ECOLE DES MINES
DE PARIS

Collège doctoral

N° attribué par la bibliothèque

|__|__|__|__|__|__|__|__|__|__|

THESE

pour obtenir le grade de

Docteur de l'Ecole des Mines de Paris

Spécialité «Informatique Temps Réel, Robotique et Automatique»

présentée et soutenue publiquement

par

Katarzyna Maria WEGRZYN-WOLSKA

le 18 décembre 2001

ETUDE ET REALISATION D'UN META-INDEXEUR POUR LA RECHERCHE SUR LE WEB DE DOCUMENTS PRODUITS PAR L'ADMINISTRATION FRANÇAISE

Directeur de thèse : Robert MAHL

Jury

M.Richard Bouché	Rapporteur
M.Jean-Jacques Girardot	Rapporteur
M.Christian Scherer	Examinateur
M.Robert Mahl	Examinateur
M.François Irigoin	Examinateur

Remerciements,

En premier lieu, je remercie M. Richard Bouché et M. Jean-Jacques Girardot d'avoir accepté d'être rapporteurs de ma thèse.

Je remercie également M. Christian Scherer pour sa participation au jury.

Je tiens tout particulièrement à remercier toutes les personnes qui, par leur aide ou leurs encouragements, m'ont permis de réaliser cette thèse :

- *M. Robert Mahl, directeur du Centre de Recherches en Informatique de l'École des Mines de Paris, directeur de ma thèse, qui m'a incité à effectuer cette recherche, m'a permis de la réaliser et l'a encadrée,*
- *M. François Irigoin pour sa participation au jury,*
- *tous les membres du CRI pour m'avoir accueilli, et en particulier :*
  - *Mmes Corinne Ancourt, Marie-Thérèse Lesage, Claire Medrala et Nadine Olivier pour leurs commentaires et leurs contributions à la correction de mon rapport,*
  - *M. Laurent Daverio pour sa disponibilité et tous ses conseils au cours de ma thèse,*
  - *Mme Jacqueline Altimira et M. Jean-Marie Kessis pour le soutien logistique qu'ils m'ont apporté,*

et à Jan ...

Table des matières

Table des matières 5

Résumé 15

Introduction 17

I Le contexte et l'existant 19

1 La recherche d'informations 21

- 1.1 La recherche d'informations en général 21
  - 1.1.1 Modèle booléen 22
  - 1.1.2 Modèle vectoriel 23
  - 1.1.3 Modèle probabiliste 24
  - 1.1.4 Modèles alternatifs 26
- 1.2 Indexation du document 26
  - 1.2.1 Le principe de l'indexation 26
  - 1.2.2 La structure de l'index 27
    - 1.2.2.1 Les fichiers directs 28
    - 1.2.2.2 Les fichiers inversés 28
  - 1.2.3 Étapes du processus d'indexation 29
    - 1.2.3.1 Découpage - extraction des mots du document . . . 30

1.2.3.2 Normalisation des mots 30

1.2.3.3 L'élimination des mots vides 31

1.2.4 Mots-clés, descripteurs et thésaurus 32

1.2.4.1 Les relations du thésaurus 33

1.3 L'espace documentaire 34

1.3.1 Différentes représentations des documents 34

1.3.1.1 Représentation vectorielle 35

1.3.1.2 Représentation par n-grammes 35

1.3.1.3 Représentation dynamique 36

1.3.2 Différentes méthodes de mesure de la pertinence 37

1.3.2.1 Méthodes textuelles 39

1.3.2.2 Méthodes des hyperliens 41

2 L'univers du Web 45

2.1 Histoire d'Internet et du Web 45

2.1.1 Histoire d'Internet 45

2.1.2 Histoire du Web 46

2.2 Les protocoles et standards utilisés 46

2.3 Les documents sur le Web 47

2.3.1 La structure des documents HTML 48

2.3.2 Localisation des documents 50

2.3.3 Fiabilité et accessibilité des liens 50

2.3.3.1 Fiabilité des liens 50

2.3.3.2 Accessibilité des liens 50

2.3.4 Quantité et qualité des documents 51

2.3.5 L'hétérogénéité 54

2.3.5.1 L'hétérogénéité des formats de documents 54

2.3.5.2 L'hétérogénéité des ressources 54
2.3.6 Le problème du multilinguisme 57
2.3.6.1 Différentes langues de documents 57
2.3.6.2 Différents jeux de caractères 57
2.3.6.3 Les projets liés au multilinguisme 58
2.4 Performances et détails de réponses sur Internet 59
2.4.1 Performances du réseau 59
2.4.2 Temps de réponse du serveur 59

3 La recherche des documents sur le Web 61

3.1 Différents outils de recherche 62
3.1.1 Les catalogues par sujets 62
3.1.2 Les moteurs de recherche 63
3.1.3 Méta-moteurs de recherche 64
3.1.3.1 Le méta-moteur sur un serveur distant. 66
3.1.3.2 Le méta-moteur client 68
3.1.4 Les bases de données avec accès *on-line* 70
3.1.5 Les portails 71
3.1.6 D'autres outils pour la recherche 71
3.1.6.1 Les robots pour rapatrier des pages sur un site : aspirateurs, *spiders* et *crawlers* 71
3.1.6.2 Les agents de surveillance 72
3.2 Modèle de recherche utilisé par les moteurs 72
3.2.1 Les différentes méthodes de recherche 72
3.2.1.1 La recherche booléenne 73
3.2.1.2 Les systèmes vectoriels 74
3.2.1.3 D'autres systèmes pondérés 75

3.2.2 Les différents modes de création de base d'index par les moteurs de recherche . . . 75
  3.2.2.1 Les moteurs basés sur les robots . . . 75
  3.2.2.2 Les moteurs basés sur les signets . . . 75
  3.2.2.3 Les moteurs basés sur la topologie des liens . . . 76
3.2.3 Problèmes liés à la récupération et à l'indexation des documents par les moteurs de recherche . . . 76
  3.2.3.1 Problèmes liés à la récupération des documents . . . 76
  3.2.3.2 La validité des liens et la mise à jour des bases d'index 79
  3.2.3.3 Les pages dynamiques accessibles à partir de formulaires . . . 80
  3.2.3.4 Le moteur de recherche face au problème du multilinguisme . . . 80
  3.2.3.5 Le moteur de recherche face à différents jeux de caractères . . . 81
3.3 Les étapes de la recherche . . . 81
3.3.1 Formulation et analyse des questions posées . . . 81
  3.3.1.1 Opérateurs booléens dans la requête . . . 81
  3.3.1.2 Troncature de mot . . . 82
  3.3.1.3 Les opérateurs de proximité . . . 82
  3.3.1.4 Syntagmes nominaux (phrases) . . . 83
3.3.2 Récupération et vérification (évaluation) des réponses obtenues 83
3.3.3 Présentation des résultats de la recherche . . . 84

II Implémentation d'un méta-moteur FIM Métaindexeur 87

4 Modèle de méta-moteur 89

4.1 Le principe . . . 89

4.2 Les avantages 89
  4.2.1 Interrogation simultanée de plusieurs moteurs 90
  4.2.2 Uniformisation des requêtes 90
  4.2.3 Présentation homogène des réponses 91
    4.2.3.1 Fonctions supplémentaires 91
    4.2.3.2 Souplesse d'adaptation des méta-moteurs 92
4.3 Les inconvénients des méta-moteurs 93
4.4 Les modules principaux 94
  4.4.1 Module d'analyse des données en entrée 94
    4.4.1.1 Analyse des données de paramétrage 95
    4.4.1.2 Analyse de la question posée par l'utilisateur 95
    4.4.1.3 Comparaison avec les questions déjà posées 96
    4.4.1.4 L'aide à la formulation de la question 98
  4.4.2 Module d'interrogation des moteurs 98
    4.4.2.1 Méthodes d'interrogation 98
    4.4.2.2 Choix des moteurs pour l'interrogation 99
    4.4.2.3 Transformation d'une question de l'utilisateur vers le format utilisé par chaque moteur interrogé 99
    4.4.2.4 Récupération des pages de réponses 100
    4.4.2.5 Vérification des liens vers les documents-réponses . . 101
    4.4.2.6 Récupération des documents-réponses 102
  4.4.3 Module d'analyse des réponses récupérées 102
    4.4.3.1 Vérification de la pertinence 103
    4.4.3.2 Présentation des résultats 104
  4.4.4 Coopération entre le méta-moteur et les moteurs de recherche interrogés 105
    4.4.4.1 L'échange d'informations 105
    4.4.4.2 Standard Formel d'utilisation des moteurs par le méta-moteur 106

5 Analyse de moteurs à interroger 109

5.1 Les moteurs interrogés 109

5.2 Mode de recherche 111

5.2.1 Type de moteur interne utilisé 111

5.3 Construction de la requête 111

5.3.1 Méthodes du protocole HTTP utilisées pour l'interrogation des moteurs 111

5.3.2 Traduction des questions 114

5.3.2.1 Recherches booléennes 114

5.3.2.2 Recherche des phrases (guillemets - simples et doubles) 115

5.3.2.3 Traduction des lettres accentuées 119

5.3.3 Paramètres optionnels inclus dans la requête 119

5.4 Mode d'affichage des réponses 123

5.4.1 Analyse de la topologie des pages de réponses et filtrage des URL 123

5.4.1.1 Différents modèles des pages de réponses 123

5.4.1.2 Filtrage d'URL 124

5.4.2 Page unique ou plusieurs pages de réponses 127

6 Réalisation d'un méta-moteur 129

6.1 Architecture 129

6.1.1 Module d'entrée des données 129

6.1.1.1 Analyse des données de paramétrage 129

6.1.1.2 Analyse de la question posée 132

6.1.2 Module d'interrogation des moteurs 132

6.1.2.1 Création d'une requête 132

6.1.2.2 Récupération de réponses 134

6.1.3 Module d'analyse des réponses récupérées 134

6.2 Présentation de FIM-Métaindexeur 134
  6.2.1 Données pour les FIM-Métaindexeur 134
  6.2.2 Formulation d'une question par l'utilisateur 137
  6.2.3 Options de la recherche 137
    6.2.3.1 Temps de connexion maximal par moteur 140
    6.2.3.2 Vérification des réponses 140
    6.2.3.3 Nombre maximum de réponses 141
    6.2.3.4 Nombre maximum de réponses par moteur 141
    6.2.3.5 Choix du serveur à interroger 142
  6.2.4 Affichage des résultats 142
    6.2.4.1 Tri des réponses 142
    6.2.4.2 La page de réponses 145
  6.2.5 La page d'aide 148
6.3 Statistiques effectuées à partir du fichier LOG 148
  6.3.1 Statistiques globales 150
  6.3.2 Statistiques sur les réponses de chaque moteur 151
  6.3.3 Le temps de réponse 151
    6.3.3.1 Nombre de réponses en fonction du temps de réponse 151
    6.3.3.2 L'estimation de différents temps de réponses 152
  6.3.4 L'heure de l'interrogation 153

III Aspects statistiques **155**

7 Statistiques et classification de moteurs **157**

7.1 La méthode appliquée 157
  7.1.1 Les statistiques effectuées 157
    7.1.1.1 Analyse des pages de réponses 159

7.1.1.2 Analyse des pages de réponses et des documents-réponses appropriés 159

7.1.2 Les requêtes de teste 160

7.1.2.1 Analyse de la structure des requêtes posées 160

7.1.2.2 Les fichiers-test contenant les requêtes pour les statistiques 161

7.1.3 La base de données 164

7.1.3.1 Les données pour les statistiques concernant les résultats de recherche (pages de réponses) 164

7.1.3.2 Les statistiques concernant les documents-réponses (vérification de réponses) 166

7.2 Statistiques effectuées sur des questions-test (TEST1) 168

7.2.1 Nombre des réponses récupérées 168

7.2.2 Temps de réponse 169

7.2.3 Temps de réponse en fonction du nombre des réponses récupérées 174

7.3 Profil thématique des serveurs 176

7.3.1 Méthode 176

7.3.2 Corrélation globale 176

7.3.3 Corrélations particulières 186

7.4 Statistiques effectuées (TEST2) 190

7.4.1 Accessibilité des documents-réponses (Codes HTTP) 190

7.4.2 Taille des documents-réponses 191

7.4.2.1 Statistique générale sur la taille 191

7.4.2.2 Taille des documents-réponses pour chaque moteur . 191

7.4.3 Pourcentage de documents existants 194

7.4.4 Existence de mots de la requête dans les documents-réponses 199

IV Conclusion et glossaire **203**

8 Conclusion **205**

8.1 Bilan 205

8.2 Perspectives 209

8.2.1 Coopération du méta-moteur avec l'outil interrogé 209

8.2.2 Veille spécialisée dans des domaines spécifiques 210

8.2.3 Autres perspectives en matière de recherche-développement futurs 210

9 Glossaire **211**

Bibliographie **225**

Table des figures **237**

Liste des tableaux **239**

Annexe A : Les technologies utilisées **245**

Annexe B : Les statistiques sur la taille des documents-réponses **249**

Résumé

Avec la croissance de la taille du Web, chercher une information spécifique devient de plus en plus difficile. Il est nécessaire de proposer des outils spécialisés afin de faciliter cette recherche. Notre objectif est de construire un système de recherche des documents de l'administration française.

Nous analysons les problèmes liés à la recherche des informations sur le Web et en particulier la technique de méta-recherche. Puis nous décrivons la réalisation d'un outil ayant pour but de retrouver les documents administratifs grâce à cette technique. L'objectif est de trouver des documents à partir d'une même question posée simultanément à plusieurs moteurs de recherche de différents serveurs de l'administration, ainsi qu'à des moteurs de recherche généraux.

Nous présentons d'abord une étude bibliographique, les critères de recherche et le choix de la méthode. Nous étudions ensuite l'implémentation du méta-moteur réalisée dans le cadre de notre recherche.

Des expériences ont été réalisées sur l'évaluation de la pertinence des réponses reçues de la part des moteurs de recherche. L'analyse, étant effectuée *on-line,* est assez rapide. En raisons de l'hétérogénéité des bases de données, des différentes méthodes d'indexation des documents, ainsi que des différents algorithmes d'analyse de la pertinence, l'évaluation et la synthèse de l'ensemble des réponses fournies par les moteurs pose des problèmes particuliers pour la vérification et la classification des documents-réponses.

Un certain nombre de solutions à ces problèmes sont fournies dans le cadre de cette thèse, qui en outre s'est intéressée à la classification des moteurs par des voies statistiques.

Introduction

La taille du Web augmente de plus en plus rapidement. En même temps, la recherche de l'information sur le Web est facilitée par un grand nombre de moteurs de recherche.

Cependant, avec la croissance du nombre de serveurs, la recherche d'informations devient une tâche de plus en plus lourde. Souvent, dans un ensemble de réponses fournies par des moteurs de recherche généraux, beaucoup ne sont pas pertinentes.

Pour un domaine de recherche bien connu comme le droit ou les services administratifs, la recherche d'une information spécifique sur les serveurs spécialisés du domaine donne des résultats plus sûrs, mais c'est aussi une tâche complexe. Les difficultés ne proviennent pas seulement du fait qu'il faut connaître ces serveurs pour effectuer la recherche, mais aussi de la lenteur de l'interrogation consécutive de plusieurs serveurs. C'est pourquoi l'interrogation en parallèle de plusieurs moteurs du domaine grâce à un méta-moteur présente un intérêt évident.

Cette recherche a pu être menée grâce à une aide du Fonds Interministériel de Modernisation de l'administration française (FIM) accordée à l'École des Mines de Paris (Centre de Recherche en Informatique).

Première partie

Le contexte et l'existant

Chapitre 1

La recherche d'informations

1.1 La recherche d'informations en général

[And98, Coo68, FPC98, HH97, JRP00, KT00, Kor97, Los97, vR79, TS98, SM83, BYRN99]

La recherche d'informations est un processus qui réunit plusieurs tâches et types d'approches.

Elle concerne la représentation, l'organisation et la récupération de différentes formes d'informations comme les documents textes, les dépêches de presse, les images, les sons, la vidéo, etc. Dans ce travail, on traitera seulement les problèmes liés aux informations textuelles.

Le problème général de la recherche documentaire est de retrouver et en même temps de décider quel document parmi tout l'ensemble des documents est pertinent et quel document ne l'est pas. Cette décision dépend de différents facteurs tels que :

- la méthode de présentation des documents,
- les algorithmes de recherche,
- les algorithmes de vérification de pertinence choisis,
- la méthode de recherche utilisée par l'utilisateur.

Le calcul de la pertinence d'un document par rapport à une requête donnée existe sous plusieurs variantes. Il existe donc beaucoup de modèles de recherche documentaire liés aux choix des méthodes de recherche et des algorithmes utilisés. La classification la plus générale, en fonction des algorithmes choisis, permet de classifier la recherche documentaire en trois familles de modèles principaux :

- modèle booléen,
- modèle vectoriel,
- modèle probabiliste.

Avant de détailler plus chacun des modèles cités ci-dessus, on va définir le modèle de recherche d'information $<<$RI$>>$ le plus général.

Définition : Le modèle de $<<$RI$>>$ est un ensemble $[D,\ Q,\ \varphi,\ \mathrm{R}\ (q_i,d_j)]$ où :

- D est la représentation des documents dans la collection (ensembles globaux),
- Q est la représentation de requêtes q_i, les informations demandées par l'utilisateur,
- φ est la fonction de relation entre les représentations des documents et des requêtes,
- $\mathrm{R}(\mathrm{q}_i,\mathrm{d}_j)$ est le score de classification des réponses $d_j \in D$ en fonction des requêtes posées $\mathrm{q}_i \in Q$.

Pour développer un modèle de récupération, le plus important est la bonne représentation de la collection des documents, et la bonne définition des besoins de l'utilisateur. Avec ces deux données on peut construire les fonctions de récupération et de classification. Dans chacun des trois modèles décrits ci-dessous ces fonctions sont différentes.

1.1.1 Modèle booléen

Le modèle booléen est un modèle très simple, fondé sur le principe de l'algèbre booléenne. Les questions de l'utilisateur sont exprimées sous forme d'expressions booléennes avec les fonctions OU, ET et SAUF.

Dans ce modèle on examine si le terme de requête est présent ou non dans les documents.

La classification du document en fonction de la question posée est une fonction à valeur binaire.

Définition : Dans le modèle booléen la requête q est une expression Booléenne. La fonction de classification des documents répondant à la question posée est définie comme suit :

$R(q_i,d_j) = 0 \text{ ou } 1$

Si $R(q_i,d_j) = 1$ le document d_j est considéré comme pertinent pour la question q_i. Dans le cas contraire le document d_j est non pertinent.

Désavantages de ce modèle :

- La recherche est basée sur le principe de décision binaire caractéristique pour les fonctions booléennes. Un document dans la collection peut donc être identifié seulement comme pertinent ou non-pertinent. Il n'y a pas de moyen d'élargir l'échelle de classification, ce qui implique que c'est un modèle beaucoup plus adapté pour la recherche de données concrètes (système de réservation de billets, etc.) que d'information textuelle.
- L'utilisateur rencontre une certaine difficulté à formuler ses besoins sous forme d'expressions booléennes.

Le modèle booléen a été adopté très largement dans le passé, mais il commence maintenant à être remplacé par des modèles différents.

1.1.2 Modèle vectoriel

[SM83, BYRN99] Dans le cas du modèle vectoriel la pertinence d'un document est déterminée par le calcul d'une fonction de similarité entre le vecteur $\overrightarrow{q}$ représentant la requête et les vecteurs $\overrightarrow{d_j}$ représentant les documents. Cette fonction de classification qui relie le document d$_j$ et la question posée q (fonction de similarité) prend en compte le poids des termes dans la requête et le document.

La valeur de ce poids est utilisée pour déterminer la fonction de similarité entre les documents dans l'ensemble des documents-réponses à la question posée.

Définition Dans le modèle vectoriel, pour la requête q, le document ($d_j \in D$) et le terme indexé k$_i$, le poids $w_{i,j}$ d'un paire (k_i,d$_j$) n'est pas binaire et il est toujours positif. Également, pour la requête les poids des termes sont calculés.

Définition Si pour la paire $[k_i,\ q]$, ou le pois $w_{i,q}$ est toujours $\geq$0, on définit le vecteur de requête $\overrightarrow{q}$ comme :

$\overrightarrow{q} = (w_{1,q},\ w_{2,q},\ldots.w_{t,q})$ où t est le nombre de tous les termes indexés dans le système. Le vecteur des documents $\overrightarrow{d}_j$ est présenté comme : $\overrightarrow{d_j} = (w_{1,j},\ w_{2,j},\ldots.w_{t,j})$.

Dans cette représentation, le document d_j et la requête q sont présentés dans un espace à t dimensions où «t» est le nombre de tous les termes indexés dans le système.

La correspondance entre la question posée et le document en cours de vérification est la corrélation entre les deux vecteurs, le vecteur $\overrightarrow{q}$ et le vecteur $\overrightarrow{d_j}$. Cette corrélation est présentée comme la distance entre ces deux vecteurs et dans le cas le plus fréquent la mesure de similarité est égale au cosinus de l'angle entre les vecteurs $\overrightarrow{q}$ et $\overrightarrow{d_j}$.

Définition La fonction de proximité (similarité) entre la question posée q et le document d_j dans un modèle vectoriel est représentée par la norme du vecteur dans l'espace à t dimensions :

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \bullet \overrightarrow{q}}{|\overrightarrow{d_j}| \times |\overrightarrow{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

où :

$|\overrightarrow{d_j}|$ est la norme du vecteur $\overrightarrow{d_j}$

$|\overrightarrow{q}|$ est la norme du vecteur $\overrightarrow{q}$.

La fonction de classification des documents (proximité entre la question posée et le document) est donc une fonction continue qui prend des valeurs comprises entre 0 et 1.

Dans le modèle vectoriel il est donc possible de classer un document comme pertinent ou non, non seulement sur une échelle à deux valeurs (modèle booléen), mais aussi selon la formule de proximité vectorielle.

Les avantages du modèle vectoriel :

- il facilite la classification automatique,
- il permet d'obtenir des classements hiérarchiques de documents par rapport à un critère de ressemblance,
- il permet d'établir un ordre de pertinence sur l'ensemble des documents retrouvés.

Par contre un problème majeur de ce modèle est le fait que ses performances sont très dépendantes de la topographie de l'espace créé lors de l'indexation. Ce modèle ne prend pas non plus en compte l'aspect subjectif qui intervient dans le jugement de la pertinence d'un document par l'utilisateur.

1.1.3 Modèle probabiliste

[DW96, HH97, KT00, LG98b, LG98c, LG99, Los00, Los97]

Dans ce modèle, à la différence des modèles précédents, pour classifier un document comme pertinent, il ne s'agit pas de trouver sa pertinence par rapport à question posée, mais d'estimer la probabilité que ce document soit un document pertinent.

Différentes approches sont possibles. Les plus souvent on calcule le poids de chaque terme à partir de sa fréquence d'apparition dans les documents pertinents et dans les documents non pertinents. Après ce calcul sur tout le corpus, on peut décider que lorsqu'un terme est présent à la fois dans une requête et dans un document, il y a une certaine probabilité que le document soit pertinent.

Dans ce modèle on suppose que la probabilité de pertinence (pour la question posée q et le document d_j) dépend seulement de la question posée et de ce document.

On présume aussi que dans la collection de documents il existe un ensemble de documents pertinents R.

La collection des documents est donc constituée de deux ensembles disjoints :

R- ensemble des documents pertinents

$\overline{R}$- ensemble des documents non pertinents.

Définition Dans le modèle probabiliste, l'ensemble des poids des termes w$_{i,j}$ est binaire : w$_{i,j}\{0,1\}$.

La question q contient les mots de l'ensemble de mots indexés, donc elle constitue le sous-ensemble de tous les termes indexés.

L'ensemble R est un ensemble connu (estimé ou deviné au départ), qui contient les documents pertinents.

Alors P($R/\overrightarrow{d_j}$) est la probabilité que le document d_j soit pertinent pour la question q, et P($\overline{R}/\overrightarrow{d_j}$) est la probabilité que le document d_j soit non pertinent. La fonction de proximité (similarité) entre la question posée et le document d_j dans un modèle probabiliste est :

$$sim(d_j, q) = \frac{P(R|\overrightarrow{d_j})}{P(\overline{R}|\overrightarrow{dj})}$$

En utilisant les règles de Bayes :

$$sim(d_j, q) = \frac{P(\overrightarrow{d_j}|R)\times P(R)}{P(\overrightarrow{d_j}|\overline{R})\times P(\overline{R})}$$

La probabilité $P(\overrightarrow{d_j}|R)$ est la probabilité de sélectionner aléatoirement le document d_j à partir de l'ensemble R (l'ensemble des documents pertinents), la probabilité $P(R)$ est la probabilité que le document d_j choisi au hasard à partir de toute la collection est pertinent. Les significations des probabilités $P(\overrightarrow{d_j}|\overline{R})$ et $P(\overline{R})$ sont analogiquement complémentaires.

Pour tous les documents dans la collection les probabilités $P(R)$ et $P(\overline{R})$ sont identiques, donc on peut représenter la fonction de similarité entre le document d_j et la question q par la formule suivante :

$$sim(d_j, q) \sim \frac{P(\overrightarrow{d_j})}{P(\overrightarrow{d_j}|\overline{R})}$$

1.1.4 Modèles alternatifs

Il existe également des différentes modèles dits «alternatifs».

Pour le modèle booléen, les plus connus sont :

- le modèle flou,
- le modèle booléen étendu (en anglais : *Extended Boolean Model*).

Pour le modèle vectoriel :

- le modèle vectoriel généralisé (en anglais : *Generalized Vector Space Model*),
- (en anglais : *Latent Semantic Indexing Model*),
- le modèle du réseau de neurones (en anglais : *Neural Network Model*).

Pour le modèle probabiliste :

- le modèle des réseaux bayésiens (en anglais : *Bayesian Network Model*),
- le modèle des réseaux d'inférence (en anglais : *Inference Network Model*),
- le modèle estimatif (en anglais : *Belief Network Model*).

1.2 Indexation du document

1.2.1 Le principe de l'indexation

L'indexation est le processus de création à l'avance de structures de données (fichiers d'index), qui contiennent des informations concernant chaque document de la collection. C'est la partie la plus importante dans le processus d'analyse documentaire. C'est elle qui conditionne la valeur d'un système documentaire. L'indexation d'un texte est la recherche dans ce texte de mots ou d'idées-clés qui le caractérisent. Les mots ou les notions correspondantes donnent une idée du contenu du texte.

Définition 1 : L'indexation consiste à décrire et à caractériser un document textuel ou visuel afin de faciliter sa recherche. Elle est l'une des opérations du système documentaire qui permettent de retrouver l'information dans l'ensemble de documents sauvegardés.

Définition 2 [Mer] : L'indexation est l'opération qui consiste à analyser et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. La transcription documentaire se fait grâce à des outils d'indexation tels que thèsaurus, classification, etc... qui permettent le choix et l'attribution de descripteurs ou mots-clés décrivant de la façon la plus exhaustive le contenu conceptuel d'un document.

Dans le première approche on peut distinguer les deux différents types d'indexation :

- l'indexation humaine - fondée sur le jugement d'un être humain, un spécialiste dans le domaine,
- l'indexation automatique - créée par ordinateur.

C'est l'indexation automatique qui va être analysée ici, et que par la suite nous appellerons indexation tout court.

1.2.2 La structure de l'index

Le processus d'indexation peut être divisé en deux étapes suivantes :

- l'analyse du document, en fonction d'un certain nombre de critères,
- l'extraction des éléments qui décident de sa pertinence et enfin le placement de ces données dans les fichiers d'index.

Définition Les fichiers d'index sont les structures de données qui contiennent les informations nécessaires pour l'indexation de chaque document de la collection.

On distingue deux méthodes de création des fichiers d'index, et donc deux méthodes d'indexation :

- l'indexation directe,
- l'indexation inversée.

1.2.2.1 Les fichiers directs

On parle d'indexation directe si les fichiers d'index donnent, pour chaque document, une liste des mots-clés contenus ou caractérisant le document. Les fichiers créés de cette façon sont appelés fichiers directs.

Organisation des fichiers directs : pour chaque document de la collection analysée, on construit la liste des mots (mots-clés) qui apparaissent dans ce document. Les mots-clés peuvent être obtenus à partir des mots du document par diverses relations (synonymie, hiérarchie, données par des thésaurus). Un exemple est présenté dans la table 1.1.

	mot$_1$	mot$_2$	...	mot$_n$
document$_1$	x			
document$_2$		x	x	
...	x			x
document$_n$		x		

TAB. 1.1 – Exemple de fichier d'indexation direct.

Dans ce modèle, pour retrouver les documents, il faut parcourir le fichier tout entier. Cela a pour conséquence que la recherche des documents qui contiennent le mot clef est assez lente.

1.2.2.2 Les fichiers inversés

Au contraire du modèle direct, pour l'indexation inversée, le processus se fait en fonction de chaque mot-clef. La structure de données créée par cette méthode est appelée fichier inversé.

Organisation des fichiers inversés : à chaque mot-clef existant on associe la liste des documents qui contiennent ce mot. Un exemple est présenté dans la table 1.2.

Cette méthode permet un accès direct à toutes les documents qui contiennent le mot cherché.

Elle est donc plus rapide et plus adaptée pour faire une recherche en fonction des mots-clés.

Exemple : Systèmes les plus connus fondés sur l'utilisation de fichiers inversés :

	document$_1$	document$_2$	...	document$_n$
mot$_1$	x			
mot$_2$			x	
...	x		x	
mot$_n$	x			x

TAB. 1.2 – Exemple de fichier d'indexation inversé.

- DIALOG, Mai 1980, *Locked Information Systems of Palo Alto, California,*
- STAIRS, *IBM Corporation,*
- BRS *(The Bibliographic Retrieval Services), State University of New York,*
- MEDLARS, *National Library of Medecine,*
- ORBIT,
- *Information Bank*,
- LEXIS.

1.2.3 Étapes du processus d'indexation

[And98, iC98, FPC98, JRP00, Kor97, Lar97, Lel98, Pyt98, vR79, SM83]

Le processus d'indexation constitue une étape première de l'analyse du contenu de document. Cette analyse est faite en fonction d'un certain nombre de critères, on procède à la sélection et à l'extraction des éléments qui décident de sa pertinence (descripteurs), puis à leur traduction en langage documentaire.

L'indexation est une étape très importante dans le processus de recherche. En effet, la qualité et la rapidité de recherche dépend beaucoup de l'indexation. Une bonne analyse de chaque document de la base est indispensable pour la construction d'un bon fichier d'index. Ce fichier devrait regrouper les termes pertinents et significatifs pour le corpus donné.

Pour extraire les termes (descripteurs) significatifs et pertinents d'un document, il faut d'abord effectuer les étapes suivantes :

- découper le document pour extraire correctement les termes,
- faire une analyse des syntagmes et des expressions idiomatiques,
- effectuer une lemmatisation des mots,
- éliminer les mots vides,

1.2.3.1 Découpage - extraction des mots du document

Pour analyser un document textuel, il faut commencer par le segmenter en ses unités élémentaires : les mots.

Cette opération s'appelle découpage.

Dans cette partie de l'indexation on fait le choix de la définition du mot. En fonction de cette définition, on découpe le document pour construire la liste de tous les mot qui apparaissent dans le document analysé.

Définition : Les mots sont les chaînes de caractères qui constituent la plus petite unité de la phrase.

La définition du mot peut être variable en fonction de différents critères. Dans un document textuel, le mot est habituellement défini comme une chaîne des caractères constituée des caractères suivants :

- lettres de l'alphabet "a-z "en majuscules ou en minuscules,
- caractères accentués appropriés à la langue du document,
- tiret "-",
- apostrophe '.

Les autres caractères (comme espace, tabulation, signes des ponctuation, etc.) sont considérés comme des séparateurs entre les mots.

1.2.3.2 Normalisation des mots

Lemmatisation Le mot est un terme générique qui couvre plusieurs approches. Au cours du processus d'indexation en essaie de retrouver la forme basique de chaque mot, ou *lemme,* et c'est cette forme-là qui sera insérée dans le fichier d'index.

Définition : La lemmatisation est l'opération de réduction des formes fléchies des mots à leur forme de base.

Le processus de lemmatisation consiste à retrouver pour un mot sa forme normalisée, qui va correspondre aux entrées présentes dans les fichiers d'index (généralement le masculin pour les noms, l'infinitif pour les verbes, le masculin-singulier pour les adjectifs, ...).

Cette approche permet d'effectuer une recherche beaucoup plus souple et rapide, et en outre le fait de sauvegarder juste une forme pour chaque mot offre un gain en termes de taille de fichier.

Par exemple, pour une recherche sur un verbe, on retrouve tous les documents contenant ce verbe dans toutes ses formes conjuguées, et pas seulement les documents contenant le verbe dans la forme entrée par l'utilisateur.

Troncature à droite

Définition : La troncature à droite est l'opération qui consiste à réduire les mots à indexer en leur racine.

La troncature constitue une approche facile à effectuer de la normalisation, mais elle est beaucoup moins pertinente et moins juste que la lemmatisation.

1.2.3.3 L'élimination des mots vides

Les mots vides sont des mots courts, communs qui n'ont pas de valeur de pertinence pour la recherche, par exemple les articles ou les conjonctions comme “le”, “un”, “de”, “et”, “ou”, “pour”, etc.

Définition : Un mot vide est un mot non significatif du point de vue de la recherche.

Exemple : les articles (le, la, les) et les prépositions (de, pour, avec, à, etc).

Pour rendre les fichiers d'index plus déterminants, il faut éliminer tous les mots non significatifs, appelés *mots vides.* Cette opération, effectuée aussi bien lors de l'indexation que de la recherche (élimination des mots vides de la question) joue un grand rôle dans la précision de la recherche et diminue le bruit dans les réponses fournies.

En fonction de la définition des mots vides on peut distinguer différentes méthodes pour les éliminer.

Méthode statique - liste prédéfinie de mots vides Dans cette méthode on construit d'abord la liste de tous les mots non significatifs pour un ensemble de documents donnés. Pendant l'indexation, on vérifie pour chaque mot du document si ce mot appartient à la liste des mots vides. Seuls les mots qui n'appartiennent pas à cette liste sont insérés dans les fichiers d'index.

Inconvénient :

- difficulté de construire un liste complète,
- manque de souplesse pour les documents multilingues.

Méthode dynamique - Fréquence d'apparition On utilise la fréquence d'apparition des mots dans l'ensemble des documents de la base.

Si un mot apparaît trop souvent, il est considéré comme trop fréquent, donc n'apportant pas d'information importante. Ce mot ne sera donc pas pris en compte lors de l'indexation et de la recherche.

Cette approche peut aussi provoquer des erreurs. Si la question posée ne contient que des mots très fréquents, la recherche ne pourra pas être effectuée.

1.2.4 Mots-clés, descripteurs et thésaurus

[Lèf00] Définition la plus simple pour les mots-clés : les mots clés sont tous les mots du texte qui n'ont pas été désignés comme mots vides.

Généralement le mot-clé ou le descripteur, c'est une entrée dans le fichier d'index, qui appartient au vocabulaire d'un domaine de connaissance spécifique.

Mot-clé

Nouvelle définition : [ATO] mot ou groupe de mots choisi librement dans le titre ou dans le corps d'un texte (ou d'un document), permettant d'en caractériser le contenu et d'en faciliter la classification et la recherche. Le mot-clé peut aussi éventuellement faire partie d'une liste fermée de mots (dits descripteurs) utilisés à des fins de description et de recherche de documents (on parle de vocabulaire contrôlé).

Descripteur

Définition : [AFN97] mot ou groupe de mots pris dans un thésaurus et choisi parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion apparaissant dans un document ou dans une demande de recherche documentaire.

Thésaurus

[Mar, Per99, Hu994]

L'idée de thésaurus est assez simple : les descripteurs sont bien regroupés en gardant les relations entre eux. On peut définir un thésaurus comme une liste normalisée et structurée de descripteurs (mot-clés) qui contiennent des relations entre eux et représentent un corpus sémantique d'un domaine. Le thésaurus [Hu994] est un outil en construction permanente.

On peut citer différentes définitions de thésaurus.

- [ATO] "ensemble de mots ou de termes (dits descripteurs) constituant un vocabulaire défini (vocabulaire contrôlé de termes), ayant entre eux des relations d'ordre sémantique (par exemple, une relation hiérarchique orientée du générique vers le spécifique) ou pragmatique, et qui s'appliquent à un ou plusieurs domaines de la connaissance."
- [ISO86] "... vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relations générique-spécifique)".

Du point de vue de sa fonction, un thésaurus est à la fois un instrument de contrôle et de référence du vocabulaire utilisé pour décrire et organiser des informations dans un domaine donné.

La norme ISO : 2788 [ISO86] définit le contenu d'un thésaurus selon trois méthodes principales :

- liste alphabétique de descripteurs avec leur relations,
- liste des champs sémantiques (systématique) organisés en domaines ou disciplines,
- présentation graphique (schéma fléché ou disposition graphique).

Les types sont accompagnés d'un index alphabétique.

Le nombre de descripteurs varie de quelques centaines de termes à 20 000, la moyenne étant de 3 000. Un descripteur est toujours au singulier et peut être un mot simple ou composé.

1.2.4.1 Les relations du thésaurus

Les thésaurus sont reliés entre eux par des relations sémantiques et sont exprimés par des signes conventionnels.

Hiérarchique Dans la relation hiérarchique le descripteur est replacé dans son environnement sémantique. On relève les termes de sens plus large, termes génériques et ceux de sens plus étroit, termes spécifiques.

Association (de voisinage) La relation d'association ou voisinage indique des analogies entre des descripteurs liés.

Synonymique (équivalence) La relation d'équivalence renvoie divers synonymes d'un concept vers le descripteur choisi.

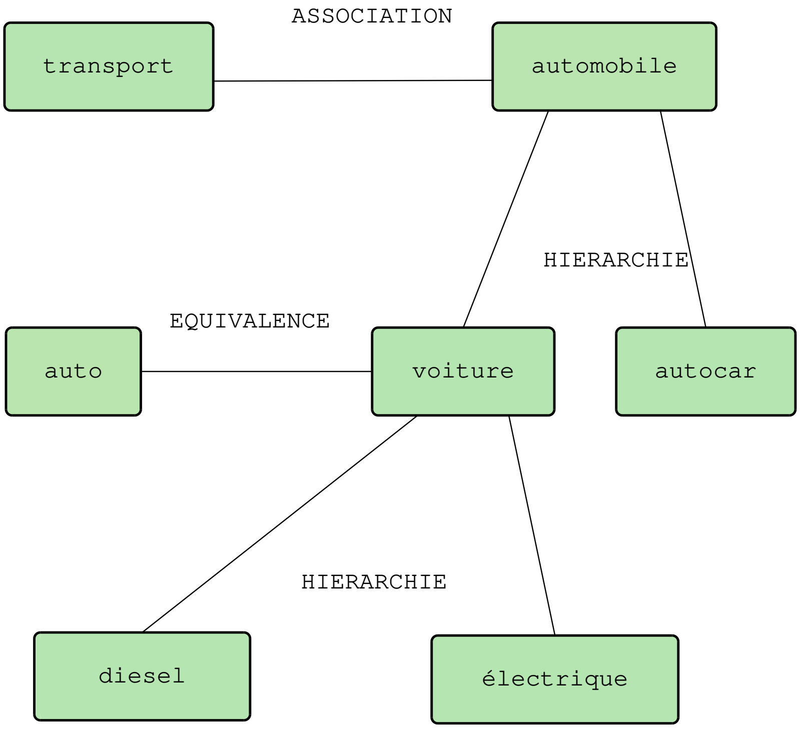


FIG. 1.1 – Exemple de trois relations sémantiques entre les descripteurs dans le thésaurus.[Lèf00]

1.3 L'espace documentaire

1.3.1 Différentes représentations des documents

[AFN97, SC, DW96, FPC98, Huy96a, JRP00, Kla98, Kor97, Lar97, Lèf00, Les96, Los00, Mer]

Les documents peuvent être présentés de différentes façons. On présente ensuite :

- la représentation vectorielle,
- la représentation par les n-grammes,
- la représentation créée dynamiquement.

1.3.1.1 Représentation vectorielle

La représentation d'un document par le modèle vectoriel a été introduite par Salton [SM83]. Dans ce modèle, le document est représenté par un vecteur qui contient autant de composantes qu'il y a de descripteurs dans le corpus. Le mots vides ne sont pas pris en compte pour la construction de ce vecteur. Chaque vecteur est constitué par les fréquences de chaque descripteur du texte.

Les limites La représentation de documents par des vecteurs de fréquences de mots est limitée :

- Si on se base sur les formes fléchies des mots pour analyser de gros corpus comme ceux d'Internet, le nombre de termes à analyser et à indexer augmente très vite vers des millions. En outre, la création des formes composées devient très irréaliste.
- Si on se base sur les formes lemmatisées, le nombre de descripteurs diminue, il est donc possible de créer des formes composées, mais leur extraction (spécifique pour chaque langue différente) nécessite de traitements informatiques très lourds.

En outre pour les langues où les mots ne sont pas faciles à découper, comme les langues agglutinantes (turc, russe, allemand dans une moindre mesure), et les écritures asiatiques où seul le contexte permet de séparer les mots, ce découpage devient encore plus lourd.

1.3.1.2 Représentation par n-grammes

Il s'agit d'une méthode indépendante de la langue pour l'indexation de grands corpus de documents.

Définition : Un n-gramme est une séquence de n-caractères consécutifs.

Cette méthode caractérise les documents par leurs fréquences de n-grammes.

Représentation de textes par des vecteurs de fréquences de n-grammes Chaque document d'un corpus est caractérisé par le vecteur des fréquences de ses n-grammes, c'est-à-dire de toutes les chaînes différentes de longueur fixe de **n** caractères.

C'est surtout pour les langues agglutinantes et les écritures asiatiques où le découpage de mot n'existe pas, que la codification de textes sous la forme de vecteurs de fréquences de n-grammes est prometteuse.

Les étapes de la codification d'un document par les n-grammes sont les suivantes :

- réduction des caractères (étape analogue à l'élimination des mots vides dans la méthode de l'indexation par les mots). Le but est de réduire au maximum le nombre de n-grammes sans perdre de n-gramme significatif. Aussi les n-grammes de fréquence faible seront éliminés.

  > Exemple : en langue française, les séquences de plusieurs caractères blancs consécutifs seront réduites à un seul espace, tous les signes de ponctuation seront remplacés par un signe, les n-grammes de fréquence 1 et 2 seront éliminés.

- extraction des différents n-grammes,

Actuellement des modèles à bigrammes, trigrammes et jusqu'à des modèles à 5-grammes sont utilisés couramment dans les systèmes de reconnaissance de la parole continue.

1.3.1.3 Représentation dynamique

[Ziz95] Dans ce modèle la représentation des documents est adaptée en temps réel en fonction de la question posée.

Le plus souvent le modèle dynamique est basé sur le principe des réseaux de neurones, où chaque document est représenté par un sous-réseau. Toutes les références pour chaque document sont symbolisées par les noeuds du réseau. L'ensemble de tous ces sous-réseaux crée un réseau initial, qui va être modifié pendant la recherche de l'utilisateur. Cette technique permet d'effectuer facilement des manipulations sur les données pour adapter la structure des réponses fournies aux utilisateurs. Malheureusement, les réseaux de neurones sont assez lents, donc le temps de réponse de cette méthode est très élevé. C'est une méthode adaptée aux applications où la qualité des réponses est un facteur plus importante que le temps dans lequel la réponse est fournie.

1.3.2 Différentes méthodes de mesure de la pertinence

[SM83, And98, HH97, JRP00, Kla98, Kor97, Lèf00, Les96, vR79, SB96]

Définition de la pertinence : Un document pertinent est un document qui répond exactement à la question posée.

La mesure de pertinence dépend de la méthode utilisée pour sa mesure. Les indices utilisés comme mesure de pertinence en général sont basés sur une approche booléenne.

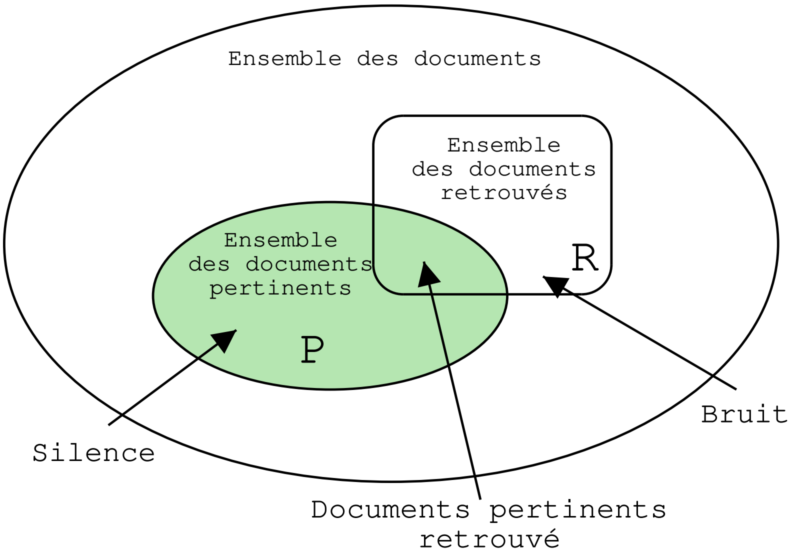


FIG. 1.2 – Exemple de représentation du bruit et du silence en recherche de l'information

L'évaluation de la pertinence dans les recherches documentaires essaie de quantifier le rappel et la précision, avec les indices de bruit et de silence. On les définit comme suit.

Dans les définitions on utilise les symboles suivants :

D_{rp} - l'ensemble de documents retrouvés et pertinents,

D_r - l'ensemble de documents retrouvés,

D_p- l'ensemble de documents pertinents.

Définition de la précision [Lel98] : ratio entre le nombre de documents pertinents trouvés et le nombre total de documents trouvés. Elle mesure le bruit et plus elle est proche de 100%, moins il y a de bruit, et meilleure est la réponse.

$Precision = \frac{Drp}{Dr}$

Définition du bruit [Lel98] : ratio entre le nombre de documents ramenés en réponse, mais qui ne sont pas pertinents par rapport à la question posée, et le nombre total de documents.

$Bruit = 1 - Precision = \frac{Dr - Drp}{Dr}$

Définition du rappel [Lel98] : ratio entre le nombre de documents pertinents trouvés et le nombre de documents pertinents présents dans la base. Plus il est proche do 100%, moins il y a de silence, et meilleure est la réponse.

$Rappel = \frac{Drp}{Dp}$

Définition du silence [Lel98] : ratio entre le nombre de documents pertinents qui n'apparaissent pas dans le résultat de la recherche et le nombre total de documents.

$Silence = 1 - Rappel = \frac{Dp - Drp}{Dp}$

Analyse de la pertinence

Il existe beaucoup d'algorithmes et de techniques pour analyser la pertinence des réponses. On peut les regrouper en deux groupes :

- méthodes textuelles,
- méthodes hypertexte.

1.3.2.1 Méthodes textuelles

Dans les méthodes textuelles, on effectue la vérification de pertinence à partir de l'informations textuelles contenues dans les documents.

La méthode la plus répandue est basée sur la fréquence et la position relative des termes dans le document.

Fréquence et position relative des mots dans le document Pour cette définition on utilisera les symboles suivants :

tf - fréquence (nombre d'occurences) d'un terme dans le document, ang. *term frequency*,

D - ensemble de tous les documents dans la collection,

N - nombre de documents dans la collection D,

J - indice du document,

I - indice du terme,

d_j - document de la collection D,

k_i - un terme dans le document d_j,

n_i - nombre de documents qui contiennent le terme k_i,

Rareté des termes dans l'ensemble des documents
La fréquence inversée *idf* (*inverse document frequency*) du mot dans le document :

Définition La fréquence inversée du document idf_i pour le mot k_i est définie par l'équation :

$$idf_i = \log \frac{N}{n_i}$$

où :

N - nombre de tous les documents dans la collection

n_i- nombre de documents qui contiennent le terme k_i.

Poids du terme dans le document. Méthode *tf-idf*

La méthode la plus connue pour calculer le poids d'un terme dans le document, c'est une méthode basée sur la mesure de la fréquence inversée *tf-idf.*

Définition Dans la méthode *tf-idf* le poids d'un document est défini comme :

$$v_{i,j} = tf_{i,j} \times \log \frac{N}{n_i}$$

où :

$v_{i,j}$ - poids du terme k_i dans le document d_j.

Taille du document

Pour ne pas privilégier les gros documents par rapport à ceux de petite taille, la mesure de fréquence d'apparition d'un terme dans le document est normalisée :

Définition Pour une collection D de N documents et s'il y a n_i documents qui contient le terme k_i, la fréquence $tf_{i,j}$, donnant le nombre de termes k_i dans le document d_j, la fréquence normalisée $f_{i,j}$ est définie par la formule suivante :

$$f_{i,j} = \frac{tf_{i,j}}{max_l tf_{l,j}}$$

où :

f_{ij} - la fréquence normalisée,

max_l - le maximum, calculé pour tous les termes du document d_j.

Si le document d_j ne contient pas le terme k_i, la fréquence normalisée $f_{i,j}$ vaut *0.*

Définition Le poids d'un terme dans un document calculé par la méthode de la fréquence inverse normalisée $f_{i,j}$ est définie comme :

$$w_{i,j} = f_{i,j} \times \log \frac{N}{n_i}$$

où :

$w_{i,j}$ - poids du terme k_i dans le document d_j.

1.3.2.2 Méthodes des hyperliens

Contrairement aux méthodes précédentes, on ne prendra pas ici en compte seulement l'information textuelle mais on fait aussi l'analyse de l'information contenue dans les liens hypertexte.

Méthode de l' "hyper information" [Mar97] Dans cette méthode, la valeur informationnelle de chaque document dépend de deux facteurs, l'information textuelle et de l'information insérée dans les liens hypertexte. On peut donc décrire la pertinence d'un document donné par le modèle suivant :

$Information(D) = TexteInfo(D) + HyperInfo(D)$

où :

D - le document analysé,

$Information(D)$ - l'information globale du document D,

$TexteInfo(D)$ - l'information textuelle du document D,

$HyperInfo(D)$ - l'information du document D contenue dans les liens.

Le poids de l'information textuelle de chaque document est généralement calculé par la méthode *tf-idf* [Section 1.3.2.1]. Par contre la valeur apportée par les liens hypertexte dépend des valeurs de tous les documents qui sont pointés par le document analysé. Ce calcul du poids apporté par les documents pointés peut descendre en profondeur jusqu'au niveau choisi. La valeur informationnelle apportée par chaque niveau de profondeur diminue avec la distance entre ce niveau et le document de base. Exemple pour des liens simple de niveau n est présenté dans la figure 1.3

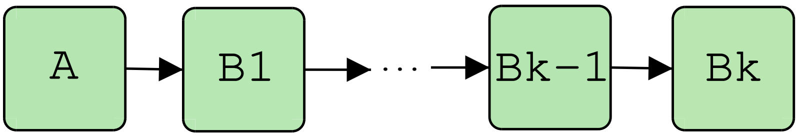


FIG. 1.3 – Méthode des hyperliens. Liens simples, niveau n.

La figure 1.4 montre un exemple de document A *(une page d'index)* qui ne contient pas l'information brute elle même, mais regroupe les liens vers d'autres pages. Dans toutes les approches précédentes, il n'y avait pas de possibilité de retrouver ces pages pendant la recherche. Pourtant les *pages d'index*, même si elles-mêmes ne contiennent

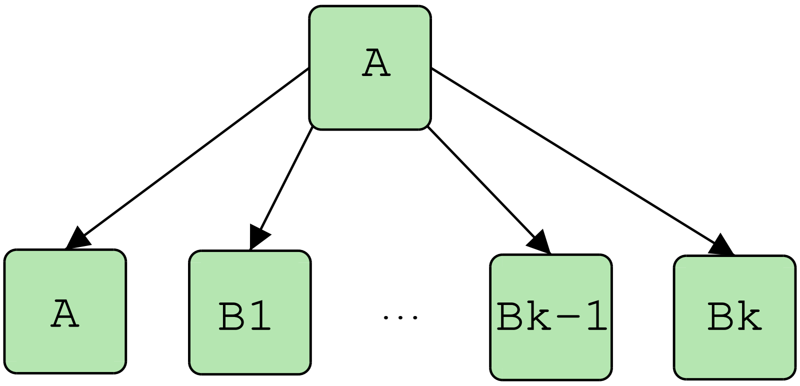


FIG. 1.4 – Méthode des hyperliens. Liens multiples, niveau 2.

pas l'information recherchée, donnent directement une vaste ressource avec les liens vers les documents, souvent pertinents et regroupés par thème. L'approche proposée par cette méthode augmente beaucoup la probabilité de retrouver parmi les réponses les *pages d'index* comme les pages pertinentes.

Analyse de la topologie des liens

[CGK98, GKR98, Gri00, Kle98]

La méthode d'analyse de la topologie des liens, est une méthode dans laquelle la pertinence d'un document est définie en fonction du nombre de liens qui pointent sur le document.

Cette approche définit la "visibilité" d'un document déterminé par sa popularité ou la popularité de son site, mais pas obligatoirement par la qualité de l'information qu'il contient.

Désavantage :

- un moindre poids de la qualité de l'information contenue dans le texte et un plus grand poids pour les liens,
- les nouveaux documents ont peu de chances d'avoir une grande visibilité, donc d'être retrouvés par le moteur

Exemple de moteur qui utilise cette méthode : Google[1].

---

[1] La méthode *PageRankTM* de Google dont le brevet est en cours d'homologation, est basé sur la même principe.

Le fonctionnement de *PageRank* est exposé dans la page Web : google "why use google" `http://www.google.fr/intl/fr/why_use.html`

"*PageRank* capitalise sur cette caractéristique exceptionnellement démocratique du Web en utilisant sa vaste structure par liens en guise d'outil organisationnel. Essentiellement, Google interprète un lien de la page A vers la page B comme un vote, de la page A, en faveur de la page B. Google détermine l'importance d'une page par les votes qu'elle reçoit. Mais Google ne se limite pas à compter le nombre de votes, ou liens ; il analyse également la page qui émet le vote. Les votes émis par des pages qui sont elles-mêmes "importantes" ont plus de poids et contribuent à rendre d'autres pages "importantes".

Ces résultats importants, hautement qualitatifs reçoivent un *PageRank* plus élevé et tiendront une place prépondérante dans les résultats. Ainsi, *PageRank* est l'indicateur général d'importance de Google et ne dépend pas d'une recherche spécifique, mais c'est plutôt la caractéristique de la page elle-même, basée sur les données d'Internet que Google analyse par le biais d'algorithmes complexes qui déterminent la structure par liens.

Évidemment, les pages importantes ne vous sont d'aucune utilité si elles ne correspondent pas à votre requête. Aussi, Google utilise des techniques sophistiquées de corrélation textuelle pour trouver les pages qui sont à la fois importantes et pertinentes en regard de votre recherche. Par exemple, quand Google analyse une page, il examine les commentaires des pages se reliant à elle."

Chapitre 2

L'univers du Web

2.1 Histoire d'Internet et du Web

2.1.1 Histoire d'Internet

L'histoire d'Internet n'est pas très ancienne. Son début date des années soixante, quand l'ARPA (*U.S Defense Department's Advanced Research Projects Agency*) a lancé le premier réseau d'information basé sur le réseau téléphonique. Ce premier réseau nommé ARPANET (c'est-à-dire le réseau de l'ARPA) construit dans un but de recherche militaire, a vite été repris par les universitaires qui y ont trouvé une occasion extraordinaire d'échanger des informations au moyen du courrier électronique.

Dans les années suivantes, l'ARPA a continué ses travaux de recherches dans le domaine du développement de protocoles de transfert de données entre les différents réseaux d'ordinateurs. Avec le développement du réseau initial dans les années 70 par les différentes universités américaines, est venu le nouveau nom, et le réseau a été nommé Internet (*l'Inter Networking*).

Dans les années 80, le réseau Internet a commencé son énorme expansion au travers des universités, des laboratoires de recherche, et des grandes entreprises du monde entier. Mais l'accès public au réseau n'était pas encore possible. C'est l'époque à laquelle sont nés les premiers fournisseurs privés d'accès à Internet.

Dans les années 1990, la facilité d'accès à Internet ainsi que la naissance du Word Wide Web, ont fait croître le réseau à une vitesse exponentielle.

2.1.2 Histoire du Web

L'histoire du Web est très récente, pourtant les concepts sur lesquels il est basé sont plutôt anciens. La première note mentionnant une “toile d'information” fut publiée en 1945 par Vannevar Bush, le conseiller de président Roosevelt. L'idée de l'hypertexte et du logiciel de navigation hypertexte fut développée par Ted Nelson en 1965. Mais il fallut attendre encore beaucoup de temps pour que ces deux concepts liés entre eux commencent à fonctionner et donnent la naissance du World Wide Web.

