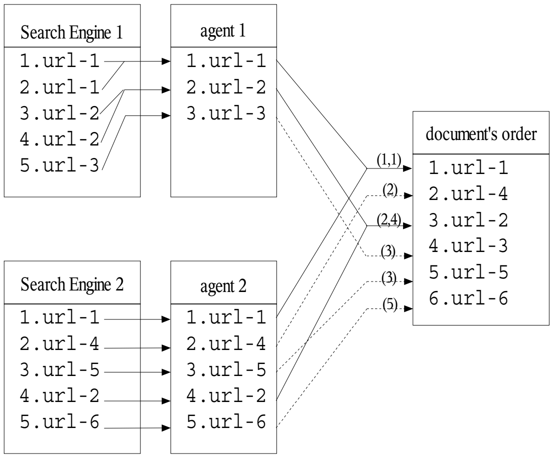
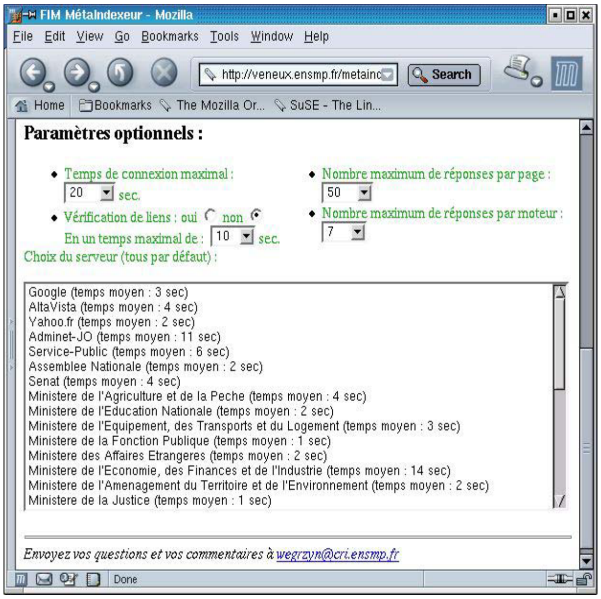


Figure 5. Results sorting algorithm

5. FIM-MetaIndexer user interface

5.1. Input data form

Our Meta-Search Engine gives a user-friendly interface with some configuration options (Figure 6).


FIM MétaIndexeur - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://veneux.ensmp.fr/metainc
Search
Home
Bookmarks
The Mozilla Or...
SuSE - The Lin...
Done

Paramètres optionnels :

- Temps de connexion maximal : 20 sec.
- Vérification de liens : oui ○ non ●
  En un temps maximal de : 10 sec.
- Nombre maximum de réponses par page : 50
- Nombre maximum de réponses par moteur : 7

Choix du serveur (tous par défaut) :

Google (temps moyen : 3 sec)
AltaVista (temps moyen : 4 sec)
Yahoo.fr (temps moyen : 2 sec)
Adminet-JO (temps moyen : 11 sec)
Service-Public (temps moyen : 6 sec)
Assemblee Nationale (temps moyen : 2 sec)
Senat (temps moyen : 4 sec)
Ministere de l'Agriculture et de la Peche (temps moyen : 4 sec)
Ministere de l'Education Nationale (temps moyen : 2 sec)
Ministere de l'Equipement, des Transports et du Logement (temps moyen : 3 sec)
Ministere de la Fonction Publique (temps moyen : 1 sec)
Ministere des Affaires Etrangeres (temps moyen : 2 sec)
Ministere de l'Economie, des Finances et de l'Industrie (temps moyen : 14 sec)
Ministere de l'Amenagement du Territoire et de l'Environnement (temps moyen : 2 sec)
Ministere de la Justice (temps moyen : 1 sec)