La première toile, d'abord seulement en mode texte, fonctionne au CERN en fin d'année 1991 avec le premier navigateur intitulé World Wide Web, deux années après la publication par Tim Berners-Lee (mars 1989) de l'article «Hypertexte et le CERN». En début d'année 1993 il existe déjà une cinquantaine de serveurs HTTP dans le monde. 1993 est est une année riche en événements très importants pour le développement du Web, en effet :

- Le CERN lance la version alpha du premier navigateur graphique pour X Window et Macintosh (janvier),
- Le NCSA édite la première version du navigateur Mosaic qui fonctionne sous X Window UNIX (février), puis sous MacIntosh et PC Windows (octobre),
- Fondation de Netscape par Clark et Andreessen (aussi fondateur de Silicon Graphics) pour développer Netscape Navigator (mars),
- Le CERN, le MIT puis l'INRIA créent le WWW Consortium pour guider à la normalisation du Web (juillet)

Depuis, le développement du Web a pris une vitesse exponentielle. Tout le monde ne parle plus que des nouvelles possibilités offertes par le Web dans touts les domaines de notre vie.

2.2 Les protocoles et standards utilisés

[FFBL$^+$99]

Le *Word Wide Web* (Toile d'Araignée Mondiale) est un gigantesque réseau d'hypermédia. La communication entre les différentes serveurs et les clients est basée sur le réseau Internet et un protocole nommé HTTP (*Hypertext Transfer Protocol*) conçu pour l'échange de documents hypertexte entre serveurs Web. Mais le Web ne se limite pas seulement à ce protocole et permet également d'accéder à des documents par différents protocoles tels que FTP (*File Transfer Protocol*), *RealAudio*, ...

2.3 Les documents sur le Web

Le Web est basé sur la notion d'hypertexte, permettant création et consultation de documents, puis connexion entre eux de manière non-linéaire, à travers des noeuds de connexion appelés liens hypertexte. Les pages hypertexte sont généralement écrites dans un langage spécial appelé HTML [Rag95] (*HyperText Mark-up Language*). En même temps, pour les documents non hypertexte, d'autres formats sont utilisés : documents en texte non formaté, documents Postscript, PDF (*Portable document Format*), ...

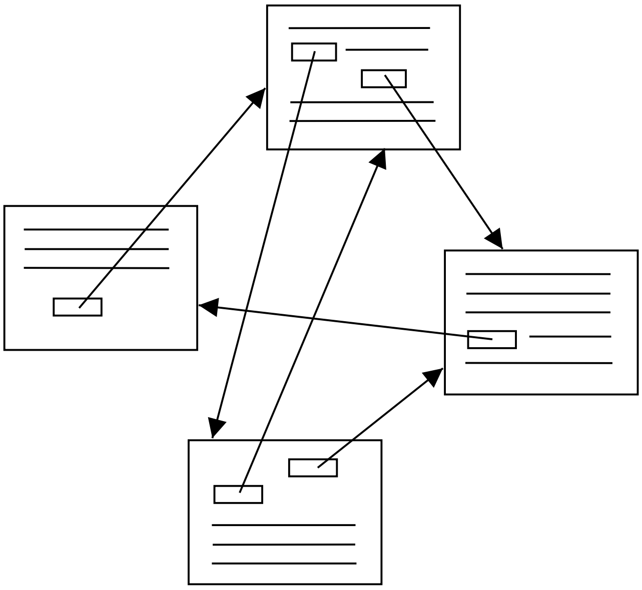


Fig. 2.1 – Exemple de structure hypertexte .

Tim Berners-Lee définit le Web comme "un système d'information réparti, coopératif, hétérogène et multimédia".

- Système d'information réparti : il s'étend dans le monde entier pour permettre une liaison entre des documents se trouvant sur des serveurs de documents situés dans différents endroits.
- Coopératif : il facilite la publication de différentes données liées les unes avec les autres,
- Hétérogène : les documents sont accessibles sur différentes plates-formes,
- Multimédia : il contient différents types de documents tels que documents textuels, images, animations, vidéos, sons, etc.

2.3.1 La structure des documents HTML

Les documents écrits en HTML sont constitués de texte ordinaire et contiennent les éléments suivants :

- le texte du document lui-même,
- des balises HTML : pour décrire la structure du document et pour insérer des liens hypertexte ou multimédias.

Bien que les navigateurs acceptent aussi des documents non-structurés, un document HTML doit toujours posséder :

- une déclaration HTML (<html></html>) qui définit le début et la fin du document,
- un en-tête (<head></head>) qui contient des informations relatives au document (son titre, l'URL de base, etc.), non affichées sur la page dans le navigateur.
- le corps (<body></body>) du document qui contient les information affichées sur la page dans le navigateur.

Exemple de structure de document HTML est présenté dans la figure 2.2.

```
<HTML>
  <HEAD>
    <TITLE>
      Exemple de structure de document HTML
    </TITLE>
  </HEAD>
    <BODY>
      Corps du document
  </BODY>
</HTML>
```

Fig. 2.2 – Structure de document HTML .

Le format d'affichage des documents HTML, défini par des balises spéciales, est constitué par différents éléments tels que titres, paragraphes, listes, tableaux, images, liens vers d'autres documents...

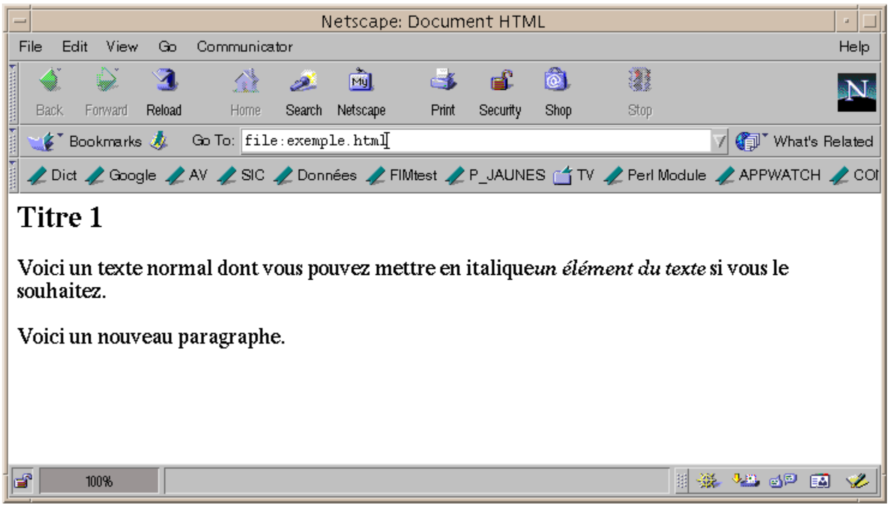
Les navigateurs ignorent les étiquettes qu'ils ne peuvent reconnaître.

Exemple 1 : Un code de page HTML simple [Figure 2.3].

Exemple 2 : Interprétation de code HTML (d'Exemple 1, [Fig 2.3]) par le navigateur [Figure 2.4].

```
<HTML>
  <HEAD>
    <TITLE>
      Document HTML</TITLE>
  </HEAD>
  <BODY>
    <H1>
      Titre 1</H1>
    <p>
      Voici un texte normal dont vous pouvez mettre en italique
      <em>
        un élément du texte</em>
      si vous le souhaitez.
    <p>
      Voici un nouveau paragraphe.
  </BODY>
</HTML>
```

FIG. 2.3 – Un code de page HTML simple .



Go To: file:exemple.html

Titre 1

Voici un texte normal dont vous pouvez mettre en italique*un élément du texte* si vous le souhaitez.

Voici un nouveau paragraphe.

FIG. 2.4 – Interprétation de code HTML dans le navigateur Netscape.

2.3.2 Localisation des documents

[Rag95] Toutes les ressources disponibles sur le Web sont accessibles via une adresse absolue appelée URL (*Uniform Resource Locator*). La forme générale l'adresse URL est la suivante :

`protocole ://serveur[ :port]/chemin/document[?requête][#partie]`

Dans cet exemple d'adresse URL on peut distinguer les parties suivantes :

- (`protocole :`) - préfixe qui indique le protocole à utiliser pour accéder au document, par exemple FTP, Gopher, HTTP, News, Wais, etc,
- (`serveur[ :port]`) le nom du serveur,
- (`/chemin/document[?requête][#partie]`) chemin d'accès au document sur ce serveur.

Exemple d'URL :

`http ://www.cri.ensmp.fr/~wegrzyn/index.html`

Cet URL indique que le document `index.htm` dans mon répertoire personnelle `/~wegrzyn` du serveur `www.cri.ensmp.fr.` est accessible via le protocole HTTP.

2.3.3 Fiabilité et accessibilité des liens

2.3.3.1 Fiabilité des liens

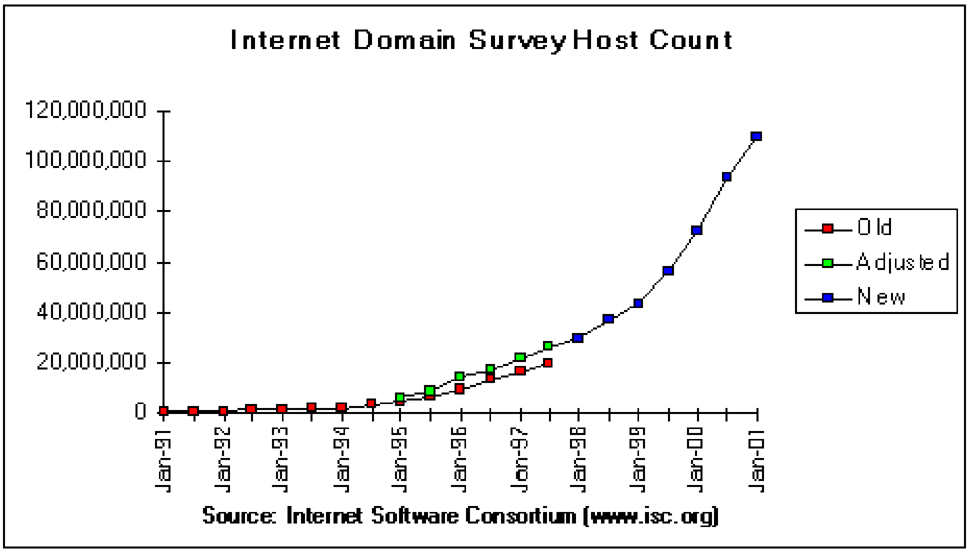
Le Web est un espace d'échanges d'information très dynamique. Les documents changent assez souvent, de nouveaux documents sont ajoutés, des documents anciens sont supprimés, les liens entre les documents peut être modifiés, et la localisation des documents peut également changer. En raison de ce caractère dynamique, on ne peut jamais être sûr que les liens et les documents sauvegardés un jour seront valables le lendemain.

2.3.3.2 Accessibilité des liens

Il n'est pas possible de retrouver tous les documents existants à partir d'un moteur de recherche. L'étude de S. Lawrence et Lee Giles [LG99] montre que seuls 16% environ des documents sont indexés par les moteurs de recherche, et que la proportion entre les documents indexés et non indexés a plutôt tendance à diminuer. Le processus d'indexation pour la page nouvelle ou une page modifiée prend beaucoup de temps, même jusqu'à un mois. Forcément, il existe beaucoup de documents qui ne sont pas faciles à accéder à cause de leur visibilité très modeste.

2.3.4 Quantité et qualité des documents

[MAR99] Même si le nombre de documents existants aujourd'hui sur le Web commence à être de plus de plus difficile à estimer, sa taille ne cesse de croître à très grande vitesse. Dans la figure 2.5 nous présentons un graphe de la croissance du Web à partir de l'année 1981[1] . Les tableaux suivantes [Tableaux : 2.1, 2.2, 2.3] présentent les dernières statistiques des taille du Web publiés par Netsizer.



Source: Internet Software Consortium (www.isc.org)

FIG. 2.5 – La croissance du Web.
Juin 2001, `http://www.isc.org`

En plus de l'utilisation du Web comme outil ouvert de diffusion de l'information, où tout le monde peut publier sans aucune vérification des sources de l'information publiée, ni aucune censure, il faut poser la question de l'évaluation de la qualité des documents. L'information peut être entâchée de différents doutes. Il est possible de

---

[1] Comparaison de l'étude de S.Lawrence et Lee Giles [LG99] pour l'estimation de la taille du Web en 1999 avec l'estimation faite par T.Bray en 1996 [Bra99] :

Taille du Web	1996	1999
en nombre des pages	200 millions	800 millions
en information brute	10 TeraOctets	15 TeraOctets

Mois	1999	2000	2001
Janvier	44 3279	70 4429	105 954
Février	48 1736	73 1777	111 472
Mars	50 6056	75 3368	116 151
Avril	53 6073	77 3942	119 372
Mai	55 9600	80 4436	122 098
Juin	57 5106	83 1374	-
Juillet	59 4623	86 1141	-
Août	61 5411	88 9090	-
Septembre	63 5226	91 8528	-
Octobre	65 8382	94 5665	-
Novembre	67 5230	96 9222	-
Décembre	69 0147	99 5656	-

TAB. 2.1 – La croissance du site Web par mois en milliers.
Juin 2001, http ://www.netsizer.com

classer les incertitudes relatives à l'information dans deux catégories : les incertitudes sur le support (date de validité, fréquence de la mise à jour) et les incertitudes sur le contenu (fiabilité, véracité).

L'incertitude peut provenir d'une falsification d'information non nécessairement malveillante ni provoquée par une modification volontaire. Cette incertitude est largement liée au problème de la source d'information. Même quand la source d'information est bien identifiée, il est possible de retrouver des documents contenants des informations fausses, par exemple dûes au piratage du site. Pour éliminer au maximum le risque de déformation de l'information, d'espionnage, de piratage, etc., des recherches sont en cours dans les domaines de la sécurisation de l'information basée sur les protocoles de communication (authentification de la machine, codage d'information), des protections d'accès aux bases de données, et des méthodes d'authentification de documents (clés électroniques)[2].

---

[2] En année 2000 Président Clinton a signé le première loi pour légalisation de signature électronique.

Le 30 mars 2001 Premier Ministre L.Jospin a signé un décret de loi portant sur la signature électronique [Décret no 2001-272 du 30 mars 2001 pris pour l'application de l'article 1316-4 du code civil et relatif à la signature électronique]

Pays	Nombre de sites [en milliers]	Nombre de utilisateurs [en milliers]
USA	79231.400	179183.00
Japan	6210.620	46143.10
Germany	4033.680	34698.50
UK	4184.690	46833.80
Canada	5696.550	25968.90
Australia	1829.650	11256.30
Finland	1007.330	3186.17
Netherlands	1875.060	11190.90
Sweden	1533.180	6591.29
France	1589.680	11407.70
Norway	589.100	4231.73
Italy	2330.630	14799.60
Taiwan	1636.110	9665.35
New Zealand	416.6130	1368.90
Spain	1044.850	7441.39
Denmark	502.206	3640.66
South Africa	288.600	1226.77
South Korea	553.566	16105.10
Brazil	797.933	12632.40
Switzerland	552.833	3213.99
Austria	700.186	4764.96
Russian Federation	314.859	9862.99
Belgium	608.353	4653.75
Poland	541.373	5381.24
Hong Kong	788.273	5407.20
Israel	286.986	
Singapore	280.066	2818.32
Czech Republic	180.546	415.89
Hungary	199.526	1086.53
Mexico	483.486	2033.81
Greece	190.433	4565.89

TAB. 2.2 – Nombre des sites et d'utilisateurs par pays.
Juin 2001, http ://www.netsizer.com/

Continent	Nombre de sites	Nombre de utilisateurs
Africa	304.473	3328.88
Asia	10037.500	79637.60
Europe	22830.700	101243.00
Oceania	2258.800	19543.70
Central America	487.840	1613.59
South America	1341.900	17648.80
North America	84927.200	205152.00

TAB. 2.3 – Nombre des sites et d'utilisateurs par continent.
Juin 2001, http ://www.netsizer.com/

2.3.5 L'hétérogénéité

Comme on l'a dit, le Word Wide Web est un gigantesque réseau d'hypermédia hétérogène. Cette hétérogénéité porte non seulement sur la diversité des documents, mais aussi sur les sources de ces documents.

2.3.5.1 L'hétérogénéité des formats de documents

Les documents disponibles sur le Web sont très variés. On trouve ici essentiellement des documents HTML (*Hyper Text Markup Language*) mais aussi des documents en texte non formaté (ASCII), des documents *PostScript* (*.ps*), des documents au format Acrobat (*.pdf*), des documents au format Word (*.doc*)[3]. Le Web ne se compose pas seulement de documents textuels, mais c'est aussi une vaste source d'information sous des formes très variées : images, fichiers audio, animations, etc.[4].

2.3.5.2 L'hétérogénéité des ressources

L'hetérogénéite du Web n'est pas seulement dûe à la diversité des catégories et des formats de documents qu'il contient. Ces différents documents sont aussi créés dans

---

[3] Les documents textuels en formats autres que HTML, comme Postscript, Acrobat, Word,... sont d'une manière générale relativement marginaux sur le Web. Les études fait par Tom Bray Measuring the Web[Bra99] [Bra99, HHMN99] et à University of California Berkeley, An Investigation of Documents from the World Wide Web [WAB+99], montre donnes suivantes :

Format de documents	Poucentage sur le Web
Postscript	1.5 %
Acrobat	0.2 %
Word	0.2%

[4] Études de classification de l'information sur le Web, par catégorie et format du fichier [WAB+99] (en nombre de fichiers)

des buts très différents. Dès le début le Web a surtout été considéré comme un grand espace d'échanges d'informations scientifiques entre universitaires. Mais sa croissance a aussi ouvert d'autres perspectives pour les acteurs, notamment économiques. Le Web commence donc souvent à être considéré comme une grande vitrine publicitaire et un espace commercial pour les différentes institutions. Les informations diffusés sur le Web sont très différentes. On peut les catégoriser en fonction de leurs fonctionnalités :

- éducative,
- culturelle (diffusion électronique de musique, vidéo, peintures, ...),
- actualité (journal, valeur sur la bourse, météo, annuaires ...),
- service (commande en ligne de billets, vente en ligne, ...),
- diffusion en ligne de radio, tv,
- espace de jeux.

En même temps la présence des éditeurs d'information sur le Web est très variée[5] :

Catégorie	Format de documents	% sur le Web
Audio	Sun audio (au)	0.7 %
	MS WAVE (wav)	0.3%
	Audio IFF (aif/aifc/aiff)	0.1%
	MIME (snd), Amiga MOD, IFF (iff), SoundBlaster (voc)	0.0%
Image	GIF (gif)	61.7%
	JPEG	7.8%
	X bitmap (xbm)	2.9%
	TIFF (tif/tiff)	0.2%
	RGB (rgb), portable pixmap (ppm), X window dump (xwd)	0%
Vidéo	MPEG (mpe/mpeg/mpg)	0.3%
	Quicktime (mov/qt)	0.2%
	MS video (avi)	0.1%
Fichiers compressés	GNU zip (gz/gzip/taz/tgz)	0.7%
	Zip (zip)	0.7%
	Compress (Z)	0.6%

[5] Le documents sur le Web proviennent de différents domaines. Les études faites durant l'année 1996 à l'University of California Berkeley[WAB+99] et celle publiée en juin 2001 par Netsizer (http ://www.netsizer.com) montrent les résultats suivants :

Documents triés par domaine sur le Web	[WAB+99] 1996	Netsizer juin 2001
other	41%	30%
com	20 %	31%
edu	27 %	6%
gov	4 %	1%
net	4 %	30%
mil	1 %	1%
org	3 %	1%
total	100 %	100%

- centres de recherche et universités (articles, documents scientifiques, revues spécialisées, rapports d'activité, informations sur les conférences et congrès, etc.),
- sites gouvernementaux et administratifs (informations officielles, documents publics, réglementations, formulaire à remplir - téléprocédure, etc.),
- institutions et entreprises (documents publicitaires, catalogues de produits, documentations techniques, offres d'emploi, etc),
- agence de presse et éditeurs (dépêches de presse, presse en ligne, etc),
- bibliothèques (catalogues, documents à télécharger, consultation de basées de données gratuites et payantes, etc.),
- associations et différentes organisations,
- sites avec portails thématiques, pages privées, etc.

Dans le tableau 2.1 on présente la diversités des domaines sur le Web.

Domaine	Nombres de sites (en milliers)
com	38325.200
net	36338.600
edu	7967.050
mil	1817.060
org	1304.570
gov	949.539
us	1846.220
jp	5883.640
de	2236.070
uk	2530.480
ca	2771.670
au	1700.670
fi	830.808
nl	1657.240
fr	1433.000
se	1042.560
no	565.168
it	2323.920
tw	1025.680
es	893.246
dk	486.407
...	...
total	122824.000

Tab. 2.4 – La nombre des Sites d'Internet.
Juin 2001,http ://www.netsizer.com/

2.3.6 Le problème du multilinguisme

2.3.6.1 Différentes langues de documents

[Leb99] La couverture mondiale du Word Wide Web oblige à diffuser les documents sur le Web dans des langues différentes. Dès le début, la principale langue d'Internet a été l'anglais, mais d'autres langues sont maintenant de plus en plus présentes. Même si l'anglais est encore prédominant, les sites bilingues ou multilingues sont de plus en plus nombreux, et ce pour des raisons aussi bien commerciales que culturelles[6]. Il est possible qu'avec le temps la répartition des langues sur le Web corresponde à la répartition des langues dans le monde. Cette diversité de langages implique un fort besoin d'outils spécialisés comme les logiciels de traduction disponibles en ligne gratuitement.[7]

2.3.6.2 Différents jeux de caractères

Chaque langue possède son alphabet spécifique. Pour tous ces alphabets, les jeux de caractères sont très différents, ce qui implique l'existence de beaucoup de systèmes différents pour l'encodage des caractères. Au départ, les informations étaient diffusées dans une seule langue à la fois. Ces systèmes étaient donc généralement adaptés à l'encodage dans un seul jeu de caractères en même temps. Même pour des langues utilisant le même alphabet (par exemple, l'alphabet latin), aucun système d'encodage ne couvrait toutes les lettres, les signes de ponctuation et les symboles techniques d'usage courant. Ainsi, par exemple, l'Union Européenne avait besoin de plusieurs systèmes d'encodage pour couvrir toutes ses langues d'usage.

Le grand inconvénient de ces systèmes d'encodage est qu'ils sont souvent incompatibles entre eux. Ainsi, deux systèmes peuvent utiliser le même code pour deux

---

[6] [Leb99]"En 1997, Babel, une initiative conjointe d'Alis Technologies et de l'Internet Society, a mené la première étude d'ensemble sur la répartition réelle des langues sur Internet. Les résultats sont publiés dans le Palmarès des langues de la Toile, daté de juin 1997 et qui devrait être réactualisé prochainement. Les pourcentages étaient de 82,3 % pour l'anglais, 4 % pour l'allemand, 1,6 % pour le japonais, 1,5 % pour le français, 1,1 % pour l'espagnol, 1,1 % pour le suédois et 1 % pour l'italien."

[7] Babelfish, un des premiers traducteurs automatiques, de la société Systran (société franco-américaine, pionnière dans le domaine de la traduction automatique) a été intégré dans le moteur AltaVista, avec possibilité de traduction automatisée de l'anglais vers cinq autres langues (allemand, espagnol, français, italien et portuguais), et vice versa. Babelfish atteint 500 à 600 000 visiteurs quotidiens pour environ un million de traductions par jour. 15.000 sites environ insèrent un lien vers Babelfish.

caractères différents, ou au contraire utiliser des codes diférents pour le même caractère. Il est donc très difficile de publier sur le même document des textes en langues différentes.

La nature multilingue du Web nécessite que les serveurs et les systèmes de diffusion d'information supportent les différents jeux de caractères en même temps. Pour satisfaire la nécessité de code unique pour tous les caractères existants un encodage spécial nommé Unicode a été développé [Hof00]. A la différence des encodage anciens, où chaque symbole était codé sur un octet, chaque caractère en Unicode est encodé sur 2 octets. Cet encodage spécifie un code unique pour chaque caractère, quelle que soit la plate-forme, quel que soit le logiciel et quelle que soit la langue utilisée. Grâce à Unicode, il est possible de publier simultanément dans différentes langues du monde entier. Le développement d'Unicode va de pair avec la norme ISO/CEI 10646. La version 3.0 d'Unicode [Con00] est identique à la norme l'ISO/CEI 10646 : 2000 [ISO86], mais contient en plus de nombreux éléments supplémentaires. Le standard Unicode est utilisé dans de nombreux systèmes d'exploitation, dans tous les navigateurs récents, et dans de nombreuses autres applications.

2.3.6.3 Les projets liés au multilinguisme

Internationalisation Ce domaine est un grand champs de recherche dans le monde. Les plus grands projets mondiaux liés à l'internationalisation sont :

Babel[8] - projet (d'Alis Technologies[9] et de l'Internet Society[10]) pour l'internationalisation d'Internet.

LISA[11] (*Localisation Industry Standards Association*) - organisation pour la localisation et l'internationalisation,

W3C Internationalisation/Localisation[12] - projet du *Word Wide Web Consortium* (W3C),

Traduction automatique et assistée par ordinateur [Leb99]

---

[8] Projet Babel : http ://www.isoc.org :8080

[9] http ://www.alis.com

[10] http ://www.isoc.com

[11] http ://www.lisa.unigu.ch

[12] http ://www.w3.org/International/

Traduction automatique (TA) C'est la traduction de textes d'une langue naturelle à une autre, effectuée totalement par l'ordinateur. Elle fut un des premiers domaines de recherche en informatique. [13]

Traduction assistée par ordinateur (TAO) Contrairement à la traduction automatique qui, sans intervention humaine, analyse le texte dans la langue-source et génère automatiquement le texte correspondant dans la langue-cible, la traduction assistée par ordinateur (TAO) exige une certaine interaction entre l'homme et la machine.

2.4 Performances et détails de réponses sur Internet

[Nak] La croissance du Web, qui s'accompagne d'une grande augmentation du trafic sur le réseau, provoque en même temps une hausse des problèmes de télécommunications.

2.4.1 Performances du réseau

Les problèmes de performance et de fiabilité du réseau sont importants.

Les solutions aux problèmes d'efficacité du réseau, tels que : réduction de l'encombrement de réseau, augmentation de la rapidité, tolérance aux pannes, etc. sont liés directement aux questions de routage et de maintenance du réseau.

2.4.2 Temps de réponse du serveur

Le temps de chargement d'une page ne dépend pas seulement de son contenu mais également des performances du réseau. Le temps de réponse d'un serveur Web dépend en même temps des paramètres stables du réseau comme la vitesse de transfert des données et la bande passante, mais aussi de paramètres très aléatoires et instantanés. La même page peut également être chargée en des temps très variables en fonction de paramètres instantanés comme le trafic sur le réseau et l'heure de communication.

---

[13] L'Association Européenne pour la Traduction Automatique (*European Association for Machine Translation* - EAMT) : “Il s'est avéré que cet objectif était difficile à atteindre. Cependant il existe aujourd'hui un certain nombre de systèmes produisant un résultat qui, s'il n'est pas parfait, est de qualité suffisante pour être utile dans certaines applications spécifiques, en général dans le domaine de la documentation technique. De plus, les logiciels de traduction, qui sont essentiellement destinés à aider le traducteur humain à produire des traductions, jouissent d'une popularité croissante auprès d'organisations de traduction professionnelles.”

Chapitre 3

La recherche des documents sur le Web

Avec la nature distribuée d'Internet et le Web, trouver des informations pertinentes est une tâche souvent difficile. Retrouver un document parmi des millions présents sur le Web représente également une difficulté. Sur Internet, il y a des millions de serveurs, chacun possédant des milliers de documents. Avec l'apparition d'outils spécialisés, la tâche est devenue plus simple, mais ne pourra jamais être une opération triviale. Chaque recherche demande une approche unique et une bonne stratégie. Il faut connaître des utilitaires et des méthodes appropriées.

L'exploration de WWW au moyen d'un navigateur est souvent une opération amusante ou stressante, mais rarement efficace. Quand on cherche une information spécifique, sans connaître l'endroit où elle se trouve, le balayage peut être long et frustrant. Dans ce cas, mieux vaut faire appel aux outils de recherche spécialisés, qui possèdent d'énormes bases de liens vers les documents du Web. Il existe beaucoup de types de serveurs sur Internet, les serveurs FTP, les Usenews, les serveurs WWW, etc. Il n'est pas possible de faire une recherche d'information en même temps sur tous les serveurs du Web. Il faut de préférence utiliser les outils spécialisés. Ces outils fournissent différents services aux utilisateurs. En fonction du type de services fournis, on peut les classer dans les catégories suivantes :

- les *browsers*, outils qui récupèrent les documents à partir des adresses données par les utilisateurs,
- les outils de recherche (catalogues, moteurs et méta-moteurs de recherche),

- les outils qui ont la possibilité d'envoyer à l'utilisateur des documents automatiquement, notamment les nouveaux documents, etc (service *PUSH*),
- les outils qui peuvent analyser et diagnostiquer le profil des utilisateurs,
- les outils qui peuvent analyser les informations contenues dans les documents récupérés par l'utilisateur.

3.1 Différents outils de recherche

Toute recherche d'information sur le Web a pour but principal de trouver les documents les plus pertinents répondant à une question posée par l'utilisateur. Mais cette tâche est complexe et nécessite l'utilisation d'outils spécialisés. Actuellement, selon la méthode utilisée, les outils de recherche se divisent en cinq groupes :

- les catalogues par sujets,
- les moteurs de recherche,
- les méta-moteurs de recherche,
- les bases de données *on-line*.
- les portails.

3.1.1 Les catalogues par sujets

catalogue	taille [Bar00b]	recherche phrases	recherche boolèenne	troncature
Librarians' Index www.lii.org	$>$ 7.000	non	AND, OR, NOT	avec *
Infomine infomine.ucr.edu	$>$ 20.000	oui	AND, OR	automatique
Academic Info www.academicinfo.net	non fournie	non	AND, OR, NOT	non
Yahoo ! www.yahoo.com	~1.000.000	oui	+,-	avec *
About.com www.about.com	~1.000.000	oui	+,-	avec *

TAB. 3.1 – Le catalogue de recherche.

[Bar00b] Les ressources Internet dans les catalogues par sujets sont organisées de manière hiérarchique : les documents sont classés dans une arborescence de thèmes et de sous-thèmes. Les catalogues sont créés avec une intervention humaine ; souvent, à partir d'une base de données créée par des robots, le classement est ensuite établi manuellement par des spécialistes des domaines. Les plus connus de ces outils à couverture mondiale sont Yahoo ![1], Librarians' Index[2], Infomine[3], Academic Info[4], About.com[5], Galaxy[6], Magellan[7], et Lycos[8].

Pour les catalogues francophones, les plus connus sont Lycos France, Yahoo France[9] et Nomade[10].

Cette méthode de catalogues classés par sujets apporte une valeur ajoutée de sélection, validation et catégorisation des ressources. Elle limite le bruit par rapport aux méthodes automatiques, mais la mise à jour est beaucoup moins rapide et la couverture beaucoup moins large. Enfin, la catégorisation de documents manuelle est souvent plus juste, mais en même temps assez dépendante des facteurs comme : la couverture géographique et la diversité linguistique, culturelle, et thématique.

Les caractéristiques des catalogues de recherche les plus connus sont détaillées dans le tableau 3.1.

3.1.2 Les moteurs de recherche

Les moteurs de recherche sont des outils permettant d'effectuer des recherches sur une requête posée par l'utilisateur. Pour modéliser un moteur de recherche, on peut définir deux modules :

- le module qui est responsable de la collecte des informations et de l'indexation en vue de construire la base d'index
- le module interrogé par l'utilisateur qui, à partir de sa base d'index, fournit les réponses.

Le fonctionnement général des moteurs de recherche est présenté dans la figure 3.1.

Les différents moteurs utilisent des méthodes et des algorithmes très variés dans leur fonctionnement. Si on veut caractériser les moteurs en fonction de la méthode utilisée

---

[1] http ://www.yahoo.com
[2] http ://www.lii.org
[3] http ://infomine.ucr.edu
[4] http ://www.academicinfo.net
[5] http ://www.about.com
[6] http ://www.galaxy.com
[7] http ://www.magelan.com
[8] http ://www.lycos.com
[9] http ://www.yahoo.fr
[10] http ://www.nomade.fr

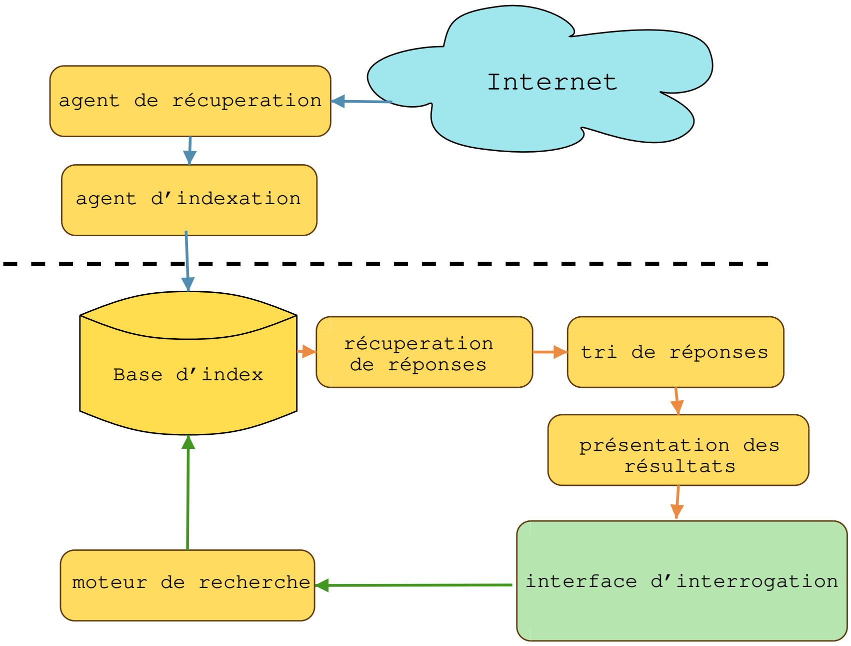


FIG. 3.1 – Le modèle de moteur de recherche.

au cours du processus de création de base d'index, on peut choisir les caractéristiques suivantes :

- les moteurs qui utilisent des robots,
- les moteurs basés sur les signets,
- les moteurs basés sur le principe de la topologie des liens.

Un même moteur pouvant également appliquer plusieurs de ces principes à la fois. Les caractéristiques des moteurs de recherche les plus connus sont détaillées dans le tableau 3.2.

3.1.3 Méta-moteurs de recherche

[Ben98, Abi99, Chu96, DBP94a, DH97a, DH97b, DBP94b, Dre96a, Dre96b, Gre00, HHN00, Ham96, KT00, LG98a, LG98b, LG98c, Laa97, Lar97, Liu98, Mik97, Not98, RC99, SE95, Sai99, Sta98, Sta99, Tan98, Tom99, Wei98, Weg96, LCK$^+$97]

Les moteurs de recherche simples sont très utiles pour la recherche d'information sur le Web, mais ils possèdent aussi de nombreuses faiblesses. Chaque moteur possède sa propre base de documents. Même si, en général, ces bases sont très importantes,

	taille[Not01]	phrase	recherche booléenne	default	truncature	limitation	classification	traduction
Google www.google.com	~ 700 milions pages indexées	oui	- OR	and		langage, number, data, domain, d'autres	pertinence	oui
Alta Vista www.altavista.com www.av.com	~ 500 milions	oui	AND, OR, NOT, +, -, ( )	or	oui	langage, number, data, domain, d'autres	pertinence	oui
Northern Light Power www.northernlight.com/ power	>300 milions	oui	AND, OR, NOT, +, -, ( )	and	oui	type de doc, data, domain, d'autres	pertinence	
Fast Search www.alltheweb.com	>600 milions	oui	OR, +, -, ( )	and	non	langage, number, domain	pertinence	
MSN Search www.msn.com.com	~500 milions	oui	AND, OR, NOT, +, ( )	and	non	langage, data, d'autres	pertinence	
iWon www.iwon.com	~500 milions	oui	AND, OR, NOT, ( )	and	oui	aucun	pertinence	
HotBot www.hotbot.com		oui	AND, OR, NOT, +, ( )	and	oui	langage, data, d'autres	pertinence	
Excite www.exite.com		oui	AND, OR, NOT, +, -, ( )	or	non	domain, langage	pertinence	
NBCi www.nbci.com		oui	-	and	oui	langage, data	pertinence	

Tab. 3.2 – Comparaison de différentes moteurs de recherche.

elles ne peuvent contenir qu'un sous-ensemble de toutes les ressources disponibles sur le Web. Leur couverture est donc insuffisante [Figures 3.2 et 3.3].

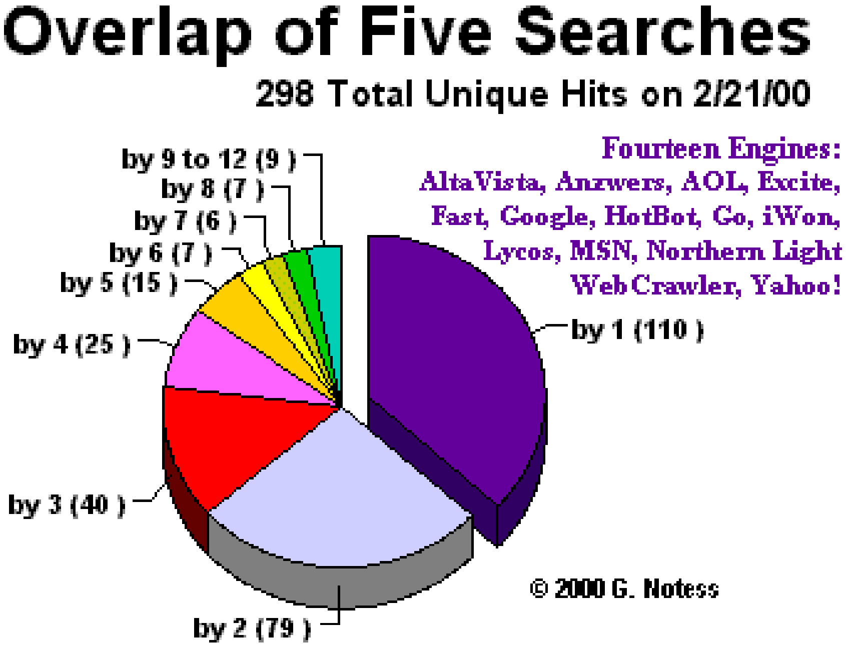


FIG. 3.2 – Comparaison de couverture des différents moteurs.

De plus, ces bases de documents sont mises à jour plus ou moins régulièrement. En conséquence, il est assez fréquent d'être dans l'obligation d'interroger plusieurs moteurs pour obtenir l'information recherchée.

Il serait donc souhaitable d'avoir un outil plus général avec une couverture plus large, et une interface uniforme pour poser la requête simultanément à plusieurs moteurs, selon les besoins, et obtenir l'affichage des résultats des recherches dans un format cohérent. Cette fonctionnalité est réalisée par les outils appelés *méta-moteurs.*

On distingue deux types de méta-moteurs : le méta-moteur sur le serveur distant et le méta-moteur client.

3.1.3.1 Le méta-moteur sur un serveur distant.

[Bar00a]

Nous proposons dans le tableau 3.3 une comparaison de méta-moteurs sur un serveur distant

méta-moteur	où cherche	combien moteurs	rêquete	résultats	temps de réponse
Debrefing www.debriefing.co	Web,	14	booléene		
Dogpile www.dogpile.com	Web, FTP, Usenet	25	booléenne, guillemets, trois sites à la fois	10/site sommaire, résultats independants	reglage max : 60 sec
ProFusion www.profusion.com	Web, Usenet	9	booléenne, phrase	liens vérifiés, dedoublonage, sommaire optionnel, max 50/site	
Highway61 www.highway61.com		6	ET , OU	résultats classés, sommaires	
Search.com www.search.com www.savvysearch.com	Web, Usenet, bases de données	700	booléenne, phrase	max 50/site, dedoublonage	
IxQuick www.ixquick.com	Web, mp3, images	14	langage naturel, booléenne, troncatures, parenthèses, (+),(-), MAJ/min,	dedoublonage, par pertinence, 10/site	
Internet Sleuth www.isleuth.com	Web, News, Usenet			titre, titre/sommaire, max 100/site	reglage max : 10-120sec
The BigHub www.isleuth.com		3000			
MetaCrawler www.metacrawler.com www.go2net.com	Usenet, files, stock quotes computer products,	12	phrase recherche avançée Power Search	dedoublonage, par pertinence, sommaire limitation par région geogr. ou par les sites (edu, com etc)	reglage t.max
MetaEureka www.metaeureka.com	Web, dépêches, News, musique, image, encyclopedies		booléenne, AND, OR, NOT, parenthèses, phrase	dédoublonnage	
MetaFind www.metafind.com		6	booléenne	par pertinence	10 -60 sec
Mamma www.mamma.com	Web, Usenet,news, stock symbols, company name, MP3, pictures	10	phrase, titre, booléenne	résumé, par pertinence dedoublonage	
SearchCaddy www.searchcaddy.com		6	phrase, troncature, booléenne	par pertinence	

TAB. 3.3 – Comparaison de quelques méta-moteurs sur serveur distant.

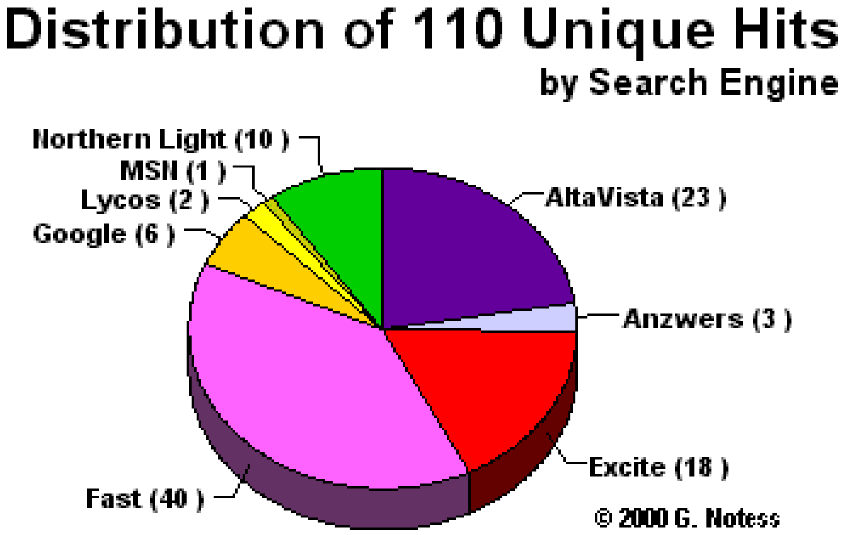


FIG. 3.3 – Comparaison de couverture de certains moteurs.

3.1.3.2 Le méta-moteur client

Un méta-moteur ne nécessite pas d'être exécuté sur une machine très puissante. Il est donc possible d'installer un client méta-moteur sur une machine quelconque. Cette solution est plus souple qu'un méta-moteur distant car elle permet de personnaliser le méta-moteur client. Des recherches très spécifiques peuvent être appliquées. On peut ajouter des modules spécifiques comme des filtres spéciaux éliminant les réponses non désirées (élimination de certains thèmes ou domaines) etc. Le méta-moteur client est donc configurable par l'utilisateur. Il peut conserver la liste des dernières questions et réponses sur sa propre machine, pour faire des mises à jour dynamiques. Un travail de veille est envisageable avec gestion de profils, surveillance des mises à jour en fonction d'un calendrier et de critères de filtrage. La plupart des méta-moteurs interrogent les moteurs généraux comme AltaVista, Excite, Lycos, InfoSeek etc. D'autres incluent également de la recherche sur USENET, ou encore sur les bases de données spécialisées. Les méta-moteurs clients sont créés généralement pour les systèmes commercialisés par Microsoft (sous Windows et Windows NT).

Nous comparons les possibilités de certains méta-moteurs clients dans le tableau suivant [Tableau 3.4].

méta-moteur	où cherche	nombre de moteurs	rêquetes	résultats
BullsEye www.intelliseek.com	Web bases de donnèes		étapes successives	résumés automatiques, (Search97 de Verity)
Copernic www.copernic.com	Web annuaires	20	booléenne, phrase, possibilité de rafiner	au fur et à mesure, par pertinence, doublonage, élimination des liens périmés
Hurricane Websearch www.gatecomm.com		12	booléene multiple	temps de réponse réglable
LexiBot www.lexibot.com/index.asp		142	moteurs paramétrables	par pertinence
NetAttaché Pro www.tympani.com/products.html			programation pour les mises à jour	
Umap www.umap.com		16	dictionnaire de mots-clés	dedoublonage, cartographie des réponses
Web Compass www.qdeck.com/qdeck/intl/		35	programation pour les mises à jour	avec résumé
Web Seeker www.bluesquirrel.com/products/seeker/webseeker.html	Web annuaires		booléenne	non triés
Jeeves www.askjeeves.com				aucune intégration, aucun rangement
Cyber411 cyber411.com		16	booléenne	seulement des titres et les urls, dedoublonage
WebFerret www.ferretsoft.com/	Web, auction, chat, ftp	20	booléenne phrase	

TAB. 3.4 – La comparaison des certains méta-moteurs client.

3.1.4 Les bases de données avec accès *on-line*

[Sul00, Bar00a]

Contrairement aux documents qui sont indexés par les moteurs et les méta-moteurs de recherche et les catalogues, les informations contenues dans les systèmes de bases de données ne peuvent pas être accédées de la même façon. Les bases de données font partie du Web “invisible”, les pages sont créées dynamiquement au fur et à mesure des demandes de l'utilisateur. Ces pages ne sont pas sauvegardées, et existent seulement temporairement pendant le temps de la consultation. Elles ne peuvent donc pas être retrouvées par les logiciels qui *aspirent* le Web, et ne peuvent pas être non plus indexées par les moteurs. La seule façon de consulter l'information contenue dans les bases de données *on-line*, est de les interroger directement via un formulaire (interface) spécial. Comme pour les méta-moteurs, il existe également des outils qui offrent la possibilité d'interroger plusieurs bases de données en même temps, à partir d'une interface homogène. Leur principe de fonctionnement est similaire à celui du méta-moteur, sauf qu'ils font leur recherche sur des ressources structurées. Nous comparons les possibilités de certains méta-moteurs clients dans le tableau 3.5.

outil	taille info général	téléchargable/ consult. on-line	l'eval. du base	rech.
The Invisible Web Catalog www.invisibleweb.com	10000 bases	oui/oui	oui	oui
Direct Search gwis2.circ.gwu.edu/~gprice/direct.htm		oui/oui depend des bases	oui	non
Internets www.internets.com/		oui/oui	non	rare
IncyWincy www.incywincy.com		oui/oui	descriptions	quelques
A Collection of Search Engines www.leidenuniv.nl/ub/biv/specials.htm		oui/ avec difficulté	quellques	non
Complete Planet www.completeplanet.com/	20000	oui/ pas direct.	non	non
Search Power www.searchpower.com/		oui/oui	non	non
Internet Oracle internetoracle.com/		non/oui	non	oui
Special Search Engines web.ntu.edu.sg/library/specialcat.htm		non/oui	descriptions	non

TAB. 3.5 – La comparaison de certains outils avec accès à des bases de données.

3.1.5 Les portails

Les portails sont des pages Web qui contiennent une multitude de liens dans un domaine spécialisé. Les liens sont classés selon différents thèmes et dans des catégories différentes, spécialisées par l'"expert" de chaque domaine. Ces experts, qui participent à la création du portail, sont des spécialistes, mais aussi souvent des agences, des associations, des hobbystes, etc.

Quelques exemples de portails mondiaux universels sont présentés dans le tableau 3.6.

portail	fenêtre de recherche
AlphaSearch www.calvin.edu/library/ searreso/internet/as/	par les sujets categories et descripteurs (alphabetical liste).
Argus Clearinghouse www.clearinghouse.net/	recherche et consultation on-line AND, OR, NOT, truncation
MM WWW Virtual Library conbio.rice.edu/vl/database	par les sujets, categories et alphabetical listes
The New Atheneum athenaeum/mguide1.htm members.spree.com/	sujets et categories.
Suite 101 www.suite101.com/	liste des sujets,

Tab. 3.6 – Exemple de portails.

3.1.6 D'autres outils pour la recherche

Il existe d'autres outils qui ne sont pas nécessairement indépendants des outils précédents mais qui y sont généralement intégrés. On peut les diviser en deux catégories :

- les robots pour *aspirer* les sites (*spiders* et *crawlers*),
- les agents de surveillance.

3.1.6.1 Les robots pour rapatrier des pages sur un site : aspirateurs, *spiders* et *crawlers*

Ce sont des agents qui suivent les hyperliens et récupèrent les pages Web. Ils peuvent également analyser les pages, extraire et régénérer les liens sur les pages aspirées, etc. Ils constituent le module essentiel de la plupart des moteurs de recherche mais existent également comme outils indépendants. Les caractéristiques de certains agents sont détaillées dans le tableau 3.7

outil		fonctionalités
Ecatch www.ecatch.com	navigateur hors ligne logiciel de veille aspirateur de pages	téléchargement, capture de sites veille sur les nouveautés analyse des pages chargées
MemoWeb www.memoweb.com/fr/indexfr.htm	aspirateur de pages	récupère différents formats images, applets, VRML, FTP
WebZIP www.spidersoft.com/webzip/home.htm	aspirateur de pages navigateur hors ligne	récupère différents formats image, MP3, video
Web Whacker www.bluesquirrel.com/whacker/	aspirateur de pages logiciel de veille navigateur hors ligne	veille sur les changements récupère différents formats accepte les filtres
Wget	aspirateur de pages	récupère différents formats

TAB. 3.7 – Les robots pour rapatrier des sites.

3.1.6.2 Les agents de surveillance

Ce sont des logiciels spécialisés, qui surveillent des sites spécifiés par l'utilisateur. Ils sont souvent intégrés dans des outils plus complexes. Généralement ils travaillent en arrière-plan, lancé pendant la nuit pour avoir des résultats le lendemain. Les principales fonctions réalisées par ces agents sont les suivantes :

- surveiller les modifications éventuelles des pages des sites indiqués,
- effectuer la mise à jour de bases de données,
- envoyer des courriers électroniques contenant les résultats des recherches (service *PUSH*),
- envoyer des signets *alarmes* en cas de panne, etc.

3.2 Modèle de recherche utilisé par les moteurs

3.2.1 Les différentes méthodes de recherche

Les modèles de recherche les plus fréquemment utilisés par les moteurs de recherche, sont :

- modèle basé sur la recherche booléenne,
- modèle basé sur les systèmes vectoriels,
- modèle basé sur d'autres systèmes pondérés.

3.2.1.1 La recherche booléenne

La recherche booléenne, qui a été décrite auparavant, [Section 1.1.1] est la méthode utilisée la plus fréquemment. Son principe est basé sur une simple comparaison de chaînes de caractères. L'utilisateur définit ses besoins par la requête sous forme d'une fonction booléenne composée de mots clés, qui sera transmise au moteur pour l'interrogation. Ensuite, le moteur recherche dans son fichier d'index les termes correspondant à ceux de la requête. Ainsi, les documents sélectionnés pour la réponse et considérés comme pertinents, sont ceux qui satisfont la fonction booléenne de la requête.

Souvent, selon le moteur interrogé, la fonction booléenne est écrite de manière différente, au niveau de la combinaison des opérateurs, des parenthèses, de la troncature des mots, etc. Par exemple, la plupart des moteurs utilisent par défaut un opérateur implicite entre les mots, qui est le "OU" logique (ang. *OR*). Ceci implique, que la recherche n'est pas limitée, et le moteur ne retourne pas en réponse une liste vide de documents. Mais il existe aussi les moteurs[11] qui utilisent l'opérateur "ET" logique (ang. *AND*) comme opérateur implicite. Ces différents modes d'interrogation montrent que la recherche booléenne n'est pas toujours facile à utiliser.

En plus, avec la méthode booléenne, il n'y a que deux états de classement possibles : le document répond ou non à la question. Donc, dans les systèmes basés seulement sur le modèle booléen, mesurer la pertinence d'un document par rapport à une question est difficile. De plus, le moteur de recherche retourne une liste de documents pertinents à la question, mais non triés entre eux.

Pour classer les réponses selon leur pertinence, la méthode basée sur le modèle booléen n'est pas suffisante. Pour corriger le manque de possibilité d'affiner au niveau de la pertinence, les moteurs booléen actuels utilisent d'autres méthodes. Dans ces méthodes, pour évaluer la pertinence, les moteurs prennent en compte le nombre d'occurence des mots de la question dans le document, présence ou absence des termes

---

[11] HotBot : http ://www.hotbot.com
Lycos : http ://www.lycos.com

de la question dans le titre du document, popularité du site auquel appartient le document, etc.

3.2.1.2 Les systèmes vectoriels

Les systèmes vectoriels [Section 1.1.2] sont basés sur le modèle de Salton [SM83].

Dans cette méthode, tous les documents du corpus possèdent une représentation vectorielle dans un espace documentaire *t-dimensionel* [Page 1.1.2] (où t est représenté par le nombre de termes dans le corpus). Pour effectuer une recherche dans un espace à t dimensions composé de n documents, il suffit donc de modéliser la requête de l'utilisateur par un vecteur de termes.

Ensuite, pour retrouver les documents répondant à la recherche et pour mesurer leur pertinence en fonction de la requête posée, on calcule un coefficient de proximité entre les vecteurs de la question et le vecteur de chaque document.

Le coefficient de proximité obtenu prend en compte la similarité des termes utilisés dans la question et dans le document analysé, tout en tenant compte aussi du poids des termes dans chacun d'eux. Une telle fonction de similarité est généralement définie comme un cosinus de l'angle entre les deux vecteurs.

Ce modèle, contrairement au modèle booléen, permet donc d'obtenir un bon classement des résultats. Il permet de calculer un degré de similarité entre la question et chaque document de la base, mais aussi entre chaque document dans le corpus de documents.

Par contre, le fait que la performance de ce système soit fortement liée à la topographie de l'espace vectoriel créé lors de l'indexation, provoque un processus de reindexation très complexe. Ce système n'est pas très bien adapté pour les moteurs de recherche sur Internet, parce que le Web est un espace documentaire de taille gigantesque, très dynamique et variable en même temps et la procédure de mise à jour des index n'arrête jamais. Il n'est donc pas conseillé d'utiliser la procédure de mise à jour de base d'index très coûteuse en temps et en ressource.

Malgré ses avantages, le système vectoriel est utilisé très rarement[12] par les moteurs de recherche sur le Web, et c'est le système booléen qui est le plus répandu.

---

[12] WebCrawler [Mot98] le moteur qui utilise la méthode vectorielle. http ://www.webcrawler.com

3.2.1.3 D'autres systèmes pondérés

D'autres systèmes[13] prennent aussi en compte la proximité sémantique entre les documents. Pour cela, ils utilisent à la fois les différentes méthodes de mesure de pertinence, comme le couplage des méthodes statistiques (calcul du poids des termes par rapport à leur apparition dans la collection de documents) et la méthode linguistique (normalisation des mots, extraction des mots composés, analyse des liens syntaxiques, etc.)

3.2.2 Les différents modes de création de base d'index par les moteurs de recherche

3.2.2.1 Les moteurs basés sur les robots

Les moteurs les plus répandus sur le Web sont ceux dans lesquels le processus de création d'une base de documents (base d'index) est effectué par les outils spéciaux appelés *robots* (aussi *spider* ou *crawler*). Ils visitent le Web, de page en page, en suivant les hyperliens pour aspirer les pages HTML du Web sur le serveur du moteur. Des millions de pages sont répertoriées de cette façon. Ensuite, les documents récupérés par le robot sont transmis à l'indexeur, qui analyse leur contenu et fait l'extraction des termes, pour associer chaque descripteur de la base à un document. Le moteur possède donc une immense base de mots-clés (descripteurs) associés à des URLs.

Généralement dans le processus de collecte des documents, plusieurs centaines, ou même plusieurs milliers de robots pour les très gros moteurs américains, agissent simultanément. Un seul de ces robots peut visiter plusieurs centaines de milliers de pages Web par jour.

3.2.2.2 Les moteurs basés sur les signets

Ces moteurs, pour récupérer les documents sur le Web, utilisent les signets personnels comme références des pages intéressantes et pertinentes. Ce type de moteur est assez spécifique, il possède une base d'index beaucoup moins large que les moteurs universels, par contre les pages indexées par ce moteur sont choisies parmi les

---

[13] SPIRIT, le moteur français commercialisé par la société TGID (http ://www.faust.ascode.fr/exposants/exps/tgid.htm).

pages, qui étaient déjà sélectionnées comme pertinentes. Généralement les sites cités dans les signets sont riches en l'information, donc sont probablement très bons pour commencer la recherche.

Les premiers outils basés sur ce principe sont déjà mis en service, par exemple : HotLinks[14] et BackFlip[15].

3.2.2.3 Les moteurs basés sur la topologie des liens

Ces moteurs examinent la topologie des liens [Paragraphe 1.3.2.2] sur les pages Web et regardent combien de sites proposent le lien vers une page donnée. Le classement des documents dépend donc du nombre de liens qui pointent sur ce document : plus ce nombre est élevé, plus la page en question est considérée comme pertinente.

Le moteur le plus connu basé sur ce principe est Google[16].

3.2.3 Problèmes liés à la récupération et à l'indexation des documents par les moteurs de recherche

La collecte des documents et la création de base d'index constituent une première étape nécessaire à la recherche d'information sur le Web. Cette étape semble être facile à effectuer, mais pose beaucoup de problèmes. Un grand nombre de documents ne sont pas facilement accessibles pour les moteurs de recherche : documents dynamiques, documents stockés dans des différents formats propriétaires, documents n'utilisant pas l'alphabet latin, etc.

3.2.3.1 Problèmes liés à la récupération des documents

Il n'est pas possible de chercher dans tous les documents existants sur le Web. Le problème n'est pas seulement de localiser ces documents, mais aussi de les récupérer,

---

[14] http ://www.hotlinks.com

[15] http ://www.backflip.com

[16] http ://www.google.com
Ce moteur fait en plus l'indexation de pages en format PDF.
D'autre part, Google est aussi le premier qui a proposé un moteur de recherche pour les service WAP (Web sur le téléphone mobile), capable d'effectuer des recherches non seulement sur des sites compatibles WAP, mais aussi sur tous les autres sites HTML connus de Google avec la conversion automatique de format HTML en WML.

les analyser, les sauvegarder en local (en entier ou en ne gardant que les parties significatives) et de faire une mise à jour régulière.

On peut estimer qu'il existe actuellement plus de 1 milliard de documents accessibles sur le Web, ce qui donne environ 15 TeraOctet de texte [Page 2.3.4]. L'exemple suivant montre les changements de taille d'une base de données [Figures 3.4 et 3.5].

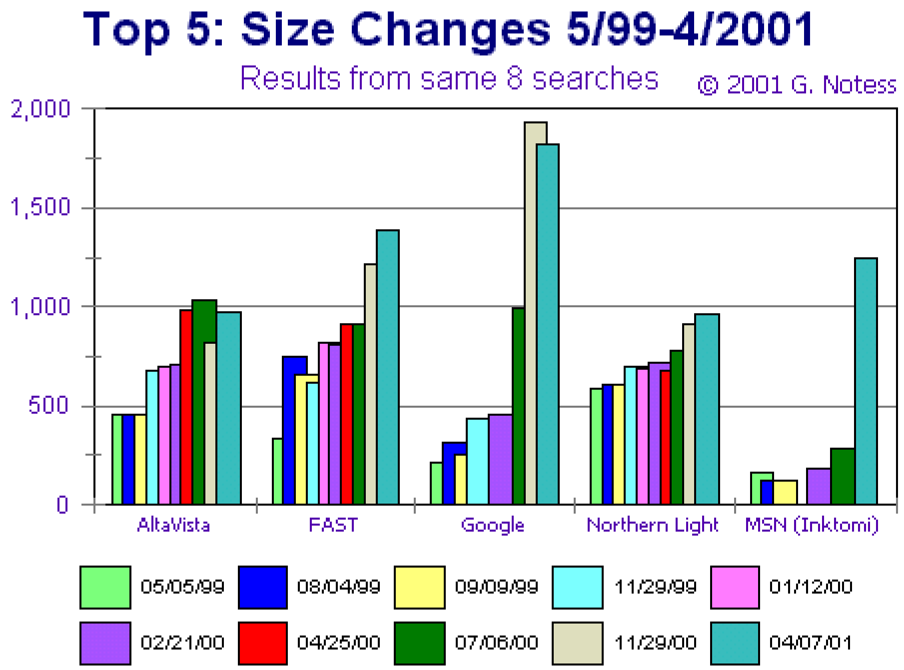


FIG. 3.4 – Statistiques de changement de taille pour les bases de données [Not01]

Les performances des réseaux ne sont pas non plus suffisantes : il faut prendre en compte que le réseau est de temps en temps saturé. Préparer une base de données initiale pour le moteur de recherche est une tâche qui peut durer plusieurs mois. La localisation des documents pose également problème. Une des méthodes pour localiser les documents sur le Web est de suivre les liens inclus dans les documents. Mais cette méthode n'est pas toujours très simple : jusqu'à 60% des documents n'ont qu'un seul lien qui pointe sur eux, et environ 90% des documents sont pointés par moins de 5 liens. La probabilité de trouver les documents est évidemment plus grande s'il existe beaucoup de liens qui pointent sur ces documents. Il existe sûrement des documents sur Web qui ne sont pointés par aucun document Web. Ces documents ne peuvent jamais être retrouvés par un robot. Il n'y a pas de moyen d'évaluer combien d'URL initiales sont nécessaires pour découvrir tous les documents sur le Web. De

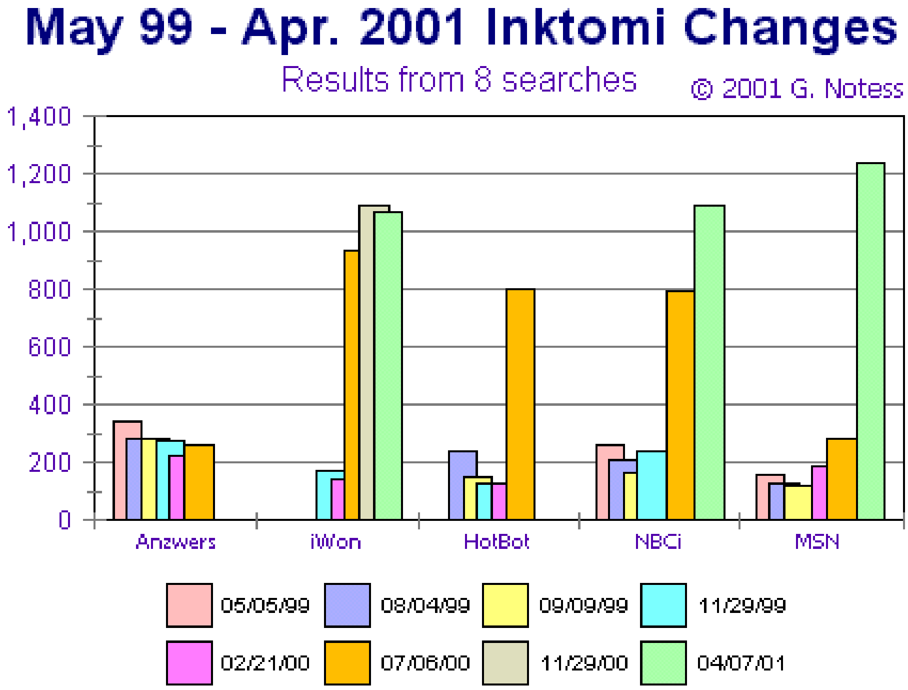


FIG. 3.5 – Statistiques du changement de taille pour les bases de données (pour les partenaires de Inktomi)[Not01]

nombreuses pages contiennent des liens vers d'autres documents, cachés sous une image cliquable. Quand l'utilisateur clique sur l'image, le navigateur renvoie la requête avec les coordonnées. L'interprétation de la requête dépend de ces coordonnées (exemple : une image représentant différents objets à choisir). Les pages pointées à partir d'images ne peuvent pas être retrouvées par les robots. D'autres documents peuvent être atteints uniquement au moyen d'un formulaire. Les champs du formulaire doivent être remplis avec des valeurs qui ne peuvent être déduites par le robot. Certaines pages sont créées dynamiquement par les cgi-scripts ou autres programmes exécutables. Sous la même URL, on trouve différentes informations dépendant des arguments passés au cgi-script. Pour les robots, il est difficile de trouver quels sont les arguments possibles et autorisés. L'accès à certains documents peut être limité par un mot de passe ou par le «*Robot Exclusion Protocole*»

3.2.3.2 La validité des liens et la mise à jour des bases d'index

Pour ne pas proposer à l'utilisateur des liens vers des documents inaccessibles, la fiabilité des liens [Section 2.3.3] est un problème assez important. Les documents sur le Web changent fréquemment : de nouveaux documents apparaissent, beaucoup sont supprimés, les liens entre les documents sont modifiés et la localisation des documents peut aussi changer. C'est pourquoi la mise à jour régulière de la base est souhaitable. Il est donc nécessaire de parcourir régulièrement l'ensemble des documents déjà indexés pour les réindexer. Avec l'augmentation énorme de puissance des ordinateurs en rapidité et en capacité pendant les dernières années, la plupart des moteurs de recherche actuels disposent d'une puissance de calcul très élevée, qui permet d'effectuer la mise à jour des bases d'index relativement fréquemment. Cependant, en raison de la croissance de la taille du Web, probablement aucune fréquence de mise à jour ne sera suffisante pour suivre tous les changements des documents. Dans le tableau 3.8, on présente deux études comparatives de moteurs de recherche en fonction de liens non valides.

Par ailleurs, la cause d'un lien non valide peut être le résultat d'un arrêt momentané ou d'une panne du serveur, la saturation du réseau etc.

Une solution, pour éviter de fournir aux utilisateurs les liens non valides, est de vérifier les liens lors de l'affichage des réponses. Pourtant c'est un processus assez coûteux en temps, en ressource, et qui en plus implique un surcroît de charge sur le réseau. Cette solution n'est donc pas utilisée par la plupart des moteurs de recherche.

Moteur	ètude par G.Notes	[Not01]	Lawrence.S et Giles.C.L [LG99]
	% liens non valides	% 400 error	% liens non valides
Alta Vista	13.7	9.3	6.7
Excite	8.7	5.7	2.7
Northern Light	5.7	2.0	9.8
Google	4.3	3.3	7.0
HotBot	2.3	2.0	2.2
Fast	2.3	1.8	-
MSN Inktomi	1.7	1.0	2.6
Anzwers	1.3	0.7	-
Lycos			14
Infoseek			5.5
Snap			2.8
EuroSeek			2.6

Tab. 3.8 – Les liens non valides. Le comparaison de moteurs de recherche .

3.2.3.3 Les pages dynamiques accessibles à partir de formulaires

L'accès aux pages dynamiques n'est pas évident. Ces pages, créées généralement via le formulaire pour chaque demande d'utilisateur, ne peuvent être indexées par les moteurs de recherche.

Pourtant, un nombre de plus en plus grand de documents est accessible de cette manière, et en outre ce sont des informations contenues dans des bases de données très intéressantes. En résumé, il reste encore une grande partie du Web non explorée par les moteurs de recherche.

3.2.3.4 Le moteur de recherche face au problème du multilinguisme

Les documents sur le Web existent dans des langues différentes [Page 2.3.6.3], et cette diversité linguistique est fortement liée à la couverture géographique.

Les moteurs qui indexent les documents spécifiques à une région géographique et linguistique peuvent utiliser un certain degré de connaissance de la langue. Cependant, pour les moteurs universels avec une couverture mondiale, le problème de multilinguisme devient assez important. Ces moteurs devront au moins avoir la possibilité de déterminer automatiquement la langue des documents récupérés pour l'indexation, ou encore mieux pouvoir effectuer tous les traitements linguistiques nécessaires pour retrouver des documents de différentes langues, à partir d'une question unique formulée dans une seule langue, ou enfin faire également une bonne traduction automatique [Page 2.3.6] dans la langue choisie par l'utilisateur (ou langue de la requête).

3.2.3.5 Le moteur de recherche face à différents jeux de caractères

Avec le multilinguisme de documents, on observe en même temps le problème de la diversité des jeux de caractères [Page 2.3.6.2]. Mais, contrairement aux problème de langue, où la principale tâche était d'identifier la langue du document, l'identification du jeu de caractères utilisé par le document est presque automatique. Une majorité des moteurs de recherche est capable de reconnaître et d'indexer le jeu de l'alphabet latin, qui appartient au code ISO-Latin1. Ceci est important pour le traitement des accents dans les langues accentuées. Mais ce n'est pas suffisant, car ce code ne couvre même pas la totalité des langues qui utilisent l'alphabet latin, sans compter que grand nombre de documents sont écrits dans des langues utilisant des alphabets totalement différents (russes, japonais, arabes, ...). Ces documents ne sont donc pas souvent référencés par les moteurs de recherche généraux.

3.3 Les étapes de la recherche

3.3.1 Formulation et analyse des questions posées

La formulation de la question est une étape très importante dans le processus de recherche, qui demande souvent une approche itérative. Le succès d'une recherche dépend beaucoup du choix des mots utilisés dans la requête. Il faut donc bien réfléchir aux mots clés utilisés, ne pas utiliser des termes qui ont une signification ambiguë, et enfin bien respecter le mode utilisé par le moteur (interrogation booléenne, liste de mots, langage naturel).

3.3.1.1 Opérateurs booléens dans la requête

Dans le mode booléen, la question est formulée par des termes séparés par des opérateurs logiques. Ces opérateurs sont exprimés différemment selon le moteur, même s'ils ont toujours la même signification logique. Par exemple, l'opérateur logique "*AND*" peut être utilisé dans le moteur comme "*and*", "et" ou "+". De même. l'opérateur logique "*OR*" peut prendre la forme de "*or*", "ou" ou espace, Enfin, le "*NOT*" logique peut être exprimé comme "*not*", "no" ou "-". Avec ces opérateurs, l'utilisateur peut écrire son expression logique très simplement.

Pour écrire une expression plus compliquée, il faut ajouter au jeu des simples opérateurs logiques, d'autres symboles comme les parenthèses. Mais cette possibilité n'est pas disponible dans la plupart des moteurs de recherche.

3.3.1.2 Troncature de mot

La troncature de mot permet de rechercher des mots contenant une chaîne de caractères indiquée dans la requête. L'opérateur de troncature (généralement "*") remplace un ensemble de caractères afin d'effectuer des recherches plus larges. On distingue trois types de troncature :

- la troncature droite,
- la troncature gauche,
- la troncature interne.

La troncature droite (la plus commune et souvent automatique) permet d'effectuer une recherche en utilisant le début d'un mot. De cette façon, on peut obtenir les différents termes, et élargir la recherche aux mots dérivés. Mais cependant, on risque d'ajouter aussi des mots n'ayant aucun rapport avec ce que l'utilisateur a choisi dans sa question.

Les troncatures gauche et interne fonctionnent de la même manière : elles permettent respectivement de rechercher des chaînes sans spécifier le début du mot, ou une partie interne du mot.

La fonction de troncature doit être utilisée avec précaution. En effet, la recherche devient beaucoup plus universelle, et en même temps peut provoquer du bruit dans les réponses obtenues.

3.3.1.3 Les opérateurs de proximité

Les opérateurs de proximité permettent de signaler au moteur de recherche s'il existe des liens syntaxiques ou sémantiques pour les mots indiqués dans la requête.

Ils permettent donc de poser plus précisément la requête et d'éliminer les réponses qui contiennent les mots de la requête, mais qui ne sont pas liés entre eux.

Par contre, les algorithmes réalisant cette opération ne permettent pas de spécifier qu'il doit exister un lien syntaxique entre les mots. Généralement, ils considèrent qu'un document est pertinent en fonction de la distance des mots indiqués dans la requête.

3.3.1.4 Syntagmes nominaux (phrases)

Beaucoup de moteurs de recherche permettent de faire la recherche d'une phrase (chaîne de caractères) exacte. Le fonctionnement interne est basé sur un "*pattern matching*" approximatif selon des algorithmes spécifiques.

3.3.2 Récupération et vérification (évaluation) des réponses obtenues

Après l'interprétation d'une requête par le moteur, l'étape suivante est l'interrogation de la base d'index du moteur afin de retrouver les résultats (les pages Web) conformes à la requête. Dans cette étape, il est essentiel de bien trier les résultats, de supprimer tous les documents trop éloignés de la question, et de classer les réponses sélectionnées dans le bon ordre. Ce classement des documents-réponses fournis par les moteurs est un aspect important pour l'utilisateur. Le plus souvent les résultats sont classés par ordre de pertinence. En effet, l'utilisateur trouve beaucoup plus vite les documents pertinents si le classement est bon. Les méthodes utilisées par les moteurs sont pour la recherche des réponses et pour leur classification très variées [Section 3.2]. Les critères le plus souvent utilisés pour classer les réponses par rapport aux mots de la question sont :

- Le nombre d'occurrences du mot dans le document[17].
- La présence du mot dans le titre du document[18].
- La présence du mot dans l'URL du document[19].
- La présence du mot dans les premières lignes du document[20].
- La fréquence inverse d'occurrences du mot dans la base[21].

---

[17] Alta Vista, Galaxy, HotBot, InfoSeek, Lycos, Magellan, WebCrawler
[18] HotBot, Lycos, Magellan
[19] Magellan
[20] Alta Vista, Lycos
[21] InfoSeek

- La popularité du site dans lequel les mots de la question sont trouvés[22]. Cette popularité est estimée par la méthode de topologie de liens [Section 1.3.2.2].

Le classement des réponses par les moteurs de recherche est essentiel pour l'internaute qui cherche des informations, mais également pour le *Webmaster*. Faire apparaître son site bien placé dans la liste des réponses fournies par les moteurs est devenu aussi une préoccupation pour le propriétaire du site. Il existe donc des techniques spécifiques pour qu'une page HTML soit présente en tête des listes de résultats des moteurs.

Les astuces les plus connues pour faire apparaître une page HTML dans les meilleurs scores fournis par les moteurs de recherche [And98], sont :

- répéter de nombreuses fois le même mot dans la rubrique "mots clés", ou dans les commentaires de la page HTML,
- inclure un texte supplémentaire contenant des mots clés très populaires en les affichant dans la même couleur que le fond d'écran du navigateur,
- créer un grand nombre de liens dirigés depuis d'autres sites en direction de cette page.

3.3.3 Présentation des résultats de la recherche

La dernière étape dans le processus de recherche d'information par le moteur est la présentation des résultats à l'utilisateur.

Les résultats devront être affichés dans la forme qui facilite l'évaluation de la pertinence des documents-réponses pour récupérer le document qui répond au mieux à la question posée sans avoir besoin de consulter toutes les réponses pour vérifier.

Les informations les plus importantes affichées sur la page des réponses sont :

- URL (*Uniform Resource Locator*) du document.
  Indique la localisation du document, le protocole qu'il utilise.
- Un lien hypertexte vers le document.
  Donne la possibilité de le consulter *on-line* directement.
- Le titre de document.
- Le rappel de la question posée par l'utilisateur.
  C'est une information importante : elle permet de vérifier comment la question

---

[22] Google

était posée, d'éliminer les éventuelles fautes de frappe, de faciliter la réformulation de la question, et identifier la question sur la page des résultats.

- Le nombre de documents retrouvés par le système.
  Cette information permet à l'utilisateur d'évaluer la pertinence de la question. Un nombre trop élevé nécessite de préciser la question et de relancer la recherche. Un nombre faible indique plutôt la nécessité de réformuler la question de manière plus générale.
- La liste des termes retrouvés.
  Elle permet à l'utilisateur de mieux évaluer les réponses par rapport aux mots de la question.
- Le score de pertinence du documents.
  Il permet d'évaluer approximativement la pertinence d'un document par rapport à un autre.
- Un extrait ou un résumé du document.
  C'est une valeur ajoutée pour faciliter l'évaluation des réponses. Le plus souvent les quelques premières lignes de document, moins fréquemment le "résumé automatique" (par exemple Excite) et encore plus rarement le résumé rédigé manuellement (par exemple AliWeb, Magellan).
- La mise en évidence des mots de la question dans les documents-réponses.
  C'est un moyen pour faciliter l'évaluation de la pertinence d'un document.
- La taille du fichier qui contient le document.
  Elle permet d'estimer la longeur du document texte, et de prévoir le temps et l'espace nécessaires pour téléchargement.
- La date de dernière mise à jour du document.
  Cette information permet d'estimer immédiatement si l'information contenue dans le document est encore actuelle. C'est une information importante pour la veille informationnelle. (exemple : Galaxy).
- La date de dernière mise à jour dans la base d'index (date de visite par le robot).
  Un autre façon d'indiquer à l'utilisateur l'actualité de la base d'index en même temps que celle du document. (Alta Vista, HotBot).

Deuxième partie

Implémentation d'un méta-moteur FIM Métaindexeur

Chapitre 4

Modèle de méta-moteur

Dans le cadre de cette thèse, a été réalisé un méta-moteur nommé FIM-Métaindexeur[1], dont l'objectif est une aide pour la recherche de documents produits et mis sur le Web par les différentes administrations françaises. Dans ce chapitre, nous décrivons le modèle général de méta-moteur, puis nous analysons les différentes possibilités de réalisation d'un méta-moteur et présentons dans les chapitres suivants l'analyse des moteurs sélectionnés pour l'interrogation, la dépendance des paramètres de ces moteurs pour le choix des méthodes d'implémentation ainsi que la réalisation détaillée de FIM-Métaindexeur.

4.1 Le principe

Le méta-moteur est un outil de recherche qui permet d'exécuter une requête simultanément sur plusieurs moteurs de recherche. Généralement, les résultats de cette recherche sont évalués et présentés sous une forme unique. Le principe de fonctionnement d'un méta-moteur de recherche est présenté dans la figure 4.1.

4.2 Les avantages

Les méta-moteurs présentent beaucoup d'avantages. Les plus importants sont :

- une couverture plus large grâce à l'interrogation de plusieurs moteurs,

---

[1] http ://veneux.ensmp.fr/metaindexeur/meta.cgi

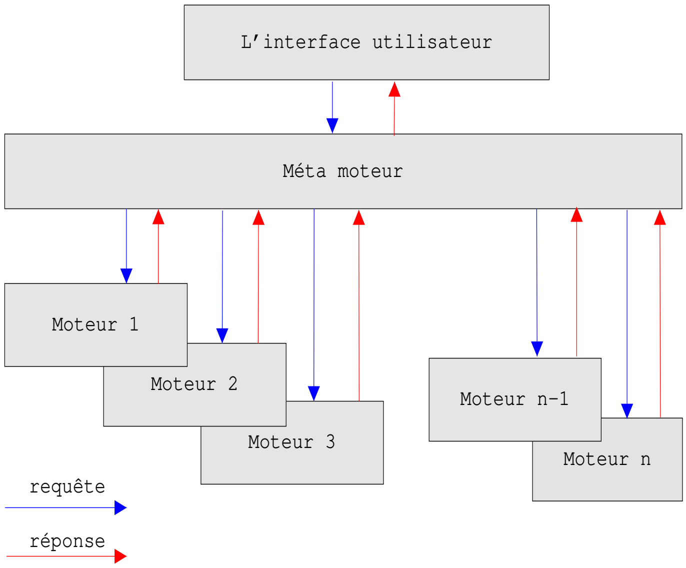


FIG. 4.1 – Schéma du méta-moteur.

- une forme de requête unique pour les différents services,
- une présentation intégrée des résultats avec classement des réponses des différents moteurs par ordre de pertinence,
- la possibilité de vérifier les liens, d'éliminer les doublons, etc.,
- l'absence de dépendance d'une seule base d'index,
- une faible exigence au niveau de la performance d'une machine (disques, CPU etc...).

4.2.1 Interrogation simultanée de plusieurs moteurs

Les méta-moteurs interrogent simultanément différents outils de recherche pour fournir les réponses pertinentes à la question posée. L'exploration des bases des différents outils de recherche par les méta-moteurs élargit leur couverture [Section 3.2 et Section 3.3], et offre donc la possibilité d'une meilleure pénétration du Web.

4.2.2 Uniformisation des requêtes

Un des avantages les plus importants des méta-moteurs de recherche est de donner la possibilité à l'utilisateur de formuler la question destinée aux différents outils sous une forme unique. Chaque moteur de recherche accepte les requêtes dans des formes très variées mais en même temps très spécifiques [Section 3.2]. Le méta-moteur a besoin de faire une analyse très approfondie de toutes les formes de requêtes acceptées par chaque moteur interrogé, pour connaître la structure des requêtes et le mode de recherche qu'il accepte. Grâce à cette connaissance, il est capable de transformer la question unique de l'utilisateur dans les formes spécifiques à chaque moteur de recherche interrogé. L'utilisation des méta-moteurs dispense l'utilisateur de connaître ces diverses formes de requêtes, qui peut ainsi, avec une seule requête standard de méta-moteur, interroger simultanément plusieurs outils.

4.2.3 Présentation homogène des réponses

Les réponses renvoyées par les moteurs interrogés sont analysées pour fournir à l'utilisateur les documents les plus pertinents, valides, sans doublons, et rangés par ordre d'importance.

Les résultats sont présentés sous une forme intégrée et homogène pour tous les moteurs. Ce mode de présentation des résultats venant de différents serveurs sous une forme homogène facilite l'évaluation des réponses par l'utilisateur. Pour choisir les meilleures réponses parmi toutes les réponses reçues, le méta-moteur doit effectuer certaines tâches supplémentaires :

- élimination des doublons,
- vérification des liens,
- évaluation et analyse de la pertinence des réponses reçues.

4.2.3.1 Fonctions supplémentaires

Élimination des doublons

Le méta-moteur vérifie si l'ensemble des réponses reçues de tous les serveurs contient des réponses identiques. Dans ce cas, il élimine toutes les réponses en double et n'affiche la référence du document qu'une seule fois sur la page des réponses. Cependant, le fait qu'un document soit cité par plusieurs moteurs peut être considéré comme significatif : dans ce cas, il est souhaitable d'attribuer un poids à ce document, matérialisé par un indice dépendant du nombre de fois où il a été cité par les moteurs. La fonction d'élimination de doublons permet à l'utilisateur de consulter la liste des documents-réponses plus rapidement, sans se préoccuper si le document est unique ou pas.

Vérification des liens Le développement rapide du Web [Sections 2.3.3 et 3.8] fait que les moteurs de recherche ne sont pas capables de mettre à jour leurs bases d'index assez rapidement. Il est donc assez fréquent que parmi les réponses fournies par les moteurs, il existe des documents dont les URLs ne sont plus valides. Le méta-moteur est capable de vérifier, pendant le temps où il effectue l'analyse des réponses, si les liens donnés par les moteurs sont valides et d'éliminer toutes les réponses qui ne sont pas accessibles. Cette fonctionnalité, très utile du point de vue de l'utilisateur, peut présenter l'inconvient de ralentir la recherche et de provoquer une charge supplémentaire du réseau.

Évaluation et analyse de la pertinence des réponses Le méta-moteur reçoit en réponse à son interrogation les pages de réponses des différents moteurs. Sur chacune de ces pages, les réponses sont généralement classées dans l'ordre d'importance, mais les algorithmes utilisés par ces moteurs pour faire ce classement sont très variés [Section 3.2]. Pour fournir à l'utilisateur la liste des documents-réponses classés dans le bon ordre, triés par ordre d'importance, le méta-moteur doit faire une évaluation de réponses reçues. Cette fonctionnalité permet à l'utilisateur de trouver plus rapidement les documents intéressants provenant de sources diverses.

4.2.3.2 Souplesse d'adaptation des méta-moteurs

La tendance actuelle de beaucoup de sites de recherche est d'offrir aux utilisateurs non seulement la recherche simple de documents HTML, mais également de fournir en même temps beaucoup d'autres services comme par exemple : réservation de billets d'avion, réservation de places pour des spectacles, achats dans les centres commerciaux, commandes d'articles, etc. Les outils basés sur le principe des méta-moteurs peuvent être très bien adaptés pour ces services. Leur principe de fonctionnement est d'interroger des sites différents et variés. Généralement, ce sont des outils assez souples et faciles à configurer. Ils peuvent donc être facilement adaptés pour la recherche dans différents domaines, la comparaison de ces offres et ensuite pour fournir les meilleures offres en fonction de critères choisis par l'utilisateur.

4.3 Les inconvénients des méta-moteurs

La méthode de recherche utilisée par le méta-moteur possède beaucoup d'avantages, mais présente en même temps certains inconvénients.

Les plus importants sont :

- le temps de réponse plus élevé que celui des moteurs de base,
- le temps consacré à l'interrogation d'un moteur est plafonné (*Time out*),
- la nécessité d'interroger des bases hétérogènes,
- la nécessité d'avoir accès à une base de données légale et confirmée,
- les problèmes de droit de publication, publicités etc.,
- le méta-moteur réalise souvent le nivellement par le bas des fonctionnalités de chaque outil, dont il élimine les particularités,
- la plupart des méta-moteurs passent seulement peu de temps dans chaque base d'index et n'exploitent souvent qu'une petite partie des résultats reçus.

Le temps de réponse plus élevé. Comme le méta-moteur interroge d'abord d'autres moteurs de recherche pour recevoir les réponses, le temps au bout duquel il peut fournir les résultats de recherche dépend totalement de la rapidité des moteurs de recherche. Pour faire l'analyse de tous les résultats reçus et présenter à l'utilisateur la liste complète des réponses, le méta-moteur est obligé d'attendre la totalité des réponses. En conséquence, le temps au bout duquel il peut afficher la liste complète des résultats est supérieur au temps de recherche du moteur le plus lent. Pour raccourcir le temps d'attente des résultats, le méta-moteur peut commencer à fournir les résultats au fur et à mesure, pour donner à la fin de l'évaluation de toutes les réponses, la liste définitive [LG98a, LG98b, LG98c].

Le temps de connexion plafonné (*Time Out*) Le fonctionnement du méta-moteur dépend totalement de l'état de chacun des outils de recherche utilisés. Pendant le temps d'interrogation par le méta-moteur, le serveur hébergeant le moteur de recherche peut être indisponible ou le temps de chargement trop lent. Dans ce cas, le méta-moteur ne pourra pas attendre la résolution de ce problème. Il est obligé de couper la communication avec le serveur. L'utilisateur a la possibilité de régler le temps (*Time Out*) entre un Temps minimal et un Temps maximal.

Le temps de réponse plus élevé et le temps de connexion plafonné sont les seuls inconvénients directement visibles par l'utilisateur. Les autres, bien que transparents pour l'utilisateur, sont assez importants pour le développement et le fonctionnement du méta-moteur.

L'interrogation limitée Chaque outil de recherche possède ses propres particularités. Pour bénéficier de toutes les capacités des outils interrogés, le méta-moteur devrait respecter (utiliser) ces particularités. Pourtant la plupart des méta-moteurs utilisent seulement les fonctionnalités de base sans exploiter des possibilités plus particulières. En outre ils passent peu de temps dans chaque base d'index et exploitent souvent une petite partie des résultats reçus.

L'accès à une base de données. Les problèmes de droit de publication, des publicités, etc. Un autre problème qui se pose au niveau du méta-moteur est celui de la consultation des bases d'index d'autres outils de recherche. L'accès à une base d'index de chaque outil devrait être autorisé, officiel et légal, et ne pas poser de problème au niveau des droits de publication, etc. D'autre part, les méta-moteurs, en affichant les résultats de recherche sur leurs pages de réponses, excluent toujours les annonces publicitaires, source principale de revenus pour les moteurs de recherche. Il est possible que le propriétaire d'un moteur particulier soit amené à interdire l'accès au méta-moteur pour protéger la propriété intellectuelle de son bien.

4.4 Les modules principaux

Le fonctionnement d'un méta-moteur est basé sur trois modules principaux :

- le module d'analyse des données en entrée,
- le module d'interrogation des moteurs,
- le module d'analyse des réponses récupérées.

Cette structure modulaire est présentée dans la figure 4.2.

4.4.1 Module d'analyse des données en entrée

Ce module réalise une fonction très importante. Pendant cette étape, on récupère toutes les information fournies par l'utilisateur (dans le formulaire HTML) via une interface spéciale et on fait l'analyse de toutes ces données : la question posée par l'utilisateur et les options choisies. La complexité de ce module dépend beaucoup du nombre de possibilités offertes par le méta-moteur.

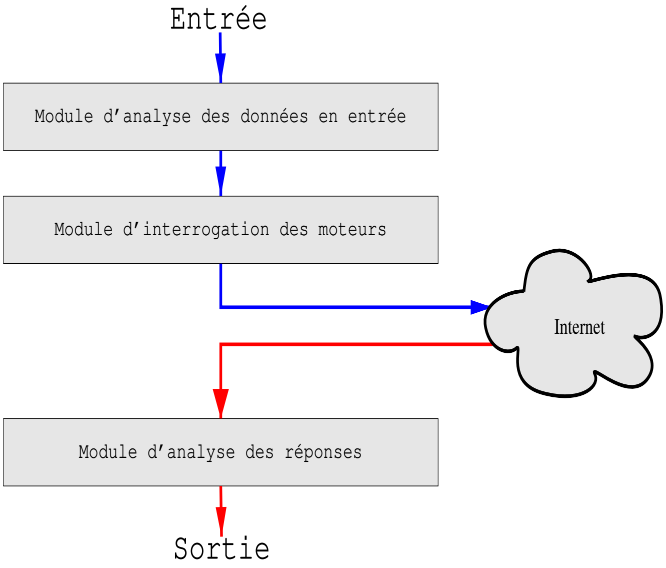


FIG. 4.2 – Les modules principaux de méta-moteur.

4.4.1.1 Analyse des données de paramétrage

La plupart des méta-moteurs donnent la possibilité de régler certains paramètres optionnels pour la recherche [Tableaux 3.3 et 3.4]. Ces paramètres sont d'abord analysés pour connaître le mode de recherche choisi par l'utilisateur.

Les données insérées dans les paramètres optionnels contiennent des informations de deux types : des informations pour le méta-moteur et les paramètres qui seront utilisés dans le processus d'interrogation et transmis au moteur en même temps que la questions posée dans une requête spéciale.

Après la récupération de tous les paramètres choisis via un formulaire, le méta-moteur fait l'analyse des données nécessaires à son fonctionnement telles que : les moteurs choisis pour l'interrogation, le temps global de recherche, les options d'affichage des résultats, etc. Il prépare la liste finale des moteurs sélectionnés pour la recherche et crée le profil de recherche général choisi par l'utilisateur. Ce profil sera ensuite utilisé dans le processus d'interrogation des moteurs choisis.

4.4.1.2 Analyse de la question posée par l'utilisateur

Dans ce module, le méta-moteur fait l'analyse de la question posée par l'utilisateur pour la soumettre au moteur dans la forme la plus adaptée. Toutes les modifications de la question posée dans ce module sont effectuées sur la forme de base insérée dans le formulaire. C'est cette forme qui sera utilisée ensuite pour préparer les requêtes adaptées à chaque moteur choisi pour l'interrogation. Le schéma général est présenté dans la figure 4.3.

Analyse sémantique : fonctions booléennes, opérateurs spéciaux
Dans un premier temps, le méta-moteur fait une analyse sémantique de la question posée par l'utilisateur. Cette analyse devrait donner au méta-moteur l'information nécessaire au niveau de la structure de la question posée. Il récupère donc toutes les fonctions booléennes et les opérateurs spéciaux utilisés et fait une analyse de toutes les relations entre ces fonctions.

Analyse lexicale : recherche multilingue, dictionnaire des synonymes
[LG98a, LG98c]

Le méta-moteur peut, en plus de l'analyse sémantique, procéder à une analyse plus profonde au niveau lexical de la question. Il utilise le plus souvent des dictionnaires de synonymes intégrés pour élargir la recherche. Dans ce cas, le méta-moteur remplace, dans la question posée, les mots par leurs synonymes et construit de cette façon une nouvelle question. Cette méthode permet de retrouver des documents pertinents, bien qu'ils ne contiennent pas les termes utilisés dans la question.

Une autre méthode très bien adaptée au méta-moteur est d'effectuer une recherche multilingue. Pour cela, le méta-moteur fait en plus des autres analyses, la traduction de la question posée dans une autre langue ou même plusieurs, et l'envoie dans cette nouvelle forme au moteur approprié à chaque langue.

4.4.1.3 Comparaison avec les questions déjà posées

[DH97a, DH97b, Dre96a]

L'ensemble des questions déjà posées avec leurs références sur les réponses pertinentes et les serveurs qui ont donné les bonnes réponses est une information très intéressante à exploiter. Ces informations sont sauvegardées dans des bases spéciales pour être utilisées ensuite pendant la recherche. Le méta-moteur recherche d'abord dans ces bases si une question similaire a déjà été posée. Dans ce cas, en fonction de ses possibilités, il peut utiliser les informations récupérées. Il peut par exemple :

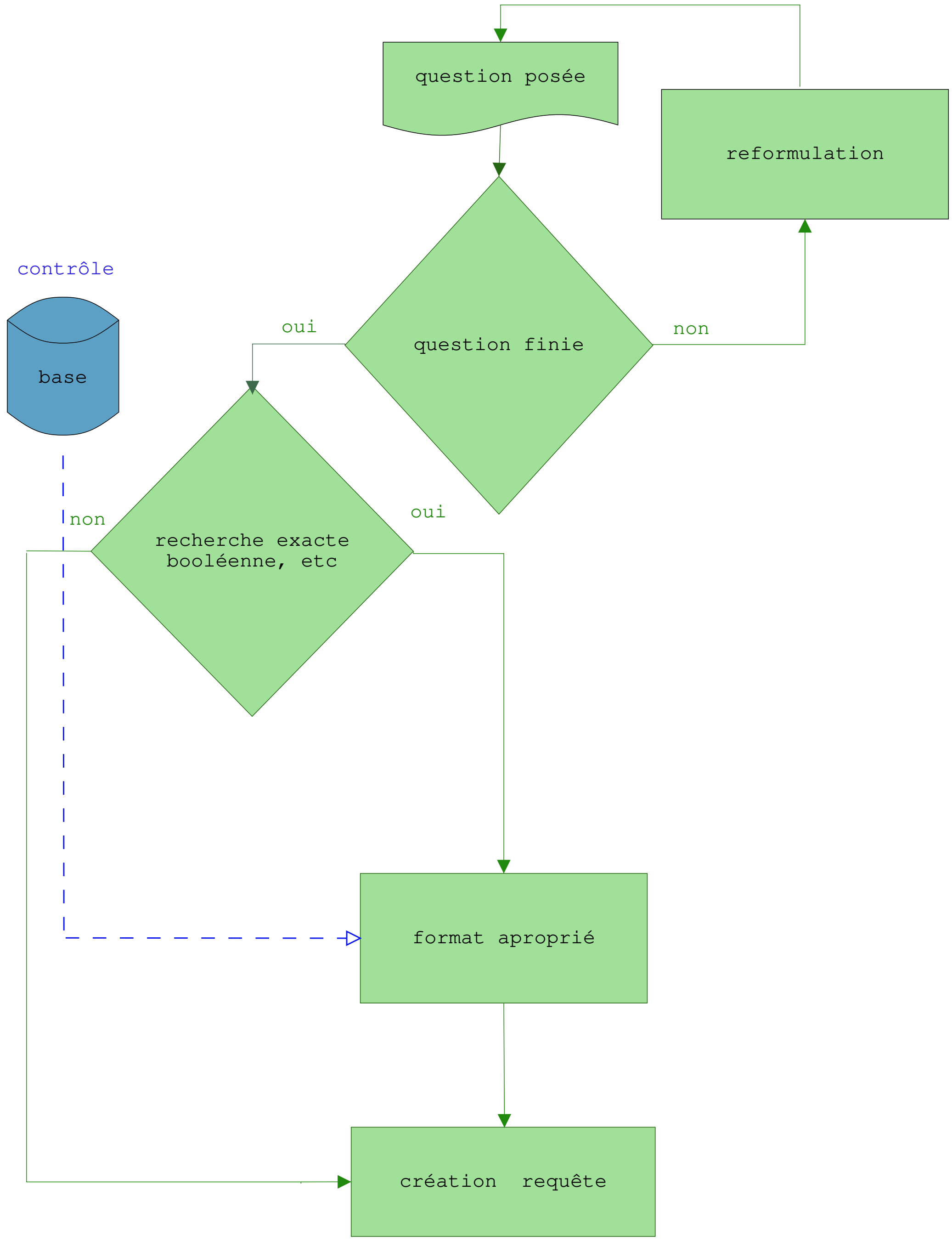


Fig. 4.3 – Analyse de la question posée. Le schéma général.

- récupérer directement les réponses sauvegardées,
- vérifier et actualiser les réponses,
- fournir au méta-moteur la liste des serveurs qui ont donné les bonnes réponses pour les questions similaires.

Cette méthode a l'intérêt de raccourcir beaucoup le temps de recherche et en même temps de diminuer la charge du réseau ; par contre, utilisée sans précaution, elle peut générer d'autres problèmes : si la base des questions déjà posées n'est pas mise à jour suffisamment fréquemment, on pourra passer à coté de réponses intéressantes.

Il est donc souhaitable de la coupler avec la méthode standard.

4.4.1.4 L'aide à la formulation de la question

[SC, Dus97a, Dus97b, Huy96b, JRP00, Jon98, Key96]

La bonne formulation d'une question est très importante dans le processus de recherche. La question devra être suffisamment précise. Pour cela, il faut choisir le terme juste, sans utiliser de mots qui peuvent avoir une signification ambiguë. Il est possible d'intégrer dans le méta-moteur un module spécifique pour aider l'utilisateur à bien exprimer sa question. Les méthodes utilisées dans le processus d'aide à la formulation de la question sont assez variées et généralement basées sur la méthode du "*feedback*" [Huy96b].

Après une analyse de la question posée, l'outil d'aide propose à l'utilisateur différents choix pour sa reformulation.

Dans cette partie de l'analyse de la question, on prend souvent en compte, non seulement la question elle-même, mais également les réponses récupérées ainsi que les données sauvegardées auparavant [Section 4.4.1.3].

4.4.2 Module d'interrogation des moteurs

4.4.2.1 Méthodes d'interrogation

Les méthodes d'interrogation des moteurs peuvent être divisées en deux groupes :

- l'interrogation séquentielle,
- l'interrogation simultanée.

Dans la méthode séquentielle, le méta-moteur interroge les moteurs les uns après les autres. C'est une méthode très lente, qui peut donner éventuellement de bons résultats pour l'interrogation d'un nombre de moteurs très restreint.

Dans la méthode simultanée, le méta-moteur interroge plusieurs moteurs en même temps. Généralement, ce sont différents agents spécialisés qui font en parallèle la recherche sur tous les moteurs.

4.4.2.2 Choix des moteurs pour l'interrogation

Le méta-moteur interroge plusieurs moteurs de recherche. La liste de moteurs à interroger peut être sélectionnée de différentes façons, ou elle peut rester fixe sans possibilité de modification.

Le choix des moteurs peut être effectué par la sélection manuelle ou la sélection automatique. Dans la sélection manuelle, l'utilisateur fait lui-même le choix des moteurs, en fonction de ses préférences ; très souvent dans ce cas, le méta-moteur possède une liste par défaut des moteurs à interroger (exemple : tous les moteurs disponibles).

Dans la méthode de sélection automatique des moteurs, le méta-moteur fait le choix en fonction de différents critères, par exemple :

- la classification thématique,
- le profil du moteur,
- la couverture géographique,
- le langage utilisé,
- le temps de réponse,
- la recherche déjà effectuée.

Dans ce cas, le méta-moteur choisit, en fonction de sa configuration, une liste des moteurs à interroger qui répondent au mieux aux critères sélectionnés.

Les méthodes de sélection automatique et manuelle ne s'excluent pas, elles peuvent être utilisées ensemble dans le même outil.

4.4.2.3 Transformation d'une question de l'utilisateur vers le format utilisé par chaque moteur interrogé

Pour envoyer la requête au moteur de recherche, il faut utiliser la forme acceptée par ce moteur. La question venant directement du formulaire, ainsi que toutes les questions modifiées (améliorations par la reformulation, questions ajoutées par utilisation de synonymes et également questions traduites dans d'autres langues [Section 4.3]) gardent leurs formes générales. Il est donc nécessaire de faire la transformation de la forme générale de la question vers le format accepté par chaque moteur interrogé.

Pour communiquer avec les moteurs de recherche, les méta-moteurs utilisent les requêtes standard du protocole HTTP. C'est dans la requête HTTP que le méta-moteur envoie la question posée par l'utilisateur et également en même temps tous les paramètres optionnels dédiés au moteur. Le format des requêtes est différent pour chaque moteur. Il faut donc que le méta-moteur sache quelle méthode du protocole HTTP est acceptée (GET ou POST), qu'il connaisse la structure des requêtes utilisées par chaque moteur, qu'il sache si le moteur accepte les fonctions booléennes, lesquelles sont acceptées, qu'il sache également passer les paramètres optionnels, etc. Toutes ces informations sont gardées par le méta-moteur dans une base spécifique.

4.4.2.4 Récupération des pages de réponses

Après avoir envoyé les requêtes aux moteurs sélectionnés, le méta-moteur attend les réponses pendant un certain temps : **le temps de réponse** [Section 4.3].

Toutes les réponses sont récupérées par le moyen du protocole HTTP comme les pages de réponses standard des moteurs de recherche. Cette forme de réponse est bien adaptée à l'utilisateur humain, mais l'est beaucoup moins pour une analyse automatique par le méta-moteur.

La récupération des seuls liens vers les documents-réponses pose au méta-moteur des problèmes particuliers. Les plus importants sont :

- le filtrage des bonnes URLs,
- la distinction entre les pages de bonnes réponses, les pages sans réponse, les pages avec un message d'erreur, etc.,
- le contrôle du nombre de réponses.

Filtrage des URLs Généralement la page standard de réponse fournie par le moteur de recherche est assez complexe : elle ne contient pas seulement la liste des résultats de la recherche, mais aussi beaucoup d'autres informations. De plus, ces informations ne sont pas figées, elles changent très fréquemment, souvent en fonction des thèmes de recherche, par exemple les publications thématiques sélectionnées automatiquement. Il est donc nécessaire de distinguer les URLs des réponses parmi toutes les autres. Pour réaliser cette fonction, le méta-moteur utilise des filtres spéciaux. Ces filtres, différents pour chaque moteur de recherche, sont créés manuellement. Les changements fréquents de la forme des pages de réponses, fournies par le moteur, demandent de faire des mises à jour des filtres assez souvent. Il est possible d'intégrer au méta-moteur un agent de veille spécial qui va transmette au méta-moteur un signal (alarme) indiquant les changements de forme des pages de réponses. Il faut ensuite actualiser le filtre approprié.

Différents modèles de pages de réponses Il est assez fréquent que les moteurs de recherche donnent les pages de réponses sous une forme différente en fonction de l'état de recherche, exemple : les pages avec un message d'erreur, les pages indiquant l'absence de réponse, etc. Le méta-moteur devra distinguer toutes ces formes, pour procéder au traitement approprié à chacun des cas.

Le nombre de réponses Généralement les méta-moteurs n'ont pas besoin de toutes les réponses venant de chaque moteur interrogé. Le choix du nombre de réponses à récupérer à partir de chaque moteur peut être fixé par un paramètre optionnel ou ce nombre est fixé par défaut. Il est donc souhaitable, pour ne pas charger inutilement le réseau, de contrôler le nombre des réponses fournies par chaque moteur interrogé. Ce contrôle dépend du mode de travail de chaque moteur :

- Le moteur interrogé envoie les réponses sur plusieurs pages. Dans ce cas, il faut contrôler le nombre de pages à récupérer.
- Les réponses sont présentées sur une seule page, mais le moteur accepte le réglage du nombre de réponses à envoyer. Dans ce cas, il faut envoyer ce nombre dans les paramètres de la requête.

4.4.2.5 Vérification des liens vers les documents-réponses

Les liens indiqués par les moteurs de recherche peuvent être non valides [Section 2.3.3 et 3.2.3.2]. Il serait donc souhaitable d'éliminer de la liste des réponses créées par le méta-moteur toutes les URLs non accessibles. Pour cela, le méta-moteur teste d'abord tous les liens. Généralement, cette procédure est réalisée par plusieurs agents de récupération travaillant en parallèle. Pour vérifier seulement la validité des liens, sans regarder le contenu des documents, les agents utilisent la requête HEAD. C'est une méthode du protocole HTTP pour ne récupérer que l'en-tête HTTP des documents ; cette méthode est bien adaptée pour diminuer la charge du réseau et suffisante pour vérifier l'existence du document.

Pour récupérer plus d'informations comme le titre du document, il est nécessaire d'utiliser la méthode GET ou POST, selon l'acceptation par le moteur. Pour ne pas récupérer la totalité de la page, il est possible de contrôler la taille du bloc à récupérer et de demander seulement une partie du document, par exemple l'en-tête du document HTML, etc.

Dans le cas où le méta-moteur effectue aussi la vérification des réponses, il est nécessaire de récupérer le document complet.

4.4.2.6 Récupération des documents-réponses

Pour faire la vérification de pertinence de réponses par rapport à la question posée, il faut analyser le contenu des documents-réponses. Il faut d'abord les rapatrier sur le serveur du méta-moteur. Cette récupération des documents est effectuée par plusieurs des agents travaillant en parallèle. Chaque agent répète le processus de récupération autant de fois que nécessaire pour atteindre le nombre de documents choisi auparavant dans la configuration du méta-moteur.

4.4.3 Module d'analyse des réponses récupérées

Les documents rapatriés par les agents de récupération sont traités par le module spécialisé d'analyse des réponses. La complexité de ce module dépend de la forme de vérification de réponses et des possibilités offertes par le méta-moteur. La réalisation la plus simple est de fournir à l'utilisateur une liste des réponses venant de tous les moteurs sans faire aucune vérification du contenu des documents. C'est une méthode rapide, avec une charge minimale pour le réseau, mais qui ne garantit pas la pertinence des réponses affichées sur la page de résultats. En même temps il est difficile de faire une bonne classification des réponses et la comparaison entre les documents en se basant seulement sur le classement et les scores des documents renvoyés par les moteurs interrogés. Il est donc souhaitable d'effectuer une vérification de pertinence des documents récupérés aussi au niveau de méta-moteur.

4.4.3.1 Vérification de la pertinence

Les indices d'évaluation de pertinence dans la recherche documentaire sont le rappel et la précision ainsi que les indices de bruit et de silence. Dans un système de recherche documentaire traditionnel, les indices de mesure de la pertinence sont le bruit et le silence [Section 1.3.2].

Dans le cas des méta-moteurs, il est impossible de connaître l'ensemble des bases d'index de tous les moteurs interrogés, alors on ne peut pas utiliser ces paramètres avec la même signification. Il est impossible de mesurer et de calculer du bruit et du silence avec exactement les mêmes définitions. Alors, le problème de l'élimination du bruit et du silence dans les résultats de recherche sur le Web doit être regardé différemment.

Le silence correspond aux documents pertinents qui n'apparaissent pas dans le résultat de la recherche. Sans la connaissance de la totalité du corpus (ensemble de toutes les bases d'index interrogées), le méta-moteur ne peut agir directement à ce niveau. Par contre, l'utilisation de méthodes pour élargir la recherche du méta-moteur (reformulation de la question, utilisation des synonymes, traduction en différentes langues, ou simplement augmentation de la couverture de recherche, etc.) doit accroître le nombre de réponses reçues, donc diminuer le silence. En contre-partie, ce moyen risque d'augmenter le bruit des réponses obtenues.

Le bruit correspond aux documents ramenés en réponse, qui ne sont pas pertinents par rapport à la question posée. Pour diminuer le bruit généré par le méta-moteur il faut effectuer une bonne vérification de la pertinence et éliminer les mauvaises réponses. Dans le cas du méta-moteur, on ne peut pas utiliser les mêmes moyens de vérification de la pertinence que pour un système centralisé avec une base d'index unique et homogène. La vérification de la pertinence par le méta-moteur présente plusieurs difficultés. On peut les classer en deux catégories : les problèmes liés aux moteurs de recherche et les problèmes liés au méta-moteur.

Problèmes liés aux moteurs de recherche Les difficultés les plus importants sont les suivantes :

- les moteurs de recherche utilisent des méthodes de recherche différentes,
- les algorithmes d'évaluation de la pertinence sont hétérogènes,
- les méthodes de classification des réponses sont différentes.

Les réponses fournies par les différents moteurs de recherche sont sélectionnées de façons différentes et leur classement est aussi très varié. Le méta-moteur doit donc prendre en compte toutes ces différences pour vérifier les réponses reçues et ne pas éliminer des réponses intéressantes (par exemple : l'élimination des réponses qui contiennent des synonymes).

Problèmes liés aux méta-moteurs Généralement l'analyse des documents récupérés est effectuée en temps réel, pendant le processus de recherche. Donc la rapidité de la méthode utilisée pour la vérification est aussi importante que son efficacité. Cependant la récupération des documents-réponses est à elle seule un processus assez coûteux en temps. Il est donc souvent nécessaire que le méta-moteur fasse un compromis entre qualité et rapidité pour choisir la méthode de vérification optimale.

4.4.3.2 Présentation des résultats

Le traitement des résultats par les différents méta-moteurs est très variable. Les réponses peuvent être affichées dans une forme plus au moins élaborée. Généralement, tout dépend des possibilités offertes par le méta-moteur. Alors, en fonction des fonctionnalités proposées par le méta-moteur, les données affichées sur les pages de réponses peuvent être très différentes. Par exemple les résultats peuvent être présentés sous les formes suivantes :

- directement comme les pages de réponses brutes pour chaque moteur,
- dans l'ordre de récupération,
- classées par source,
- avec élimination des doublons,
- avec élimination des liens non valides,
- triées par l'ordre de pertinence, etc.

Les résultats de la recherche effectuée (définitifs ou temporaires) peuvent être affichés immédiatement, au fur et à mesure de leur récupération ou après un traitement sur l'ensemble des réponses obtenues.

Différents modes d'affichage La présentation des pages de réponses fournies par les différents méta-moteurs est très variée. Les résultats de recherche sont présentés sous une forme textuelle ou graphique.

La forme la plus utilisée est la forme textuelle : les résultats sont présentés sur une liste ou peuvent être regroupés dans un tableau.

La forme graphique offre des possibilités différentes, par exemple la cartographie des réponses. La carte graphique peut représenter les réponses dans des formes très variées et contenir des informations supplémentaires.

Exemple : Umap[2] et Kartoo[3] sont deux méta-moteurs qui affichent les résultats sous forme graphique. Dans Umap, la représentation graphique est une mosaïque de morceaux colorés, qui représente les mots regroupés dans des zones de proximité. Kartoo par contre, crée une carte thématique avec la sélection des meilleurs sites choisis parmi les réponses reçues.

4.4.4 Coopération entre le méta-moteur et les moteurs de recherche interrogés

4.4.4.1 L'échange d'informations

Les méta-outils actuels interrogent les moteurs de recherche en leur envoyant la même requête que celle qui est utilisée par chaque moteur sur sa page de formulaire de recherche. En retour de cette requête, il obtient les pages de réponses standards. Il n'y a aucune information supplémentaire échangée entre le moteur et le méta-moteur. Pourtant, l'ajout de quelques informations spécifiques dans le flux de données des deux côtés de l'échange pourrait améliorer fortement le fonctionnement du méta-moteur, comme par exemple l'insertion, du côté du méta-moteur, d'informations précisant la demande, telles que le nombre et le format des réponses, ou également l'envoi au moteur interrogé des données élaborés par le méta-moteur concernant la qualité des réponses.

Il serait donc envisageable d'élaborer une méthode (protocole) spéciale pour échanger des informations entre le méta-moteur et les moteurs. Ce genre de "protocole" pourrait utiliser le protocole HTTP pour réaliser les transferts de données. Il est évidemment nécessaire que ce "protocole" soit implémenté de chaque côté. Pour cela, il faut mettre des clients applicatifs spécifiques de chaque côté, qui vont gérer le service de la communication entre le méta-moteur et les moteurs interrogés. Pour questionner le moteur de recherche, contrairement à la méthode d'interrogation par

---

[2] http ://www.umap.com

[3] http ://www.kartoo.com

la requête standard, le “protocole” spécial permettra au méta-moteur d'exprimer sa demande avec plus de précision, par exemple en demandant explicitement le nombre de réponses dont il a besoin. Du coté des réponses envoyées par le moteur, il suffit de filtrer l'information brute et d'éliminer les données inutiles pour le méta-moteur qui sont affichées normalement sur la page de réponses standard.

La possibilité de communiquer entre le méta-moteur et les moteurs avec des moyens spécialisés aurait beaucoup d'avantages pour le fonctionnement du méta-moteur, mais également pour les moteurs et pour la charge du réseau.

Les principaux avantages, côte méta-moteur, sont :

- la facilité de récupération des réponses,
- l'élimination d'une partie par le filtrage des données sur la page des réponses,
- la possibilité d'avoir l'information concernant des changements sur le moteur (les pages de réponses changent assez fréquemment, et il est difficile de suivre ce changement par le méta-moteur),
- la possibilité d'automatiser l'insertion d'un nouveau moteur, de modifier un moteur déjà intégré (processus qui nécessite normalement une intervention manuelle).

Côte moteurs de recherche interrogés :

- la diminution de la charge du moteur au niveau des services standards,
- la possibilité de donner au méta-moteur des directives spéciales.

Côte réseau :

- diminution de la charge du réseau.

Un de premier protocole de ce genre, protocole *START (Stanford Protocol Proposal for Retrieval and Search)* a été proposé par les scientifiques le centre de recherche de l'Universite Stanford en 1997 [GCGMP97].

4.4.4.2 Standard Formel d'utilisation des moteurs par le méta-moteur

Les méta-moteurs sont de plus en plus développés et utilisés. Une proposition a été faite pour définir un standard d'autorisation d'accès par un méta-moteur [Dre96b]. Il n'y a pas d'obligation de l'utiliser, donc on ne peut pas garantir que tous les méta-moteurs vont le respecter. Ce n'est qu'une proposition, offrant pour la communauté du Web le moyen de faciliter la coopération entre les propriétaires de bases de données et les méta-moteurs. L'utilisation de ce standard permettrait aux dévelopeurs de méta-moteurs de suivre les mêmes règles. La réalisation de ce standard pourrait être dans une forme analogue au "*Robot Exclusion Standard*", c'est à dire sous forme d'un fichier, placé sur la machine où tourne le moteur de recherche, contenant des directives utilisables par les dévelopeurs et les utilisateurs de méta-moteurs. Il pourrait contenir les informations suivantes :

- les horaires souhaitables pour interroger la base de données,
- le nombre maximal de requêtes posées par jour (l'heure etc.),
- le temps maximal d'occupation du moteur de recherche,
- des instructions concernant l'autorisation de modifier le format des pages reçues,
- des instructions spéciales.