Envoyez vos questions et vos commentaires à wegrzyn@cri.ensmp.fr

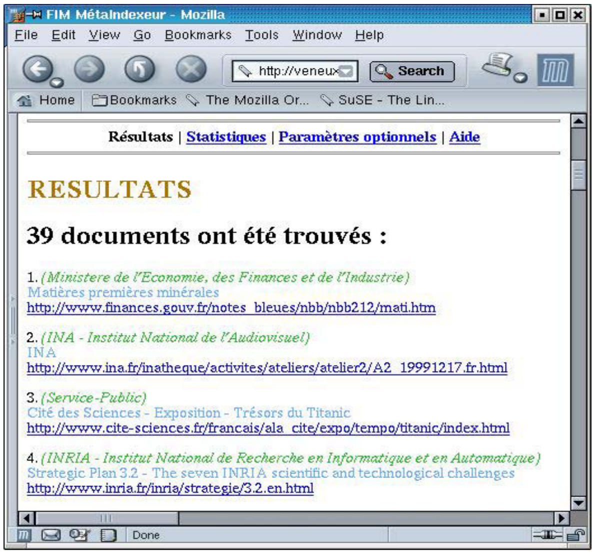
Figure 6. FIM optional configuration parameters page of presentation results

The user has the choice of selecting some optional parameters such as: connection time-out, checking existence of links (on/off), global searching time-out, choice from the list of servers to be interrogated, number of answers and maximum number of answers for each server interrogated.

5.2. The presentation of results

The results' page (Figure 7) presents the list of search results and some statistical estimates such as average: response time, minimal response time, the percentage of right and non relevant responses, and the percentage of interrogations without connection.

For each answer some more details such as: server name, document title and document addresses (URL) are presented.



Résultats | Statistiques | Paramètres optionnels | Aide

RESULTATS

39 documents ont été trouvés :

1. *(Ministere de l'Economie, des Finances et de l'Industrie)*
Matières premières minérales
http://www.finances.gouv.fr/notes_bleues/nbb/nbb212/mati.htm

2. *(INA - Institut National de l'Audiovisuel)*
INA
http://www.ina.fr/inatheque/activites/ateliers/atelier2/A2_19991217.fr.html

3. *(Service-Public)*
Cité des Sciences - Exposition - Trésors du Titanic
http://www.cite-sciences.fr/francais/ala_cite/expo/tempo/titanic/index.html

4. *(INRIA - Institut National de Recherche en Informatique et en Automatique)*
Strategic Plan 3.2 - The seven INRIA scientific and technological challenges
http://www.inria.fr/inria/strategie/3.2.en.html

Figure 7. FIM page of results

6. Analysis, statistics and classifications of Search Engines

We performed experiments to evaluate the answers received from the selected Search Engines and to classify the profile of the French Civil Service websites.

Our experiment method was based on statistical analyzis of the data obtained form the answers given by the Search Engines queried. All of the experiments were done using the FIM-MetaIndexer.

The next sections describe some statistical evaluation of our results.

6.1. Evaluation of the test sets

Two sets of standard queries were prepared. The first with 1000 queries and the second with 50 queries. The queries were selected from the log file of the Adminet[9] server, the popular French Civil Service Search Engine.

All query sets were carefully chosen to represent the true user question form. Tests proved that selected queries have a similar structure to that of the Adminet's questions, for example: a mean query length of 2.7 words; a mean word length of 7,5 characters. These values are common for the Web log statistics.

Two experiments were carried out with the results saved into the local FIM-MetaIndexer database: "TEST1" - all servers were queried using the first set of queries (1000 queries) and answers (hyper links) were saved locally; "TEST2" – all servers were queried using the second set of queries (50 queries), all the links were followed up locally. The results were then analyzed using statistical methods.