Chapitre 5

Analyse de moteurs à interroger

5.1 Les moteurs interrogés

Notre FIM-Métaindexeur utilise pour les recherches 32 moteurs, dont 3 sont généraux, 2 sont des portails administratifs et 27 sont des moteurs particuliers des sites de l'administration française.

La liste de tous ces moteurs interrogés avec leurs adresses (URL) est maintenant présentée. Ils sont divisés en trois catégories. Les moteurs généraux [Tableau 5.1], les portails administratifs français [Tableau 5.2].

	moteur	url
1	Google	*http ://www.google.com*
2	AltaVista	*http ://www.altavista.com*
3	Yahoo.fr	*http ://fr.yahoo.com*

TAB. 5.1 – Les moteurs généraux interrogés par FIM-Métaindexeur.

	moteur	url
1	Adminet-JO	*http ://www.admi.net/jo*
2	Service-Public	*http ://www.service-public.fr*

TAB. 5.2 – Les portails administratifs interrogés par FIM-Métaindexeur.

Les moteurs de recherche sur les sites de l'administration française [Tableau 5.3].

	moteur	url
1	Assemblée Nationale	*http ://www.Assemblee-nationale.fr*
2	Sénat	*http ://www.senat.fr*
3	Ministère de l'Agriculture et de la Pêche	*http ://www.agriculture.gouv.fr*
4	Ministère de l'Education Nationale	*http ://www.education.gouv.fr*
5	Ministère de l'Equipement, des Transports et du Logement	*http ://index.equipement.gouv.fr*
6	Ministère de la Fonction Publique	*http ://www.fonction-publique.gouv.fr*
7	Ministère des Affaires Etrangères	*http ://www.france.diplomatie.fr*
8	Ministère de l'Economie, des Finances et de l'Industrie	*http ://www.minefi.gouv.fr*
9	Ministère de l'Aménagement du Territoire et de l'Environnement	*http ://www.environnement.gouv.fr*
10	Ministère de la Justice	*http ://www.justice.gouv.fr*
11	Ministère de la Recherche	*http ://www.recherche.gouv.fr*
12	Bibliothèque Publique d'Information	*http ://www.bpi.fr*
13	Bibliothèque Nationale de France	*http ://www.bnf.fr*
14	Commissariat général du Plan	*http ://www.plan.gouv.fr*
15	Secrétariat d'Etat au Commerce Extérieur	*http ://www.commerce-exterieur.gouv.fr*
16	Cour de Cassation	*http ://www.courdecassation.fr*
17	INRA - Institut National de la Recherche Agronomique	*http ://www.inra.fr*
18	INRIA - Institut National de Recherche en Informatique et en Automatique	*http ://www.inria.fr*
19	INA - Institut National de l'Audiovisuel	*http ://www.ina.fr*
20	CNRS-annuaire	*http ://www.urec.cnrs.fr*
21	EDF	*http ://www.edf.fr*
22	Institut Curie	*http ://www.curie.fr*
23	Laboratoire Central des Ponts et Chaussées	*http ://www.lcpc.inrets.fr*
24	Universite de Valenciennes	*http ://www.univ-valenciennes.fr*
25	Université de Caen Basse-Normandie	*http ://www.unicaen.fr*
26	Université La Rochelle	*http ://www.univ-lr.fr*
27	Association Agropolis	*http ://www.agropolis.fr*

TAB. 5.3 – Les moteurs de recherche sur les sites de l'administration française interrogés par FIM-Métaindexeur.

5.2 Mode de recherche

5.2.1 Type de moteur interne utilisé

Les sites interrogés par FIM-Métaindexeur sont de types très différents. Les méthodes de recherche utilisées par les moteurs varient elles aussi. La plupart d'entres elles se basent sur les résultats d'autres moteurs que nous appelons *internes*. Parmi tous les moteurs interrogés, plusieurs d'entre eux utilisent, comme moteur *interne,* des logiciels disponibles gratuitement ou payants. D'autres ont développé leur propre moteur *interne*. Dans le tableau 5.5, sont regroupés les moteurs interrogés en fonction du moteur *interne* utilisé lors de la recherche.

5.3 Construction de la requête

Chaque moteur accepte une forme différente de requêtes [Section 3.3.1]. Il est donc nécessaire pour le méta-moteur de connaître ces différentes formes. On présente, dans cette section, les éléments les plus importants de la construction d'une requête. Ce sont des données indispensables à la traduction des questions de l'utilisateur dans une forme acceptable par le moteur, à la création de la requête contenant la question traduite et les paramètres optionnels choisis par l'utilisateur et enfin les directives supplémentaires nécessaires pour procéder à la recherche.

5.3.1 Méthodes du protocole HTTP utilisées pour l'interrogation des moteurs

Les moteurs interrogés utilisent pour les recherches standard, via un formulaire HTML, une des deux principales méthodes du protocole HTTP : GET ou POST. Ils acceptent par contre, le plus souvent, les deux en interrogation. FIM-Métaindexeur utilise pour l'interrogation des moteurs la méthode GET. La méthode POST n'est utilisée que dans les cas où le moteur n'accepte pas de requête GET. Le tableau 5.6 présente la liste des moteurs interrogés accompagnés de la méthode qu'ils utilisent pour l'interrogation.

moteur interrogé	moteur *interne*
Google	Google
AltaVista	AltaVista
Yahoo.fr	-
Adminet-JO	-
Service-Public	-
Assemblée National	-
Sénat	-
Ministère de l'Agriculture et de la Pêche	-
Ministère de l'Education Nationale	htdig
Ministère de l'Equipement, des Transports et du Logement	-
Ministère de la Fonction Publique	-
Ministère des Affaires Etrangères	-
Ministère de l'Economie, des Finances et de l'Industrie	Excite
Ministère de l'Economie, des Finances et de l'Industrie	Spirit V2
Ministère de l'Aménagement du Territoire et de l'Environnement	-
Ministère de la Justice	OnePage
Ministère de la Recherche	htdig
Bibliothèque Publique d'Information	htdig
Bibliothèque Nationale de France	-
Commissariat général du Plan	-
Secrétariat d'Etat au Commerce Extérieur	-
Cour de Cassation	-
INRA - Institut National de la Recherche Agronomique	htdig
INRIA - Institut National de Recherche en Informatique et en Automatique	AltaVista
INA - Institut National de l'Audiovisuel	Alkaline
CNRS-annuaire	Search
EDF	-
Institut Curie	Excite
Laboratoire Central des Ponts et Chaussées	htdig
Universite de Valenciennes	Wais
Université de Caen Basse-Normandie	htdig
Université La Rochelle	-
Association Agropolis	htdig

Tab. 5.5 – Les moteurs utilisés comme moteur de recherche sur les sites interrogés.

moteur interrogé	méthode
Google	GET
AltaVista	GET
Yahoo.fr	GET
Adminet-JO	GET
Service-Public	GET
Assemblée National	GET
Sénat	GET
Ministère de l'Agriculture et de la Pêche	GET
Ministère de l'Education Nationale	GET
Ministère de l'Equipement, des Transports et du Logement	GET
Ministère de la Fonction Publique	GET
Ministère des Affaires Etrangères	GET
Ministère de l'Economie, des Finances et de l'Industrie	GET
Ministère de l'Aménagement du Territoire et de l'Environnement	POST
Ministère de la Justice	GET
Ministère de la Recherche	GET
Bibliothèque Publique d'Information	GET
Bibliothèque Nationale de France	GET
Commissariat général du Plan	GET
Secrétariat d'Etat au Commerce Extérieur	GET
Cour de Cassation	POST
INRA - Institut National de la Recherche Agronomique	GET
INRIA - Institut National de Recherche en Informatique et en Automatique	GET
INA - Institut National de l'Audiovisuel	GET
CNRS-annuaire	GET
EDF	GET
Institut Curie	GET
Laboratoire Central des Ponts et Chaussées	GET
Universite de Valenciennes	GET
Université de Caen Basse-Normandie	GET
Université La Rochelle	GET
Association Agropolis	GET

TAB. 5.6 – La méthode du protocole HTTP utilisée pour l'interrogation des moteurs de recherche.

5.3.2 Traduction des questions

Chaque moteur de recherches utilise un format spécifique pour exprimer dans une requête HTTP la question posée par l'utilisateur (Section 3.3.1). Il est donc nécessaire d'effectuer pour chacun de ces moteurs la traduction de la question provenant du formulaire du méta-moteur dans un format accepté par ce moteur. Dans cette section, on présente les méthodes utilisées par FIM-Métaindexeur pour la traduction d'une question. Cette traduction est analysée sous les aspects suivants :

- utilisation et forme d'une expression booléenne,
- possibilité de recherche d'une phrase,
- acceptation des mots accentués dans la question,
- méthode de support des accents

5.3.2.1 Recherches booléennes

Le méta-moteur doit choisir comment effectuer la traduction pour que la forme acceptée par le moteur exprime au mieux la question posée par l'utilisateur. Les expressions booléennes peuvent s'exprimer sous des formes variées, leur traduction pose quelques problèmes. D'abord, il faut vérifier si la fonction utilisée dans la question de l'utilisateur est acceptable par le moteur et la ré-exprimer avec des opérateurs appropriés au ce moteur. Lorsque cette fonction n'existe pas sur le moteur interrogé, il faut transformer la question en fonction des possibilités offertes par ce moteur.

- La conjonction.

  FIM-Métaindexeur accepte les opérateurs suivants : "AND", "ET" et "+". Chaque opérateur de la question posée par l'utilisateur sera, au moment de la création de la requête, remplacé par un opérateur accepté par le moteur interrogé.

  En général, si le moteur ne dispose pas de conjonction, les opérateurs logiques (AND, OU) doivent être supprimés, ou, dans certains cas particuliers, remplacés par l'espace. Donnons un exemple :

  Dans la question : *recherche and information,* l'opérateur AND sera supprimé *(recherche information),* mais dans la question *recherche+information* , le "+" sera remplacé par un espace *(recherche information).*

  D'autre part, le nombre d'espaces entre les mots d'une question n'est pas significatif. Afin d'éliminer une vérification et accélérer le processus d'analyse des questions, FIM-Métaindexeur remplace systématiquement tous les opérateurs de conjonction par un espace.

- L'alternatif.

  FIM-Métaindexeur accepte les opérateurs suivants : "OR", "OU" .

  Chacun de ces opérateurs doit être au moment de la création de la requête remplacé par un opérateur accepté par le moteur interrogé. Dans le cas où le moteur interrogé n'accepte pas l'alternatif, des requêtes contenant chacune un des composants de la fonction boolèenne sont envoyées séparément.

- La négation.

  FIM-Métaindexeur accepte les opérateurs suivants : "NOT", "SAUF" et "-".

  Ici encore, chaque opérateur sera remplacé par un opérateur interprétable par le moteur interrogé. Si le moteur n'accepte pas de négation, le composant contenant la négation est supprimé de la question. Par exemple, dans la question :

  commerce +intérieur -extérieur, si le moteur ne dispose pas de la négation, il faut envoyer cette question sous fa forme :

  commerce +intérieur.

5.3.2.2 Recherche des phrases (guillemets - simples et doubles)

Parmi tous les moteurs interrogés par FIM-Métaindexeur, il y en a certains qui permettent d'effectuer des recherches de phrases (expression exacte). Cette fonction est généralement réalisée par insertion de la phrase entre guillemets. Selon le mode de recherche, les moteurs acceptent les guillemets doubles, simples ou les deux. Dans le tableau 5.8, on présente la table du type de recherche des expressions exactes possibles pour tous les moteurs interrogés.

Tout d'abord, il faut vérifier si l'expression utilisée dans la question de l'utilisateur est acceptable par le moteur et ensuite la ré-exprimer en fonction d'opérateurs compréhensibles par ce moteur.

Pour traduire la recherche d'une expression exacte par FIM-Métaindexeur, l'utilisateur doit insérer sa question entre guillemets. FIM-Métaindexeur accepte autant

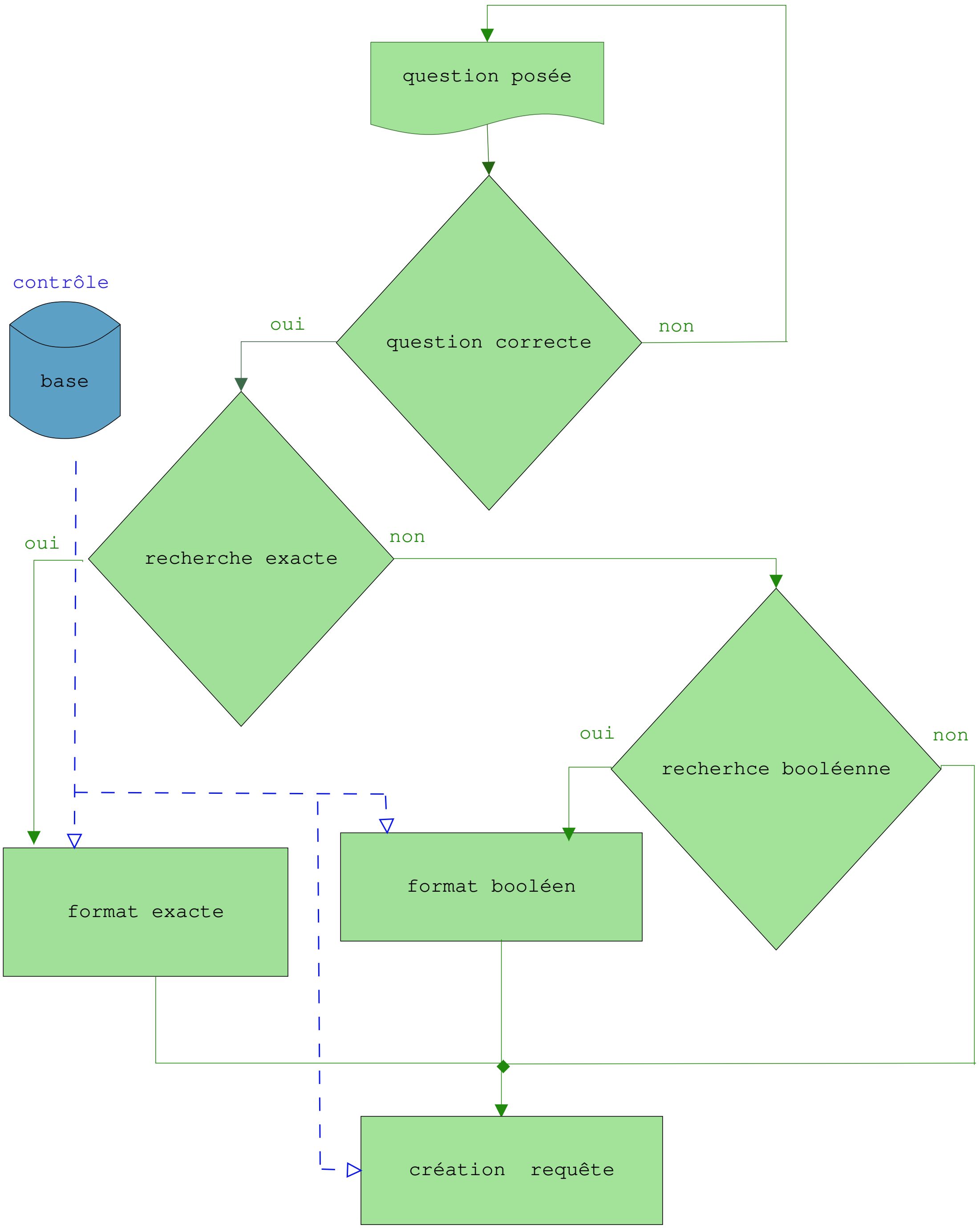


FIG. 5.1 – Traduction de question posée vers la requête spécifique de chaque moteur.

moteur interrogé	table de la corespondance pour les fonction boolèen							
	+	-	ET	OU	SAUF	OR	AND	NOT
Google	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
AltaVista	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Yahoo.fr	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Adminet-JO	+	-	+	rien	-	rien	+	-
Service-Public	*l'espace*	-	AND	OR	NOT	OR	AND	NOT
Assemblée National	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Sénat	+	-	ET	OU	SAUF	OR	AND	NOT
Ministère de l'Agriculture et de la Pêche	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Ministère de l'Education Nationale	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Ministère de l'Equipement, des Transports et du Logement	+	rien	ET	OU	SAUF	OU	ET	SAUF
Ministère de la Fonction Publique	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Ministère des Affaires Etrangères	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Ministère de l'Economie, des Finances et de l'Industrie	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Ministère de l'Aménagement du Territoire et de l'Environnement	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Ministère de la Justice	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Ministère de la Recherche	AND	NOT	AND	OR	NOT	OR	AND	NOT
Bibliothèque Publique d'Information	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Bibliothèque Nationale de France	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Commissariat général du Plan	+	SAUF	ET	OU	SAUF	OU	ET	SAUF
Secrétariat d'Etat au Commerce Extérieur	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Cour de Cassation	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
INRA - Institut National de la Recherche Agronomique	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
INRIA - Institut National de Recherche en Informatique et en Automatique	*l'espace*	-	*l'espace*	rien	-	rien	*l'espace*	-
INA - Institut National de l'Audiovisuel	*l'espace*	-	+	rien	-	rien	+	-
CNRS-annuaire	*l'espace*	rien	*l'espace*	rien	rien	rien	*l'espace*	rien
EDF	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Institut Curie	*l'espace*	rien	*l'espace*	rien	rien	rien	*l'espace*	rien
Laboratoire Central des Ponts et Chaussées	*l'espace*	-	ET	OU	SAUF	OR	AND	NOT
Universite de Valenciennes	*l'espace*	NOT	AND	OR	NOT	OR	AND	NOT
Université de Caen Basse-Normandie	AND	-	AND	OR	NOT	OR	AND	NOT
Université La Rochelle	*l'espace*	-	AND	OR	NOT	OR	AND	NOT
Association Agropolis	*l'espace*	rien	AND	OR	NOT	OR	AND	NOT

TAB. 5.7 – La table de la correspondance pour les fonctions booléennes.

moteur interrogé	guillemets	
	double	simple
Google	oui	oui
AltaVista	oui	non
Yahoo.fr	oui	oui
Adminet-JO	oui	
Service-Public	oui	
Assemblée National	oui	
Sénat	oui	
Ministère de l'Agriculture et de la Pêche		
Ministère de l'Education Nationale		
Ministère de l'Equipement, des Transports et du Logement		
Ministère de la Fonction Publique		
Ministère des Affaires Etrangères	oui	oui
Ministère de l'Economie, des Finances et de l'Industrie		non
Ministère de l'Aménagement du Territoire et de l'Environnement	oui	oui
Ministère de la Justice		non
Ministère de la Recherche		
Bibliothèque Publique d'Information		
Bibliothèque Nationale de France		
Commissariat général du Plan	oui	oui
Secrétariat d'Etat au Commerce Extérieur	oui	oui
Cour de Cassation	oui	oui
INRA - Institut National de la Recherche Agronomique		
INRIA - Institut National de Recherche en Informatique et en Automatique	oui	non
INA - Institut National de l'Audiovisuel		
CNRS-annuaire		
EDF		
Institut Curie	oui	non
Laboratoire Central des Ponts et Chaussées		
Universite de Valenciennes	oui	non
Université de Caen Basse-Normandie	oui	non
Université La Rochelle		
Association Agropolis	oui	non

Tab. 5.8 – Recherche des phrases. Acceptation de guillemets.

les guillemets simples que les doubles. Pour créer la requête dédiée au moteur interrogé, il faut remplacer la forme utilisée par l'utilisateur avec la forme acceptée par ce moteur. Lorsque le moteur ne dispose pas de cette possibilité, les guillemets sont supprimés, et la question est envoyée sans recherche d'expression exacte.

5.3.2.3 Traduction des lettres accentuées

Le traitement des accents est une tâche assez délicate. La recherche d'une même expression avec ou sans accents peut, suivant les moteurs interrogés, donner des résultats différents. Cela dépend de la façon dont le moteur traite les mots accentués au cours du processus de recherche, mais également pendant l'indexation de sa base de données. De plus, il existe plusieurs manières de coder des lettres accentuées [Section 2.3.6.2]. Il est donc nécessaire de procéder à un traitement (filtrage) spécifique pour chaque moteur interrogé. On peut distinguer deux sortes de traitements, un pour l'interrogation du moteur et un deuxième pour l'analyse des documents réponses. Pendant le processus d'interrogation, la question est envoyée sous une forme adaptée au moteur. Le tableau 5.9 présente les possibilités de recherche avec ou sans mots accentués pour les moteurs interrogés.

Pour effectuer la vérification de la pertinence des URL documents-réponses et obtenir une forme homogène, il est nécessaire de procéder à un filtrage spécifique des mots accentués de la question et des documents analysés.

5.3.3 Paramètres optionnels inclus dans la requête

Tous les moteurs interrogés possèdent des modes de recherche différents. Afin de donner la possibilité aux utilisateurs de choisir leur configuration, des paramètres optionnels inclus dans le formulaire de la recherche sont ajustables. Il n'est pas possible pour le méta-moteur de tous les exploiter et en général le méta-moteur utilise le mode standard de la recherche. Il existe quelques paramètres qui sont utilisés par FIM-méta-indexeur :

- Mode de recherche.

  ce paramètre est utilisé pour donner à l'utilisateur la possibilité de sélectionner un mode de recherche particulier, par exemple, le choix entre une recherche booléenne et une recherche exacte, etc. L'utilisateur fait la sélection du mode de recherche en même temps que la formulation de

moteur interrogé	accent
Google	oui
AltaVista	oui
Yahoo.fr	oui
Adminet-JO	oui
Service-Public	oui
Assemblée National	oui
Sénat	oui
Ministère de l'Agriculture et de la Pêche	oui
Ministère de l'Education Nationale	oui
Ministère de l'Equipement, des Transports et du Logement	oui
Ministère de la Fonction Publique	oui
Ministère des Affaires Etrangères	oui
Ministère de l'Economie, des Finances et de l'Industrie	oui
Ministère de l'Aménagement du Territoire et de l'Environnement	oui
Ministère de la Justice	oui
Ministère de la Recherche	oui
Bibliothèque Publique d'Information	oui
Bibliothèque Nationale de France	oui
Commissariat général du Plan	oui
Secrétariat d'Etat au Commerce Extérieur	oui
Cour de Cassation	oui
INRA - Institut National de la Recherche Agronomique	oui
INRIA - Institut National de Recherche en Informatique et en Automatique	oui
INA - Institut National de l'Audiovisuel	oui
CNRS-annuaire	
EDF	oui
Institut Curie	oui
Laboratoire Central des Ponts et Chaussées	oui
Universite de Valenciennes	oui
Université de Caen Basse-Normandie	oui
Université La Rochelle	oui
Association Agropolis	

TAB. 5.9 – Acceptation des accents.

sa question dans le formulaire (HTML) de recherche. Ensuite toutes ces données sont transmises au moteur.

Fim-Métaindexeur envoie ce paramètre dans la requête, à tous les moteurs interrogés qui utilisent cette forme de configuration.

- Nombre de réponses.

  La possibilité de fixer le nombre de réponses à récupérer parmi celles des moteurs interrogés est très utile. Elle permet de contrôler plus finement le processus d'interrogation. Le nombre de réponses est un paramètre qui peut exister sous deux formes : comme précédemment pour donner à l'utilisateur la possibilité d'effectuer un choix (paramètre visible) ou comme un paramètre invisible pour l'utilisateur standard, mais inclus dans la requête. Comme tous deux sont envoyés au moteur via une requête HTTP, il est possible de les utiliser dans la requête envoyée par FIM-Métaindexeur.

- Nombre de réponses par page.

  Les moteurs de recherche affichent les listes des documents-réponses sous deux formes différentes : sur une seule page ou plusieurs pages. Parmi les moteurs qui affichent les réponses sur plusieurs pages, beaucoup offrent la possibilité de régler le nombre de réponses par page, par le biais d'un paramètre de la requête. Cette possibilité permet à Fim-Métaindexeur de contrôler au mieux le nombre de pages de réponses à récupérer (et souvent, dans ce cas, de récupérer les réponses à partir d'une seule page) et de diminuer ainsi la charge du réseau.

Les paramètres optionnels acceptés par les moteurs interrogés et utilisés par FIM-Métaindexeur pour contrôler le processus d'interrogation sont présentés dans le tableau [Tableau 5.10].

Si le mode "par défaut" utilisé par le moteur interrogé, est le mieux adapté à l'interrogation par Fim-Métaindexeur, il n'y a pas de changement de la configuration de la recherche standard.

moteur interrogé	parametres optioneles utilisés		
	mode de recherche	nombre de réponses	nombre de pages
Google	standard	-	plusieurs
AltaVista	standard	-	plusieurs
Yahoo.fr	standard	-	plusieurs
Adminet-JO	standard	oui	une seule
Service-Public	standard	-	plusieurs
Assemblée National	standard	-	une seule
Sénat	standard	-	plusieurs
Ministère de l'Agriculture et de la Pêche	standard	-	plusieurs
Ministère de l'Education Nationale	booléenne	-	plusieurs
Ministère de l'Equipement, des Transports et du Logement	standard	-	plusieurs
Ministère de la Fonction Publique	standard	-	une seule
Ministère des Affaires Etrangères	standard	-	plusieurs
Ministère de l'Economie, des Finances et de l'Industrie	simple	-	une seule
Ministère de l'Aménagement du Territoire et de l'Environnement	standard	-	une seule
Ministère de la Justice	standard	oui	configurable
Ministère de la Recherche	booléenne	oui	configurable
Bibliothèque Publique d'Information	booléenne	oui	configurable
Bibliothèque Nationale de France	standard	oui	configurable
Commissariat général du Plan	standard	-	une seule
Secrétariat d'Etat au Commerce Extérieur	standard	oui	configurable
Cour de Cassation	standard	-	une seule
INRA - Institut National de la Recherche Agronomique	booléenne	oui	configurable
INRIA - Institut National de Recherche en Informatique et en Automatique	standard	-	une seule
INA - Institut National de l'Audiovisuel	simple	-	une seule
CNRS-annuaire	standard	-	plusieurs
EDF	standard	-	une seule
Institut Curie	simple	-	une seule
Laboratoire Central des Ponts et Chaussées	booléenne	-	plusieurs
Universite de Valenciennes	standard	-	une seule
Université de Caen Basse-Normandie	booléenne	oui	configurable
Université La Rochelle	standard	-	une seule
Association Agropolis	booléenne	oui	configurable

TAB. 5.10 – Les paramètres optionnels utilisés par le Fim-Métaindexeur.

5.4 Mode d'affichage des réponses

Afin d'utiliser les résultats des recherches effectuées par les moteurs interrogés, FIM-Métaindexeur récupère les pages de réponses standard. Les résultats sont affichés sous des formes très variées [Chapitre 5 et la Section 5.3.3]. Pour extraire automatiquement la liste de toutes les réponses à partir des pages de résultats fournis par les moteurs, il est nécessaire de bien analyser auparavant le mode d'affichage des réponses.

5.4.1 Analyse de la topologie des pages de réponses et filtrage des URL

En premier lieu, après avoir reçu la page des réponses, FIM-Métaindexeur doit distinguer le type de page qu'il a reçue : page de bonnes réponses, page sans réponse, page avec un message d'erreurs, etc. [Section 4.4.2.4]. En fonction de cette analyse, il effectue un traitement approprié.

5.4.1.1 Différents modèles des pages de réponses

Les pages de réponses des moteurs interrogés par FIM-Métaindexeur sont présentées sous une forme différente selon "l'état" des recherches. On peut distinguer différents types de pages comme : les pages avec un message d'erreurs, les pages indiquant le manque de réponse, les pages qui contiennent des bonnes réponses, etc. FIM-Métaindexeur est capable de distinguer toutes ces formes différentes de pages. Pour réaliser cette fonctionnalité, il possède une base de données contenant les informations nécessaires pour distinguer tous les types possibles de pages de réponses pour chacun des moteurs interrogés.

Selon la page de réponses reçue FIM-Métaindexeur réagit différemment. Il peut renvoyer une requête pour les pages contenant des messages d'erreurs temporaires ou contenant un code HTTP indiquant un changement d'URL, soit arrêter la communication avec un moteur dont les pages sont sans réponses ou dont les pages indiquent un message d'erreur persistant ou enfin continuer l'analyse pour les pages qui contiennent les bonnes réponses.

5.4.1.2 Filtrage d'URL

Les pages de réponses fournies par les moteurs de recherche sont assez complexes. Elles contiennent non seulement la liste des résultats des recherches, mais aussi beaucoup d'autres informations. De plus ces informations ne sont pas statiques. Elles changent très fréquemment, souvent en fonction du thème de recherche, par exemple : des publications thématiques sélectionnées automatiquement. Il est nécessaire de bien distinguer les URL de réponses parmi toutes les autres. Tous ces problèmes font que la récupération des URL indiquant les documents-réponses à partir de la page de réponses standard est une tâche difficile, mais très importante.

Pour développer cette fonctionnalité du méta-moteur, différentes solutions ont été essayées.

Sélection automatisée. Une solution “idéale” permettrait au méta-moteur d'effectuer la sélection des bonnes URL automatiquement, sans aucune intervention humain. Dans ce cas, le méta-moteur serait indépendant de toutes modifications effectuées sur les pages de réponses venant des moteurs interrogés. De plus, une telle solution faciliterait l'insertion d'un nouveau moteur à interroger. Pour implémenter cette solution, nous avons essayé des méthodes basées sur la comparaison entre pages de réponses dont la structure était connue auparavant, avec la page de réponses reçue.

Pour s'assurer que la récupération de liens vers les documents-réponses est indépendante des modifications effectuées sur la forme de la page de réponses, le plus important est d'effectuer toutes les opérations nécessaires pour la sélection en temps réel, c'est à dire en même temps que la recherche pour la question posée par l'utilisateur.

L'idée principale de cette méthode est assez simple. Pour chacun des moteurs interrogés, il faut d'abord envoyer une requête-test nous fournissant une page de références, ensuite analyser cette page de références, et enfin la comparer avec la page de réponses à la question posée par l'utilisateur. L'hypothèse pratique que l'on aimerait vérifier est la suivante :

Hypothèse pratique : Il existe une page de références (cas test) telle que : l'ensemble de tous les liens hypertexte (URL) de la page de réponses à la question de l'utilisateur, qui sont différents des liens hypertexte (URL) de la page de références (cas test), représente l'ensemble des liens recherchés, c'est à dire l'ensemble des liens répondant à la question.

Cette hypothèse a été testée sur les différents types de pages de références suivants :

- Page de références correspondant à une requête vide.

  > Cette méthode donne de bons résultats pour les moteurs de recherche, qui affichent les pages de réponses sous une forme identique, aussi bien pour la requête vide que pour la requête complète. Parmi tous les moteurs interrogés par FIM-Métaindexeur, il existe des moteurs, qui distinguent des formats de pages de réponses différents selon les types de requêtes. Cette méthode ne peut pas être universelle.

- Page de références correspondant à une question précise déterminée à l'avance.

  > Comme précédemment, les résultats varient selon les moteurs interrogés. Pour les moteurs qui affichent des pages de réponses sous un même format, sans aucune modification dynamique, les résultats sont corrects. Dans les autres cas, cette méthode génère quelques erreurs, elle ne peut donc pas être choisie comme méthode universelle.

- Couplage des deux méthodes.
  Pour que le couplage des deux méthodes puisse donner de bons résultats, il faut que pour chacun des moteur interrogés on sache quel est le type de pages de références à utiliser, et fournir ainsi une bonne complémentarité des bons résultats. Cette idée est illustrée dans la figure 5.2. Il faut ajouter cependant que certains moteurs interrogés modifient non seulement la forme de la page des réponses, mais peuvent également changer au cours du temps leur mode d'affichage des réponses, ce qui entraîne une source supplémentaire d'erreurs et peut modifier la proportion de bons résultats acquise au départ.

Finalement, les tests effectués avec les méthodes présentées, ont montré que la sélection automatique, sans intervention humaine n'est pas souhaitable.

Sélection à partir de données préparées manuellement

Ici les liens vers les documents-réponses sont filtrés à partir de la page de réponses automatiquement, en suivant des règles spécifiques pour chacun des moteurs de recherche. Il est nécessaire de choisir des critères précis, qui vont permettre de distinguer les bons liens parmi toutes les URL existantes sur la page de réponses. Au cours de cette étape, l'intervention humaine est indispensable. Pour chaque moteur de recherche, des pages de réponses différentes ont été analysées, pour sélectionner

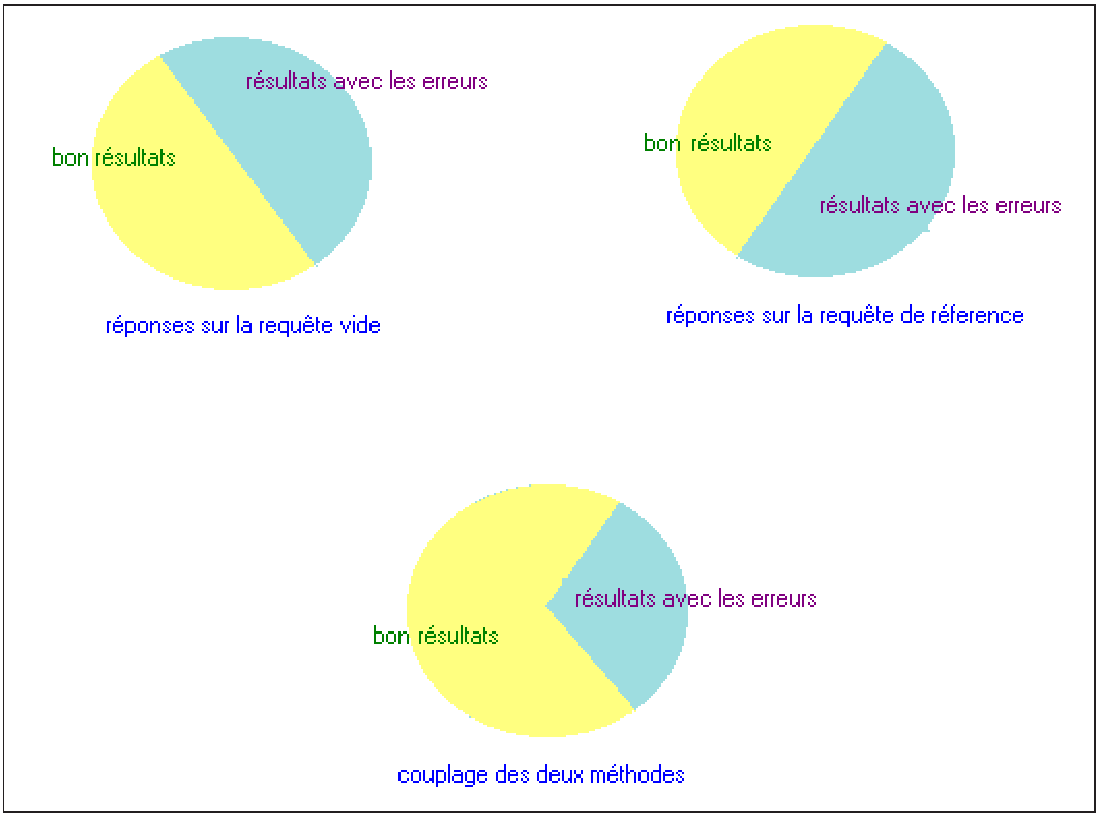


FIG. 5.2 – Couverture de moteurs qui donnent des résultats corrects pour les différentes méthodes automatiques.

des critères appropriés. A partir de ces critères, on génère des filtres de récupération, FIM-Métaindexeur fait ensuite l'extraction des liens pour les documents-réponses.

Cette méthode est fiable et donne des résultats corrects. Cependant, elle est très dépendante de toutes les modifications éventuelles effectuées sur les moteurs. Toute modification nécessite une mise à jour du filtre associé.

Méthode choisie
Finalement, pour développer FIM-Métaindexeur, la méthode à base de filtres préparés manuellement a été sélectionnée. Afin de retrouver uniquement les bonnes URL parmi toutes les URL existantes, FIM-Métaindexeur utilise des filtres spéciaux, différents pour chacun des moteurs de recherche interrogés. Le seul inconvénient de cette méthode est de dépendre des modifications éventuelles effectuées sur le moteur interrogé. Étant donné que la forme des pages fournies par le moteur change assez fréquemment, la mise à jour de ces filtres doit être effectuée souvent, d'où une maintenance régulière nécessaire.

5.4.2 Page unique ou plusieurs pages de réponses

FIM-Métaindexeur n'a pas besoin de toutes les réponses provenant de chacun des moteurs interrogés. En fonction du **Nombre de réponses** à récupérer, sélectionné par l'utilisateur dans le formulaire, FIM-Métaindexeur doit récupérer la quantité de réponses correspondantes. Pour diminuer au maximum le temps de récupération et également diminuer la charge du réseau, il est souhaitable de contrôler le nombre de réponses. Ce contrôle dépend du mode de travail du moteur. Parmi tous les moteurs interrogés par FIM-Métaindexeur, il existe des moteurs où le nombre des réponses à envoyer est configurable parmi les paramètres optionnels, mais il existe également des moteurs où ce nombre est fixé par défaut. L'analyse du mode d'envoi des réponses de chaque moteur a été présentée dans le tableau 5.10.

- Si le moteur interrogé envoie les réponses sur plusieurs pages, il faut contrôler le nombre des pages à récupérer, et répéter le processus de chargement de pages jusqu'à l'obtention du nombre de réponses suffisant.
- Si les réponses sont présentées sur une seule page, et que le moteur accepte le réglage du nombre de réponses à envoyer, il faut demander le nombre de pages correspondant dans les paramètres de la requête.

Chapitre 6

Réalisation d'un méta-moteur

6.1 Architecture

Le schéma général de FIM-Métaindexeur est présenté dans la figure 6.1

Suivant le modèle décrit dans le chapitre 5, l'architecture de FIM-Métaindexeur est basée sur trois modules principaux :

- module d'entrée des données,
- module d'interrogation des moteurs,
- module de traitement de réponses.

6.1.1 Module d'entrée des données

Ce module est responsable de la récupération et de l'analyse de toutes les informations fournies par l'utilisateur (la question et les paramètres optionnels) dans le formulaire HTML de FIM-Métaindexeur.

6.1.1.1 Analyse des données de paramétrage

FIM-Métaindexeur permet de fournir certains paramètres optionnels, dont les informations dédiées pour FIM-Métaindexeur et les informations utilisées par les moteurs interrogés.

Le fonctionnement de ce module est présenté dans la figure 6.2.

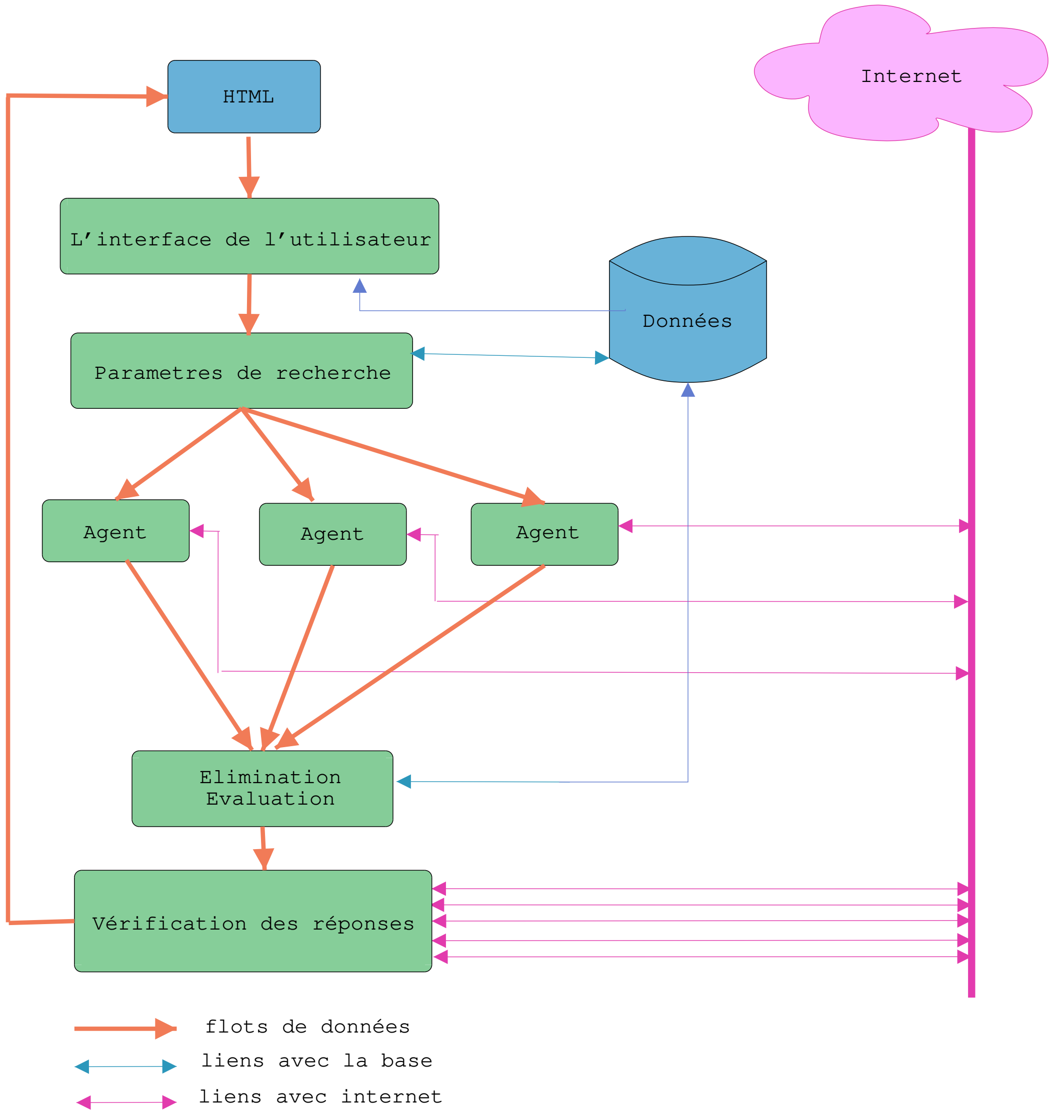


FIG. 6.1 – Schéma d'interfonctionnement des modules et des données de FIM-Métaindexeur.

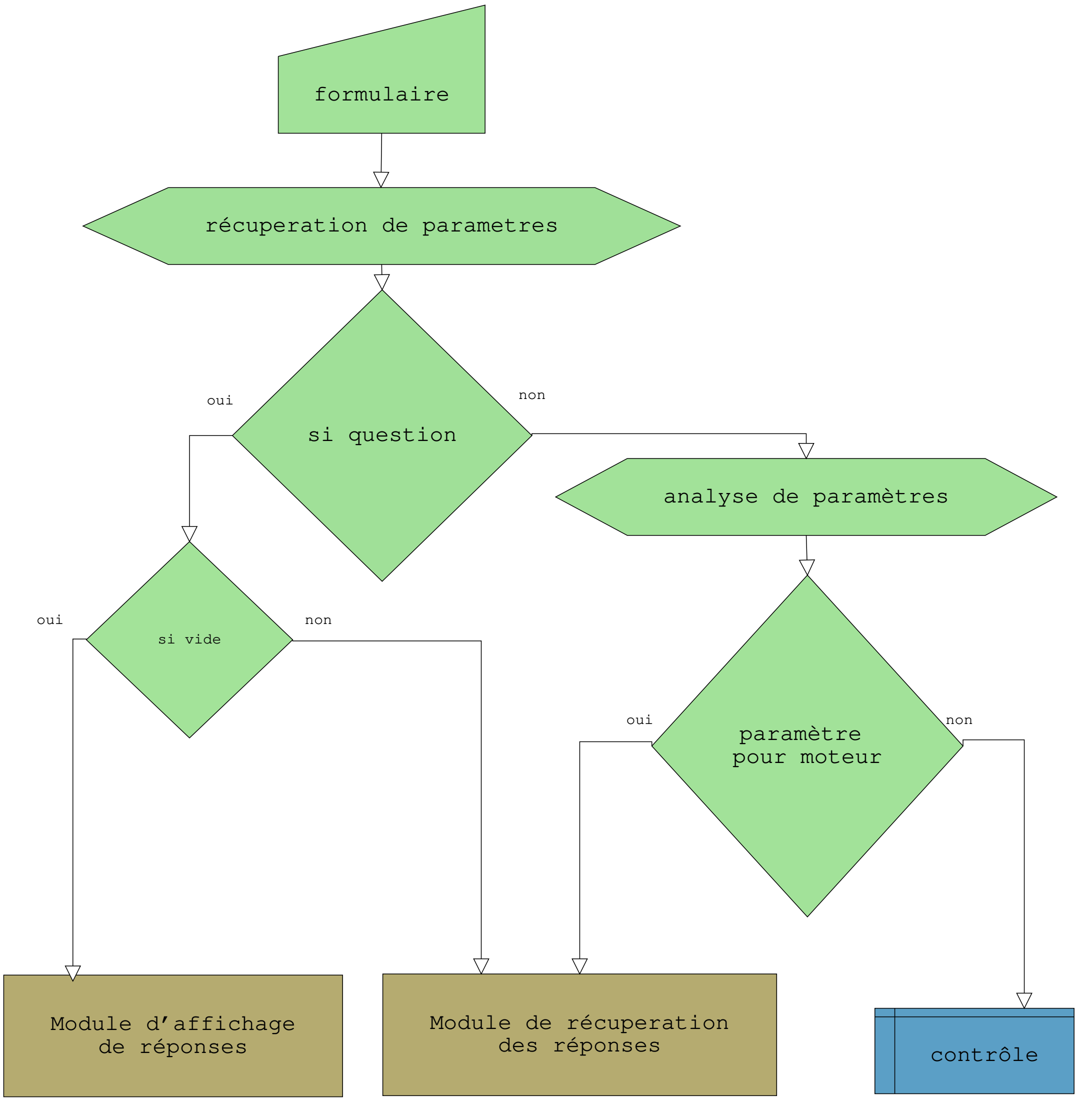


FIG. 6.2 – Module d'entrée des données.

Après la récupération de tous les paramètres choisis par l'utilisateur, FIM-Métaindexeur analyse les données nécessaires pour son fonctionnement :

- les moteurs choisis pour l'interrogation,
- le temps global de recherche,
- le nombre des résultats à récupérer,
- le mode de vérification de liens.

6.1.1.2 Analyse de la question posée

Dans ce module, FIM-Métaindexeur analyse la question posée par l'utilisateur. En premier lieu, FIM-Métaindexeur analyse l'existence de la totalité des fonctions booléennes et des opérateurs spéciaux, et les relations entre ces fonctions, puis les transmet sous une forme générale au module d'interrogation des moteurs.

6.1.2 Module d'interrogation des moteurs

Toutes les processus de ce module sont réalisés en parallèle par des agents spéciaux spécifiques à chaque moteur. Les algorithmes suivants sont présentés pour un tel agent.

6.1.2.1 Création d'une requête

- Choix des moteurs.

  La liste des moteurs à interroger, selon les choix de l'utilisateur, est fournie par le module d'analyse de données en entrée.

- Préparation de la question et des paramètres optionnels.

  Avant de questionner le moteur, FIM-Métaindexeur crée la requête spécifique à chaque moteur. Cette requête contient la question et les paramètres optionnels dédiés au moteur approprié. Pour cela, il récupère les informations concernant chaque moteur sauvegardées dans sa base de données. Selon ces données, il transforme la question générale dans le format accepté par chaque moteur interrogé et prépare les paramètres optionnels.

L'algorithme général de création d'une requête pour le moteur interrogé est présenté dans la figure 6.3.

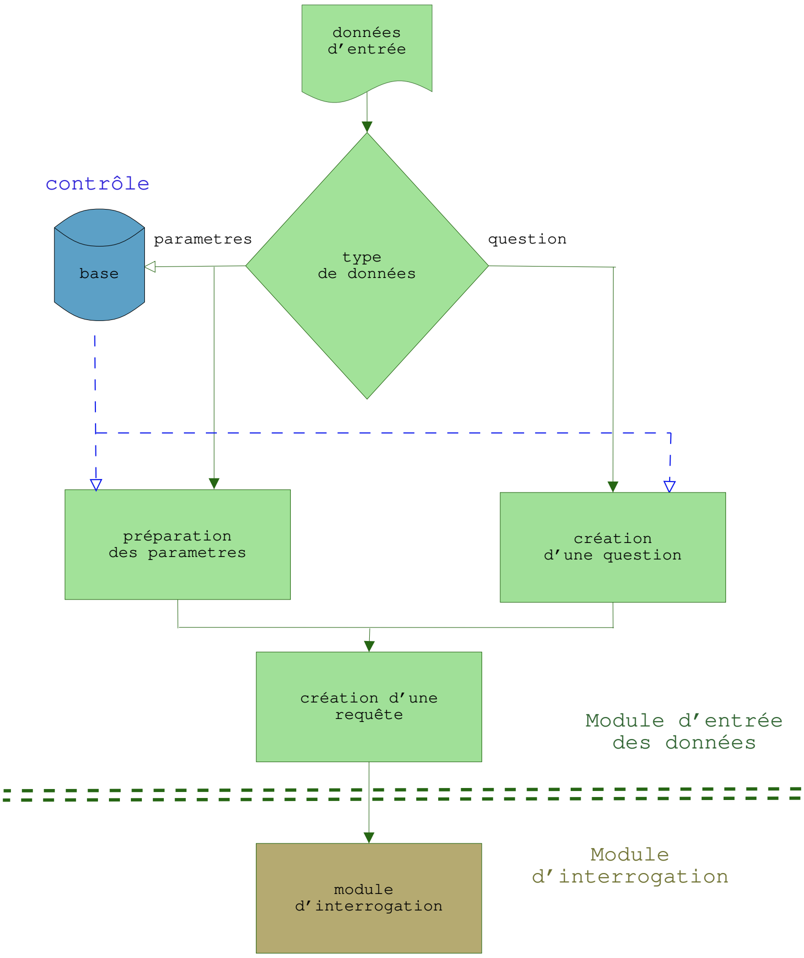


FIG. 6.3 – L'algorithme de création d'une requête.

6.1.2.2 Récupération de réponses

FIM-Métaindexeur contrôle le processus de récupération de réponses en fonction des signaux de contrôle reçus : il attend les réponses pendant un temps plafonné par le paramètre **Temps de réponse maximal,** et récupère le nombre des réponses plafonné par les paramètres **Nombre maximum de réponses** et **Nombre maximum de réponses par moteur.** Pour toutes les réponses récupérées il réalise les opérations suivantes :

- distinction entre page avec les bonnes réponses, page sans réponses, page avec un message d'erreur, etc.,
- filtrage des URL vers les documents réponses (le cas des pages avec de bonnes réponses) : pour réaliser cette fonction FIM-Métaindexeur utilise des filtres spéciaux, créés manuellement.

L'algorithme de la procédure de récupération des réponses est présenté dans la figure 6.4.

6.1.3 Module d'analyse des réponses récupérées

Dans ce module FIM-Métaindexeur réalise les opérations suivantes :

- vérification des liens vers les documents-réponses,

> FIM-Métaindexeur teste les liens selon le choix de l'utilisateur : paramètre **Vérification de l'existence des documents** et **Temps de vérification maximal**.

- tri des réponses,
- affichage des résultats,

L'algorithme de cette procédure est présenté dans la figure 6.5.

6.2 Présentation de FIM-Métaindexeur

6.2.1 Données pour les FIM-Métaindexeur

Pour effectuer l'interrogation de différents moteurs de recherche, FIM-Métaindexeur a besoin d'avoir les données concernant tous les moteurs sélectionnés. Toutes les

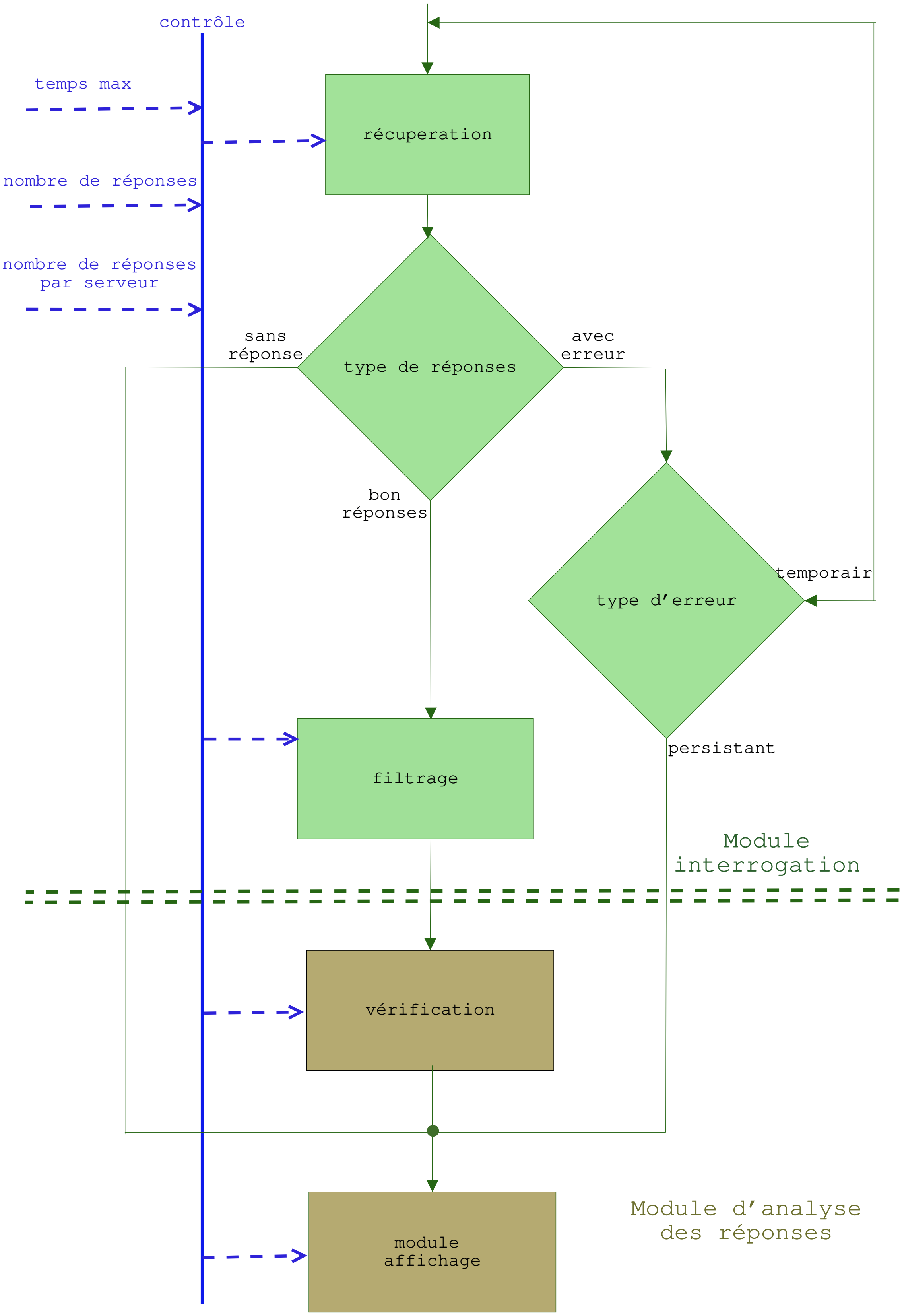


Fig. 6.4 – Algorithme de récupération de réponses.

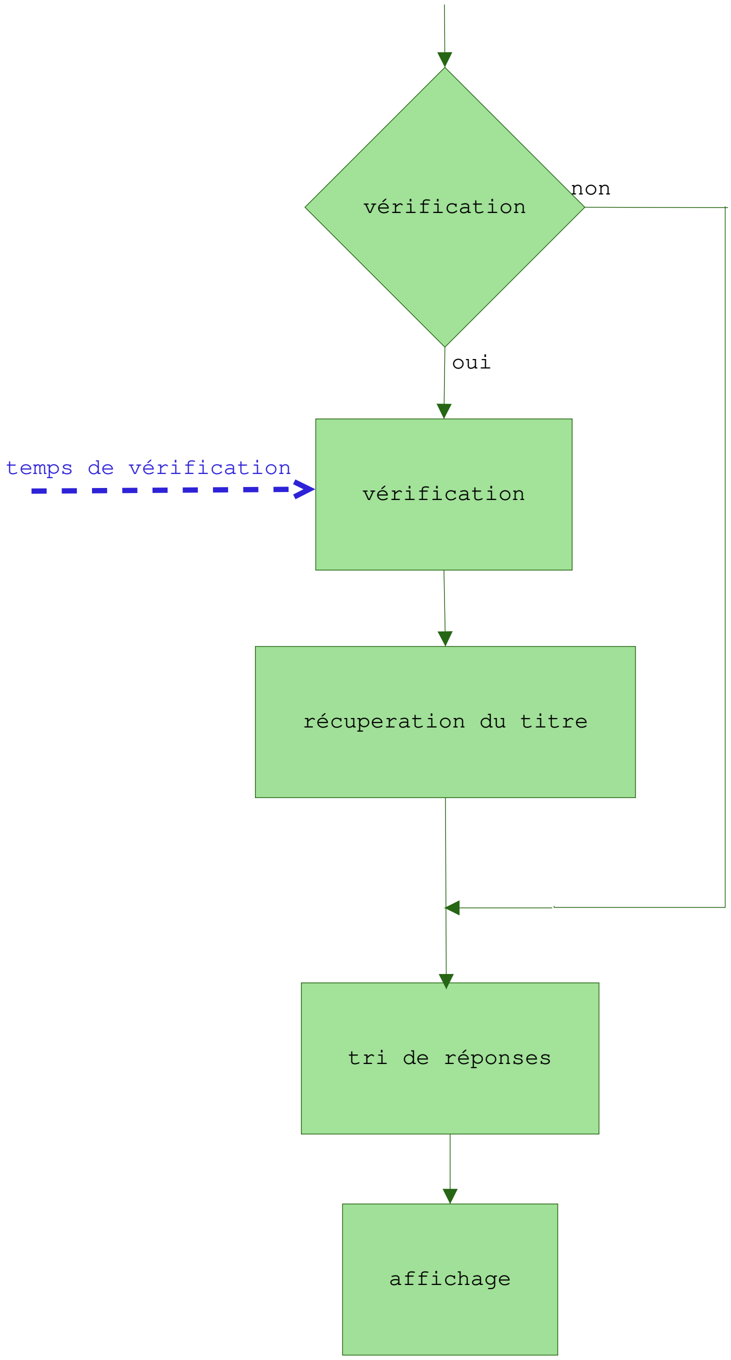


FIG. 6.5 – Algorithme d'analyse des réponses récupérées.

informations nécessaires à FIM-Métaindexeur pour interroger un moteur de recherche sont préparées auparavant et sauvegardées dans une base de données spécifique.

Pour ajouter un nouveau moteur de recherche dans la liste de moteurs sélectionnés à l'interrogation, ou pour modifier les paramètres d'un moteur déjà existants dans la liste, il faut bien analyser ce moteur. La méthode d'analyse des moteurs et toutes les données nécessaires à FIM-Métaindexeur pour l'interrogation ont été présentées dans le chapitre précédent [Chapitre 5] Pour FIM-Métaindexeur, toutes les données concernant un moteur peuvent être insérées directement dans la base, ou à l'aide d'un interface spécifique sous la forme d'un formulaire HTML. La page du formulaire pour insertion des données concernant le moteur interrogé est présentée dans la figure 6.6.

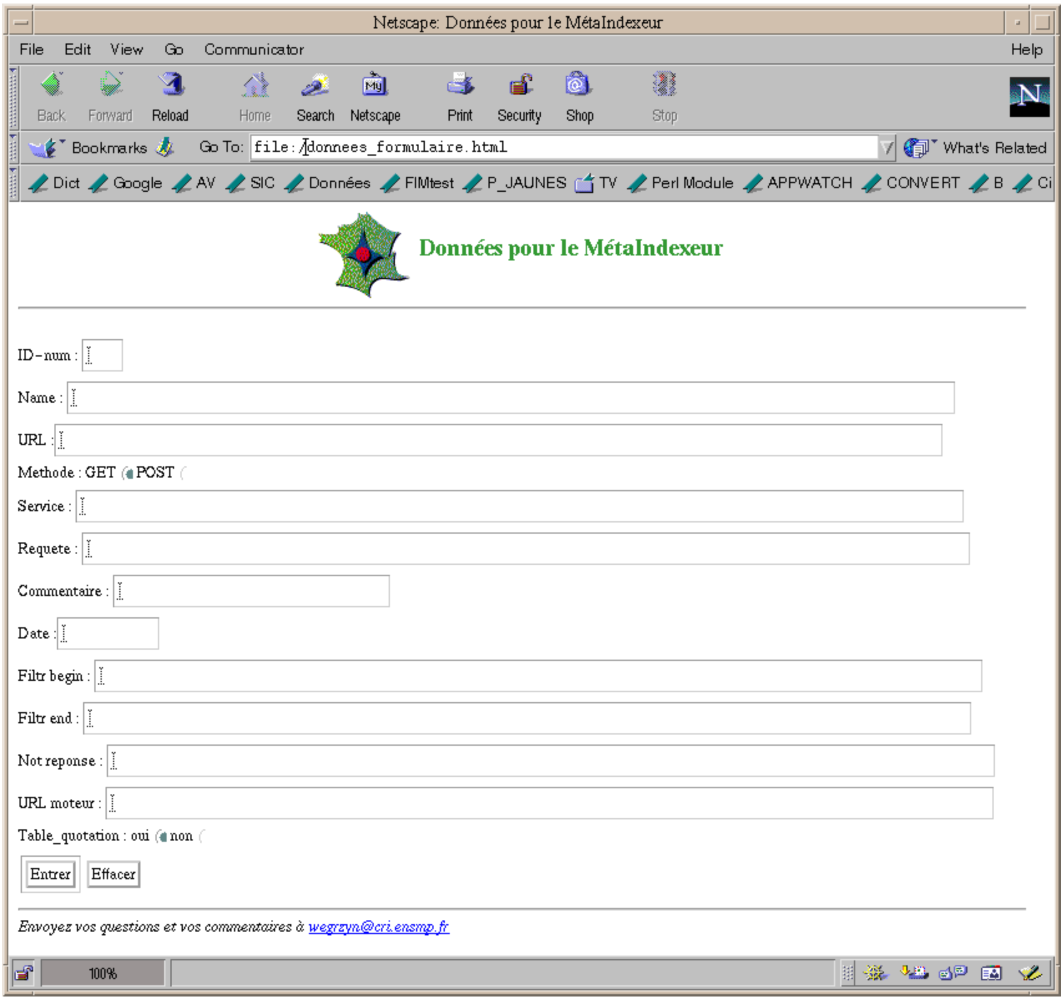
6.2.2 Formulation d'une question par l'utilisateur

La question (requête de l'utilisateur) est entrée dans le formulaire HTML présenté dans la figure 6.7. FIM-Métaindexeur accepte toutes les expressions booléennes (AND, OR, NOT, ET, OU, SAUF, "+" et "-") [Section 5.3.2.1], les mots accentués [Section 5.3.2.3] et il permet également d'effectuer une recherche sur une expression exacte (par insertion de cette expression entre guillemets doubles ou simples) [Section 5.3.2.2]. La question donnée sous cette forme universelle est transformée plus tard dans la forme appropriée à chaque moteur.

6.2.3 Options de la recherche

FIM-Métaindexeur se connecte à différents moteurs de recherche pour fournir la réponse souhaitée. Il est possible de configurer la connexion à ces moteurs et le mode de l'interrogation avec les paramètres optionnels suivants :

- temps de connexion maximal,
- vérification des réponses,
  - vérification des liens,
  - temps de vérification maximal,
- nombre maximum de réponses,
- nombre maximum de réponses par moteur,
- choix du serveur.



file:/donnees_formulaire.html

Données pour le MétaIndexeur

ID-num :

Name :

URL :

Methode : GET ● POST ○

Service :

Requete :

Commentaire :

Date :

Filtr begin :

Filtr end :

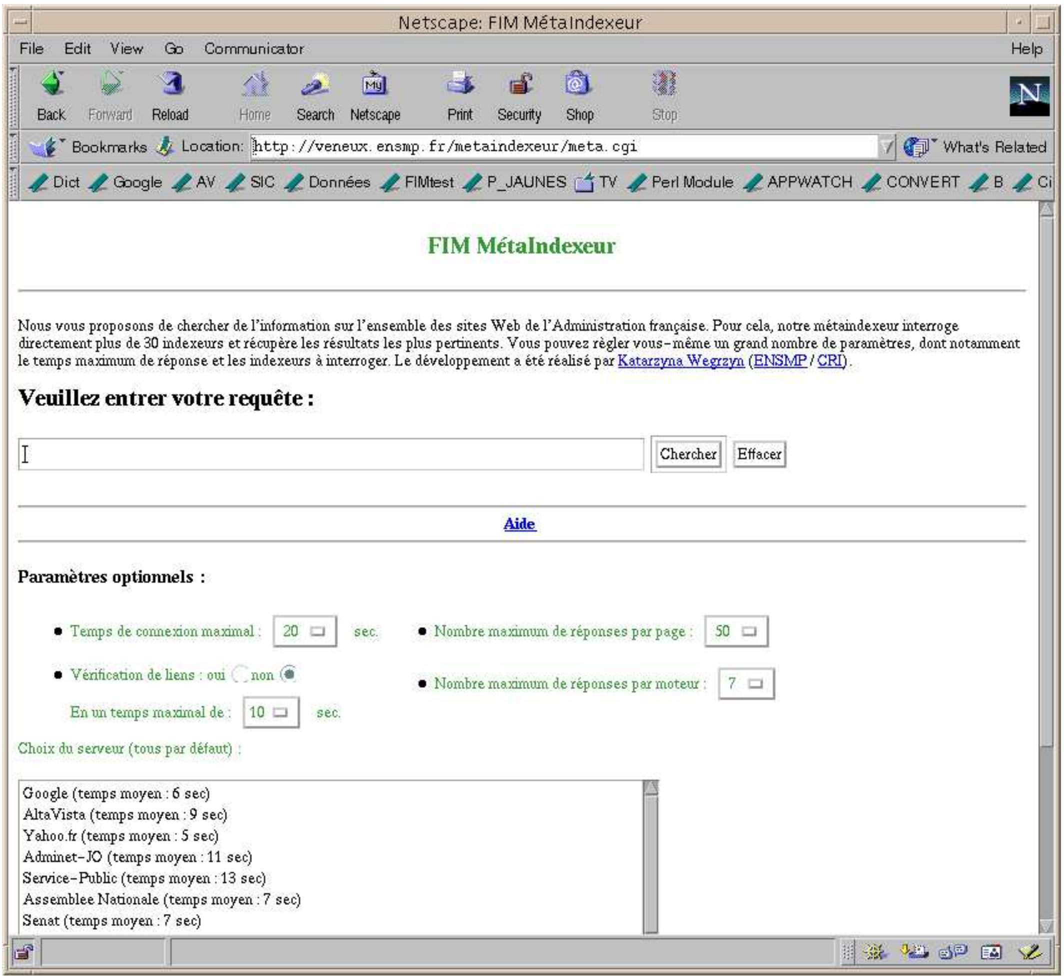
Not reponse :

URL moteur :

Table_quotation : oui ● non ○

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

FIG. 6.6 – Formulaire pour insertion de paramètres d'un moteur.



FIM MétaIndexeur

Nous vous proposons de chercher de l'information sur l'ensemble des sites Web de l'Administration française. Pour cela, notre métaindexeur interroge directement plus de 30 indexeurs et récupère les résultats les plus pertinents. Vous pouvez règler vous-même un grand nombre de paramètres, dont notamment le temps maximum de réponse et les indexeurs à interroger. Le développement a été réalisé par Katarzyna Wegrzyn (ENSMP / CRI).

Veuillez entrer votre requête :

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 : 6 sec)
AltaVista (temps moyen : 9 sec)
Yahoo.fr (temps moyen : 5 sec)
Adminet-JO (temps moyen : 11 sec)
Service-Public (temps moyen : 13 sec)
Assemblee Nationale (temps moyen : 7 sec)
Senat (temps moyen : 7 sec)

FIG. 6.7 – FIM-Métaindexeur. La page de formulaire.

La figure 6.7 présente une partie de la page du formulaire de FIM-Métaindexeur contenant les paramètres optionnels à sélectionner et à envoyer à FIM-Métaindexeur en même temps que la question.

6.2.3.1 Temps de connexion maximal par moteur

Ce paramètre spécifie le temps maximal pendant lequel FIM-Métaindexeur attend la réponse de chaque moteur. Les valeurs possibles à choisir sont : 10, 20, 40, 60, 80, 100, 120, 150, 180 secondes. La valeur sélectionnée par défaut est 20 secondes. Au-delà du temps de connexion maximal sélectionné, la communication avec le moteur est arrêtée. Dans ce cas, FIM-Métaindexeur ne pourrait pas fournir à l'utilisateur de réponse venant de ce moteur. Cette fonctionnalité est réalisée par le signal d'alarme qui se déclenche à la fin du temps de connexion maximal choisi, et arrête la procédure de communication avec le moteur donné.

6.2.3.2 Vérification des réponses

FIM-Métaindexeur offre deux modes de travail possible : avec ou sans vérification de réponses. La vérification est effectuée actuellement seulement au niveau de l'existence de documents-réponses. Pour cette fonctionnalité dans FIM-Métaindexeur, il y a deux paramètres à sélectionner dans le formulaire [Figure 6.7] :

- Vérification ou non de l'existence des documents
- Temps de vérification maximal

L'utilisateur peut choisir la vérification : par défaut, le mode sans vérification est sélectionné. C'est le mode, qui permet de recevoir la liste de réponses la plus rapide, avec le chargement de réseau minimal.

Vérification de l'existence des documents

Ce paramètre indique à FIM-Métaindexeur s'il doit tester la validité des liens renvoyés par chaque moteur : il ralentit bien évidement la recherche, et c'est la raison pour laquelle l'utilisateur peut choisir ou non cette option (et régler le temps de vérifications selon son besoin).

La vérification de l'existence des liens vers les documents-réponses renvoyés par chaque moteur interrogé est réalisée par plusieurs agents de récupération travaillant en parallèle.

De plus, parmi les moteurs de recherche interrogés, certains n'affichent pas le vrai titre de document sur leur page de réponses ; ce titre n'est donc pas accessible à FIM-Métaindexeur à partir des pages de réponses. Les agents de récupération profitent donc pour effectuer la récupérations du titre de chaque document accessible, pendant le processus de la vérification de la validité des liens.

Cette fonctionnalité est réalisée par la requête GET ou POST, selon l'acceptation par le moteur. Pour ne pas récupérer la totalité de la page, donc diminuer le chargement du réseau, FIM-Métaindexeur demande seulement l'en-tête HTML du document-réponse.

Temps de vérification maximal
Ce paramètre spécifie le temps maximal pendant lequel FIM-Métaindexeur teste la validité de chaque lien renvoyé par les différents moteurs de recherche. Les valeurs possibles à choisir sont : 5, 10, 20, 30, 40, 50 secondes. La valeur sélectionnée par défaut est 10 secondes. Au-delà du temps de vérification maximal sélectionné, FIM-Métaindexeur passe au lien suivant. Le document dont l'URL n'est pas accessible à la fin du temps de vérification maximal sélectionné, est considéré comme un document non existant, et ne sera pas affiché sur la liste de réponses de FIM-Métaindexeur.

6.2.3.3 Nombre maximum de réponses

Ce paramètre permet de limiter le nombre total de réponses renvoyées par FIM-Métaindexeur. Les valeurs possibles à sélectionner sont : 20, 40, 50, 60, 80, 100 et 150 documents. 50 étant la valeur sélectionnée par défaut. Pour raccourcir le temps de la recherche et diminuer le chargement du réseau, FIM-Métaindexeur contrôle le nombre de documents (existants ou non vérifiés, selon l'option choisie, [Par. 6.2.3.2] durant le processus de récupération. Lorsqu'il a récupéré la quantité choisie venant de tous les moteurs connectés, il arrête le processus de récupération.

6.2.3.4 Nombre maximum de réponses par moteur

Ce paramètre permet de limiter le nombre de réponses renvoyées par chaque moteur auquel se connecte FIM-Métaindexeur. Les valeurs possibles à sélectionner sont : 2, 5, 7, 10 réponses par moteur : 7 la valeur sélectionnée par défaut. Comme pour le paramètre précédent (nombre maximum de réponses total), ici FIM-Métaindexeur doit contrôler la quantité de documents à récupérer par chaque moteur connecté. La réalisation de cette fonctionnalité a été détaillée dans le chapitre précédent [Section 5.3.3 et Section 5.4.2].

6.2.3.5 Choix du serveur à interroger

Il est possible de spécifier à quels moteurs de recherche on souhaite se connecter pour effectuer la recherche.

Cette fonctionnalité est proposée à l'utilisateur dans la page du formulaire, au moyen d'une liste à sélection multiple : par défaut (aucune sélection), tous les moteurs listés seront interrogés.

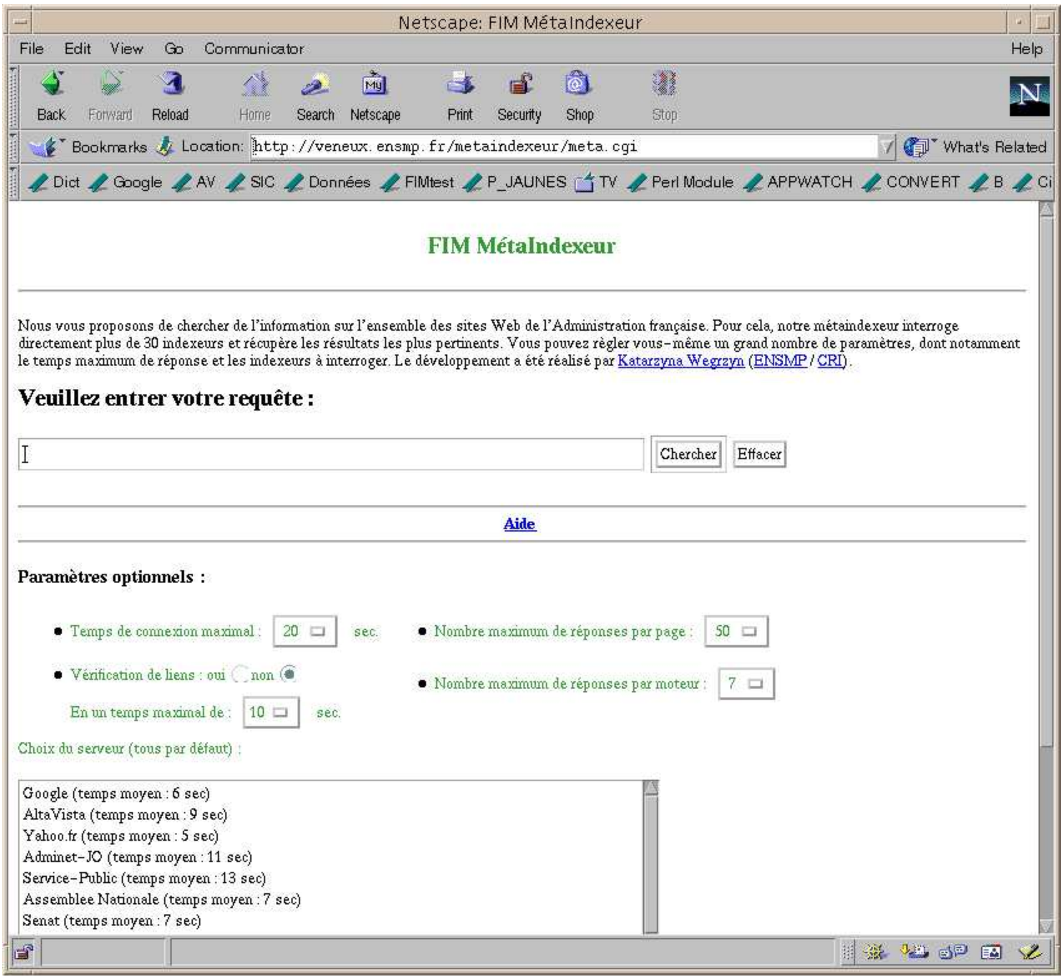
Le temps moyen figurant au regard de chaque moteur fournit une estimation du temps de réponse attendu, calculée à partir de l'historique de toutes les requêtes adressées à ce moteur par FIM-Métaindexeur. Cette estimation du temps moyen peut contenir des valeurs inexactes, dues à l'effet de signal d'alarme. On ne peut pas connaître la vraie valeur du temps nécessaire pour la communication au-delà du temps de connexion maximal. Ainsi, pour les moteurs où la communication avec FIM-Métaindexeur est arrêtée assez souvent, la vraie valeur de temps moyen de communication sera plus élevée.

6.2.4 Affichage des résultats

Toutes les réponses ramenées par les agents de récupération, avec ou sans vérification selon l'option choisie par l'utilisateur, sont transmises au module d'affichage des résultats de FIM-Métaindexeur. Dans ce module, toutes les réponses multiples (doublons) sont supprimées, les résultats des recherches venant de tous les moteurs interrogés sont triés, pour ensuite être affichés sous une forme homogène sur une page de réponses de FIM-Métaindexeur.

6.2.4.1 Tri des réponses

FIM-Métaindexeur fait le tri des réponses selon le nombre de réponses identiques parmi les résultats reçus de tous les moteurs, la pertinence du document et le poids (importance) du moteur qui a fourni cette réponse. Pour cela FIM-Métaindexeur suit les règles suivantes :



FIM MétaIndexeur

Nous vous proposons de chercher de l'information sur l'ensemble des sites Web de l'Administration française. Pour cela, notre métaindexeur interroge directement plus de 30 indexeurs et récupère les résultats les plus pertinents. Vous pouvez règler vous-même un grand nombre de paramètres, dont notamment le temps maximum de réponse et les indexeurs à interroger. Le développement a été réalisé par Katarzyna Wegrzyn (ENSMP / CRI).

Veuillez entrer votre requête :

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 : 6 sec)
AltaVista (temps moyen : 9 sec)
Yahoo.fr (temps moyen : 5 sec)
Adminet-JO (temps moyen : 11 sec)
Service-Public (temps moyen : 13 sec)
Assemblee Nationale (temps moyen : 7 sec)
Senat (temps moyen : 7 sec)

FIG. 6.8 – FIM-Méta indexeur. Sélection du serveur.

- Toutes les réponses multiples venant de différents moteurs, sont considérées comme des réponses très pertinentes. Leur poids de pertinence parmi toutes les réponses reçues, est augmenté par un coefficient spécifique. Ce coefficient est calculé dynamiquement pour chaque réponse multiple, selon la nombre de répétitions de réponses analysées.

> Sur la page de réponses de FIM-Métaindexeur les réponses multiples ne sont affichées qu'une seule fois (élimination des doublons), par contre leur poids de pertinence est corrigé.

- FIM-Métaindexeur ne mesure ou n'évalue pas lui-même la pertinence des documents-réponses. Il préserve les mesures de pertinence données par chaque moteur interrogé.
- FIM-Métaindexeur fait le tri des toutes les réponses, selon les données connues (poids de pertinence calculé pour les documents multiples, poids fourni par les moteurs interrogé et poids donné au moteur) pour fournir la liste de documents-réponses dans un ordre de pertinence. La méthode pour trier les documents-réponses est illustrée dans la figure 6.9.

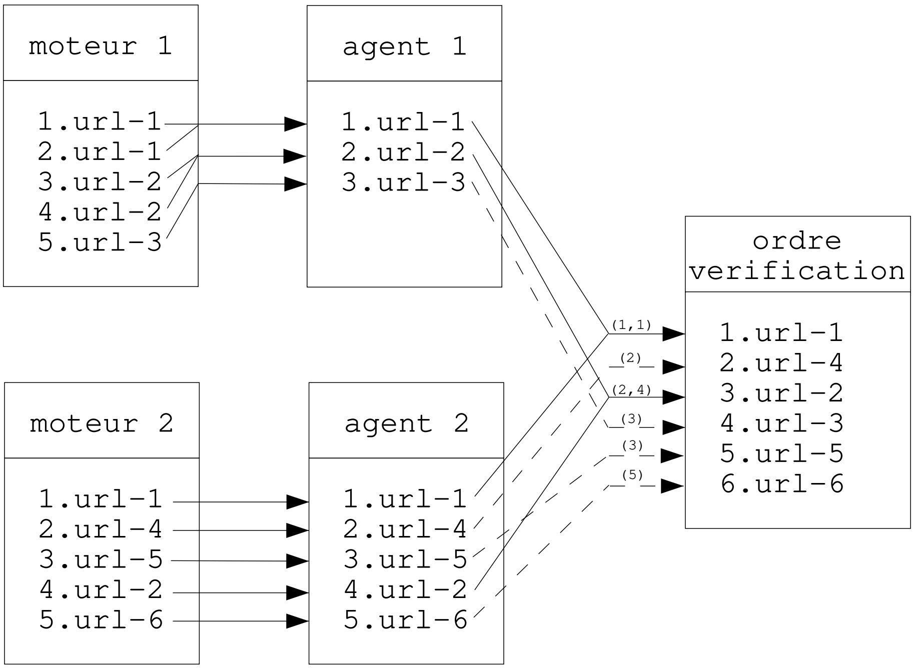


FIG. 6.9 – Tri des réponses.

6.2.4.2 La page de réponses

FIM-Métaindexeur fournit en même temps sur sa page de réponses une liste de tous les résultats des recherches, et également les statistiques effectuées sur cette recherche.

Les réponses
Les résultats de la recherche sont affichés par ordre de pertinence décroissante. Le nombre de réponses fournies est plafonné par la quantité sélectionnée pendant la configuration, paramètre **Nombre maximum des réponses** [Section 6.2.3.3]. Le nombre de réponses venant de chaque moteur est plafonné par le paramètre **Nombre maximum de réponses par serveur** [Section 6.2.3.4]

Pour chaque résultat donné, sont indiqués le nom du moteur qui a fourni ce résultat, ainsi que le titre et l'adresse (URL) de la réponse trouvée.

- Nom du moteur :

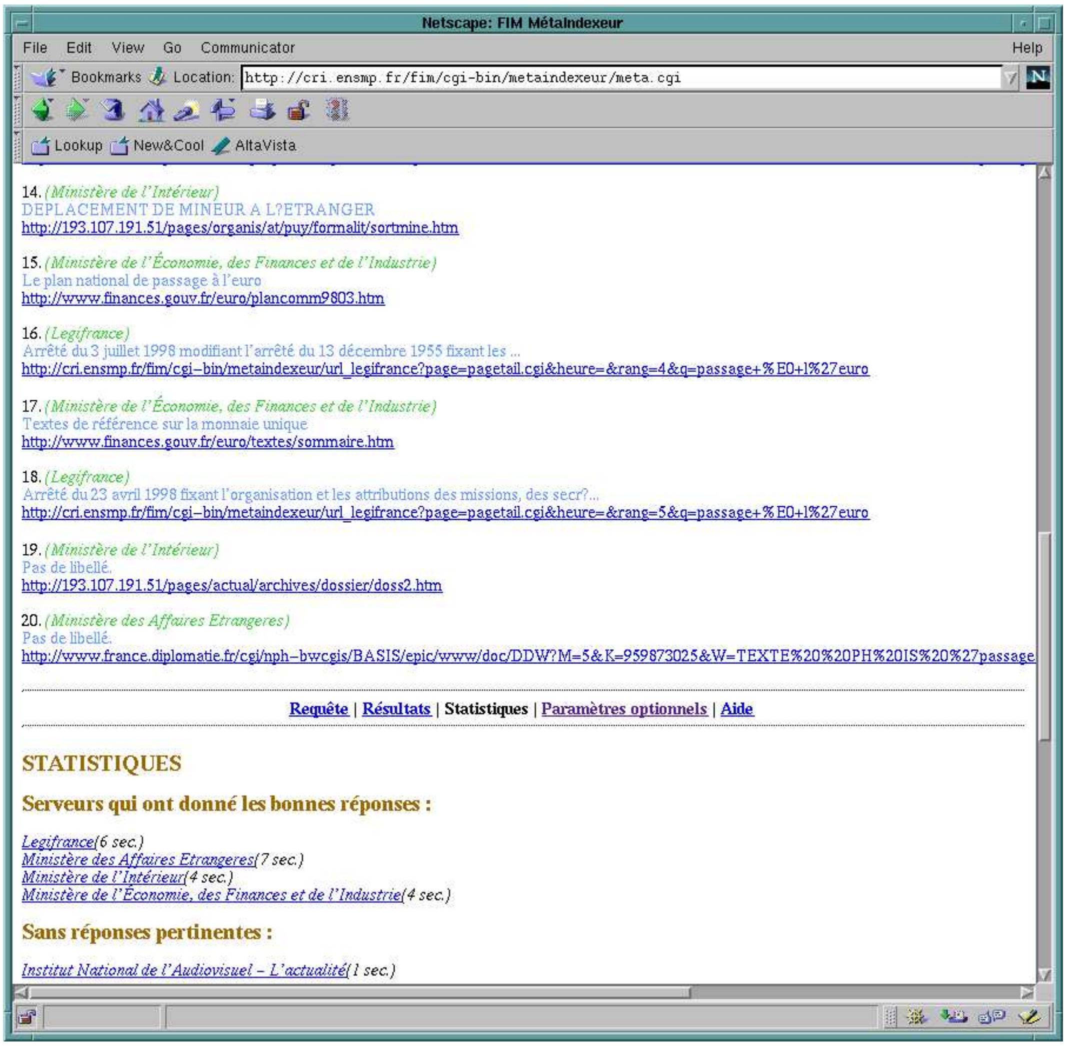
  Cette information indique à l'utilisateur le moteur qui a fourni la réponse. Dans le cas de réponses multiples, les noms de tous les moteurs donnant cette réponse sont affichés.

- Titre du document :

  Il existe les deux possibilités : le moteur interrogé fournit le titre exact du document-réponse (titre indiqué dans les balises HTML : Titre), ou donne la référence dans une forme différente. Dans ce cas, si l'option **Vérification de l'existence des documents** [Section 6.2.3.2] est sélectionnée, FIM-Métaindexeur récupère lui-même le titre exact du document, et l'affiche sur la page de réponses. Par contre pour tous les documents-réponses dont le vrai titre n'est pas connu, c'est la référence donnée par le moteur qui a renvoyé cette réponse qui sera affichée comme titre.

- Adresse (URL) :

  Elle sera présentée comme un lien cliquable, donnant l'accès direct au document. Bien sûr, dans le cas où l'option **Vérification de l'existence des documents** [Section 6.2.3.2] n'était pas choisie il est possible que ce lien ne soit pas valide.



14. *(Ministère de l'Intérieur)*
DEPLACEMENT DE MINEUR A L?ETRANGER
http://193.107.191.51/pages/organis/at/puy/formalit/sortmine.htm

15. *(Ministère de l'Économie, des Finances et de l'Industrie)*
Le plan national de passage à l'euro
http://www.finances.gouv.fr/euro/plancomm9803.htm

16. *(Legifrance)*
Arrêté du 3 juillet 1998 modifiant l'arrêté du 13 décembre 1955 fixant les ...
http://cri.ensmp.fr/fim/cgi–bin/metaindexeur/url_legifrance?page=pagetail.cgi&heure=&rang=4&q=passage+%E0+l%27euro

17. *(Ministère de l'Économie, des Finances et de l'Industrie)*
Textes de référence sur la monnaie unique
http://www.finances.gouv.fr/euro/textes/sommaire.htm

18. *(Legifrance)*
Arrêté du 23 avril 1998 fixant l'organisation et les attributions des missions, des secr?...
http://cri.ensmp.fr/fim/cgi–bin/metaindexeur/url_legifrance?page=pagetail.cgi&heure=&rang=5&q=passage+%E0+l%27euro

19. *(Ministère de l'Intérieur)*
Pas de libellé.
http://193.107.191.51/pages/actual/archives/dossier/doss2.htm

20. *(Ministère des Affaires Etrangeres)*
Pas de libellé.
http://www.france.diplomatie.fr/cgi/nph–bwcgis/BASIS/epic/www/doc/DDW?M=5&K=959873025&W=TEXTE%20%20PH%20IS%20%27passage

Requête | **Résultats** | **Statistiques** | **Paramètres optionnels** | **Aide**

STATISTIQUES

Serveurs qui ont donné les bonnes réponses :

Legifrance(6 sec.)
Ministère des Affaires Etrangeres(7 sec.)
Ministère de l'Intérieur(4 sec.)
Ministère de l'Économie, des Finances et de l'Industrie(4 sec.)

Sans réponses pertinentes :

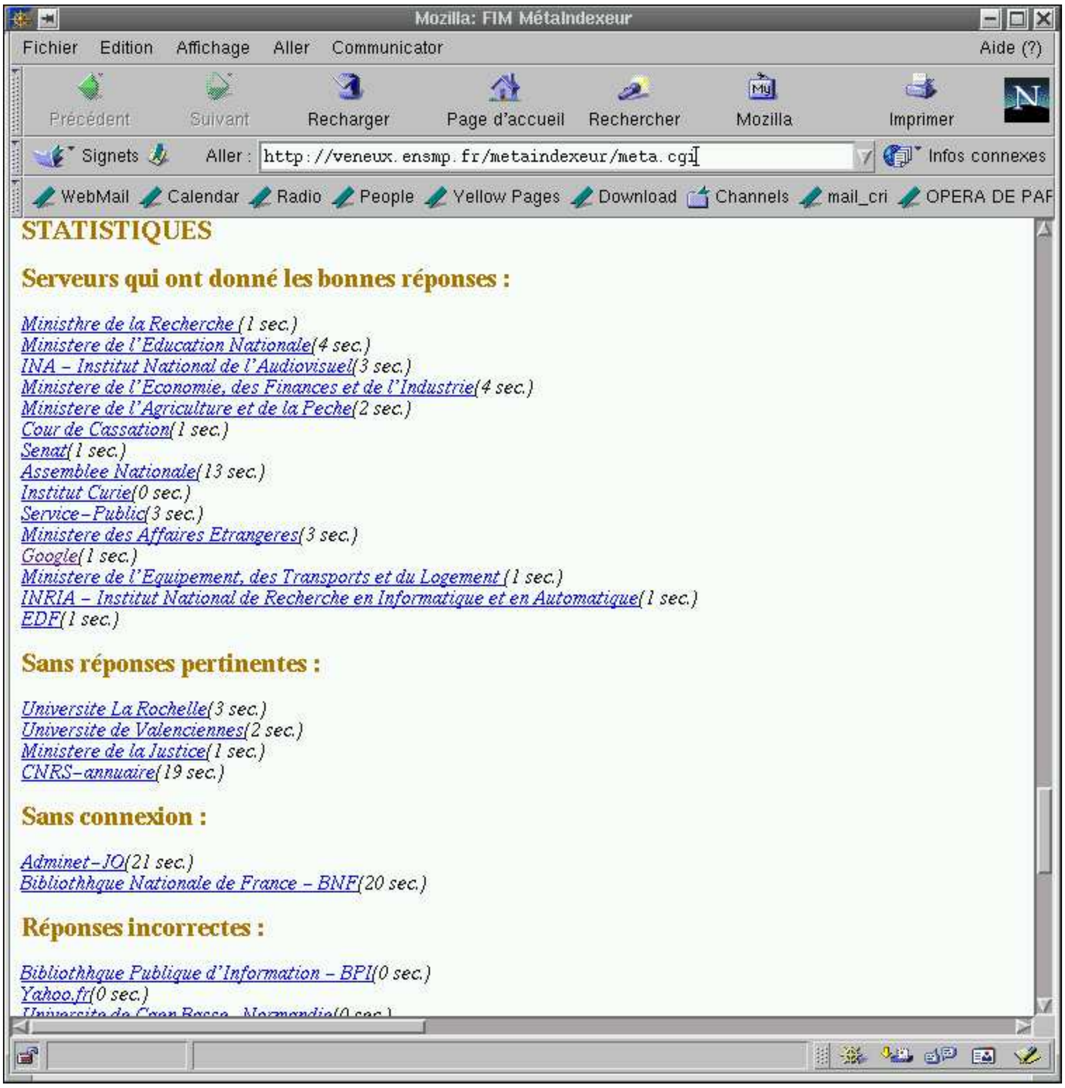
Institut National de l'Audiovisuel – L'actualité(1 sec.)

FIG. 6.10 – FIM-Métaindexeur. La page des réponses.

Le titre du document et son adresse ont été séparés volontairement (au lieu d'afficher un titre de document cliquable avec adresse cachée sous le lien) pour permettre à l'utilisateur de connaître facilement la source physique du document-réponse.

Statistiques sur les réponses obtenues

FIM-Métaindexeur fournit un certain nombre de statistiques concernant la recherche qui vient d'être effectuée. La fraction de page de réponses contenant des statistiques est présentée dans la figure 6.11.



http://veneux.ensmp.fr/metaindexeur/meta.cgi

STATISTIQUES

Serveurs qui ont donné les bonnes réponses :

Ministhre de la Recherche (1 sec.)
Ministere de l'Education Nationale(4 sec.)
INA – Institut National de l'Audiovisuel(3 sec.)
Ministere de l'Economie, des Finances et de l'Industrie(4 sec.)
Ministere de l'Agriculture et de la Peche(2 sec.)
Cour de Cassation(1 sec.)
Senat(1 sec.)
Assemblee Nationale(13 sec.)
Institut Curie(0 sec.)
Service–Public(3 sec.)
Ministere des Affaires Etrangeres(3 sec.)
Google(1 sec.)
Ministere de l'Equipement, des Transports et du Logement (1 sec.)
INRIA – Institut National de Recherche en Informatique et en Automatique(1 sec.)
EDF(1 sec.)

Sans réponses pertinentes :

Universite La Rochelle(3 sec.)
Universite de Valenciennes(2 sec.)
Ministere de la Justice(1 sec.)
CNRS–annuaire(19 sec.)

Sans connexion :

Adminet–JO(21 sec.)
Bibliothhque Nationale de France – BNF(20 sec.)

Réponses incorrectes :

Bibliothhque Publique d'Information – BPI(0 sec.)
Yahoo.fr(0 sec.)
Universite de Caen Basse - Normandie(0 sec.)

FIG. 6.11 – FIM-Métaindexeur. La page de résultats, les statistiques.

Toutes les données statistiques sont affichées dans les catégories suivante :

- Les listes de réponses sont regroupées selon le type de réponses fournies (Bonnes réponses, Interrogation sans connexion, Réponses non pertinentes, Réponses avec message d'erreur). Les noms des moteurs sont affichés selon les groupes auxquels ils appartiennent :

  - moteurs ayant fourni les bonnes réponses,
  - moteurs n'ayant pas fourni de réponses pertinentes,
  - moteurs auxquels FIM-Métaindexeur n'est pas parvenu à se connecter dans le délai indiqué,
  - moteurs ayant renvoyé un message d'erreur.

  > Chaque nom de moteur est affiché comme un lien cliquable vers la page de formulaire de recherche de ce moteur.

- Temps de réponse de chaque moteur.

  > Les noms des moteurs interrogés sont affichés avec indication de leur propre temps de réponse.

- Temps de réponse global.

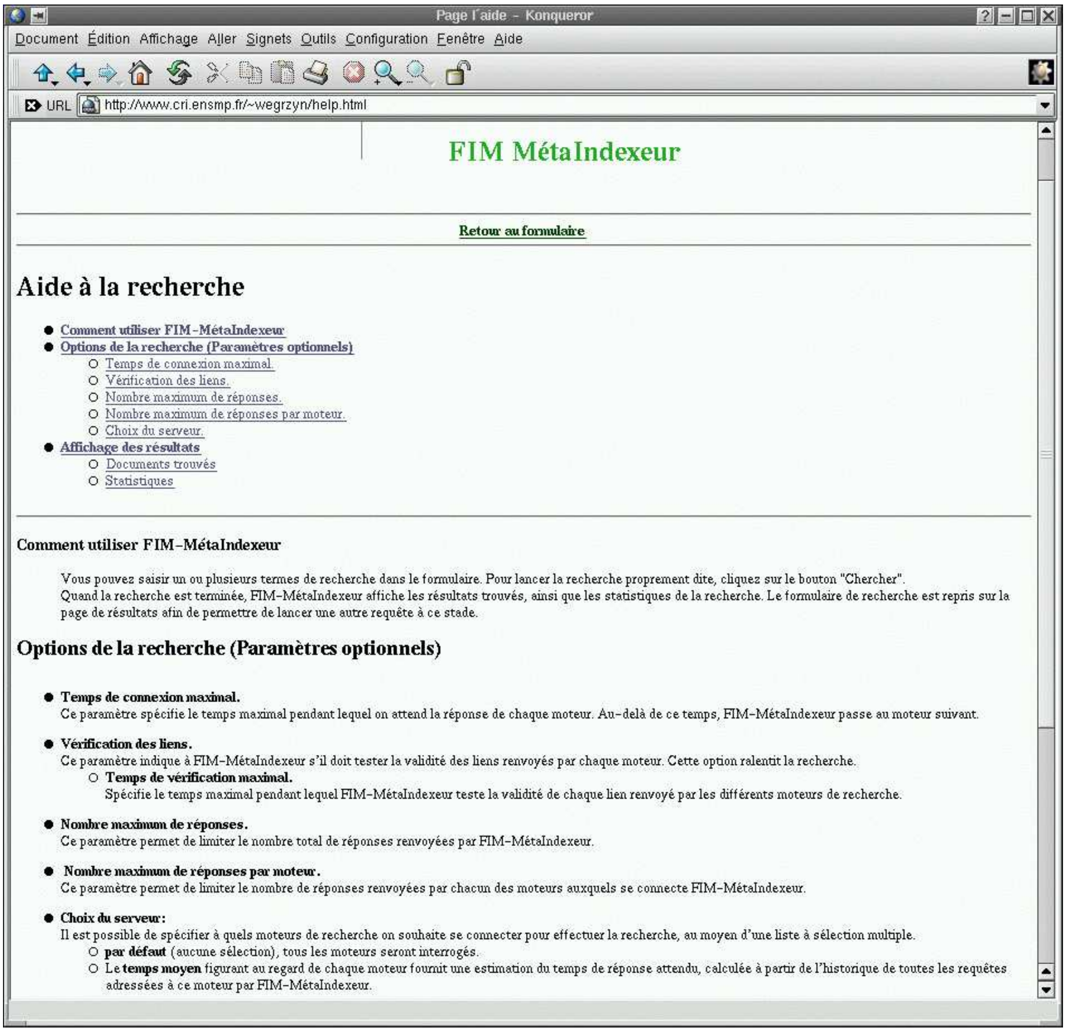
  > FIM-Métaindexeur affiche également sur la page de réponses le temps de réponse total écoulé pour obtenir les résultats de la recherche.

6.2.5 La page d'aide

Au cours de l'utilisation de FIM-Métaindexeur (page de formulaire ou page de réponses) il est toujours possible d'accéder à une page d'aide donnant des informations supplémentaires sur le fonctionnement de FIM-Métaindexeur. Cette page d'aide est présentée dans la figure 6.12.

6.3 Statistiques effectuées à partir du fichier LOG

Toutes les informations concernant les recherches effectuées par FIM-Métaindexeur sont sauvegardées dans un fichier de LOG spécial. A partir de ces données, différentes statistiques peuvent être effectués :



http://www.cri.ensmp.fr/~wegrzyn/help.html

FIM MétaIndexeur

Retour au formulaire

Aide à la recherche

- Comment utiliser FIM-MétaIndexeur
- Options de la recherche (Paramètres optionnels)
    - Temps de connexion maximal.
    - Vérification des liens.
    - Nombre maximum de réponses.
    - Nombre maximum de réponses par moteur.
    - Choix du serveur.
- Affichage des résultats
    - Documents trouvés
    - Statistiques

Comment utiliser FIM-MétaIndexeur

Vous pouvez saisir un ou plusieurs termes de recherche dans le formulaire. Pour lancer la recherche proprement dite, cliquez sur le bouton "Chercher".
Quand la recherche est terminée, FIM-MétaIndexeur affiche les résultats trouvés, ainsi que les statistiques de la recherche. Le formulaire de recherche est repris sur la page de résultats afin de permettre de lancer une autre requête à ce stade.

Options de la recherche (Paramètres optionnels)

- **Temps de connexion maximal.**
  Ce paramètre spécifie le temps maximal pendant lequel on attend la réponse de chaque moteur. Au-delà de ce temps, FIM-MétaIndexeur passe au moteur suivant.
- **Vérification des liens.**
  Ce paramètre indique à FIM-MétaIndexeur s'il doit tester la validité des liens renvoyés par chaque moteur. Cette option ralentit la recherche.
    - **Temps de vérification maximal.**
      Spécifie le temps maximal pendant lequel FIM-MétaIndexeur teste la validité de chaque lien renvoyé par les différents moteurs de recherche.
- **Nombre maximum de réponses.**
  Ce paramètre permet de limiter le nombre total de réponses renvoyées par FIM-MétaIndexeur.
- **Nombre maximum de réponses par moteur.**
  Ce paramètre permet de limiter le nombre de réponses renvoyées par chacun des moteurs auxquels se connecte FIM-MétaIndexeur.
- **Choix du serveur:**
  Il est possible de spécifier à quels moteurs de recherche on souhaite se connecter pour effectuer la recherche, au moyen d'une liste à sélection multiple.
    - **par défaut** (aucune sélection), tous les moteurs seront interrogés.
    - Le **temps moyen** figurant au regard de chaque moteur fournit une estimation du temps de réponse attendu, calculée à partir de l'historique de toutes les requêtes adressées à ce moteur par FIM-MétaIndexeur.

FIG. 6.12 – Page d'aide pour FIM-Métaindexeur

- statistiques globales,
- statistiques sur les réponses de chaque moteur,
- statistiques sur le temps de réponse,
- statistiques sur l'heure d'interrogation,

6.3.1 Statistiques globales

Parmi toutes les réponses renvoyées par les moteurs interrogés, beaucoup ne sont pas satisfaisantes. Les pourcentages statistiques des différents types de réponses reçues sont présentés dans la figure 6.13. Ces statistiques montrent que seulement 34% des résultats de recherche renvoyés par les moteurs interrogés appartiennent à l'ensemble de réponses intéressantes pour FIM-Métaindexeur : c'est parmi ces réponses que FIM-Métaindexeur cherche les résultats à afficher sur sa page de réponses.

Cette analyse montre également que grâce à FIM-Métaindexeur la majorité des réponses sans intérêt peut être éliminée : on considère alors que le bruit des réponses est diminué d'une manière significative.

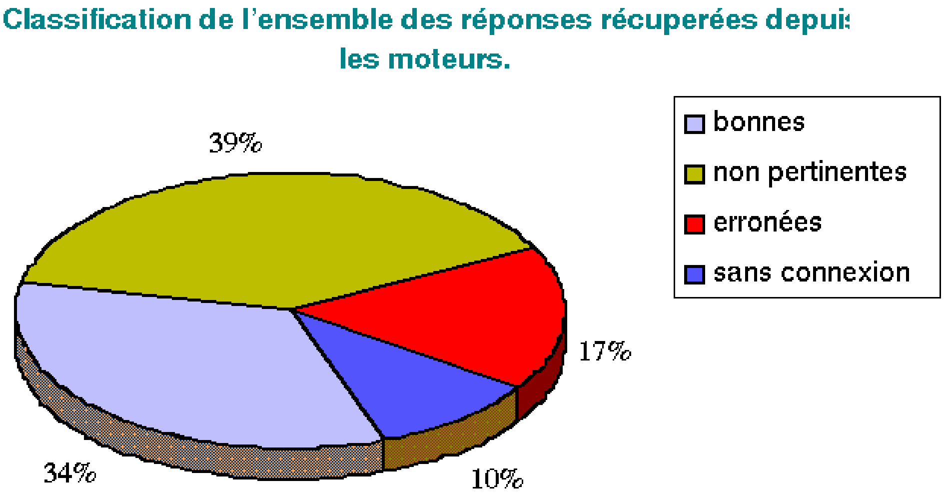


FIG. 6.13 – Les statistiques globales.

6.3.2 Statistiques sur les réponses de chaque moteur

Pour chaque moteur de recherche, on présente les statistiques concernant les ratio des différents types de réponses données. Ces statistiques présentées dans la figure 6.14 donnent des résultats regroupés en quatre catégories :

- pourcentage des interrogations sans connexion,
- pourcentage des bonnes réponses,
- pourcentage des réponses non pertinentes,
- pourcentage des réponses avec message d'erreur.

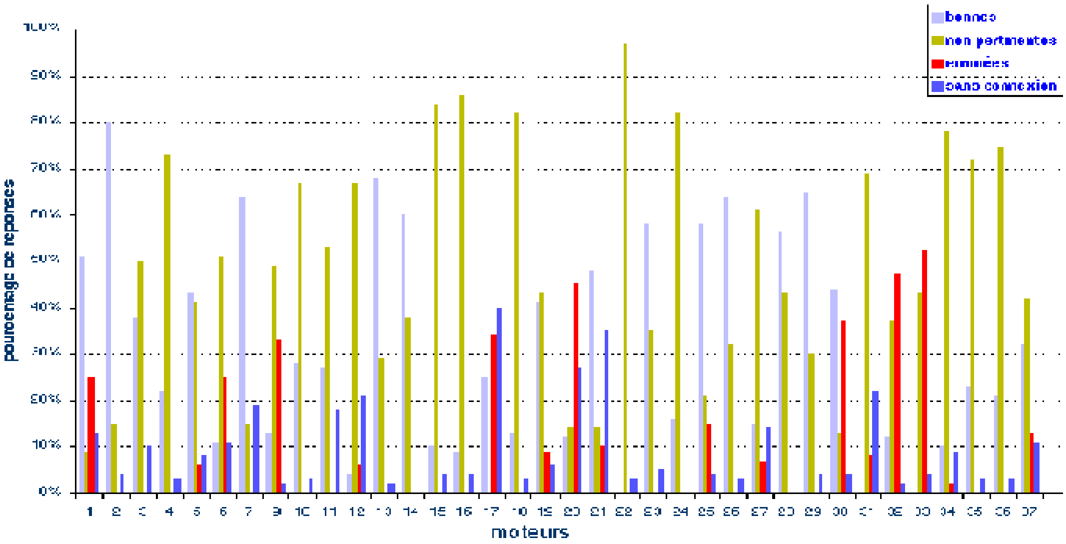


FIG. 6.14 – FIM Métaindexeur. Statistiques des réponses.

6.3.3 Le temps de réponse

Les données concernant les temps de réponse de tous les moteurs de recherche interrogés sont présentées dans deux contextes différentes :

- nombre de réponses en fonction du temps de réponse,
- estimation des différents temps de réponse (minimal, maximal et moyen).

6.3.3.1 Nombre de réponses en fonction du temps de réponse

La figure 6.15 présente les statistiques effectuées pour chaque moteur, concernant la quantité de réponses reçues dans les différentes périodes de temps. Ce graphe montre que la plupart des réponses sont fournies dans un délai inférieur à 15 secondes. Par contre, on observe également une augmentation assez importante du nombre des réponses pour les temps de réponse égaux à 20 secondes. Cette plus grande concentration des réponses dans le temps égal à 20 secondes est expliquée par le fait que cette valeur du temps contient toutes les réponses plus lentes qui sont coupées par le signal d'alarme déclanché à la fin du temps **Temps de connexion maximal** sélectionnée [Section 6.2.3.1], choisi par défaut à 20 secondes.

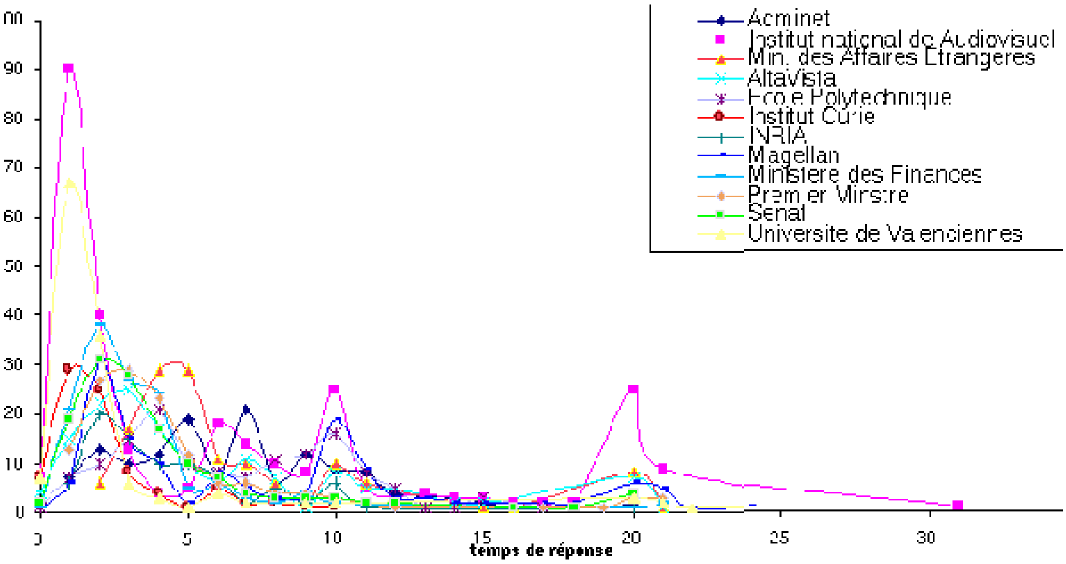


FIG. 6.15 – Nombres de réponses en fonction de temps de réponse.

6.3.3.2 L'estimation de différents temps de réponses

La figure 6.16 présente les statistiques concernant le temps de réponse donné par chacun des moteurs de recherche interrogés. Ces statistiques contiennent les valeurs calculées pour les temps différents :

- temps de réponse minimal,
- temps de réponse moyen,
- temps de réponse maximal.

Les valeurs de temps de réponses moyen et maximal ont été calculées dans les deux cas différents :

- pour toutes les recherches effectuées, même celles qui ont été coupées par le signal d'alarme,

– pour les seules recherches, qui n'ont pas été arrêtées par le signal d'alarme.

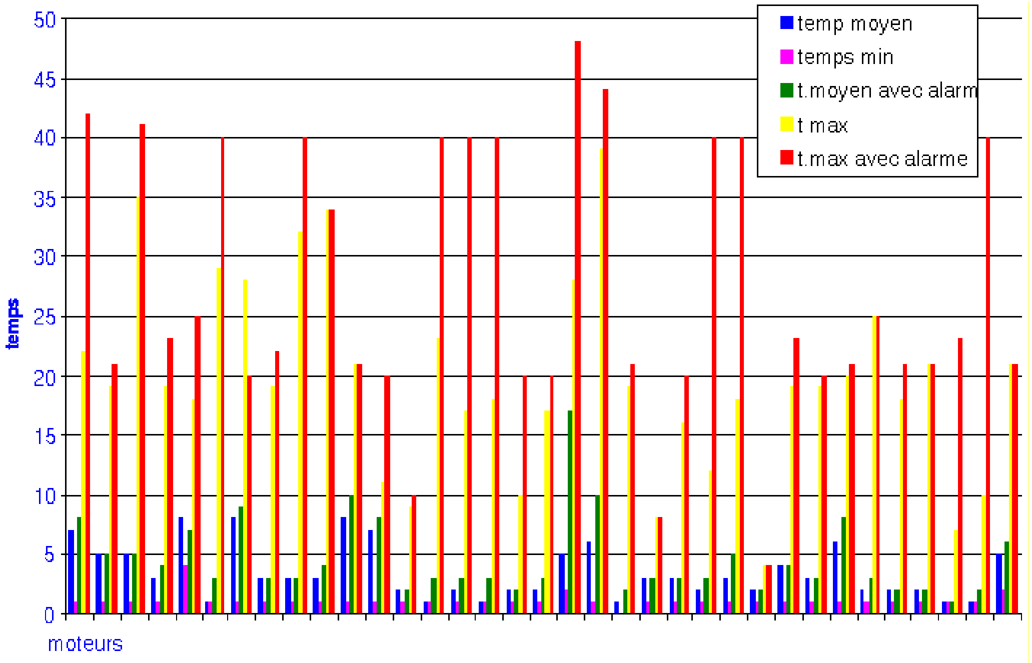


FIG. 6.16 – Statistiques sur les temps de réponse.

6.3.4 L'heure de l'interrogation

Cette statistique montre l'évolution de deux paramètres :

- le chargement de FIM-Métaindexeur
- le temps de réponse en fonction de l'heure d'interrogation.

Les graphes sont présentés dans la figure 6.17. On observe ainsi que FIM-Métaindexeur est plus particulièrement interrogé dans les heures ouvrables (heures françaises), avec une assez forte augmentation en début d'après-midi. Le fait que le chargement de FIM-Métaindexeur diminue presque totalement durant la nuit était prévu, puisque FIM-Métaindexeur dédié à retrouver les documents de l'administration française est consulté en majorité par les internautes français.

On observe aussi que le temps de réponse de FIM-Métaindexeur est assez constant. On suppose donc que le temps de réponse n'est pas fortement lié au chargement de FIM-Métaindexeur.