6.2. Response time statistics (TEST1*)*

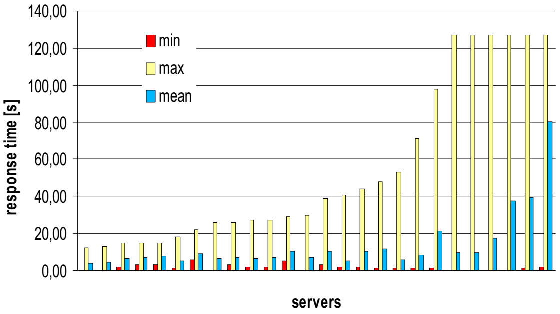


Figure 8. Response time

Different statistics concerning the response time were calculated. We present three of them: response time for each sever (Figure 8), the thematic classification of the Search Engines, and the correlation between the response time and the number of responses.

6.2.1. Thematically classifications. We tried to analyzed the thematic profile of the French Civil Service websites. With the specially prepared, thematically selected questions set, we analyzed the dependance between the theme of the question and the quantity of relevant answers to it. Then we analyzed the thematically correlation between the pairs of selected websites of French Civil Service. Figure 9 presents the graph of thematic dependance of the

[9] Adminet (http://www.adminet.fr)

analyzed Web sites with the different correlation value (correlation ≥ 0.5).

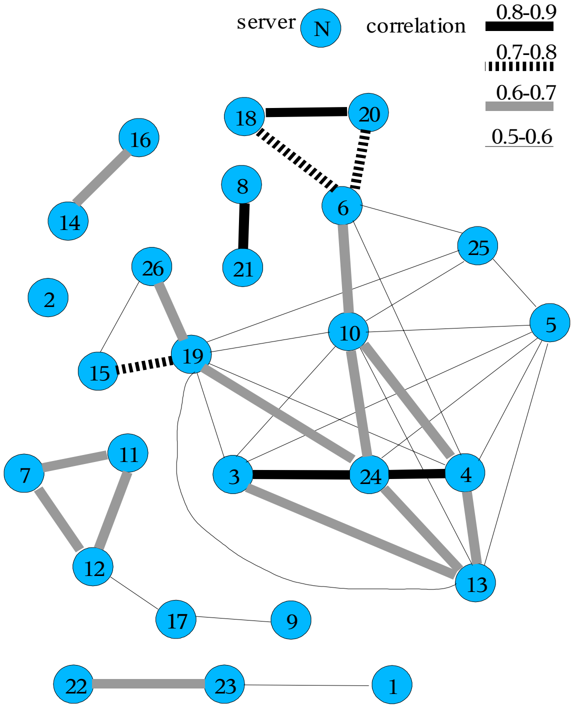


Figure 9. Thematic profile of governmental websites : dependance graph

We observed an large number of the pairs of servers grouped thematically with significant correlation (Figure 10).

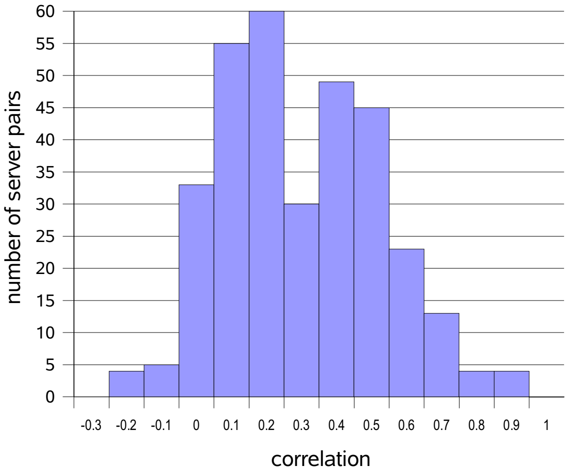


Figure 10. Correlation for the different pair of the websites

6.2.2. Correlation between the response time and the number of responses. The high positive correlation confirms the natural dependance between the measured values. We observed for some servers the high negative correlation. The negative correlation (-0.65;-0.4) was observed on servers using the cache mechanism that explains very fast access for frequently used sets of answers (Figure 11).