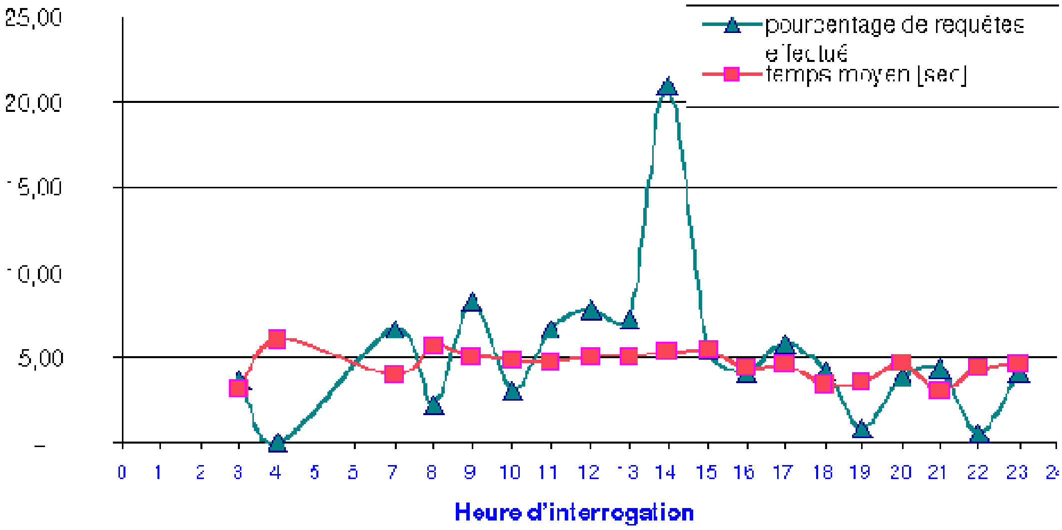


Fig. 6.17 – Statistique de temps de réponse en fonction de l'heure

Troisième partie

Aspects statistiques

Chapitre 7

Statistiques et classification de moteurs

Dans le cadre de cette thèse, à côté des analyses de moteurs de sites de l'administration, nécessaires pour procéder à l'interrogation par le FIM-Métaindexeur [Chapitre 6], nous avons réalisé certains tests, pour ensuite présenter les statistiques appropriées. Ces analyses permettent de caractériser et de classifier lese moteurs interrogés.

7.1 La méthode appliquée

Pour réaliser la classification des moteurs, nous avons effectué des tests spéciaux sur un ensemble représentatif de questions dans le domaine de l'administration.

7.1.1 Les statistiques effectuées

Le but de cette analyse est de réaliser des statistiques sur les moteurs de recherche et sur les serveurs de l'administration française en vue de les caractériser ou de les classifier. Toutes les analyses et les statistiques ont été réalisées pour 26 serveurs [Tableau 7.1].

Chacun de ces moteurs a été interrogé séparément (avec, pour chaque moteur, le même ensemble de questions préparées), et tous les résultats de recherche nécessaires pour les statistiques ont été sauvegardés dans une base de données relationnelle. Cette

id_serveur	serveur
1	Adminet-JO
2	Assemblée Nationale
3	Association Agropolis
4	Bibliothèque Publique d'Information - BPI
5	CNRS-annuaire
6	Commissariat général du Plan
7	Cour de Cassation
8	EDF
9	INA - Institut National de l'Audiovisuel
10	INRA - Institut National de la Recherche Agronomique
11	INRIA - Institut National de Recherche en Informatique et en Automatique
12	Institut Curie
13	Laboratoire Central des Ponts et Chaussées
14	Ministère de l'Economie, des Finances et de l'Industrie
15	Ministère de l'Education Nationale
16	Ministère de l'Equipement, des Transports et du Logement
17	Ministère de la Fonction Publique
18	Ministère de la Justice
19	Ministère de la Recherche
20	Ministère des Affaires Étrangères
21	Secrétariat d'Etat au Commerce Extérieur
22	Sénat
23	Service-Public
24	Université de Caen Basse-Normandie
25	Université de Valenciennes
26	Université La Rochelle

TAB. 7.1 – Serveurs analysés pour les statistiques.

solution donne l'accès aux données et permet la manipulation de données insérées dans la base de façon très souple. Deux catégories de tests ont été effectuées :

- uniquement pour les pages de réponses,
- sur les pages de réponses et également les documents-réponses.

7.1.1.1 Analyse des pages de réponses

Pour effectuer toutes les analyses nous avons envoyé un grand nombre de requêtes à chaque moteur analysé. L'ensemble des requêtes à envoyer a été préparé auparavant : un fichier-test (TEST1) contenant environ 1000 questions [Section 7.1.2]. Les pages de réponses aux questions envoyées (TEST1) venant de chaque moteur ont été analysées. Pour diminuer l'occupation du moteur interrogé, le temps de récupération et la charge du réseau, le nombre de réponses était plafonné à 1000. Les données récupérées étaient ensuite utilisées pour créer les différents statistiques :

- Nombre des réponses récupérées
- Temps de réponse
  - minimum, maximum, moyen.
- Temps de réponse en fonction du nombre des réponses récupérées.
- Profil thématique des serveurs. Corrélation entre les serveurs.

7.1.1.2 Analyse des pages de réponses et des documents-réponses appropriés

Ce test est plus ppoussé. Nous avons réalisé une analyse non seulement des pages de réponses, mais également des documents-réponses. Pour cela chaque moteur été interrogé avec le même ensemble de 50 questions-test préparées auparavant (TEST2) [Section 7.1.2]. Toutes les pages de réponses ont été analysées, pour ensuite récupérer les documents-réponses. La vérification de pertinence et le classement des réponses ont été effectués sur les 50 premières réponses de chaque moteur. Ensuite nous avons réalisé les statistiques suivantes :

- Accessibilité des documents-réponses
- Taille des documents-réponses
  - Taille moyenne générale
  - Taille pour chaque moteur
- Temps de réponse en fonction de la taille des documents-réponse

  - minimum, maximum, moyen,
- Pourcentage de documents existants
- Existence des mots de la requête dans le document-réponse

7.1.2 Les requêtes de teste

7.1.2.1 Analyse de la structure des requêtes posées

Avant de créer le fichier de requêtes-test représentatif, nous avons effectué une recherche visant à étudier la structure de “vraies” requêtes posées par les internautes. Pour cela nous avons analysé un grand ensemble de questions récupérées à partir de fichier LOG du site Adminet[1], dédié à la recherche administrative, et hébergé dans notre centre de recherche. Nous avons analysé un ensemble contenant plus de 3000 questions. Différentes statistiques ont été effectuées.

Nombre de mots dans la question

Cette statistique montre combien une question contient de mots. Le tableau [Tableau 7.2] présente le nombre de questions de chaque longueur (la longueur des questions etant exprimée en nombre de mots).

nombre de mots	1	2	3	4	5	6	7	8	9	10
nombre de questions	443	1014	717	410	172	56	20	10	3	0
nombre de mots	11	12	13	14	15	16	17			
nombre de questions	1	0	0	0	0	0	1			

TAB. 7.2 – Nombre de mots dans la question.

Pour les questions ont été calculées les longueurs minimale, maximale est moyenne :

minimale = 1

maximale = 17

moyenne = 2.71

Longueur de mots utilisés dans les requêtes.

Le tableau [Tableau 7.3] montre les longueurs des mots des requêtes.

---

[1] http ://www.adminet.fr

nombre de caractères	1	2	3	4	5	6	7	8	9	10
nombre de mots	99	506	389	687	492	884	867	821	785	823
nombre de caractères	11	12	13	14	15	16	17	18	19	20
nombre de mots	468	408	223	126	76	76	15	8	10	5
nombre de caractères	21	22	23	24	25	26	27	28	29	30
nombre de mots	3	3	0	1	1	1	2	2	0	1
nombre de caractères	31	32	33	34	35	36	37	38	39	40
nombre de mots	0	0	1	1	0	0	0	0	0	1
nombre de caractères	41	42	43	44	45	46	47	48		
nombre de mots	0	0	0	0	0	0	0	1		

TAB. 7.3 – Longueur des mots dans les requêtes.

Pour tous les mots de chaqcune des requêtes les longueurs minimale, maximale et moyenne ont été calculées.

minimale = 1

maximale = 48

moyenne = 7.51

7.1.2.2 Les fichiers-test contenant les requêtes pour les statistiques

On se basant sur les statistiques effectuées à partir de l'ensemble des questions de fichier LOG de site Adminet, deux fichiers-test ont été créés, l'un contenant environ 1000 requêtes (TEST1) et le deuxième contenant seulement 50 requêtes (TEST2). Pour montrer que les deux fichiers-test sont représentatifs, les mêmes statistiques ont été effectuées. Toutes les statistiques ont été effectuées pour ces deux fichiers-test (TEST1 et Test2)

Nombre de mots dans la question

Le tableau suivant présente le nombre de questions de chaque longueur dans le deux fichiers-test (la longueur de la question est exprimée en nombre de mots). Le tableau 7.4 donne la statistique pour le fichier TEST1 et le tableau 7.5 pour le fichier TEST2.

Les longueurs des questions : minimale, maximale et moyenne ont été calculées également (avec comparaison des valeurs pour Adminet) [Tableau 7.6]. Les données montrent que le nombre de mots moyen d'une question est égal à 2,7 pour tous les fichiers-tests (la littérature indique en effet que c'est 2,4 pour les moteurs de recherche en général).

nombre de mots	1	2	3	4	5	6	7	8
nombre de question	153	319	234	138	66	15	6	2

Tab. 7.4 – Nombre de mots dans la question (Test1).

nombre de mots	1	2	3	4	5	6
nombre de question	13	18	9	7	3	5

Tab. 7.5 – Nombre de mots dans la question (Test2).

Les caractéristiques comparatives concernant les longueurs des questions pour les fichiers-tests (test1, test2 et Adminet) sont présentées dans la figure 7.1.

fichier	minimum	maximum	moyen
Test1	1	8	2.70
Test2	1	6	2.70
Adminet	1	17	2.71

Tab. 7.6 – Longueur des questions : minimale, maximale et moyenne pour les fichiers test1,test2 et pour les questions d'Adminet.

Longueur des mots utilisés dans les requêtes.

Le tableau 7.7 présente les statistiques sur la longueur des mots de la requête (exprimée en nombre de caractères) pour le fichier Test1, le tableau 7.8 pour le fichier Test2.

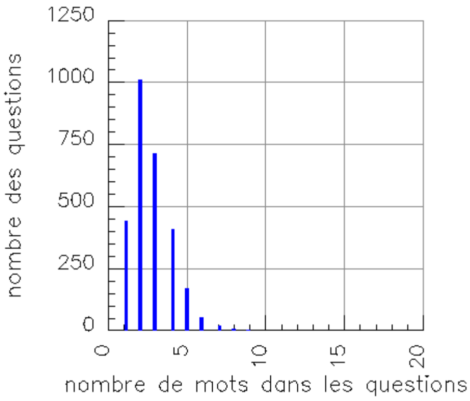
nombre de caractères	1	2	3	4	5	6	7	8	9	10
nombre de mots	31	193	122	237	191	338	301	279	245	250
nombre de caractères	11	12	13	14	15	16	17	18	19	20
nombre de mots	126	99	45	30	18	4	5	3	1	0
nombre de caractères	21	22	23	24	25	26	27	28	29	30
nombre de mots	0	0	1	2	0	0	0	1	0	0
nombre de caractères	31	32	33	34	35	36	37	38	39	40
nombre de mots	0	0	0	0	0	0	0	0	0	0
nombre de caractères	41	42	43	44	45	46				
nombre de mots	0	0	0	0	0	1				

Tab. 7.7 – Longueur des mots dans les requêtes (Test1).

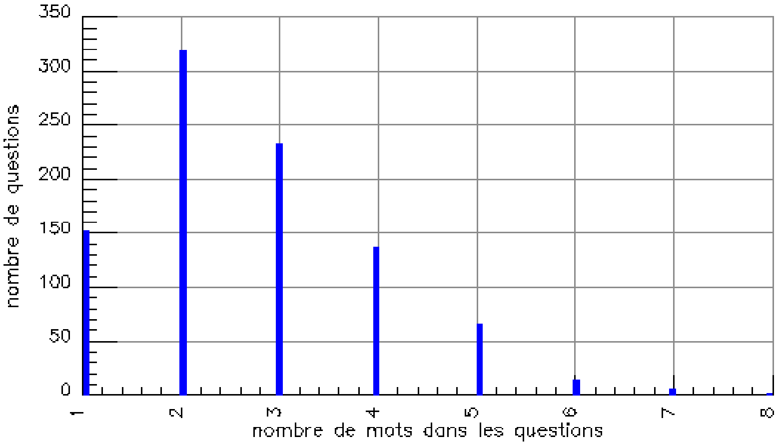
nombre de caractères	1	2	3	4	5	6	7	8	9	10
nombre de mots	2	17	7	9	11	19	25	24	15	7
nombre de caractères	11	12	13	14	15	16	17	18	19	20
nombre de mots	4	5	4							

Tab. 7.8 – Longueur des mots dans les requêtes (Test2).

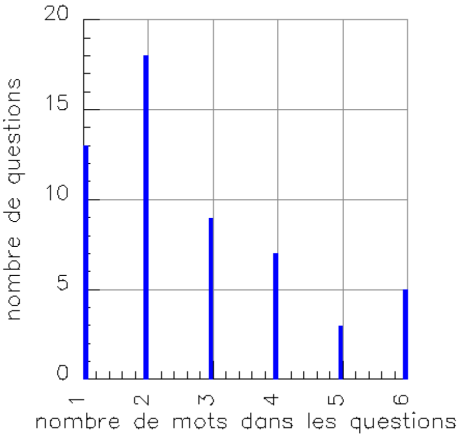
Les longueurs des mots minimale, maximale et moyenne ont été calculées également [Tableau 7.9].



Adminet



TEST1



TEST2

FIG. 7.1 – Statistiques sur la longueur de la question pour les questions de fichiers-tests (TEST1 et TEST2) et les questions d'Adminet.

fichier	minimum	maximum	moyen
TEST1	1	46	7.12405866032501
TEST2	1	13	6.64429530201342
Adminet	1	48	7.51648777964567

TAB. 7.9 – Longueurs des mots : minimale, maximale et moyenne.

Les caractéristiques comparatives concernant la longueur des mots dans les questions de fichiers-tests (TEST1, TEST2 et ADMINET) sont présentées dans la figure 7.2.

Les statistiques effectuées montrent que les requêtes des fichiers-tests ont la même structure que les requêtes du site Adminet. On peut donc les considérer comme des ensembles de requêtes représentatives.

7.1.3 La base de données

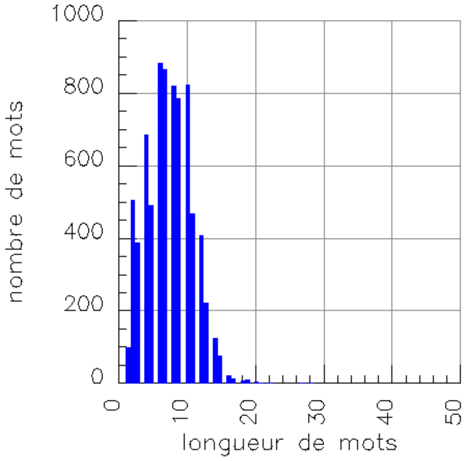
Tous les résultats de nos tests ont été sauvegardés dans la base relationnelle. Selon la méthode de test et les fichiers de requêtes utilisés (TEST1 ou TEST2) [Section 7.1.2], les données sont divisées en deux catégories :

- les données concernant les résultats de recherche (pages de réponses)
- les données concernant les documents-réponses (vérification de réponses)

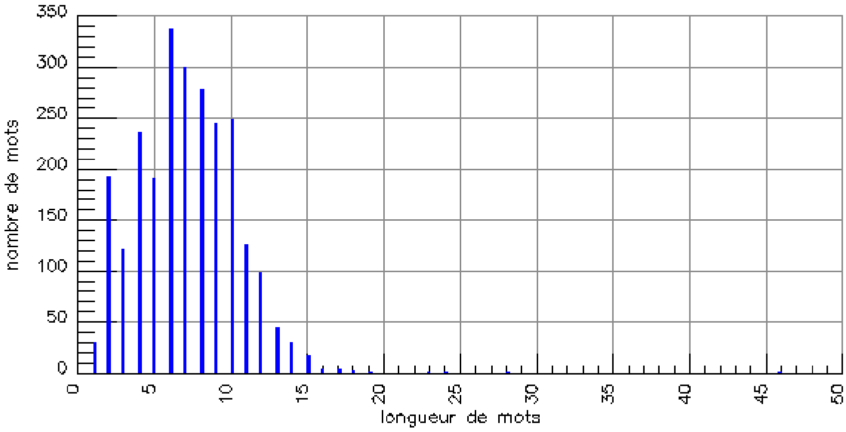
7.1.3.1 Les données pour les statistiques concernant les résultats de recherche (pages de réponses)

La base a été créée à partir des réponses de chaque moteur interrogé sur les questions du fichier-test TEST1 (environ 1000 questions). Pour chaque résultat de recherche (réponse d'un moteur à une question), les informations suivantes ont été sauvegardées dans la base (table STAT_REPONSE) :

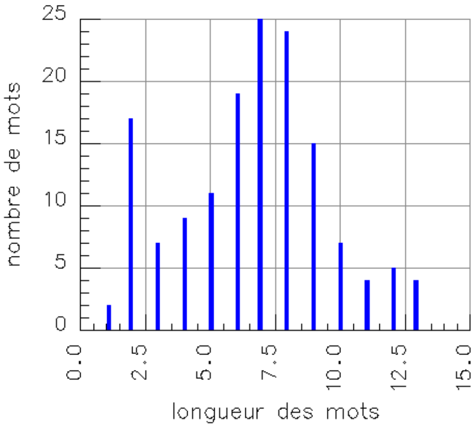
- numéro identifiant le résultat (*id_ reponse*),
- nom du serveur qui fournit cette réponse (*server*),
- question (requête) envoyée (*question*),
- temps nécessaire pour fournir la première page de réponses (*time_ reponse*),
- nombre de réponses fournies sur la *question* (*nombre_ reponses*),
- code de retour : code HTTP de page de réponse (*code_ http*),
- nombre de pages de réponses (*nombre_pages*),
- temps de réponse global : pour toutes les pages de réponses à la question *(time_ reponse_ complete)*,



Adminet



TEST1



TEST2

FIG. 7.2 – Statistiques sur la longueur des mots dans les questions pour les fichiers-tests (TEST1 et TEST2) et les questions d'Adminet.

- date d'interrogation (*date*),
- l'heure d'interrogation (*heure*).

La structure de cette table de la base de données est illustrée dans le tableau suivant [Tableau 7.10].

STAT_REPONSE		
id_reponse	INT	Auto NotNull
server	VARCHAR	NotNull
question	VARCHAR	NotNull
time_reponse	TINYINT	
nombre_reponses	SMALLINT	
code_http	SMALLINT	NotNull
nombre_pages	TINYINT	
time_reponse_complete	INT	
date	DATE	
heure	ITIME	

TAB. 7.10 – La structure de base de données. Table STAT_REPONSE.

7.1.3.2 Les statistiques concernant les documents-réponses (vérification de réponses)

Les tables suivantes contiennent les données pour les statistiques de pertinence des réponses. Elles ont été créées à partir des réponses de chaque moteur interrogé sur les questions du fichier-test TEST2 (50 questions). Pour chaque résultat de recherche (réponse d'un moteur sur une seule question), les trois tables suivantes ont été remplies dans la base : tables REPONSE, DOCUMENT et la table PAR_MOT.

La première table (REPONSE) contient les informations concernant la page de réponses. Sa structure est donc similaire à celle de la table STAT_REPONSE décrite dans la section précédente. La table REPONSE contient les informations suivantes :

- numéro identifiant le résultat (*id_reponse*),
- nom du serveur qui fournit cette réponse (*server*),
- question (requête) envoyée (*question*),
- temps pour fournir la première page de réponses (*time_reponse*),
- nombre de réponses fournies à la question (*nombre_reponses*),

La structure de cette table de la base de données est illustrée dans le tableau 7.11.

La deuxième table (DOCUMENT) contient les informations :

REPONSE		
id_ reponse	INT	Auto NotNull
server	VARCHAR	NotNull
question	VARCHAR	NotNull
time_ reponse	TINYINT	
nombre_ reponses	SMALLINT	
code_ http	SMALLINT	NotNull

TAB. 7.11 – La structure de base de données. Table REPONSE.

- numéro identifiant le document (*id_ doc*),
- l'adresse du document (*url_ doc*),
- numéro identifiant le résultat indiquant ce document (*id_ reponse*),
- taille du document : partie *body* de document HTML (*taille_ body*),
- nombre d'occurrences de *question exacte*[2] dans le corps *(body)* de document (*question_ in_ body*),
- nombre d'occurrences de *question exacte* dans le *titre* de document (*question_ in_ titre*),
- nombre d'occurrences de *question exacte* dans le *meta* de document (*question_ in_ meta*),
- code de retour : code HTTP de page de réponse (*code_ http*).

La table DOCUMENT est liée avec la table REPONSE par le champ *id_ reponse.* La structure de cette table est illustrée dans le tableau 7.12.

DOCUMENT		
id_ doc	INT	Auto NotNull
url_ doc	VARCHAR	NotNull
id_ reponse	MEDIUMINT	NotNull
taille_ body	MEDIUMINT	
question_ in_ meta	MEDIUMINT	
question_ in_ titre	MEDIUMINT	
question_ in_ body	MEDIUMINT	
code_ http	SMALLINT	NotNull

TAB. 7.12 – La structure de base de données. Table DOCUMENT.

La dernière table (PAR_ MOT) contient les informations concernant les occurrences des mots de la requête dans le document-réponse :

- numéro identifiant le mot de la requête (*id_ mot*),
- numéro identifiant le document (*id_ doc*),

---

[2] "*question exacte*" defini une chaine de caractères exacte dans le corps de cette partie du document, exemple recherche de phrase exprimée dans les guillemets

- mots de la requête : les mots vides sons éliminés (*mot*),
- nombre d'occurrences du mot dans le *titre* de document (*in_ titre*),
- nombre d'occurrences du mot dans le *meta* de document (*in_ meta*),
- nombre d'occurrences du mot dans le *(body)* de document (*in_ body*),
- toutes les positions du mot dans le *titre* de documents (*position_ titre*),
- toutes les positions du mot dans le *meta* de documents (*position_ meta*),
- toutes les positions du mot dans le corps *(body)* de documents (*position_ body*).

La table PAR_MOT est liée avec la table DOCUMENT par le champ *id_ doc.* La structure de cette table est illustrée dans le tableau 7.13.

PAR_MOT		
id_ mot	INT	Auto NotNull
id_ doc	INT	NotNull
mot	VARCHAR	NotNull
in_ titre	INT	
in_ meta	INT	
in_ body	INT	
position_ titre	VARCHAR	
position_ meta	VARCHAR	
position_ body	TEXT	

TAB. 7.13 – La structure de base de données. Table PAR_MOT.

7.2 Statistiques effectuées sur des questions-test (TEST1)

7.2.1 Nombre des réponses récupérées

Les deux graphes suivants [Figure 7.3] montrent les statistiques concernant le nombre de réponses obtenues sur les 550 requêtes du fichier-test (TEST1) depuis chacun des moteurs interrogés. Les serveurs sont identifiés sur le graphe par le numéro *id_ serveur* correspondant de la liste de serveurs interrogés [Tableau 7.1]. La couleur des données passe de la couleur rouge à la couleur bleue avec la diminution du nombre de réponses :

- la couleur rouge signifie le nombre maximal,
- la couleur bleu signifie le nombre minimal.

On peut observer :

- les moteurs qui donnent un nombre de réponses très varié selon la requête envoyée : des serveurs assez thématiques,
- les moteurs qui donnent toujours un nombre de réponses similaire, les serveurs plus universels,

Certains moteurs, classés plus tôt comme universels, donnent toujours le nombre maximal de réponses (plafonné à 1000 dans notre implémentation pour des raison d'ordre pratique).

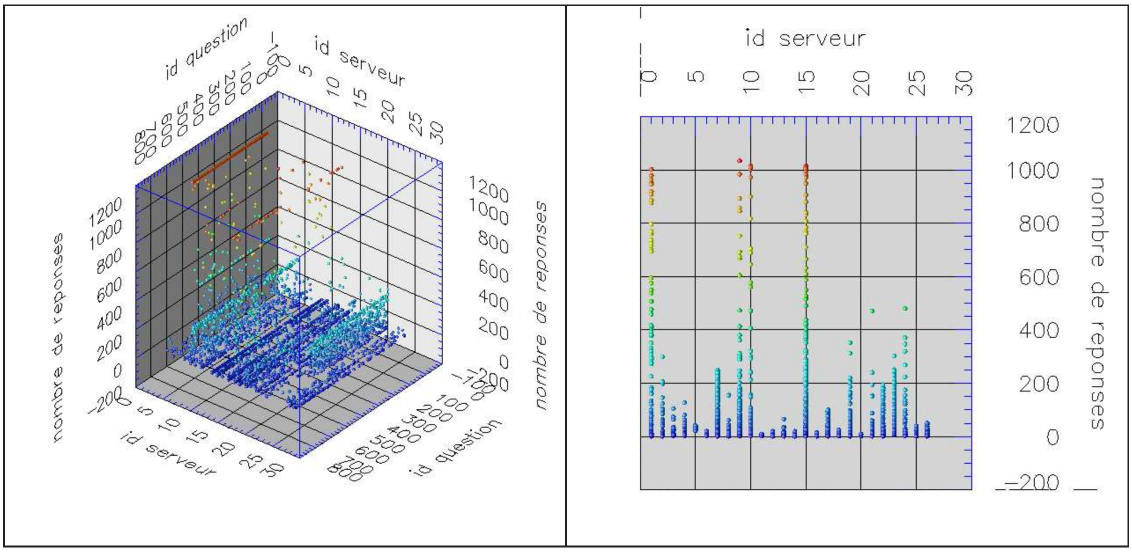


FIG. 7.3 – Nombre de réponses selon la requête, pour chacun des moteurs.

7.2.2 Temps de réponse

Pour toutes les interrogations effectuées nous avons mesuré le temps de fourniture de la réponse : temps de réponse. Pour chaque moteur, les temps de réponses minimal, maximal et moyen ont été calculés. Les interrogations (connections) avec les moteurs n'ont été pas limitées au niveau du temps de réponse. Contrairement aux statistiques pour le temps moyen effectué sur le FIM-Métaindexeur [Page 152, Section 6.3.3.2 et la page 142 du paragraphe 6.2.3.5], dans ce test, le temps moyen de communication (temps de réponse moyen) est exact. Les figures suivantes montrent les valeurs statistiques de tous les temps de réponses calculés pour chacun des moteurs :

- temps de réponse minimal (Figure 7.4),
- temps de réponse maximal (Figure 7.5),

– temps de réponse moyen (Figure 7.6).

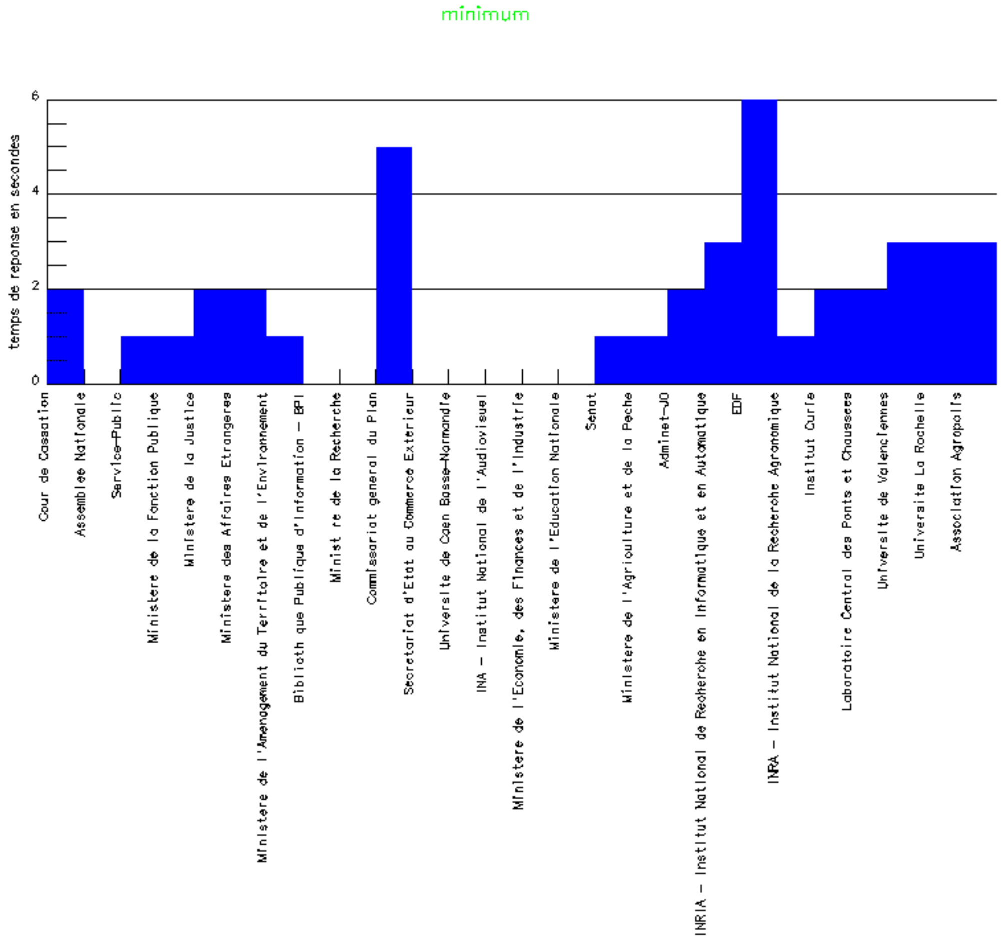


FIG. 7.4 – Le temps de réponse minimal.

Toutes les valeurs de temps de réponses (minimum, maximum et moyen) pour chacun des moteurs regroupées ensemble, sont présentées dans la figure 7.7.

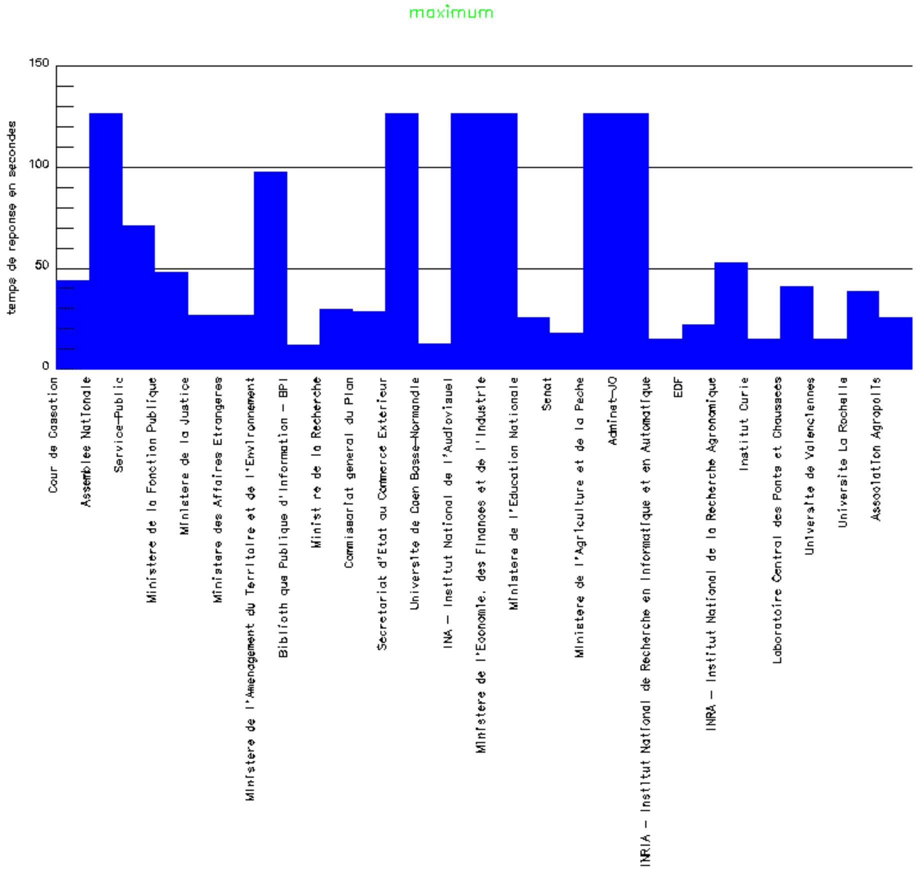


FIG. 7.5 – Le temps de réponse maximal.

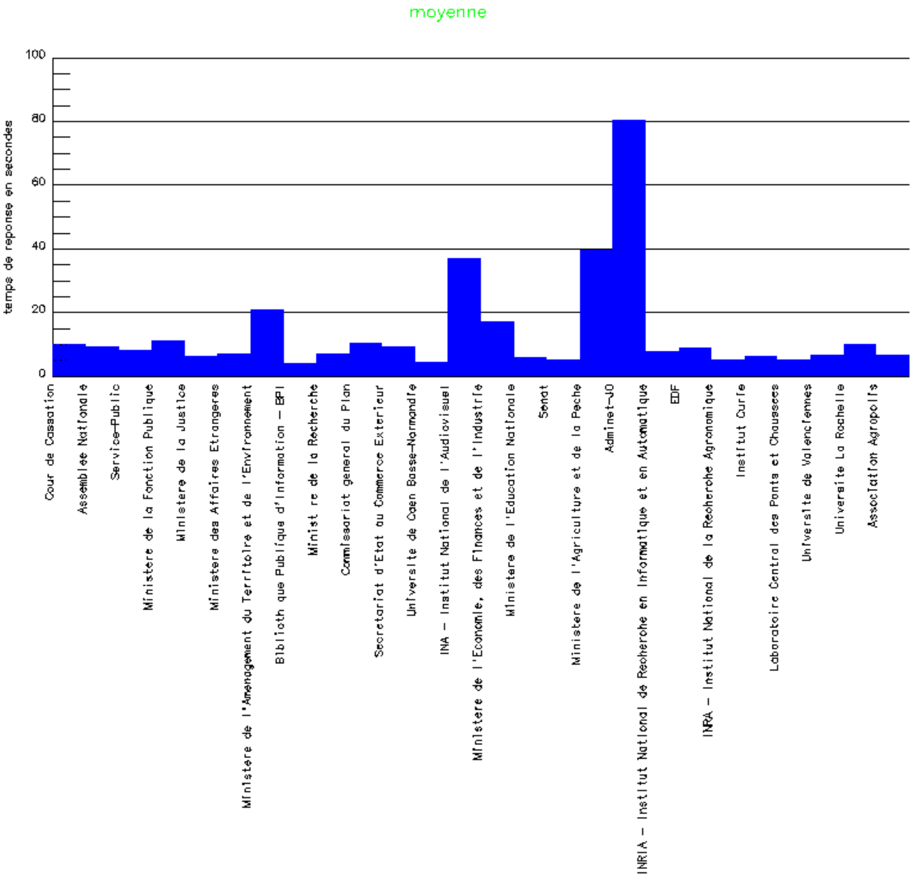


FIG. 7.6 – Le temps de réponse moyen.

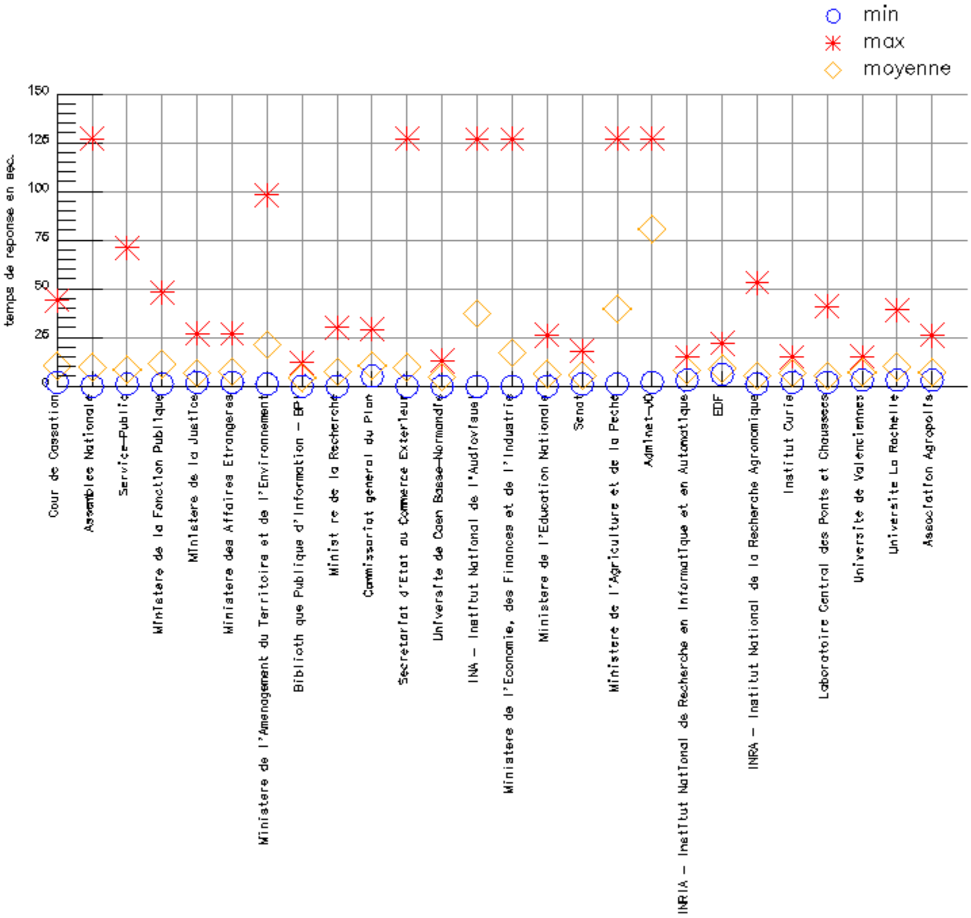


FIG. 7.7 – Les temps de réponses : minimal, maximal, moyen.

7.2.3 Temps de réponse en fonction du nombre des réponses récupérées

Les réponses sont fournies par les moteurs avec des délais de réponse très variés. Les différences de ces temps sont dûes à différents raisons :

- fonctionnement interne du moteur,
- charge du moteur,
- charge du réseau, etc.

La dépendance du temps de réponse vis à vis de la charge de moteur et de la charge de réseau est aléatoire et ne permet pas de caractériser un moteur. Par contre, la dépendance du temps de réponse vis à vis du fonctionnement interne du moteur sera un bon paramètre caractérisant le moteur.

Les statistiques effectuées concernent la dépendance entre le temps de réponse de chaque moteur et le nombre de réponses fournies par ce moteur.

Pour trouver cette dépendance, pour chacun des moteurs nous avons analysé la corrélation entre deux valeurs : le temps et le nombre de réponses.

Les résultats sont présentés dans la figure 7.8.

On observe que plusieurs moteurs possèdent un coefficient de corrélation assez important, car pour ces moteurs le temps de réponse dépend bien du nombre de questions fournies. On peut supposer que le temps pour trouver les réponses dans leurs bases d'index dépend aussi de la question posée.

On observe également des coefficients de corrélation négatifs pour les deux moteurs suivants :

- Ministère de l'Economie, des Finances et de l'Industrie (-0.63)
- Ministère de l'Equipement, des Transports et du Logement (-0.38).

Cette situation paradoxale, peut être expliquée différemment :

- Limitation du nombre de réponses (cas de Ministère de l'Economie, des Finances et de l'Industrie où le nombre de réponses est toujours égal à 20). Pour les questions ayant beaucoup de réponses, le moteur arrête la recherche lorsqu'il trouve le nombre maximal de réponses et les envoie rapidement : le temps de réponse est minimal. Dans le cas où le moteur ne trouve pas le nombre maximal de réponses immédiat, il est obligé de continuer la recherche jusqu'à bout : d'où un temps de réponse plus grand.

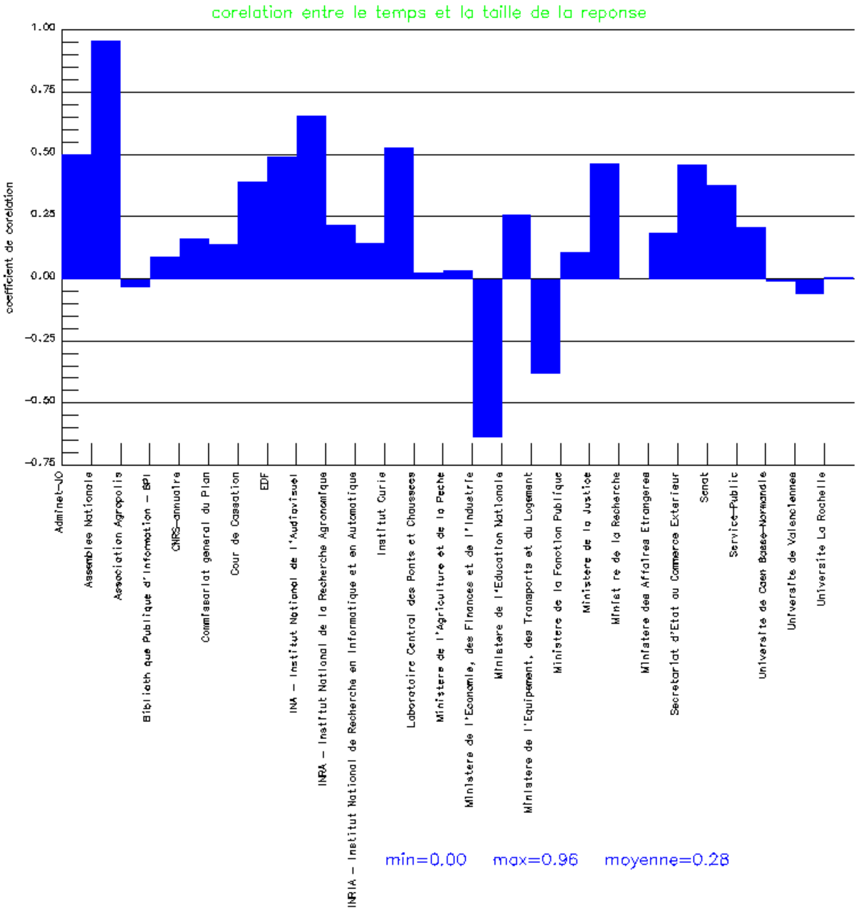


FIG. 7.8 – Corrélation entre le temps de réponse et nombre de réponses calculée pour chacun des moteurs.

- Utilisation du *cache* (base des questions déjà posées). Si le moteur connaît un nombre suffisant de réponses à la questions posée (question fréquente et typique, donc beaucoup de réponses), il peut les fournir immédiatement sans effectuer de recherche : le temps de réponse est minimal. Au cas où le moteur ne possède pas les réponses dans son *cache (*question posée rarement, moins de réponse*s),* il est obligé d'effectuer le recherche : d'où un temps de réponse plus grand.

7.3 Profil thématique des serveurs

7.3.1 Méthode

Nous avons essayé d'analyser les différents profils thématiques des serveurs d'administration interrogés. Pour cela nous avons cherché une dépendance entre la quantité de réponses et le thème de la question posée. Nous souhaitions en effet tester l'hypothèse suivante :

Hypothèse : Deux serveurs ont un profil thématique proche s'il existe une corrélation significative entre le nombre de réponses données par les moteurs de recherche de ces serveurs sur un grand ensemble de questions différentes.

Pour vérifier cette l'hypothèse, et trouver les serveurs proches thématiquement, nous avons effectué une analyse des réponses aux questions du fichier-test TEST1 (environ 1000 questions : [Section 7.1.1.1]). Les paragraphes suivants présentent les résultats de cette analyse.

7.3.2 Corrélation globale

Nous estimons le degré de la dépendance thématique de serveurs selon la corrélation du nombre de réponses. Le tableau 7.14 présente l'importance de la dépendance de deux populations selon différentes valeurs du coefficient de corrélation.

Pour retrouver les dépendances entre les serveurs interrogés, nous avons calculé la corrélation entre les 26 serveurs du tableau 7.15.

Les résultats globaux de ces calculs pour tous les serveurs ont montré qu'il existe un certain nombre de serveurs ayant une dépendance assez importante. Les premiers résultats sont présentés graphiquement dans la figure 7.9. La couleur des données passe du rouge du bleu avec la diminution du coefficient de corrélation :

Coeff. de corrélation	Corrélation	Dependance
moins de 0.20	très petite	pas de dépendance significative
[0.20 0.40]	petite	dépendance existante
[0.40 0.70]	normale	dépendance assez importante
[0.70 0.90]	haute	dépendance importante
[0.90 1.00]	très haute	dépendance très forte

Tab. 7.14 – Le degré de la dépendance selon le coefficient de corrélation. [CKCP81]

id_ serveur	serveur
1	Adminet-JO
2	Assemblée Nationale
3	Association Agropolis
4	Bibliothèque Publique d'Information - BPI
5	CNRS-annuaire
6	Commissariat général du Plan
7	Cour de Cassation
8	EDF
9	INA - Institut National de l'Audiovisuel
10	INRA - Institut National de la Recherche Agronomique
11	INRIA - Institut National de Recherche en Informatique et en Automatique
12	Institut Curie
13	Laboratoire Central des Ponts et Chaussées
14	Ministère de l'Economie, des Finances et de l'Industrie
15	Ministère de l'Education Nationale
16	Ministère de l'Equipement, des Transports et du Logement
17	Ministère de la Fonction Publique
18	Ministère de la Justice
19	Ministère de la Recherche
20	Ministère des Affaires Étrangères
21	Secrétariat d'Etat au Commerce Extérieur
22	Sénat
23	Service-Public
24	Université de Caen Basse-Normandie
25	Université de Valenciennes
26	Université La Rochelle

Tab. 7.15 – La liste de serveurs analysés.

- la couleur rouge signifie la corrélation maximale,
- la couleur bleue signifie la corrélation minimale.

Les serveurs sont identifiés sur le graphe par le numéro correspondant de la liste précédente.

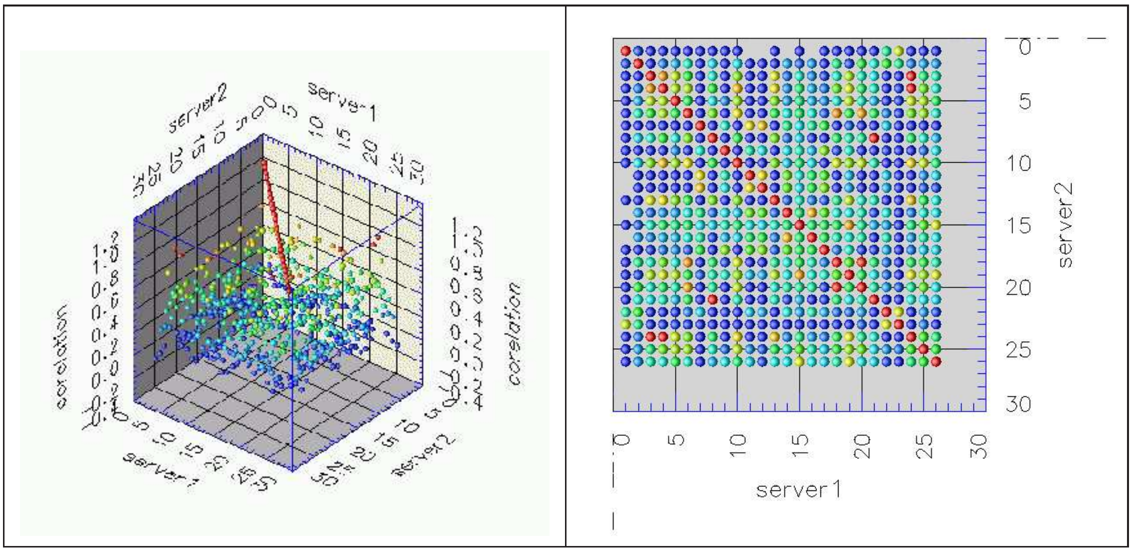


FIG. 7.9 – La corrélation thématique de serveurs d'administration.

Cette présentation des corrélations montre l'existence de serveurs dépendants, par contre il est difficile de tirer des conclusions. Les serveurs sont placés sur les axes de graphes de façon aléatoire, ce qui ne permet pas facilement de retrouver les serveurs ayant des comportements analogues (proches thématiquement).

Nous avons essayé donc de créer une présentation graphique plus significative.

Suivant les règles de dépendance des deux populations présentées dans le tableau 7.14, on présume que les paires de serveurs pour lesquelles le coefficient de corrélation est supérieur à 0.4 sont proches thématiquement (dépendance importante). Nous avons restreint l'analyse aux paires de serveurs avec corrélation supérieure à 0.5. Grâce à ces paires de serveurs, nous avons construit un graphe de dépendance. Les noeuds présentent les serveurs, liés entre eux selon leurs coefficient de corrélation. La couleur et la longueur des liens représentent le coefficient de corrélation. Ce graphe est présenté à la figure 7.10.

On peut observer qu'un grand nombre de serveurs sont liés assez fortement entre

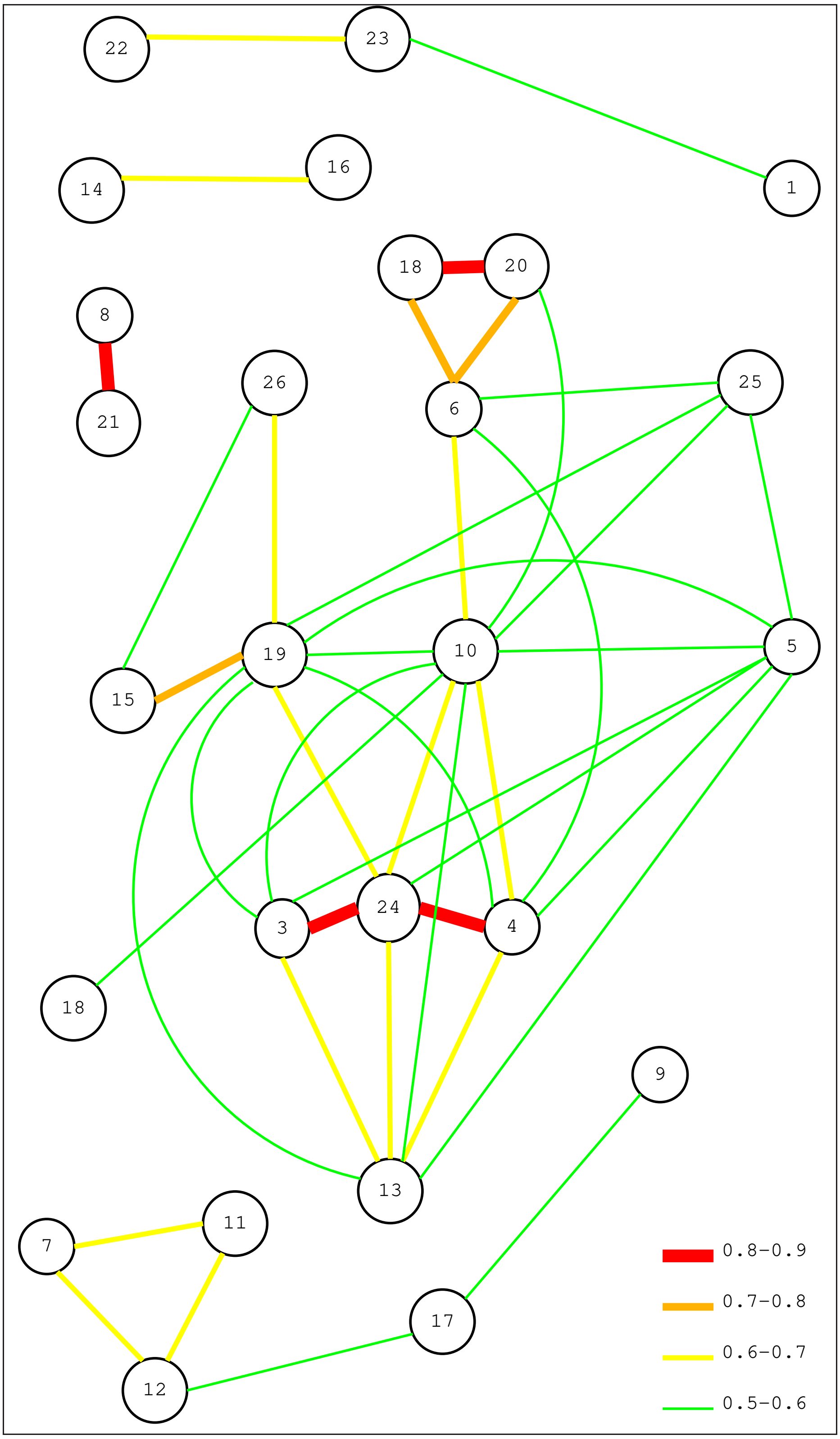


FIG. 7.10 – Le graphe de dépendance pour les paires de serveurs avec la corrélation ≥ 0.5.

eux. En plus, on observe qu'il existe également des serveurs liés, sans connexions avec les autres, et qui ont pourtant des comportements très voisins.

L'analyse de ce graphe a facilité le regroupement (tri) des serveurs proches en essayant d'améliorer la représentation graphique en donnant des numéros voisins à des serveurs fortement corrélés. Cette renumérotation a été faite manuellement. Nous avons ainsi obtenu une nouvelle liste de serveurs triés selon leurs corrélations : les serveurs sont présentés dans l'ordre du tableau 7.16.

nouveau numéro	serveur	ancien numéro
1	Adminet-JO	1
2	Service-Public	23
3	Senat	22
4	Assemblee Nationale	2
5	Universite de Valenciennes	25
6	CNRS-annuaire	5
7	Ministere de la Justice	18
8	Ministere des Affaires Etrangeres	20
9	Commissariat general du Plan	6
10	INRA - Institut National de la Recherche Agronomique	10
11	Bibliothèque Publique d'Information - BPI	4
12	Universite de Caen Basse-Normandie	24
13	Association Agropolis	3
14	Laboratoire Central des Ponts et Chaussees	13
15	Ministère de la Recherche	19
16	Ministere de l'Education Nationale	15
17	Universite La Rochelle	26
18	EDF	8
19	Secretariat d'Etat au Commerce Exterieur	21
20	Ministere de l'Economie, des Finances et de l'Industrie	14
21	Ministere de l'Equipement, des Transports et du Logement	16
22	INA - Institut National de l'Audiovisuel	9
23	Ministere de la Fonction Publique	17
24	Institut Curie	12
25	INRIA - Institut National de Recherche en Informatique et en Automatique	11
26	Cour de Cassation	7

TAB. 7.16 – Liste de serveurs regroupés.

Les figures 7.11 jusqu'à 7.17 sont ainsi dérivées de la figure 7.9 en utilisant le nouvel ordre de classement des serveurs.

Les serveurs sont identifiés sur le graphe par leur nouveau numéro présenté dans le tableau 7.16. En couleur rouge sont présentées les données dont la coefficient de corrélation satisfait la marge choisie (corrélation importante). La couleur de données passe au bleu avec la diminution de valeur du coefficient de corrélation.

Corrélation > 0.9 La corrélation avec un coefficient > 0.9, est présentée dans la figure 7.11.

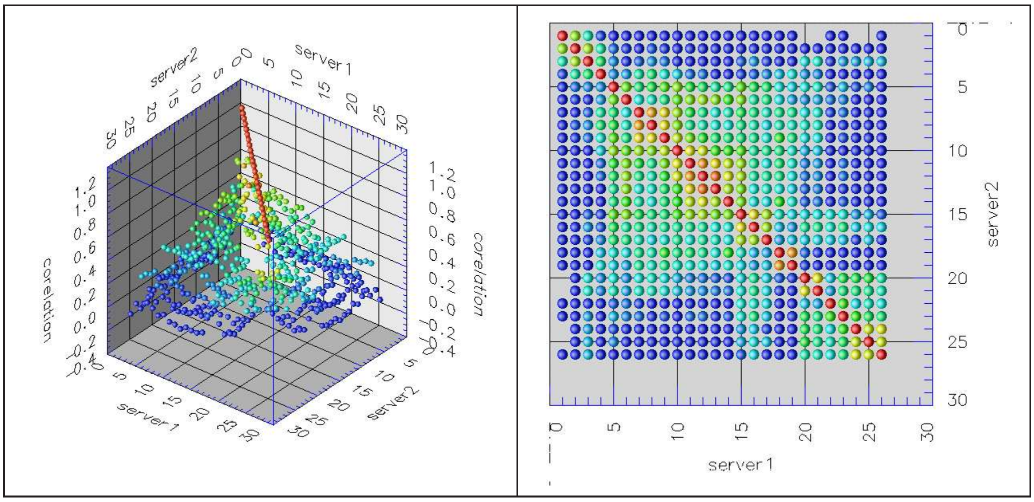


FIG. 7.11 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.9

Sur le graphe on observe qu'il n'existe pas de paire de serveurs dont le coefficient de corrélation est supérieur à 0.9. Les points de couleur rouge forment une diagonale, puisqu'il présentent les coefficients de corrélation des serveurs avec eux-mêmes, soit 1.

Corrélation > 0.8 Le tableau 7.17 montre que quatre paires de serveurs possèdent le coefficient de corrélation supérieur à 0.8. Leurs corrélation est présentée dans la figure 7.12.

Corrélation > 0.7 Le tableau 7.18 regroupe les paires de serveurs avec un coefficient de corrélation supérieur à 0.7. La figure 7.13 présente leurs corrélations.

serveur 1	serveur 2	corrélation
Ministere de la Justice (7)	Ministere des Affaires Etrangeres (8)	0.801
Bibliothèque Publique d'Information - BPI(11)	Universite de Caen Basse-Normandie(12)	0.815
Universite de Caen Basse-Normandie(12)	Association Agropolis (13)	0.856
EDF (18)	Secretariat d'État au Commerce Exterieur (19)	0.813

Tab. 7.17 – Tableau des paires de serveurs avec une corrélation > 0.8

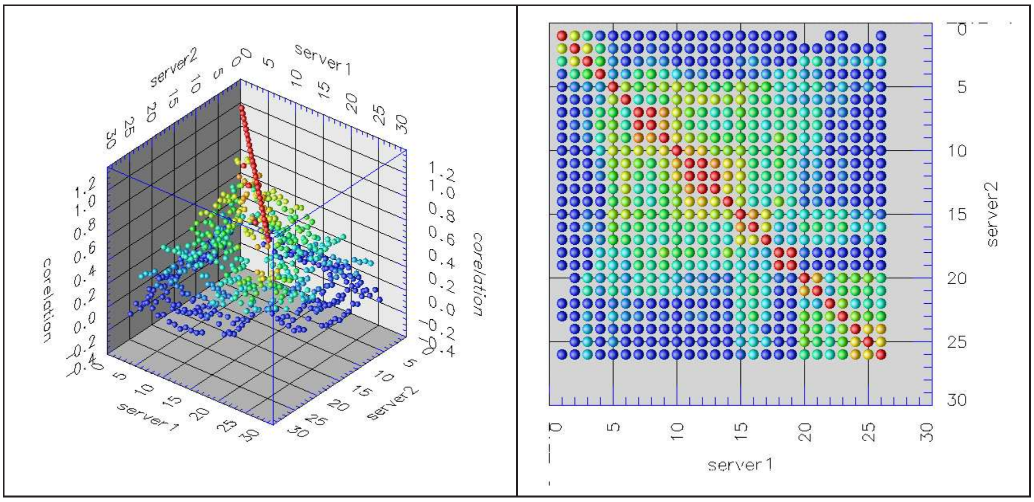


Fig. 7.12 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.8

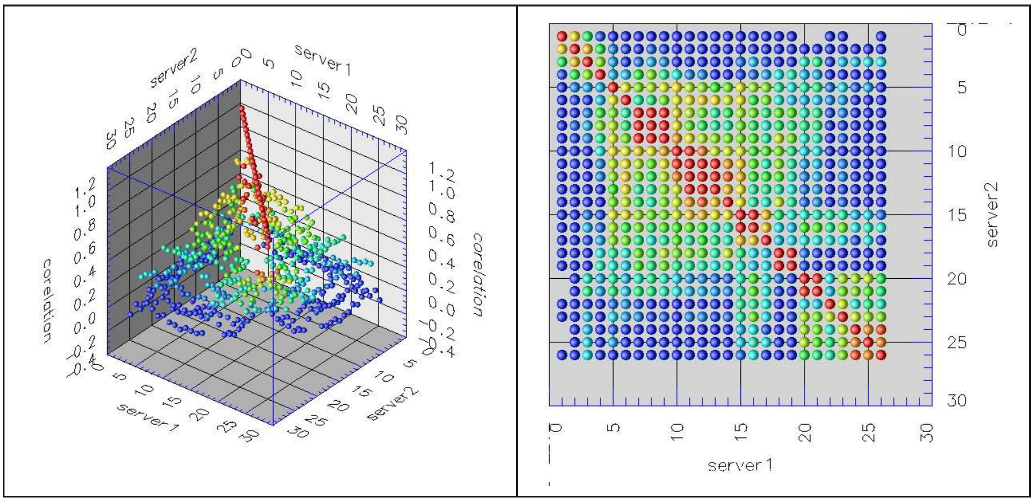


Fig. 7.13 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.7

serveur 1	serveur 2	corrélation
Ministere de la Justice (7)	Commissariat general du Plan (9)	0.732
Ministere des Affaires Etrangeres (8)	Commissariat general du Plan (9)	0.714
Bibliothèque Publique d'Information - BPI(11)	Association Agropolis (13)	0.729
Ministère de la Recherche (15)	Ministere de l'Éducation Nationale (16)	0.716

Tab. 7.18 – Tableau des paires de serveurs avec une corrélation > 0.7

Corrélation > 0.6 Le tableau 7.19 regroupe les paires de serveurs avec un coefficient de corrélation supérieur à 0.6. Leurs corrélation est présentée dans la figure 7.14.

serveur 1	serveur 2	corrélat.
Service-Public (2)	Senat (3)	0.603
Commissariat general du Plan (9)	Inst. National de la Recherche Agronomique (10)	0.632
Inst. National de la Recherche Agronomique (10)	Universite de Caen Basse-Normandie(12)	0.660
Inst. National de la Recherche Agronomique (10)	Bibliothèque Publique d'Information - BPI(11)	0.695
Bibliothèque Publique d'Information - BPI(11)	Laboratoire Central des Ponts et Chaussees (14)	0.639
Universite de Caen Basse-Normandie(12)	Laboratoire Central des Ponts et Chaussees (14)	0.680
Universite de Caen Basse-Normandie(12)	Ministère de la Recherche (15)	0.609
Association Agropolis (13)	Laboratoire Central des Ponts et Chaussees (14)	0.606
Ministère de la Recherche (15)	Universite La Rochelle (17)	0.602
Min. de l'Economie, des Finances et de l'Ind. (20)	Min. de l'Equipement, Transports et Logement(21)	0.695
Institut Curie(24)	Cour de Cassation(26)	0.686
Institut Curie(24)	INRIA (25)	0.647
INRIA (25)	Cour de Cassation(26)	0.650

TAB. 7.19 – Tableau des paires serveurs avec une corrélation > 0.6

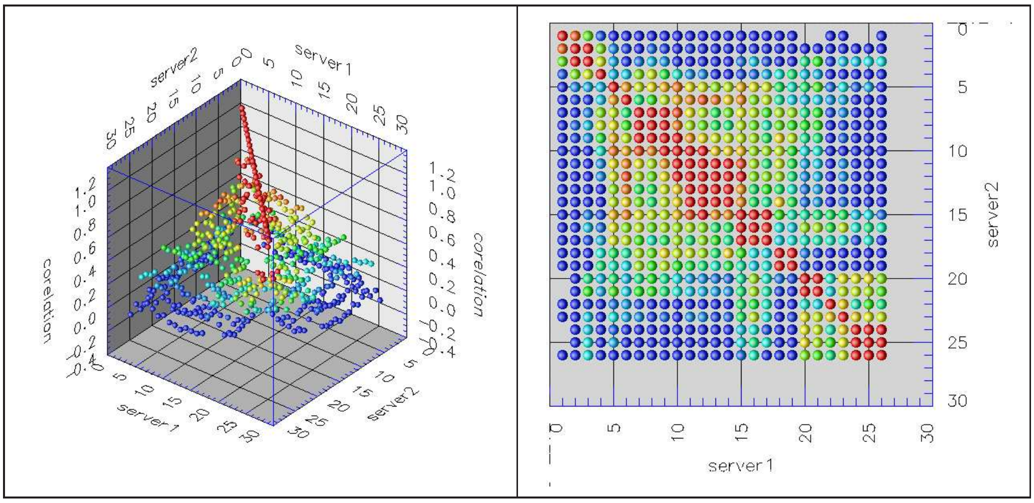


FIG. 7.14 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.6

Corrélation > 0.5 Le tableau 7.20 regroupe les serveurs avec le coefficient de corrélation supérieur à 0.5. La corrélation avec un coefficient supérieur à 0.5 est présentée dans la figure 7.15.

Corrélation > 0.4 La figure 7.16 présente la corrélation supérieure à 0.4.

Corrélation > 0.3 La corrélation supérieure à 0.3 est présentée dans la figure 7.17.

serveur 1	serveur 2	corrélation
Adminet-JO (1)	Service-Public (2)	0.577
Universite de Valenciennes (5)	CNRS-annuaire (6)	0.543
Universite de Valenciennes (5)	Commissariat general du Plan (9)	0.525
Universite de Valenciennes (5)	Inst. National de la Recherche Agronomique (10)	0.568
Universite de Valenciennes (5)	Ministère de la Recherche (15)	0.527
CNRS-annuaire (6)	Inst. National de la Recherche Agronomique (10)	0.531
CNRS-annuaire (6)	Bibliothèque Publique d'Information - BPI(11)	0.557
CNRS-annuaire (6)	Universite de Caen Basse-Normandie(12)	0.586
CNRS-annuaire (6)	Association Agropolis (13)	0.566
CNRS-annuaire (6)	Laboratoire Central des Ponts et Chaussees (14)	0.504
CNRS-annuaire (6)	Ministère de la Recherche (15)	0.557
Ministere de la Justice (7)	Inst. National de la Recherche Agronomique (10)	0.589
Ministere des Affaires Etrangeres (8)	Inst. National de la Recherche Agronomique (10)	0.561
Commissariat general du Plan (9)	Bibliothèque Publique d'Information - BPI(11)	0.546
Inst. National de la Recherche Agronomique (10)	Association Agropolis (13)	0.506
Inst. National de la Recherche Agronomique (10)	Laboratoire Central des Ponts et Chaussees (14)	0.500
Inst. National de la Recherche Agronomique (10)	Ministère de la Recherche (15)	0.518
Bibliothèque Publique d'Information - BPI(11)	Ministère de la Recherche (15)	0.587
Association Agropolis (13)	Ministère de la Recherche (15)	0.549
Laboratoire Central des Ponts et Chaussees (14)	Ministère de la Recherche (15)	0.555
Ministere de l'Education Nationale (16)	Universite La Rochelle (17)	0.597
INA - Institut National de l'Audiovisuel(22)	Ministere de la Fonction Publique(23)	0.508
Ministere de la Fonction Publique(23)	Institut Curie(24)	0.503

TAB. 7.20 – Tableau des paires serveurs avec une corrélation > 0.5

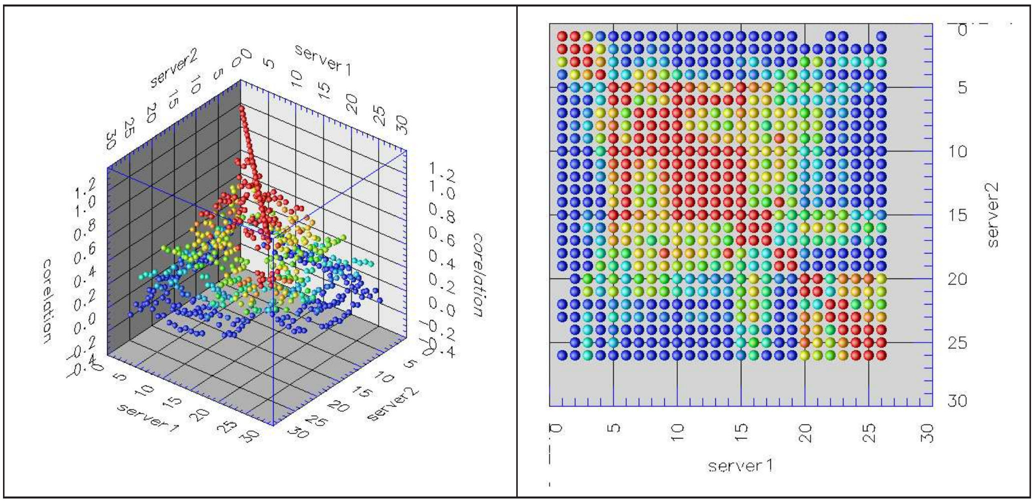


FIG. 7.15 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.5

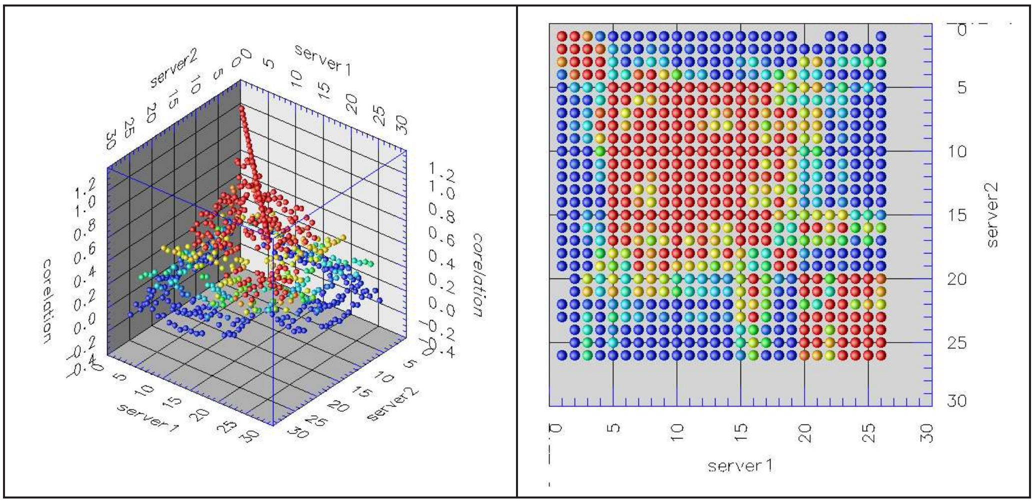


Fig. 7.16 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.4

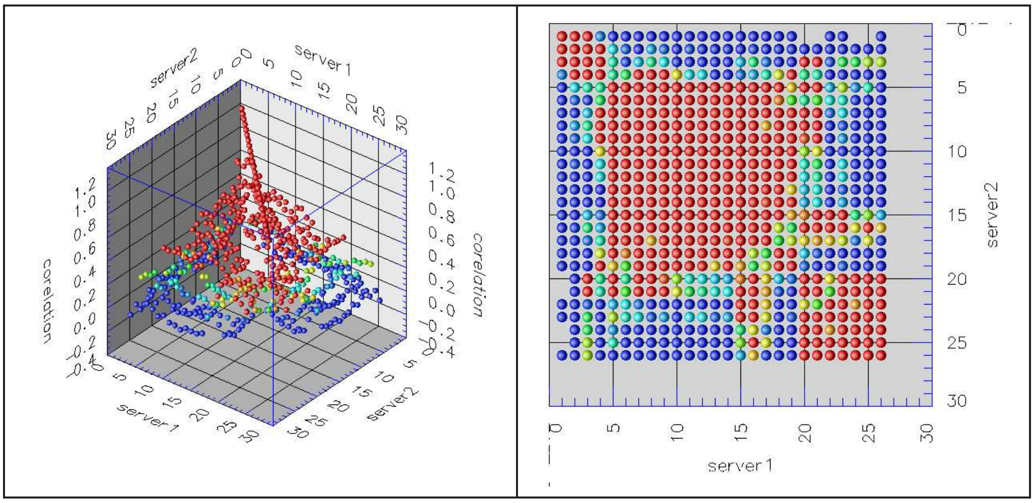


Fig. 7.17 – Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.3

La figure 7.18 présente un histogramme des corrélations de paires de serveurs.

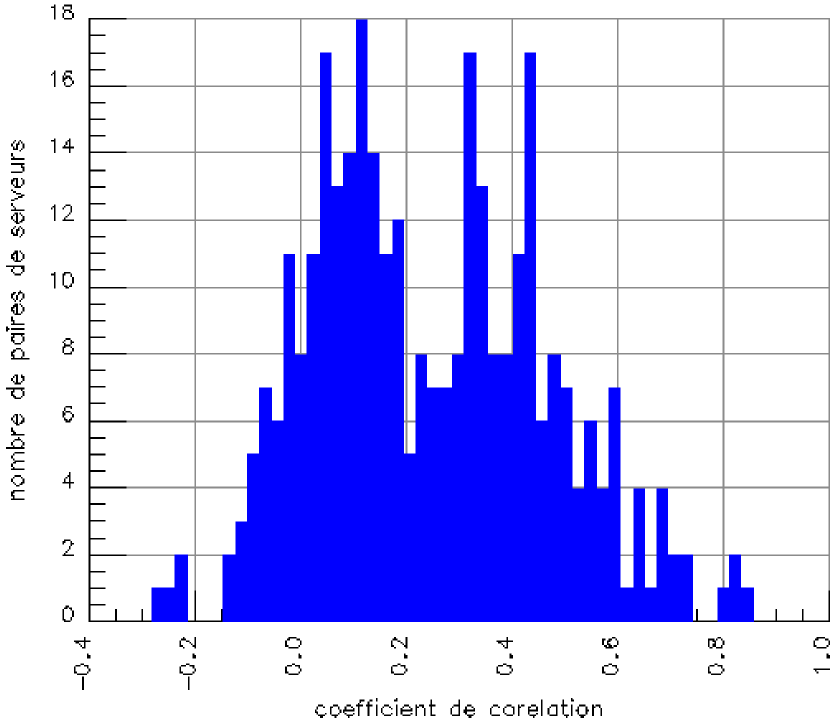


FIG. 7.18 – Histogramme des corrélations de paires de serveurs.

7.3.3 Corrélations particulières

Nous avons inspecté visuellement les corrélations d'un certain nombre de serveurs avec tous les autres. Le cas des serveurs EDF, Ministère de la Justice et Université de Caen Basse-Normandie nous ont paru intéressant à montrer graphiquement en raison des proximités qu'ils illustrent. Les figures suivantes [Figure 7.19, Figure 7.20 et Figure 7.21] présentent les corrélations de ces serveurs avec tous les autres.

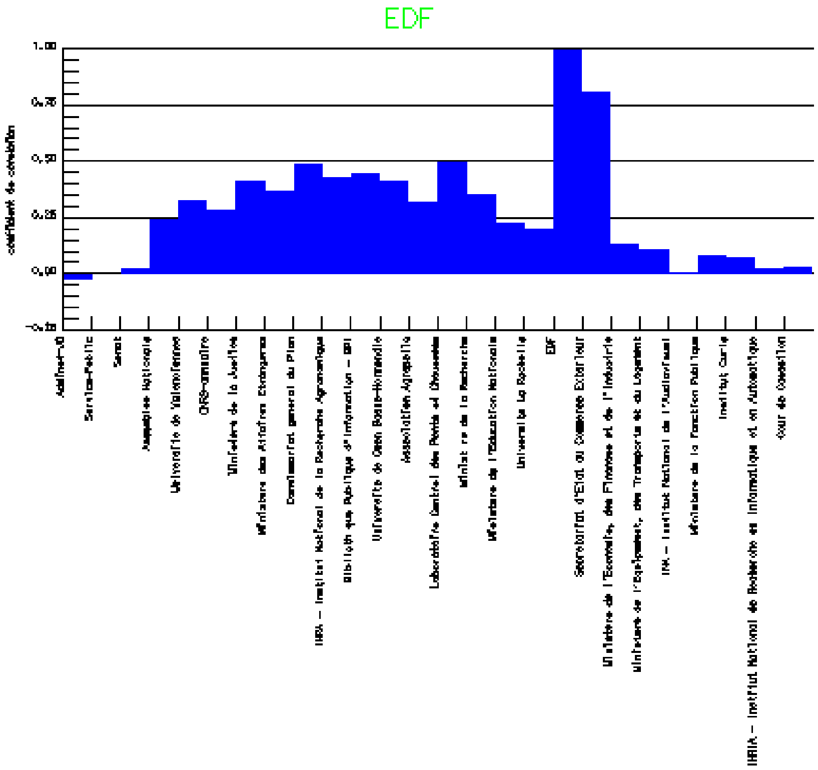


FIG. 7.19 – Corrélation pour le serveur d'EDF.

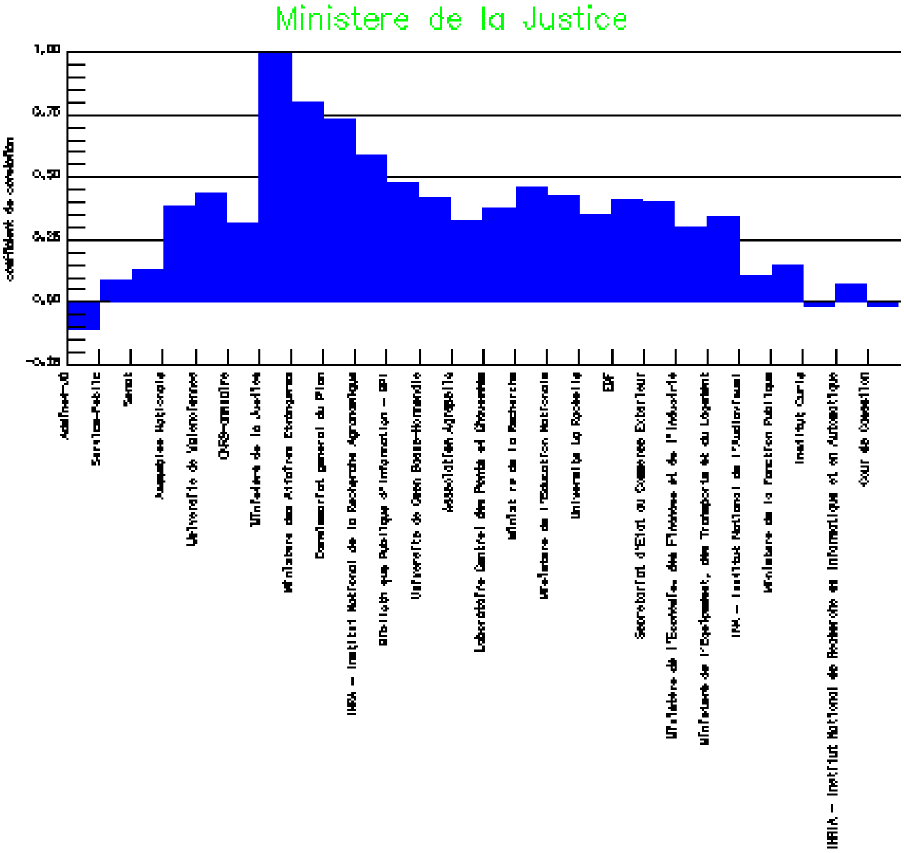


Fig. 7.20 – Corrélation pour le serveur de Ministère du la Justice.

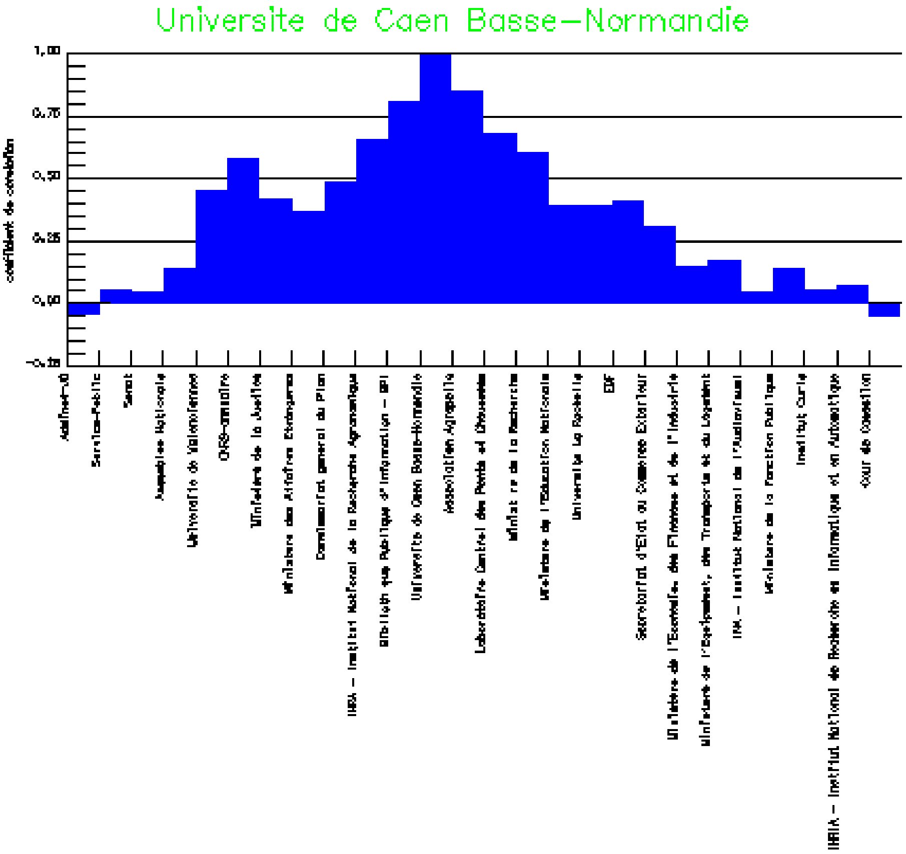
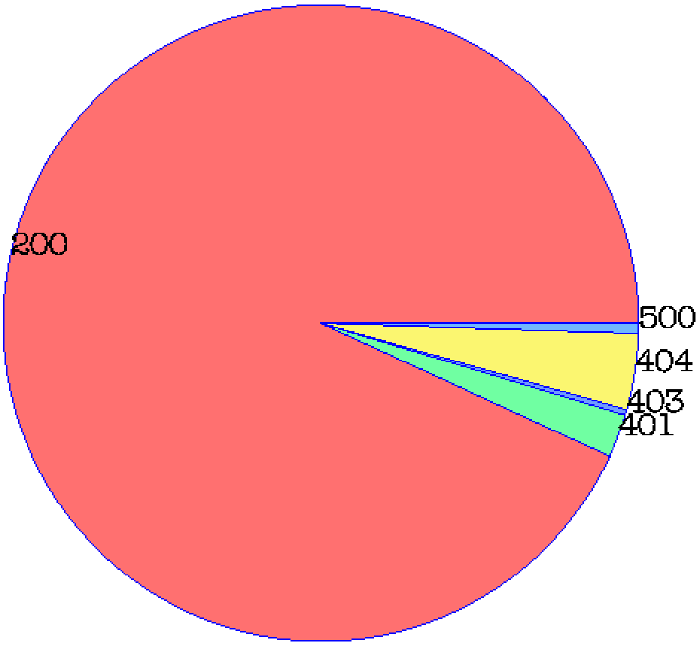


FIG. 7.21 – Corrélation pour le serveur de l'Université de Caen.

7.4 Statistiques effectuées (TEST2)

7.4.1 Accessibilité des documents-réponses (Codes HTTP)

Cette statistique a été effectuée sur la disponibilité effective des 50 premiers documents-réponses sont fournis par les moteurs [Section 7.1.1.2].



code HTTP	nombre de réponses	pourcentage
200	18343	93.131%
401	434	2.203%
403	56	0.284%
404	762	3.869%
500	101	0.513%

FIG. 7.22 – Proportions de différents codes de retour (code HTTP).

L'analyse de cette statistique donne les résultats suivants :

– une majorité (plus de 93%) de documents-réponses récupérés sans aucun problème : code de retour 200,

- environ 3.9 % des documents-réponses ont retourné le code HTTP 404 (le document demandé n'a pas pu être trouvé sur le serveur : lien incorrect),
- parmi tous les moteurs interrogés, il existe des moteurs qui insèrent sur leur page de réponse à coté des URLs de documents accessibles pour tous les utilisateurs, les URLs vers les documents-réponses accessibles seulement pour leurs utilisateurs abonnés possédant le mot de passe approprié (environ 2.5 %). Ce sont les réponses avec le code de retour 401, 403 :
  - les réponses avec code HTTP de retour 401 (requête non autorisée ; l'information est protégée, réessayez avec une authentification correcte) : pour les moteurs de Cour de Cassation, Ministère des Affaires Étrangères, INRA- Institut National de la Recherche Agronomique, Université de Valenciennes
  - les réponses avec code HTTP de retour 403 (accès non autorisé) : pour les moteurs de Cours de Cassation et Ministère des Affaires Étrangères
- le pourcentage de mauvais retours dûs à une erreur interne (code HTTP 500 : erreur qui est causé dans 95% des cas par un script défaillant et dans 5% par un dysfonctionnement ou d'une surcharge du serveur) est très faible (environ 0.5 %).

7.4.2 Taille des documents-réponses

7.4.2.1 Statistique générale sur la taille

Les analyses suivantes été effectuées sur tous les documents-réponses venant de tous les moteurs. La figure 7.23 présente la statistique effectuée sur la taille de documents-réponses.

Pour rendre la fonction de taille de documents-réponses plus détaillée que celle présentée auparavant [Figure 7.23], nous avons divisé l'axe de données (la taille de document) en trois compartiments comme suit :

- de 0 jusqu'à 5 000 mots dans le document [Figure 7.24],
- de 5 000 jusqu'à 50 000 mots dans le document [Figure 7.25],
- de 50000 jusqu'à 350 000 mots dans le document [Figure 7.26],

7.4.2.2 Taille des documents-réponses pour chaque moteur

Dans cette section on présente les analyses de la taille des documents-réponses effectuées pour chacun des moteurs séparément. Ces analyses concernent :

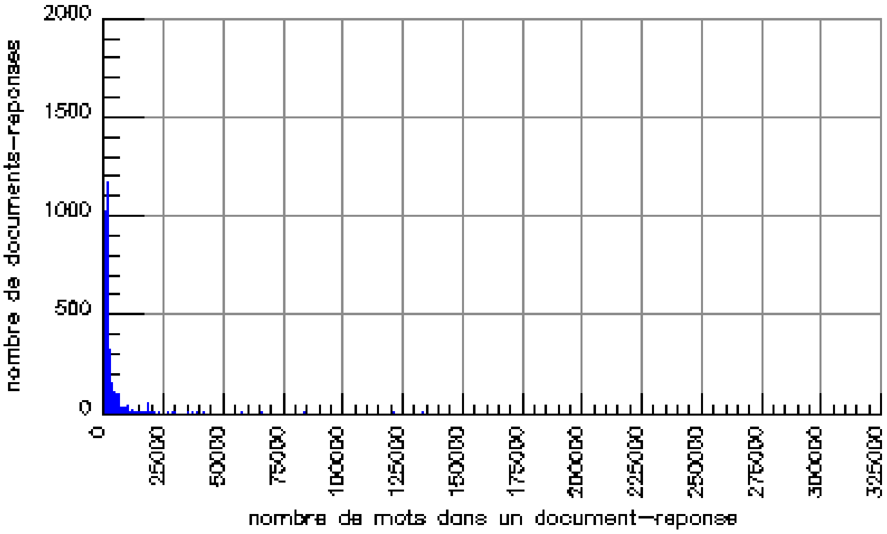


FIG. 7.23 – Taille (en nombre de mots) de documents-réponses pour tous les documents.

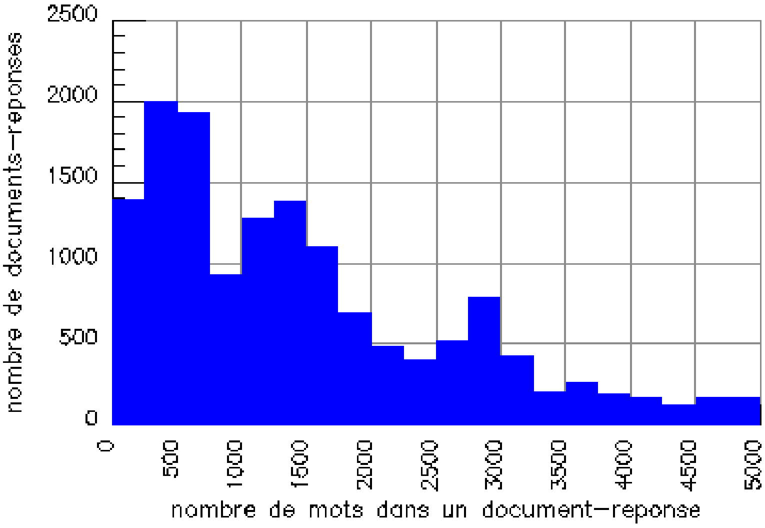


FIG. 7.24 – Taille des documents-réponses. Compartiment [0 - 5000] mots.

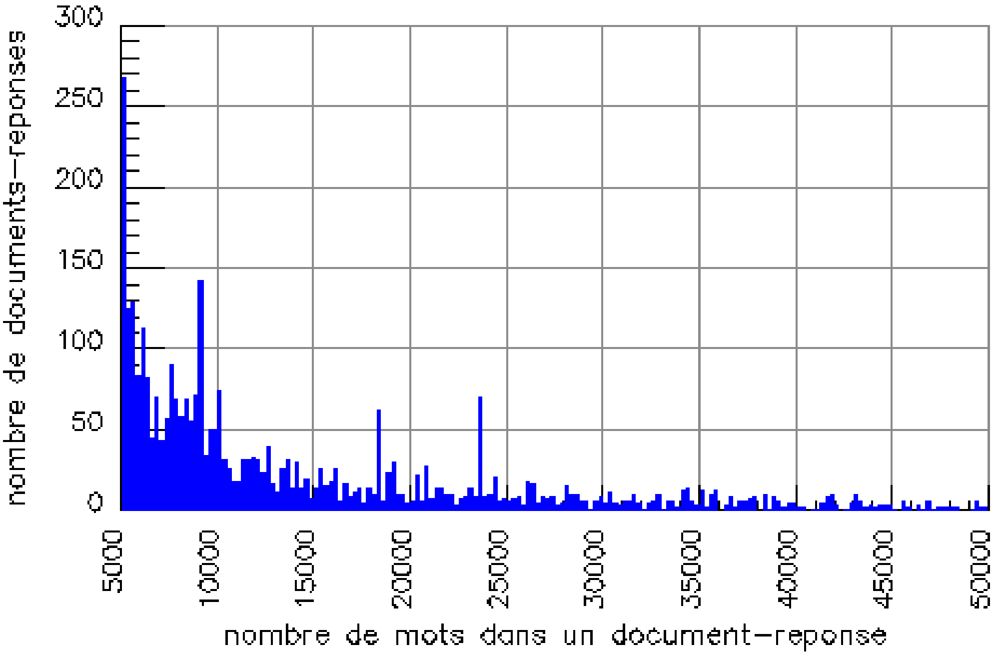


FIG. 7.25 – Taille des documents-réponses. Compartiment [5000 - 50 000] mots.

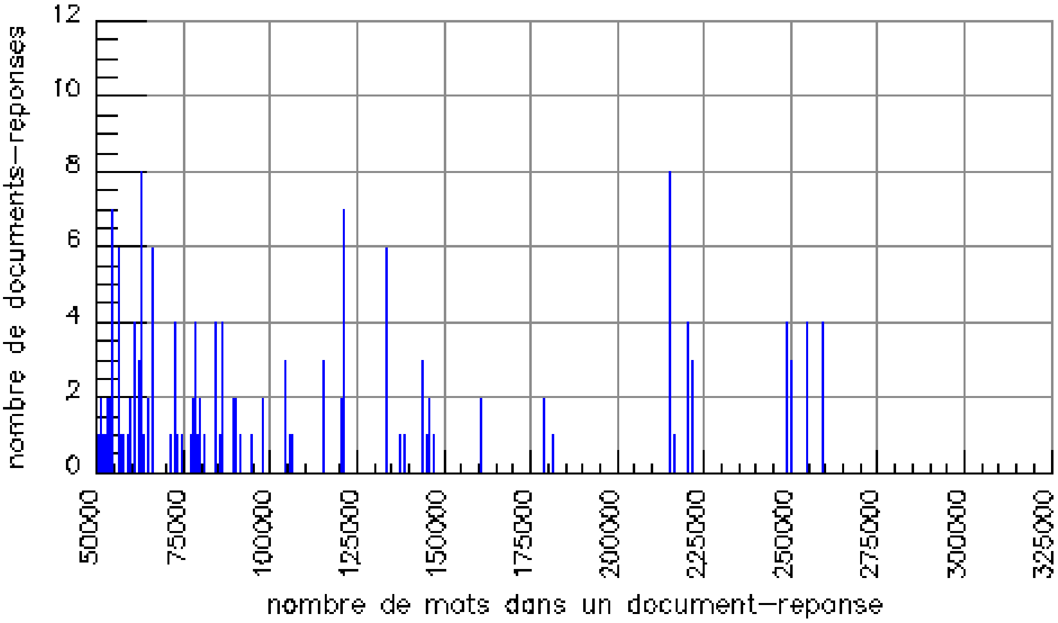


FIG. 7.26 – Taille des documents-réponse. Compartiments [50 000 - 350 000]

- une statistique graphique de taille de documents (exprimée en nombre des mots),
- estimation d'une valeur minimale, maximale et moyenne,

Les résultats d'exemples pour deux serveur sont présentés dans les figures suivantes [Figure 7.27 et 7.28]. Tous les autres résultats d'analyse et les statistiques réalisées sont présentés dans les figures de l'annexe.

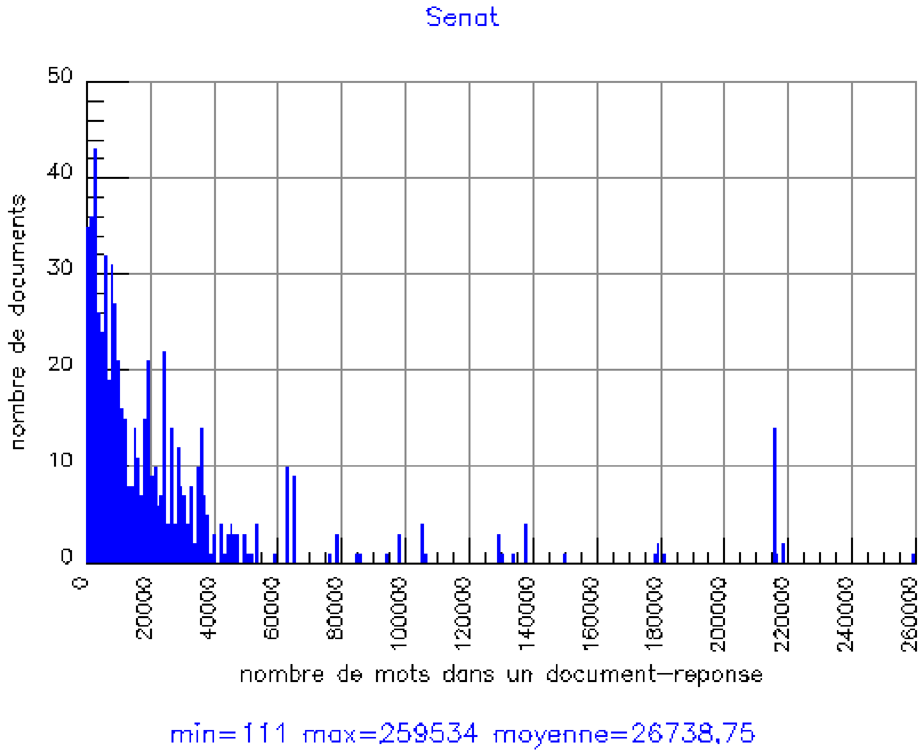


FIG. 7.27 – Taille de documents-réponses pour le serveur de Sénat.

7.4.3 Pourcentage de documents existants

Les statistiques suivantes concernent la validité de liens vers les documents réponses fournis par chacun des moteurs sur leurs pages de réponses. Pour chaque moteur nous avons calculé les valeurs suivantes selon la requête envoyée :

- Rendement maximal de liens valides [Figure 7.29],

  On observe que parmi tous les moteurs, quatre possèdent toujours les liens non valides sur leurs pages de réponses (le rendement maximal <

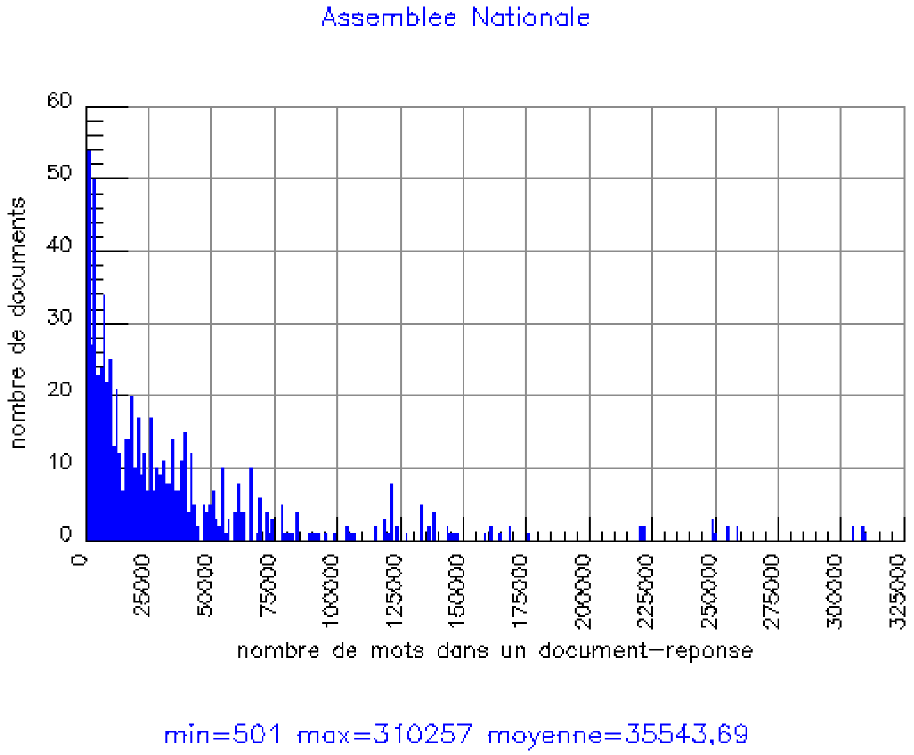


FIG. 7.28 – Taille de documents-réponses pour le serveur de Assemble Nationale.

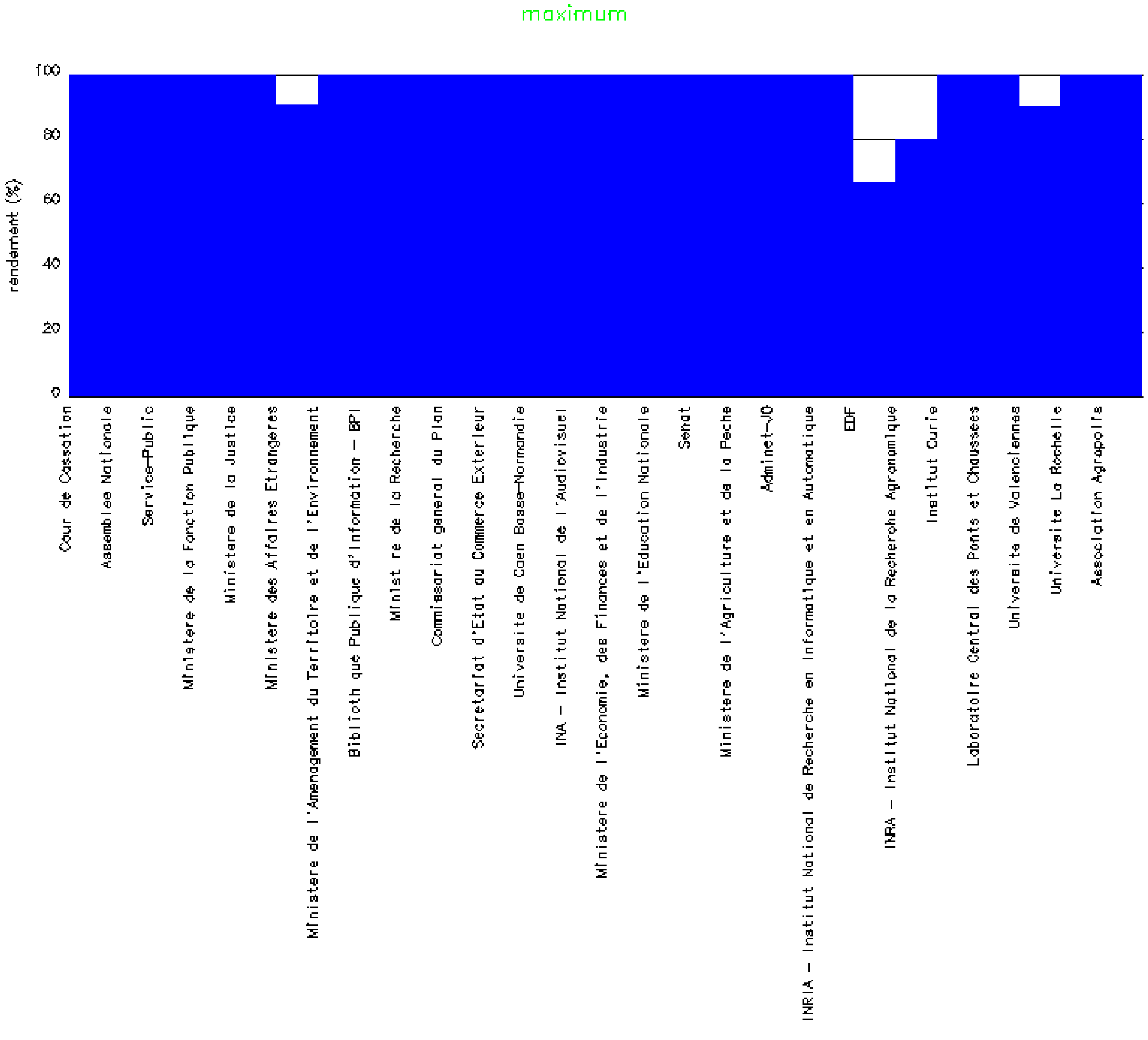


FIG. 7.29 – Rendement maximal de liens valides.

100 %) : Ministère des Affaires Étrangères, EDF, INRA et Université des Valenciennes

– Rendement minimal de liens valides [Figure 7.30],

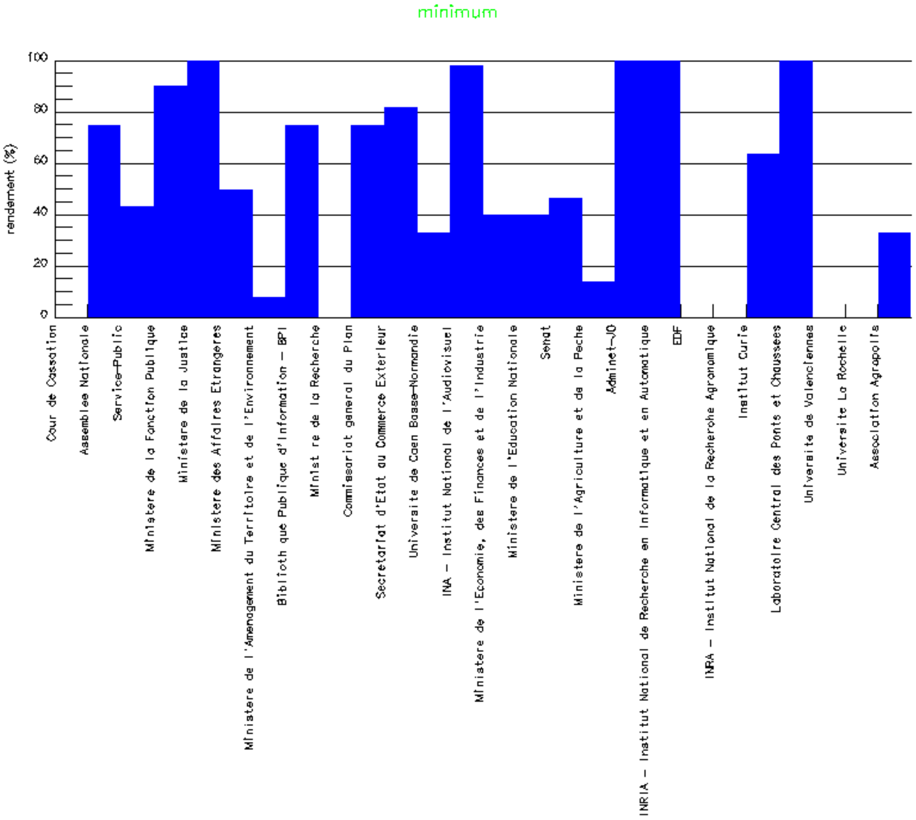


FIG. 7.30 – Rendement minimal de liens valides.

On observe que parmi tous les moteurs, quatre (Adminet-JO, INRIA, Ministère de la Justice, Laboratoire Central des Ponts et Chaussées) fournissent toujours les documents-réponses accessibles (liens existants et valides). Pour toutes les requêtes-test envoyées, il n'avait pas de pages de réponses avec les liens non valides (le rendement minimal = 100 %). Malheureusement on observe également un certain nombre de moteurs (Ministère de la Recherche, EDF, INRA, Université de Valenciennes, Université de la Rochelle) qui dans le cas extrême, présentent sur leurs pages de réponses seulement des documents non accessibles.

- Rendement moyen de liens valides [Figure 7.31],

  On observe que pour une majorité de moteurs le rendement moyen est supérieur à 75%.

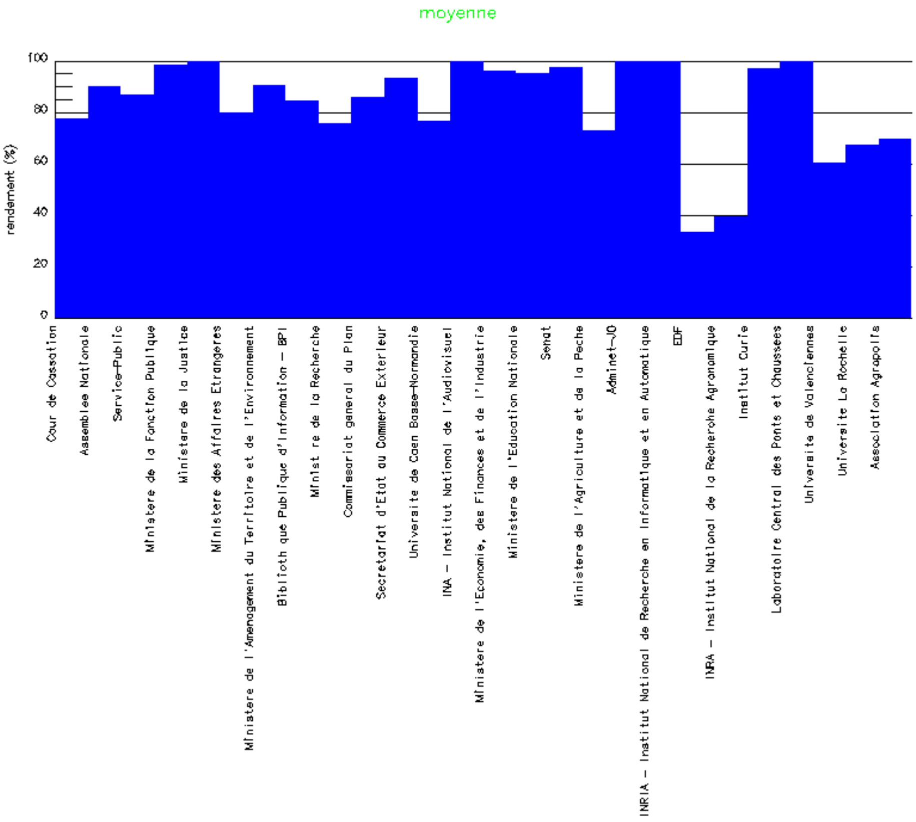


FIG. 7.31 – Rendement moyen de liens valides.

Le comparaison générale de tous les rendements est présentée dans la figure 7.32.

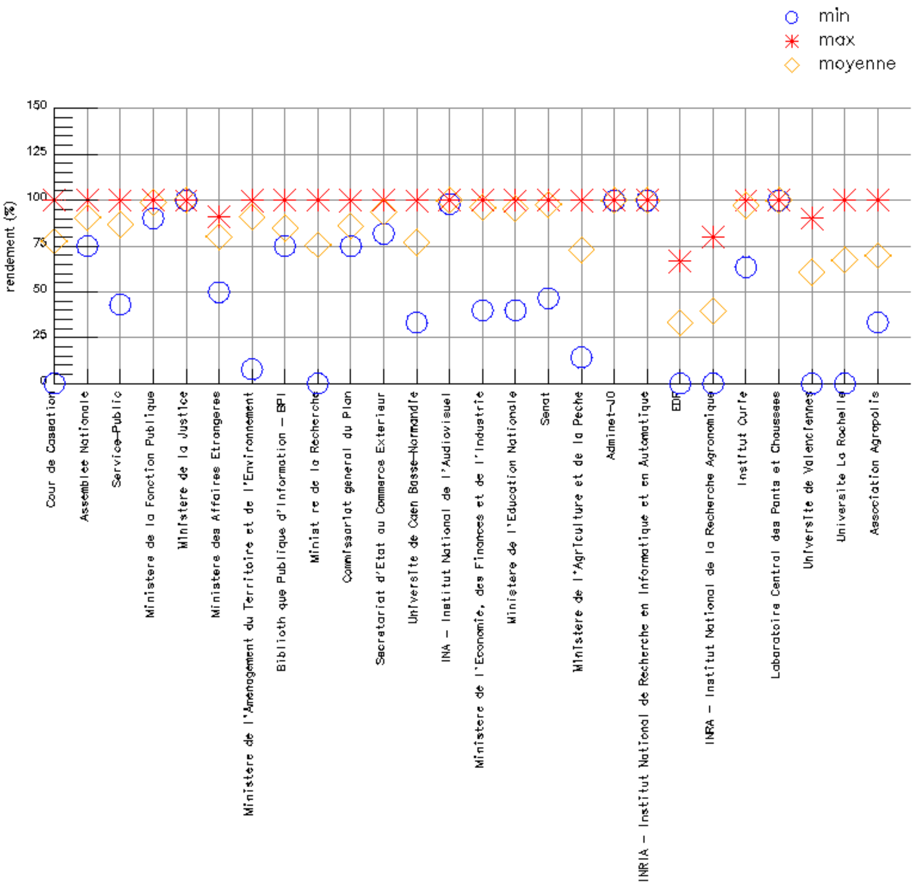


FIG. 7.32 – Rendement de liens valides.

7.4.4 Existence de mots de la requête dans les documents-réponses

Les statistiques suivantes concernent la vérification si les mots de la requête ont été contenus dans le document-réponses. La statistique a été faite pour les trois cas :

- occurrence de mots dans le titre HTML de document (entre les balises <TITLE> et <\TITLE>),
- occurrence de mots dans la partie qui contient les *meta* caractères (les mots entre les balises <META> et <\META>),
- occurrence de mots dans le *body* de document (les mots après la balise <BODY> jusqu'à la fin du document),

Les figures suivantes illustrent pour chacun des serveurs le pourcentage de réponses qui ne possèdent pas de mots de la requête sur la page de document-réponse fournie.

La figure 7.33 présente les cas où les mots de la requête n'existent ni dans le *title,* ni dans le *meta* , ni dans le corps *(body).*

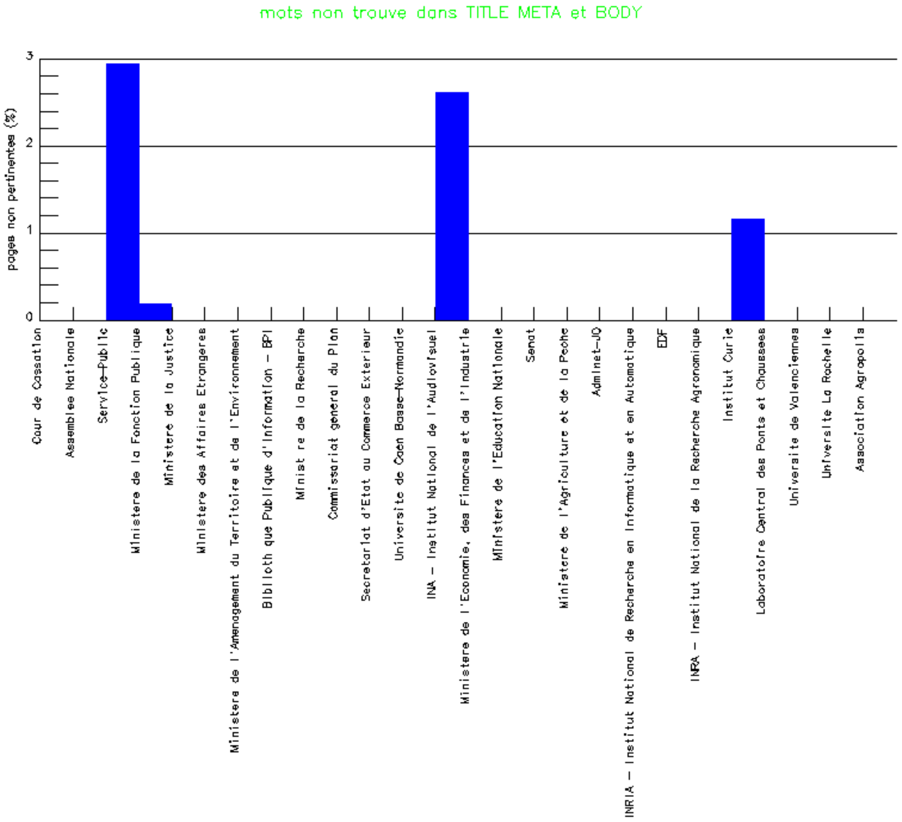


FIG. 7.33 – Existence de mots de la requête dans le document complet.

La figure 7.34 présente les cas où les mots de la requête n'existent pas dans le corps *(body)* de document-réponse. La figure 7.35 présente les cas où les mots de la requête n'existent pas dans le titre *(title)* de document-réponse.

Pour les documents-réponses dont les mots de la requête n'apparaissent pas, deux cas sont possibles :

- ils ne sont pas pertinents,
- ils sont pertinents, mais retrouvés par un algorithme qui n'est pas basé sur l'occurrence de mots de la requête, mais par exemple sur l'utilisation de synonymes.