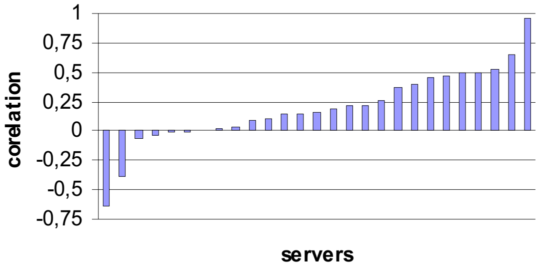


Figure 11. Correlation between the response time and the number of response

6.3. Test of the response-document accessibility (TEST2)

We verified response-document accessibility, size, and relevance. Our relevance-checking algorithm was based on verification of request-word existence in the answer-documents. We checked it in three-separate parts of the documents: in the title; the body and between the "meta" tags.

This test proved that about 93% answers arrived satisfactorily (Returns HTTP Code 200 OK). There was a small percentage of bad answers due to the incorrect question formulation (Returns HTTP from Code 4XX family): 4% Code 404; 3% Code 400 and Code 403 together. About 0.5% answers were wrong due to the Internal Server Error (Code 500).

7. FIM-MetaIndexer Meta-Search system statistics

Internet users used the FIM-MetaIndexer and the log was collected over several months. We used this log to calculate some statistics.

7.1. General statistics

Many answers returned by the questioned Search Engines were not pertinent. Our statistics (Figure 12) show that only 34% of the results returned by the Search Engines interrogated were of interest.

The FIM-MetaIndexer creates its own results-page from these answers. Our analysis shows that usage of the FIM-MetaIndexer decreases "information noise" significantly because it is able to eliminate the majority of the non-relevant answers.

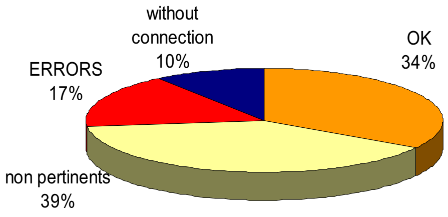


Figure 12. Response classification

7.2. Number of answers in a fixed period of time

These tests concerned the number of answers received during different time periods. The graph presented (Figure 13) shows that the majority of answers were provided within a period of time of under 15 seconds.

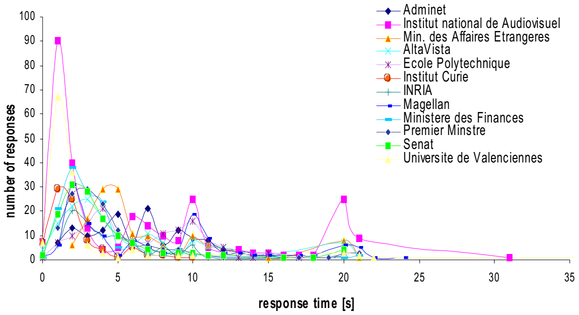


Figure 13. Number of responses received in the different time periods

We observed a significant growth in numbers of answers for the response time equal to 20 seconds. This can be explained by the fact that the default value of the "Search Server wait" time-out of FIM-MetaIndexer is equal to 20 seconds. Users rarely modified this parameter and the majority of responses slower than this time-out were stopped.

7.3. Usage of the FIM-MetaIndexer

The FIM-MetaIndexer is particularly queried during working hours (in France), with a strong increase in peak activity to about 20% in the beginning of the afternoon (Figure 14).

FIM-MetaIndexer usage decreases during the night. This would seems to be normal, as the FIM-MetaIndexer is purpose-built for the French Civil Service, and is therefore mainly used by the French community.

The statistics (Figure 14) show the user activity (the number of requests in percentage) and the response time of the FIM-MetaIndexer according to the time of the day. The FIM-MetaIndexer response time is more or less constant and equal to about 5s.

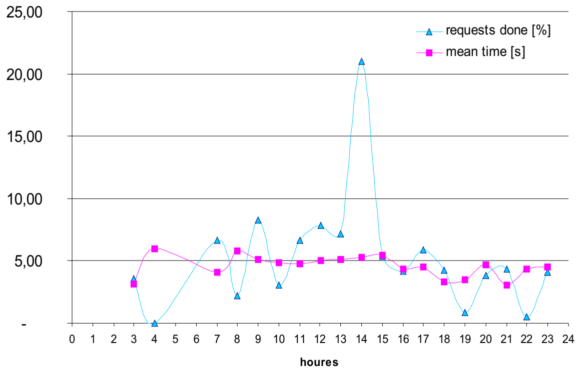


Figure 14. FIM-MetaIndexer response time

8. Conclusion and future work

The FIM-MetaIndexer is a Meta-Search Engine used to search for documents produced by the French Civil Service. Its first version was available in February 1998. Since, FIM-MetaIndexer has been used by various kinds of users and it has been a well-known and effectively used search tool.

Some new functions could be included in the FIM-MetaIndexer like a bi-directional co-operation module for Search Engines and an economic and technological survey module.

9. References

[1] A. H. Alsaffar, J. S. Deogun, V. V. Raghavan, and H. Sever. Concept-based retrieval with minimal term sets. In Z. W. Ras and A. Skowron, editors, *Foundations of Intelligent Systems: Eleventh Int'l Symposium, ISMIS'99 proceedings*, Warsaw, Poland, Jun. 1999, pp. 114--122.

[2] T. Bray,. Measuring the web: *Proceeding of Fifth International World Wide Web Conference,*1999.

[3] L. Chen,K. Sycara,. Webmate, A personal agent for browsing and searching: *Second International Conference on Autonomous Agents*, 1998, pp. 132-139. ACM SIGART, ACM Press.

[4] D. Green, The evolution of web searching: *Online Information Review,* 24(2), 2000, pp. 124-137.

[5] M. Henzinger, A. Heydon, and M. Najork Measuring index quality using random walks on the Web: *Proceeding of the _the International Word Wide Web Conference*, 1999, pp. 213-225.

[6] S. Lawrence, and G. Lee, Inquirus, the Neci meta search engine In: *Computer Networks and ISDN System.,* 1995, v30, pp. 1-7.

[7] S. Lawrence and G. Lee, Searching the world wide web: *Science, 280(5360),* 1998, pp. 38-6.

[8] A. Nicholson, A proposal for categorisation and nomenclature for web search tools: *Journal of Internet Cataloging*, 2000, 2(3/4), pp. 9-28.

[9] A. Sainul, Meta search engines: effective tool for information retrieval: *6th National Convention for Automation of Libraries in Education and Research (CALIBER 99),* Nagpur, India, 1999*,* pp. 362-369.

[10] E. Selberg and O. Etzioni, Multi-service Search and Comparison Using the MetaCrawler: *Proceeding of Fourth World Wide Web Conference*, Boston, 1995, MA.

[11] T. Stanley, Meta-search engines: where are the limits? In: *Proceeding of the Second International Online Information Meeting*, London, 1999, pp. 297-300.

[12] K. Wegrzyn, Etude et réalisation d'un robot pour la recherche d'information sur le Web: *Rapport DEA E/193/CRI*, Ecole des Mines de Paris and Université d'Evry-Val-d'Essone, 1996.

[13] K. Wegrzyn, Etude et réalisation d'un meta-indexeur pour la recherche sur le Web de documents produits par l'administration française: *Thesis A/339/CRI*, Ecole des Mines de Paris, 2001

[14] A. Weiss, The evolution of world wide web search tools: *Proceedings of the Second International Online Information Meeting*, London, 1998, pp. 289-295.

[15] J. T. Yao, Y.Y., Yao, Web-base Support Systems: *Proceedings of the Workshop on Applications, Products and Services of Web-based Support Systems (WSS'03),*Halifax, Canada,Oct 13, 2003, pp. 1-5.

[16] J. T. Yao, Y.Y., Yao, Web-base Information Retrieval Support Systems: building research tools for scientists in the new information age, *Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)*, Halifax, Canada, Oct 13-17, 2003, pp. 570-573.