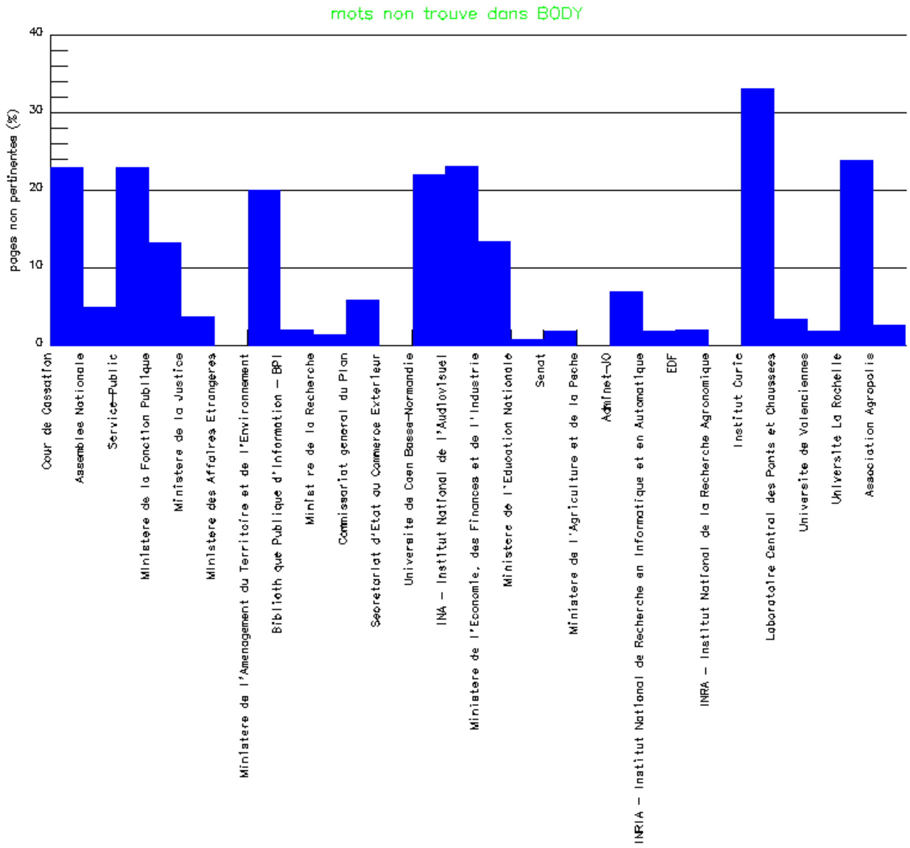


FIG. 7.34 – Existence de mots de la requête dans le corps *(body)* du document .

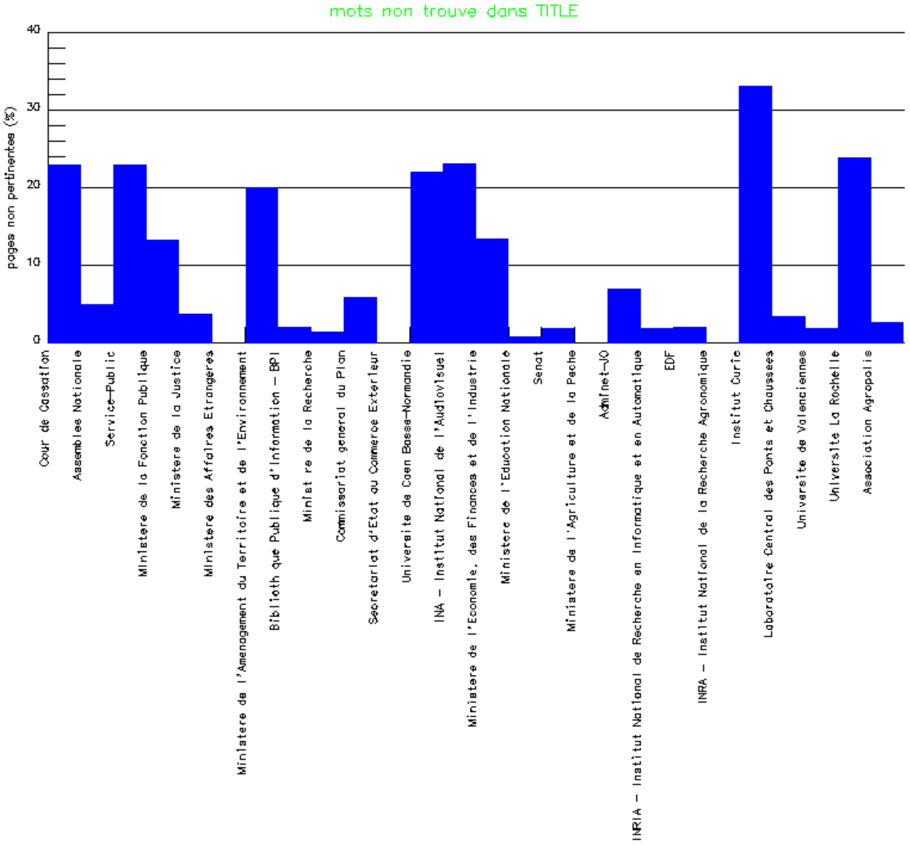


FIG. 7.35 – Existence de mots de la requête dans le *title* du document .

Quatrième partie

Conclusion et glossaire

Chapitre 8

Conclusion

8.1 Bilan

FIM-Métaindexeur, développé dans le cadre de cette thèse, est un méta-moteur pour la recherche des informations produites par l'administration française. La première version été mise en service sur le Web au début de l'année 1998. Depuis, FIM-Métaindexeur est utilisé pour la recherche de documents administratifs par un public très varié. Voici en exemple quelques références, qui indiquent que FIM-Métaindexeur est une bonne source de recherche.

La première référence, présentée dans la figure 8.1, est une page[1] de Vincent Ramette, responsable de l'Informatique du Cabinet Jeantet & Associés et également membre de l'équipe de Juriconnexion.

Le deuxième exemple montre que FIM-Métaindexeur est référencé comme un outil de recherche également dans le monde universitaire. La Faculté des Sciences Juridiques, Politiques et Sociale de Université de Lille cite FIM-Métaindexeur sur la page regroupant les sites intéressants pour la recherche dans le domaine des sciences politiques et du droit[2] et également sur la page de leurs références bibliographiques[3]. La figure 8.2 présente une page de référence pour les sites intéressants en sciences politiques.

---

[1] http/ :www.iblj.com/site/centre.htm
La page regroupant les sites Web favoris de Vincent Ramette, responsable de l'Informatique du Cabinet Jeantet & Associés, membre de Juriconnexion (www.juriconnexion.org)

[2] http ://www2.univ-lille2.fr/droit/CRAPS/enseignants/villalba/scpoweb.htm

[3] http ://www2.univ-lille2.fr/droit

http://www.iblj.com/site/centre.htm

Foreign Law Collections on the Web
Private International Law Homepage
Globelaw, International and Transnational Law

Parlement – France

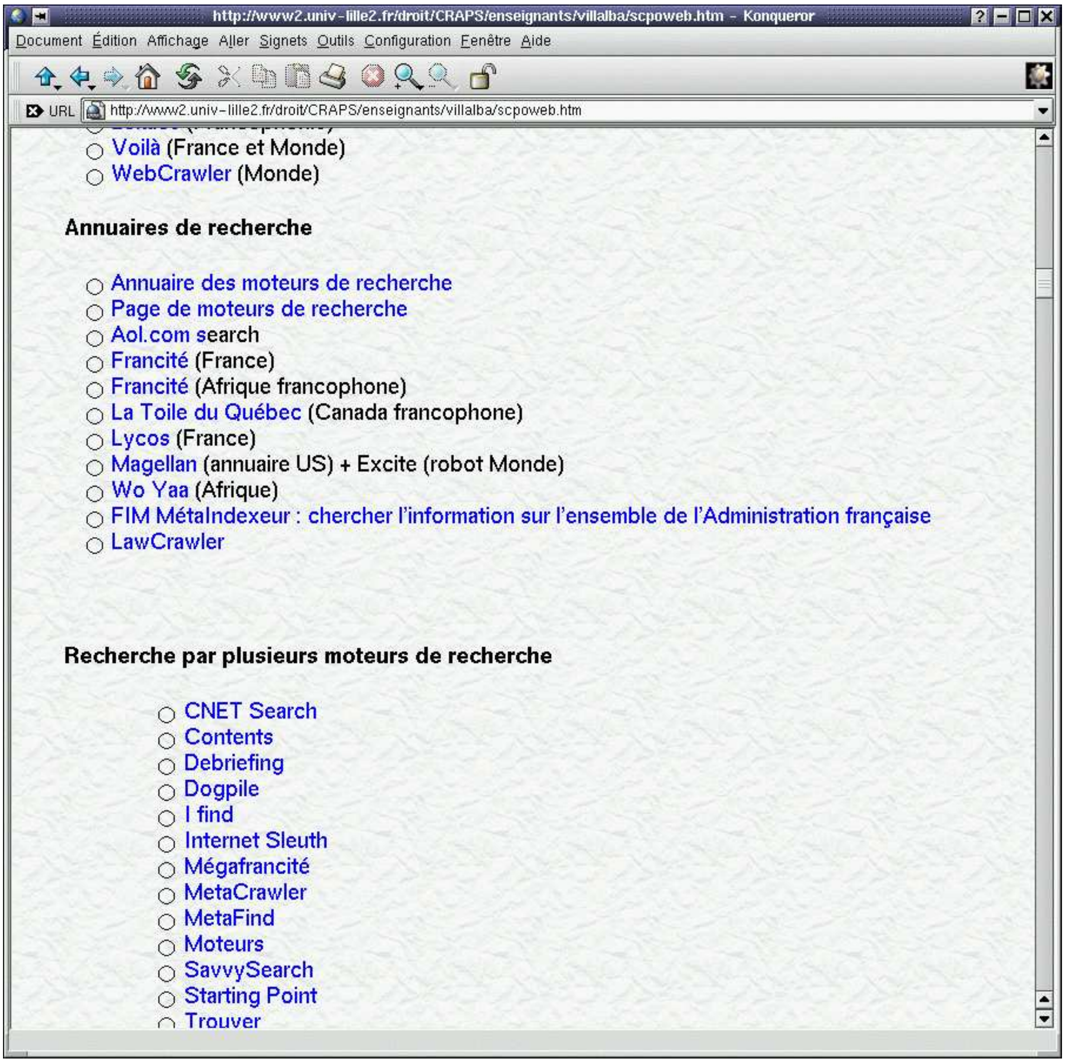
L'actualité à l'Assemblée
Assemblée nationale
Sénat

Pays France, portails et moteurs juridiques

Bulletin mensuel d''information européenne
www.en–droit.com
AbcJuris.Com– Le portail juridique francophone
Adminet
Admiroutes
Rabenou
DROIT
FIM MétaIndexeur
Bibelec
Ressources en droit

http://www.grolier.fr/cyberlexnet/coop.htm

Fig. 8.1 – La page de sites favoris de Vincent Ramette.



○ Voilà (France et Monde)
○ WebCrawler (Monde)

Annuaires de recherche

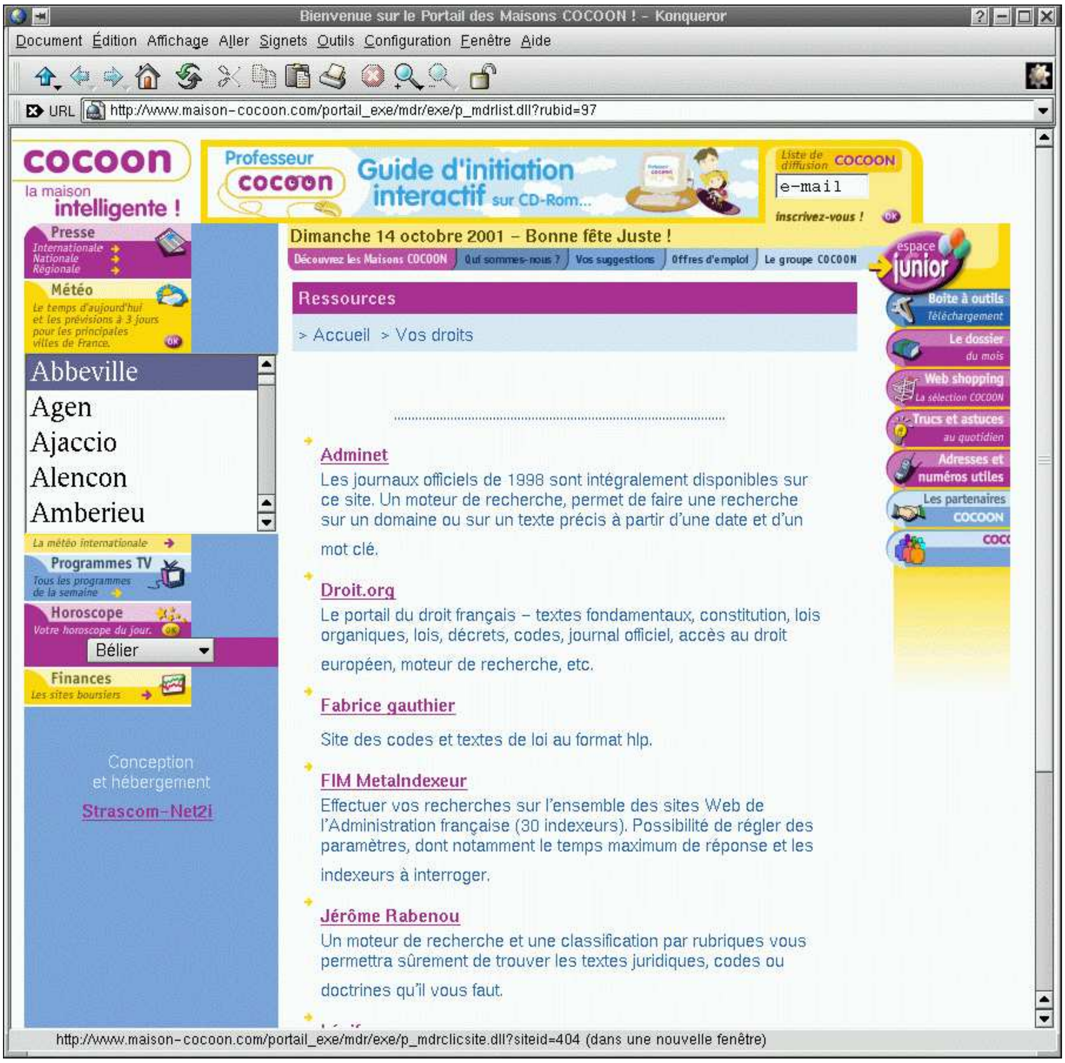
○ Annuaire des moteurs de recherche
○ Page de moteurs de recherche
○ Aol.com search
○ Francité (France)
○ Francité (Afrique francophone)
○ La Toile du Québec (Canada francophone)
○ Lycos (France)
○ Magellan (annuaire US) + Excite (robot Monde)
○ Wo Yaa (Afrique)
○ FIM MétaIndexeur : chercher l'information sur l'ensemble de l'Administration française
○ LawCrawler

Recherche par plusieurs moteurs de recherche

○ CNET Search
○ Contents
○ Debriefing
○ Dogpile
○ I find
○ Internet Sleuth
○ Mégafrancité
○ MetaCrawler
○ MetaFind
○ Moteurs
○ SavvySearch
○ Starting Point
○ Trouver

Fig. 8.2 – La page des références dans le domaine des sciences politiques du département de Droit de Université de Lille.

Le dernier exemple, présenté dans la figure 8.3, est une page de ressources pour la catégorie *"Vos Droits"* sur le site de conseil de société CoCoon[4]. On voit sur cette page FIM-Métaindexeur cité parmi d'autres sites-ressources, ce qui prouve que FIM-Métaindexeur est proposé à l'utilisation pour le grand public.



http://www.maison-cocoon.com/portail_exe/mdr/exe/p_mdrlist.dll?rubid=97

Dimanche 14 octobre 2001 – Bonne fête Juste !

Ressources

> Accueil > Vos droits

Adminet
Les journaux officiels de 1998 sont intégralement disponibles sur ce site. Un moteur de recherche, permet de faire une recherche sur un domaine ou sur un texte précis à partir d'une date et d'un mot clé.

Droit.org
Le portail du droit français – textes fondamentaux, constitution, lois organiques, lois, décrets, codes, journal officiel, accès au droit européen, moteur de recherche, etc.

Fabrice gauthier
Site des codes et textes de loi au format hlp.

FIM MetaIndexeur
Effectuer vos recherches sur l'ensemble des sites Web de l'Administration française (30 indexeurs). Possibilité de régler des paramètres, dont notamment le temps maximum de réponse et les indexeurs à interroger.

Jérôme Rabenou
Un moteur de recherche et une classification par rubriques vous permettra sûrement de trouver les textes juridiques, codes ou doctrines qu'il vous faut.

http://www.maison-cocoon.com/portail_exe/mdr/exe/p_mdrclicsite.dll?siteid=404 (dans une nouvelle fenêtre)

FIG. 8.3 – La page de CoCoon, le méta-moteur qui référence FIM-Métaindexeur pour la recherche dans le domaine du droit.

---

[4] http ://www.maison-cocoon.com/portail_exe/mdr/exe/p_mdrlist.dll ?rubid=97
La société Coocon fondée en avril 2000, filiales du Groupe Caapaction numéro deux de la construction de maisons individuelles en France, propose une autre manière de concevoir la maison individuelle. Maison-Cocoon.com propose : un moteur de recherche pour naviguer en toute simplicité sur des sites riches en informations ; des services dédiées à tous les membres de la famille ; des dossiers pour conseiller, proposer et rassurer ; un espace shopping ; un coin junior ; etc.

Tous les exemples cités prouvent que FIM-Métaindexeur est devenu un outil de recherche connu et effectivement utilisé.

Pendant cette thèse, différentes versions de développement ont été analysées et testées. Nous avons analysé les différentes solutions de fonctionnalités [Chapitres 5 et 6]. Le choix de moteurs à interroger a été modifié en fonction de l'analyse et des statistiques effectuées [Chapitre 7]. La version actuelle mise en service sur le Web est la dernière version développée et testée.

Pourtant il faut ajouter qu'il existe d'autres possibilités de fonctionnement de méta-outils que l'on peut envisager de réaliser. Il serait donc souhaitable de continuer la recherche dans les directions présentées dans la section suivante [Section 8.2].

Au cours de cette thèse nous avons également élaboré des méthodes d'analyse et de classification des moteurs de recherche, qui nous ont permis de réaliser un grand nombre de statistiques sur les moteurs de recherche des sites de l'administration française.

8.2 Perspectives

De nombreuses voies s'ouvrent dans les domaines de la recherche de l'information grâce à des méta-moteurs de recherche, notamment les suivantes :

- coopération du méta-moteur avec l'outil interrogé,
- utilisation du méta-moteur comme outil d'accompagnement pour la veille dans différents domaines.

8.2.1 Coopération du méta-moteur avec l'outil interrogé

Une meilleure coopération du méta-moteur avec l'outil interrogé pourrait améliorer fortement le fonctionnement du méta-moteur. Il serait donc envisageable d'élaborer une méthode spécifique dédiée à l'échange d'informations entre le méta-moteur et les outils interrogés. L'utilisation du protocole HTTP peut être envisagée pour réaliser les transferts de données entre applications-clientes spécifiques de chaque coté de la connexion.

8.2.2 Veille spécialisée dans des domaines spécifiques

Elle permet aux utilisateurs de se positionner dans différents domaines d'activité, de suivre leurs évolutions ou de faire des prévisions dans différents domaines : prévision économique (exemple : la bourse), commerciale, juridique, scientifique, stratégique, technologique, politique etc. Les outils offrant ces possibilités sont actuellement très recherchés. Ils peuvent aider à prendre les décisions stratégiques, à justifier un choix de programme, ou à prévoir la nécessité de réorienter ses activités. Pour être efficace, la veille d'information fait appel à l'ensemble des métiers de l'information scientifique et technique. Or, comme le méta-outil effectue la recherche sur des sources nombreuses et distribuées, il est bien adapté pour la veille de l'information.

Pour spécialiser un méta-moteur, il est nécessaire de disposer d'un thésaurus et d'un dictionnaire de règles, voire d'une syntaxe ou d'une sémantique spécifiques.

8.2.3 Autres perspectives en matière de recherche-développement futurs

Parmi les sujets qui nécessitent de nouveaux travaux, citons les suivants :

- classement automatique des moteurs de recherche grâce à des dictionnaires de questions et à des outils statistiques ; le classement automatique permet à la fois de déceler et d'utiliser la qualité des réponses et, par voie de conséquence, de classer un moteur dans un domaine particulier (dans une hiérarchie de type Yahoo !),
- réflexion sur les méthodes de tri des réponses conformes aux besoins des utilisateurs ; combinaison des critères de notoriété des sites (méthode Google), de notoriété des moteurs, d'analyse des réponses ou des titres des documents,
- utilisation des *feed-back* d'utilisateurs pour améliorer la recherche,
- amélioration de la description des primitives acceptées par les moteurs de base afin d'optimiser la traduction des questions et l'utilisation des paramètres dans les requêtes transmises à ces moteurs.

Chapitre 9

Glossaire

Agent de recherche [Car99] (agent intelligent, robot, *spiders*, *wanderers*, Web *worm*) Assistant électronique personnalisé, qui peut être paramétré. C'est un logiciel, qui accomplit un certain nombre de tâches répétitives à partir des règles de fonctionnement qui définissent son architecture. Il peut surveiller un thème à partir de différents filtres et de différentes sources, communiquer avec d'autres agents, observer un environnement que le veilleur lui commande d'explorer, etc. Il existe plusieurs niveaux de recherche :

- recherche générale : un premier survol
- recherche avancée : une observation plus fine
- recherche sectorielle : spécialisée dans un secteur
- recherche d'alertes : à la recherche de nouveautés.

Ambiguïté [dL00] Caractère de ce qui présente plusieurs sens possibles. Les langues naturelles sont par nature ambiguës. Cette ambiguïté enrichit la langue, pensons aux calembours, aux phrases suscitant des quiproquos.

- **Ambiguïté lexicale** [dL00] se situe au niveau du mot ; elle est de nature catégorielle, ou sémantique (homonymie, polysémie, ..). Un logiciel de traduction, un correcteur seront performants si l'ambiguïté est résolue. Le mot marche est une ambiguïté catégorielle : la marche d'escalier ; il marche (verbe marcher)
- **Ambiguïté syntaxique** [dL00] se situe au niveau de la structure des énoncés. On écoute les bruits de la fenêtre = on écoute les bruits que fait la fenêtre ou les bruits provenant par la fenêtre. Pour l'esprit humain, cette phrase replacée dans son contexte ne pose pas de difficulté de compréhension. Pour un programme de

traductions assistée par ordinateur (TAO), d'analyse de contenu, etc., il faudra d'abord lever cette ambiguïté.
- **Ambiguïté sémantique** [dL00] se situe au niveau du sens des mots et occasionne plusieurs représentations logiques d'un énoncé. Toutes les filles de la famille X aiment un homme = Elles aiment un homme différent ou un même homme ?
- **Ambiguïté pragmatique** [dL00] est levée par des connaissances extra-linguistiques. Le touriste était dans l'avion et il n'a pas décollé : il = le touriste ou l'avion ?

Analyseur [ATO] Terme générique qui désigne, en Analyse de texte "un programme ou un ensemble de programmes informatiques fournissant des renseignements analytiques sur des mots donnés ou sur un ou plusieurs textes. Les informations fournies par un analyseur peuvent être - d'ordre numérique : des indices numériques sur la répartition d'un mot donné dans un texte, par exemple - d'ordre symbolique - des représentations graphiques mettant en évidence la structure syntaxique ou le sens des énoncés.". C'est la finalité de l'analyse d'un texte qui justifie la mise en oeuvre d'un type d'analyseur ou la combinaison de plusieurs analyseurs.

- **Analyseur linguistique** [ATO] "Programme ou ensemble de programmes informatiques visant à produire des représentations (sous forme symbolique ou graphique) caractéristiques des phénomènes linguistiques (morphologie, syntaxe, sémantique et pragmatique) dans un texte, et cela dans le but d'en mettre en évidence le ou les sens. On trouve les analyseurs linguistiques aussi bien dans le domaine du traitement des données linguistiques écrites que dans celui du traitement de la parole.
- **Analyseur statistique** [ATO] Présent dans de nombreuses applications TIL, ce programme ou ensemble de programmes informatiques décompose le texte en une suite de signaux numériques. Le texte est alors considéré comme un ensemble de phénomènes dont les occurrences peuvent faire l'objet d'une analyse statistique et mathématique. Des informations lexicales, stylistiques (informations de structure) sont ainsi disponibles. L'analyseur statistique peut avoir une - approche quantitative : indices sur les données brutes (nombre d'occurrences, ..) - approche quantitative et qualitative : indices sur des données catégorisées
- **Analyseur lexicométrique ou lexico-statistique** [ATO] l'analyseur effectue des calculs sur les mots pris hors de leur contexte. L'approche est de type quantitatif. L'analyse lexicométrique permettra, par exemple, de vérifier la richesse objective d'un texte en en comparant le vocabulaire à des listes de référence.

Autoroutes de l'information [Dut00a] Structure constituée par des moyens de télécommunications et d'informatique interconnectés, qui permet d'offrir à un très grand nombre d'usagers de multiples services, en général à débit élevé, y compris des services audiovisuels. Note : On rencontre aussi le terme « inforoute », qui n'est pas recommandé.

Bande passante [Rac99] (*bandwith*) Largeur de bande d'un canal de transmission servant à transmettre un signal. Son amplitude est mesurée en mégahertz (MHz) (sa capacité est exprimée en bits par seconde ou bps). De ses caractéristiques dépend la qualité du signal transmis et reçu, donc le nombre d'informations qui peuvent y être véhiculées (voir baud). Les messages vocaux sont véhiculés sur des bandes étroites, tandis que les messages multimédias exigent des bandes larges.

Baud [Rac99] Unité pour mesurer en bps la vitesse de transmission entre deux modems (du nom de son inventeur Baudot). BPS : unité de mesure de la vitesse de transfert des bits par seconde (voir T1 et T3).

Base de données [Dut00b] Ensemble de données organisé en vue de son utilisation par des programmes correspondant à des applications distinctes et de manière à faciliter l'évolution indépendante des données et des programmes.

- **Base de données lexicales** [dL00] Ensemble de mots organisés selon, par exemple, leur catégorie grammaticale, afin de permettre leur utilisation et leur traitement par un programme informatique d'analyse de texte. La BDL va servir de référence lors de la catégorisation d'un texte.
- **Base de données textuelles** [dL00] désigne un ensemble de textes structurés à des fins de consultation, d'exploitation, d'analyse par des utilisateurs ou des programmes d'analyse de texte par ordinateur
- **Base de données relationnelle** [dL00] - classent les champs et les enregistrements sous la forme d'une table. Ces tables sont mises en relation les unes avec autres par des champs reliés entre eux.

Bibliométrie [dlsetI][Dut00b] définie en 1969 comme "l'application des mathématiques et des méthodes statistiques aux livres, articles et autres moyens de communication" (Pritchard). La biblométrie concerne des applications d'outils de statistique descriptive pour analyser notamment la distribution et la fréquence des données bibliographiques et/ou textuelles (notices, indexation (mots clés,), ...).

Bruit [AFN97] désigne toute réponse non pertinente à une recherche documentaire (AFNOR) Le taux de bruit est le % exprimant le rapport entre le nombre de documents non pertinents extraits et le nombre total de documents extraits. Ce terme désigne également, en Analyse de Texte par ordinateur, les résulats non désirés ou les fausses réponses parmi les résultats fournis par un programme informatique d'analyse de texte.

Catégorisation [dL00] Cette procédure consiste à associer à un mot, à un groupe de mots, ou à tout objet relevant d'un texte (signes typographiques, segments de texte, caractères spéciaux), des informations d'ordre linguistique : les catégories grammaticales (nom, verbe, ...), les traits sémantiques (humain, animal, ...), les traits narratifs (argument, contre-argument, ...) et/ou d'ordre sociologique : le domaine d'emploi, la réalité sociale sous-jacente, ...

CGI Common Gateway Interface [dL00, Nak] Protocole standard qui définit l'interface entre un serveur WEB (encore appelé serveur HTTP) et des programmes susceptibles de s'exécuter sur le serveur. Ces programmes peuvent être écrits à l'aide de tout type de langage (C/C++, Perl, Java...).

Chaîne de caractères [ATO] suite ou séquence de caractères alphanumériques, c'est-à-dire les lettres de l'alphabet, les chiffres et les caractères spéciaux (par exemple, les signes de ponctuation ou les symboles @, /, £, etc.), précédée et suivie d'un espace.

Corpus [Les96] Un corpus est un ensemble de productions linguistiques (langue écrite ou langue parlée) qui partagent les mêmes conditions de production, et qui seraient donc comparables entre elles.

Co-occurrence [SC] Une co-occurrence est un groupe de mots apparaissant fréquemment ensemble.

Cueillette [Car99] *(data mining, narrowcast search*, exploration de données) Première étape d'un processus de veille qui consiste à extraire et à rassembler le plus de données brutes, afin de mieux comprendre un environnement donné.

Descripteur

- [dL00] Désigne un mot ou groupe de mots (dit aussi mot-clé) choisi dans une liste fermée de mots reconnus comme étant le vocabulaire technique d'un domaine de spécialité donné (on parle de vocabulaire contrôlé ou lexique contrôlé), servant à représenter ou à caractériser sans ambiguïté le contenu d'un texte (ou partie d'un texte) ou d'un document (ou partie d'un document) afin d'en faciliter la classification et la recherche.

- [AFN97] Mot ou groupe de mots retenu dans un thésaurus et choisi parmi un ensemble de termes équivalents pour représenter sans ambiguité une notion apparaissant dans un document ou dans une demande de recherche documentaire.

Dictionnaire de synonymes, homonymes, antonymes [dL00] le logiciel fournit pour un mot une liste de synonymes (signification très proche), d'antonymes (contraires), d'homonymes (même prononciation, sens différent).

Document [ATO] entité identifiée et structurée contenant, entre autres, textes, tableaux, images et sons, pouvant être un objet d'étude, de traitement manuel ou électronique (par exemple, l'archivage), et d'échange entre des utilisateurs.

En ligne (*on-line*) Services accessibles à distance par Internet.

Fenêtre [Dut00a] Partie rectangulaire de l'écran d'un ordinateur à l'intérieur de laquelle sont affichées les informations relatives à une activité déterminée. Note : Plusieurs fenêtres peuvent êtres ouvertes simultanément ; elle peuvent être juxtaposées ou se recouvrir totalement ou partiellement.

Foire aux questions [Dut00a] Abréviation : *F.A.Q.* Synonymes : fichier des questions courantes, questions courantes. Rubrique présentant par sujets les questions les plus fréquemment posées par les utilisateurs, accompagnées des réponses correspondantes. Note : La foire aux questions a, en particulier, pour but de faciliter l'intégration des internautes novices dans un groupe de discussion et de diminuer le nombre des messages diffusés dans le réseau.

Freeware [Rac99] (graticiel) Logiciel gratuit souvent accessible par les inforoutes. Ne pas confondre avec *shareware* pour lequel l'auteur demande une contribution (*shareware*)

Fréquence [ATO] le nombre de fois qu'apparaît un mot donné dans un texte. La fréquence peut être exprimée en nombre absolu ou en pourcentage.

HTML *Hypertext markup language* : Langage pour la création de liens hypertextes. Ce langage de description de pages hypertextes est utilisé pour créer des pages sur le web. Il se compose d'une suite de signes ASCII dans laquelle sont inclues des commandes spéciales concernant le formatage des pages, les caractères, les caractères spécifiques pour les liaisons entre des pages, etc.

HTTP *Hypertext transfer protocol* : protocole de transfert d'hypertextes. Protocole de communication qui assure la transmission de fichiers hypertextes sur le web.

Hypertexte [dL00] Présentation d'un ensemble de textes et d'autres documents (graphique, base de données, vidéos, ...) qui permet une consultation non linéaire. Un renvoi (hyperlien) permet d'afficher un autre document situé ou non sur l'ordinateur. L'hypertexte constitue le fondement des services multimédias sur Internet.

Index [Car99] Liste des éléments, sous forme de mots-clés, contenus dans un domaine donné. Il peut devenir une liste de références permettant de localiser ces éléments dans un serveur.

Indexation [ATO] Procédure destinée à décrire et à caractériser le contenu d'un texte ou d'un document (ou d'une partie d'un texte ou d'un document) à l'aide de mots (dits mots-clés ou descripteurs) représentant les concepts ou informations dans ce document, et ce, à des fins de classification et pour en faciliter la recherche. De fait, un processus d'indexation comporte, d'une part, la reconnaissance et l'extraction des concepts informatifs (ou mots porteurs d'informations) et, d'autre part, la traduction de ces concepts dans un langage (dit langage documentaire) approprié à la classification et à la recherche de documents.

Information [Car99] Élément de sens perceptible. Ensemble cohérent qui constitue pour l'utilisateur une unité de connaissances. L'information entre dans la formation et l'acquisition de la connaissance, celle-ci étant la compréhension que l'on a d'une situation ou d'un événement.

Information textuelle [ATO] L'information ou connaissance communiquée par du code linguistique écrit, soit du texte. En d'autres termes, il s'agit d'un objet pourvu d'une signification propre (par exemple, un mot ou un groupe de mots), supporté par une forme particulière de codage, c'est-à-dire l'écriture et a fortiori le texte dans toutes ses dimensions (par exemple, une dimension consiste en la mise en relief du mot par le biais des caractères italiques). De ce fait, l'objet (le mot ou groupe de mots) devient le véhicule d'une information ou d'une connaissance donnée. En conséquence, l'information textuelle relève, d'une part, d'une organisation physique du code écrit qui la supporte et, d'autre part, d'un contenu signifiant ou informationnel qui peut faire l'objet d'analyses (par exemple, analyses linguistique, sémiotique, philosophique, etc.) ou de gestion (auquel cas on parle de gestion de l'information textuelle ; par exemple, la catégorisation du contenu ou du sens par des mots-clé).

Informatique documentaire [ATO] Désigne l'utilisation des moyens informatiques à des fins d'assistance, ou de résolution de problèmes, dans les domaines de la documentation et du traitement de l'information. Ainsi, que ce soit au niveau des analyses des textes ou des documents (par exemple, l'analyse du contenu des textes à des fins de repérage de mots porteurs d'informations) ou au niveau de l'activité documentaire (par exemple, la catégorisation du contenu des textes), on trouve de nombreux outils technologiques permettant la recherche, l'extraction, l'archivage et la gestion d'informations, l'analyse et le traitement de la documentation, etc.

Internaute [Dut00a] Utilisateur de l'internet. Note : On rencontre aussi le terme « cybernaute ».

Internet [Dut00a] Réseau mondial associant des ressources de télécommunication et des ordinateurs serveurs et clients, destiné à l'échange de messages électroniques, d'informations multimédias et de fichiers. Il fonctionne en utilisant un protocole commun qui permet l'acheminement de proche en proche de messages découpés en paquets indépendants. Note : L'acheminement est fondé sur le protocole I.P.(*Internet Protocol*), spécifié par l'Internet Society (I.S.O.C.). L'accès au réseau est ouvert à tout utilisateur ayant obtenu une adresse auprès d'un organisme accrédité. La gestion est décentralisée en réseaux interconnectés. Anglais : *Internet network, Internet, Net.*

Informations structurées [NoT] Les données structurées sont stockées dans des BD hiérarchiques, relationnelles, orientées objet, bibliographiques ou autres. Les informations sont ainsi identifiables et leurs chemins d'accès sont déterminés à l'avance par le concepteur de la BD. Les documents structurés sont des documents électroniques "représentés selon un format structuré, c'est-à-dire un format qui utilise des balises pour décrire la structure logique des documents. Souvent, la structure logique d'un document sera sa division en parties, chapitres, sections, etc., de même que certaines autres unités telles des notes de bas de page ou des références bibliographiques.". La norme SGML permet de définir des documents structurés.

Informations non structurées [dL00] Les données non structurées sont des informations qui ne sont ni classées, ni identifiées, comme par exemple les documents sur votre ordinateur ou sur le WEB. Les documents à gérer sont composés de données textuelles, numériques, sonores, graphiques, et des images (fixes ou animées). Ce type d'information représente environ les neuf dixièmes de l'information utilisée dans les organisations.

Inforoute [Mme99] Syn : réseau de réseaux, Internet, autoroute de l'information, information *highway.*Terme conçu pour décrire l'interconnexion entre les ordinateurs et les réseaux de télécommunications à travers le monde.

Ingénierie linguistique [Per99] (*language engineering*) est l'application de la connaissance des langues à l'élaboration de systèmes informatiques capables de reconnaître, de comprendre, d'interpréter et de produire du langage humain sous toutes ses formes.

Intelligence économique [Car99] (compétitive intelligence, business intelligence) Ensemble des actions coordonnées de cueillette, d'analyse-synthèse et de diffusion d'informations en vue de son exploitation par des acteurs économiques.

Interactif [Per99, Vos98] Qualifie les matériels, les programmes ou les conditions d'exploitation qui permettent des actions réciproques en mode dialogué avec des utilisateurs ou en temps réel avec des appareils. On parle d'interaction quand un programme réagit à une entrée de l'utilisateur et, sur la base de cette entrée, continue dans une direction particulière.

Interface [Vos98] Une interface désigne généralement "un composant qui sert à mettre en relation deux systèmes et qui contient des données sous une forme compréhensible pour ces deux systèmes.".

- **Interface logiciel** permet, par exemple, de convertir tous les documents d'une base de données en un format acceptable par le traitement de texte.
- **Interface utilisateur** assure le dialogue entre l'homme et la machine. On y retrouve l'interface graphique, l'interface textuelle, l'interface en langage naturel (écrit ou oral).

Internet [ISO00] Réseau informatique mondial composé d'un grand nombre de réseaux nationaux, régionaux, locaux et privés qui sont reliés par les protocoles de transmission de la famille TCP-IP (*Transmission Control Protocol - Internet Protocol*).

Internet Documentation [NOR] Toute la documentation est disponible en version électronique (gratuitement).

- **FYI :** *For Your Information* document d'information
- **Internet Draft** document de travail ; sans aucun statut officiel ; libre de rédaction ; peut devenir une RFC
- **RFC :** *Request For Comments* - document portant sur les standards relatifs aux divers protocoles de communication, de transmission, de transfert et d'interconnexion du réseau Internet ; tous les RFC ne deviennent pas des standards : documents à titre historique, informatif ou expérimental un RFC est une norme

de facto (ou standard) élaborée par une organisation ou un consortium autre qu'un organisme de normalisation

- **RFC SETS** - ensembles de normes pour faciliter la recherche d'informations.

Lemme [dL00] (forme canonique) désigne la forme de référence d'un mot, c'est-à-dire la forme du mot sans les marques (dites marques de flexion) qui l'actualisent dans le discours.

Lemmatisation [dL00] consiste à donner à un mot (accordé, conjugué) une forme canonique (forme de base = le lemme) pour, entre autres, qu'il puisse entrer dans un dictionnaire.

Linguistique [Per99] la science du langage, qu'elle étudie à travers la diversité des langues naturelles parlées sur la Terre. Le statut scientifique de la linguistique implique un certain nombre de contraintes sur la méthode. En général, on procède par la proposition de modèles qu'on essaie de tester contre des données pour les infirmer. A la lumière des faiblesses découvertes, on modifie le modèle pour le tester encore, et ainsi de suite.

Logiciel de navigation [Dut00a,] Synonyme : navigateur, n.m. Dans un environnement de type internet, logiciel qui permet à l'utilisateur de rechercher et de consulter des documents, et d'exploiter les liens hypertextuels qu'ils comportent. Anglais : *browser*.

Menu [Car99] Ensemble de commandes affichées sous la forme d'une liste ou d'une carte de connaissance servant de mécanismes d'accès à l'information.

Menu déroulant [Car99] Présentation d'un menu sous la forme d'une liste de sous-commandes qui descend à l'écran.

Métamoteur [Car99] Outil hybride permettant d'utiliser simultanément plusieurs répertoires et moteurs de recherche.

Mot [ATO] un caractère ou une suite de caractères (dite chaîne de caractères) précédé et suivi d'un espace, ayant une signification dans la réalité ou à l'égard d'un texte (par exemple, dans une phrase). Mais la notion de mot demeure complexe et largement discutée dans la littérature scientifique. Ainsi, au niveau de l'écrit, est considéré comme mot, la chaîne de caractères pomme ; mais, théoriquement l'ensemble des signes suivants, `pomme de terre` (nous parlons alors de groupe de mots) peut également être pris pour un seul mot. Dans la mesure où pour un ordinateur, le mot, quelle que soit sa définition théorique, reste un objet sans signification (au sens où l'humain l'entend) qu'il doit manipuler, le choix d'une définition dépend donc de l'utilisateur. Ce choix s'opère en fonction de plusieurs facteurs, par exemple, en fonction des objectifs de la recherche, du type des analyses et des interprétations à effectuer, du degré de raffinement recherché dans ces analyses, de la complexité des traitements, et autres.

Mot-clé [ATO] un mot ou groupe de mots choisi librement dans le titre ou dans le corps d'un texte (ou d'un document), permettant d'en caractériser le contenu et d'en faciliter la classification et la recherche. Le mot-clé peut aussi éventuellement faire partie d'une liste fermée de mots (dits descripteurs) utilisée à des fins de description et de recherche de documents (on parle de vocabulaire contrôlé).

Moteur de recherche [Car99] Outil pour des recherches précises grâce à des mots clés spécifiques. Il est constitué de deux éléments : un robot (*spider*) qui visite les sites et un système d'indexation qui, à partir de filtres, analyse leurs contenus. Certains outils utilisent des opérations booléennes comme *AND, OR, NOT*, etc. Il y a quatre types de moteurs de recherche : par mots-clés, par thèmes, par cartes, par méta-sites.

Navigateur (Logiciel de navigation) [Dut00a] Dans un environnement de type Internet, logiciel qui permet à l'utilisateur de rechercher et de consulter des documents, et d'exploiter les liens hypertextuels qu'ils comportent. Anglais : *browser.*

Occurrence [ATO] un élément linguistique ou un mot, toutes les fois qu'il apparaît dans un texte. Ainsi, l'apparition du mot informatique dans un texte constitue une occurrence du mot informatique ; de même, ordinateur et ordinateurs constituent deux occurrences du mot ordinateur. Les mots ou éléments linguistiques qui figurent en même temps aux côtés de l'occurrence dans le texte sont les cooccurrences, ou sont dits ses cooccurents.

Page d'accueil [Dut00a] Anglais : *home page.*

- Page de présentation d'un site sur la toile mondiale.
- Page de tête affichée par un logiciel de navigation.

Répertoire de recherche [Car99] (annuaire, index) Sélection d'informations regroupées par catégories et sous catégories (genre page jaune) permettant une recherche par couches successives : générales, avancées, sectorielles, d'alertes.

Serveur [Dut00a] Système informatique destiné à fournir des services à des utilisateurs connectés et, par extension, organisme qui exploite un tel système. Note :

Un serveur peut par exemple permettre la consultation et l'exploitation directe de banques de données.

SGML [ATO] (*Standard Generalized Markup Language*) un langage informatique dit langage de balisage, devenu une norme internationale (ISO 8879 :1986), permettant de codifier la structure logique d'un document (par exemple, les titres, sous-titres, paragraphes, introduction, conclusion, images, etc.) quels que soient les logiciels qui ont servi à la création du document. En d'autres termes, le format SGML permet de définir à la fois la structure interne d'un document à des fins de recherches sur son contenu, et sa mise en forme pour une restitution (ou affichage) à l'écran ou une impression sur papier.

Silence l'absence totale de résultat, suite à une analyse supposée faite par un programme informatique d'analyse de texte.

Signet [Dut00a] Moyen d'accéder rapidement à une adresse universelle préalablement stockée en mémoire par l'utilisateur. Anglais : *bookmark*.

Site de la toile [Dut00a] site sur la toile. Anglais : *website, web site*.

Syntagme est formé d'une ou de plusieurs chaînes de caractères séparées par des blancs. Un syntagme peut contenir des mots vides mais il doit comporter au moins un mot significatif.

Temps réel [Rac99] Processus informatique qui s'effectue immédiatement ; condition nécessaire aux applications interactives.

Terme

- **linguistique (terminologie)** [ATO] un mot ou groupe de mots qui ne s'applique qu'à un et un seul objet ou concept, et ce, dans un domaine donné. Autrement dit, il s'agit d'un mot qui désigne de façon univoque un objet ou un concept dans un domaine, qui, généralement est un domaine de spécialité. De fait, dans l'idéal terminologique, un terme n'a pas d'équivalent (ou synonyme) dans le domaine désigné. Un terme formé d'un seul mot (par exemple, avion) est dit terme simple ou uniterme, alors que celui constitué de plusieurs mots est appelé terme complexe ou multi-terme (par exemple, avion à réacteur), bien que le mot terme soit couramment employé dans les deux cas. On parle également d'unité terminologique.
- **Sc. de l'information (documentation)** [ATO] un mot ou groupe de mots en ce qu'il est succeptible de décrire ou de caractériser sans ambiguïté le contenu d'un texte ou d'un document, et cela, à des fins de classification et de repérage du texte ou du document. Autrement dit, il s'agit d'un mot dont le sens n'est

pas nécessairement univoque, donc qui peut avoir des équivalents (ou synonymes), mais qui, avant tout, catégorise le mieux le contenu du texte ou du document.

Texte une séquence ou suite de caractères appartenant à une langue naturelle (on parle de lexique), respectant les règles de fonctionnement de la langue (dites grammaire) et de structuration du document dans cette langue. En d'autres termes, le texte est une collection de mots organisés non seulement sur le plan linguistique, mais aussi selon une norme de présentation donnée, pour être objet d'interprétation de nature intelligente chez l'humain.

Thésaurus

- [ATO] “ensemble de mots ou de termes (dits descripteurs) constituant un vocabulaire défini (vocabulaire contrôlé de termes), ayant entre eux des relations d'ordre sémantique (par exemple, une relation hiérarchique orientée du générique vers le spécifique) ou pragmatique, et qui s'appliquent à un ou plusieurs domaines de la connaissance.”
- [Hu994] Les relations entre les termes représentent un corpus sémantique d'un domaine et tiennent compte de l'évolution du domaine concerné. Le thésaurus est donc un outil en construction permanente. Comme le signalait Hudon, le thésaurus se doit d'être un instrument de travail éminemment flexible et adaptable.
- [ISO86] “... vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relations générique-spécifique)”.

Thésaurus électronique [Mar] L'usage de l'informatique favorise la manipulation d'un thésaurus : le terme recherché est entouré des termes reliés qui sont tous des liens hypertextuels qui nous ramènent à leurs réseaux. Selon la norme ISO [ISO86], le contenu d'un thésaurus peut être représenté en trois modes principaux :

* présentation alphabétique ;

* systématique (organisation en domaines ou disciplines et organisation par facettes, ou la combinaison des deux) ;

* présentation graphique (schéma fléché ou disposition graphique). Les deux derniers types sont accompagnés d'un index alphabétique.

Traitement automatique des langues naturelles (TIL)[dL00] Les differents domaines TIL s'intéressent aux langues naturelles, parlées et écrites. Pour que les ordinateurs puissent analyser, générer, traduire, interroger, traiter, manipuler des textes, de nombreuses connaissances sur le langage naturel sont requises : la prononciation, l'orthographe, la signification, l'emploi des mots ; la combinaison des mots pour donner un sens à une phrase, etc.

Toile d'araignée mondiale (TAM) [Dut00a] toile mondiale, toile, Abréviation : T.A.M. Dans l'internet, système, réparti géographiquement et structurellement, de publication et de consultation de documents faisant appel aux techniques de l'hypertexte. Voir aussi : hypertexte, internet. Anglais : *World Wide Web*, *WWW.*

Traduction automatique (TA) Traduction automatique par ordinateur de textes d'une langue naturelle à une autre.

Traduction assistée par ordinateur (TAO) Traduction assistée par ordinateur (TAO) C'est une traduction effectuée avec intervention humaine.

Traitement de l'information [Rac99] En informatique : traitement électronique des données à l'aide d'un langage de programmation. En linguistique : traitement sémantique du contenu.

Unité textuelle [ATO] tout élément ou signe constitutif d'un texte. Par exemple, tous les mots du texte suivant sont des unités textuelles : Jean a mangé la soupe. En somme, il s'agit de tout caractère ou chaîne caractères reconnus et traitables par le système d'analyse de texte en regard du texte et seulement par rapport au texte ; contrairement aux unités lexicales ou aux lexèmes qui sont plutôt des éléments du lexique. Autrement dit, est désignée comme unité textuelle les éléments qui sont soumis à des opérations de description ou de catégorisation, mais seulement à l'égard de leur contexte.

Veille [Car99] Maîtrise de l'information nécessaire à la surveillance des environnements sociopolitiques et économiques d'un environnement. Elle suppose une politique d'intériorisation du milieu pour créer des avantages compétitifs ; c'est-à-dire une prise en compte par l'institution de tous les éléments extérieurs susceptibles d'interagir avec ses activités.

- **commerciale** Veille s'intéressant aux fournisseurs et aux client.
- **concurrentielle** Veille évaluant les forces et les faiblesses du concurrent et modifiant les stratégies commerciales.
- **économique** Veille analysant ce que font les concurrents en cherchant à réduire le niveau d'incertitude commerciale.

- **intégrée** (voir modèle intégré) Veille intègrant les trois niveaux d'activités d'une société tertiaire : technologique, économique et stratégique.
- **juridique** Veille concernant l'évolution de la réglementation et de la législation.
- **scientifique** Veille concernant surtout la recherche fondamentale.
- **stratégique** (pour l'État) Veille qui analyse le contexte de l'environnement socio-politique afin de développer des politiques publiques, et à plus long terme, un plan de société.
- **technologique** Veille qui répond aux questions : qui fait quoi, où, comment et avec quelles normes

Veilleur [Car99] Traqueur de signaux annonciateurs du futur.

Vocabulaire contrôlé [dL00] (ou mots-clé ou descripteur) désigne une liste fermée de mots reconnus comme étant le vocabulaire technique d'un domaine de spécialité donné. Il sert à : la description du contenu d'un texte l'indexation la recherche d'information. L'utilisation du vocabulaire contrôlé sert ainsi à limiter les problèmes relatifs aux traitements et à la recherche d'informations pouvant résulter d'une catégorisation ou d'une classification des documents en langage naturel ou vocabulaire libre. On parle aussi de lexique contrôlé ou de liste de termes normalisés.

Bibliographie

[Abi99] P S Abideen. Meta search engines : effective tool for information retrieval. In P. S. G. Kumar, Information C. P. Vashishth. Ahmedabad, and Library Network Centre (INFLIBNET, editors, *Proceedings of the Sixth National Convention for Automation of Libraries in Education and Research, Nagpur, India, 18-20 February 1999.*, pages 362–9, 1999.

[AFN97] AFNOR. Vocabulaire de la documentation. Technical report, Association Française de Normaliztion, 1997.

[And98] Olivier Andrieu. *Trouver l'information sur l'Internet.* Eyrolles, 61, Bld Saint-Germain, 75240 Paris, 1998.

[Aqu01] Stephen Aquino. Search engines ready to learn. *Technology Review*, AVR 2001.

[ATO] *Glossaire ATO - Service d´Analyse de Textes par Ordinateur de l´Université du Québec (UQAM).*

[Bal99] J L Balas. An extensible constructor tool for the rapid, interactive design of query synthesizers. *Computers in Libraries ;*, 19(5) :34–37, MAI 1999.

[Bar00a] Joe Barker. *Meta-Search Engines*, 2000, *http ://www.lib.berkeley.edu/TeachingLib/Guides/Internet/MetaSearch.html.*

[Bar00b] Joe Barker. *Selected Subject Directories*, 2000, *http ://www.lib.berkeley.edu/statsTeachingLib/Guides/Internet/SubjDirectories.html.*

[BD99] Jean-Sylvain Boutel and Jean-Pierre Damiano. Le développement des agents de recherche. *L'Informatique Professionnelle*, 170(1) :35–44, JAN 1999.

[Ben98] Albert Benschop. *Meta Search Engines*, 1998, *http ://www.pscw.uva.nl/sociosite/SEARCH/search3.html.*

[Ber98] Petr Berka. *Intelligent systems on the Internet*, 1998, *http ://lisp.vse.cz/ berka/ai-inet.htm.*

[BKP^{+}98] Michelle Baldonado, Seth Katz, Andreas Paepcke, Chen-Chuan K. Chang, Hector Garcia-Molina, and Terry Winograd. An extensible constructor tool for the rapid, interactive design of query synthesizers. In *DL'98*, pages 19–28, Pittsburgh, June 1998.

[Bla98] K Blakeman. Essential web search techniques. *Business Information Searcher*, 8(3) :16–18, 1998.

[Bra99] Tom Bray. Measuring the web. In *Proceedings of Fifth International World Wide Web Conference*, 1999.

[Bur99] R Burgin. The monte carlo method and the evaluation of retrieval system performance. *JASIS*, 50(2) :181–191, FEB 1999.

[BYRN99] Ricardo Baeza-Yates and Berthier Ribero-Neto. *Modern Information Retrieval*, volume 1. ACMPRESS, 1999.

[Car95] Jeremy Carl. *Protocol Gives Sites Way To Keep Out The 'Bots*, 1995, *http ://www.internetworld.com/print/1995/11/01/news/nobots.html.*

[Car99] Michel Cartier. *Multimédium-Lexique : La cueillette de l'information*, 1999, *http ://www.mmedium.com/cgi-bin/veille/lexique.cgi ?zone=3.*

[CCCGMP97] Kevin Chen-Chuan Chang, Hector Garcia-Molina, and Andreas Paepcke. *Boolean Query Mapping Across Heterogeneous Information Sources (Extended Version)*, 1997, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs7-1583/postscript.*

[CGK98] Soumen Chakrabarti, David Gibson, and Jon Kleinberg. *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text*, 1998, *http ://www7.conf.au/programme/fullpapers/1898/com1898.html.*

[Chr96] Tom Christiansen. *Perl WWW Documentation*, MAR 1996.

[Chu96] Heting Chu. Search engines for the world wide web : A comparative study and evaluation methodology. In *Proceedings of ASIS*. American Society for Information Science, 1996.

[CKCP81] Ernest Czogala, Jaroslaw Korzen-Chmielowski, and Witold Pedrycz. *Rachunek Prawdopodobienstwa i Statystyka Matematyczna.* Skrypty Uczelniane, 1981.

[CMW98] James R. Chen, Nathalie Mathé, and Shawn Wolfe. Collaborative information agents on the world wide web. In *DL'98*, pages 279–280, June 1998.

[Con00] The Unicode Consortium. *The Unicode Standard, Version 3.0*, 2000, *http ://www.unicode.org/unicode/uni2book/u2.html.*

[Coo68] William S. Cooper. Expected search length : A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *JASIS*, 19(1) :30–41, 1968.

[CS98] Liren Chen and Katia Sycara. Webmate, a personal agent for browsing and searching. In Katia Sycara and Michael Wooldridge, editors, *Second International Conference on Autonomous Agents*, pages 132–139. ACM SIGART, ACM Press, MAI 1998.

[DBP94a] Paul De Bra and Reinier Post. *Information retrieval in the Word-Wide Web : making Client-based searching feasible*, 1994, *http ://pc-paul.win.tue.nl/ reinpost/www94/www94.html.*

[DBP94b] Paul De Bra and Reinier Post. *Searching for Arbitrary Information in the WWW : the Fish-Search for Mosaic*, 1994, *http ://wwwis.win.tue.nl/ debra/wwwf94/article.html.*

[DD99] Ruxandra Domenig and Klaus R. Dittrich. An overview and classification of mediated query systems. *SIGREC*, 28(3) :63–72, SEP 1999.

[DH97a] Daniel Dreilinger and Adele E. Howe. Experiences with selecting search engines using metasearch. *TOIS*, 15(3) :195–222, JUI 1997.

[DH97b] Daniel Dreilinger and Adele E Howe. *SavvySearch : A meta-Search Engine that learns which search Engines to Query*, 1997, *http ://daniel.www.media.mit.edu/people/daniel/papers/ss-aimag.ps.gz.*

[DHKP94] Paul DeBra, Jan Houben, Geert, Yoram Kornatzky, and Reinier Post. *Information Retrieval in Distributed Hypertexts*, 1994, *http ://pc-paul.win.tue.nl/ reinpost/www94/www94.html.*

[dL00] Uniwersité de Liège. *Lexique*, 2000, *http ://www.owil.org/fr_lexique.htm.*

[dlsetI] Institut de l'information scientifique et technique (INIST). *Le programme de recherche en infométrie*, *http ://www.inist.fr.*

[Dre96a] Daniel Dreilinger. Description and evaluation of a meta-search agent. Master's thesis, CSU, Fort Collins, Colorado, OCT 1996.

[Dre96b] Daniel Dreilinger. *Formal Standards for Meta-Search*, 1996, *http ://www.w3.org/Search/9605-Indexing-Workshop/Papers/Dreilinger@CSU.html.*

[Dus97a] Oliver M Duschka. *Query Planning and Optimization in Information Integration.* PhD thesis, STANFORD, 1997.

[Dus97b] Oliver M. Duschka. *Query Planning and Optimization in Information Integration*, 1997, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs7-1598/postscript.*

[Dut00a] Deschamp Duthil. *Glossaire informatique des termes relatifs à l'informatique publiés au Journal officiel du 16 mars 1999*, 2000, *http ://www-rocq.inria.fr/qui/Philippe.Deschamp/RETIF/19990316.html.*

[Dut00b] Deschamp Duthil. *Lexique informatique officieux de la commission ministérielle de terminologie informatique*, 2000, *http ://www-rocq.inria.fr/ deschamp/CMTI/glossaire.html.*

[DW96] Karen M. Drabenstott and Marjorie S. Weller. A comparative approach to system evaluation : Delegating control of retrieval tests to an experimental online system. In *Proceedings of ASIS*. American Society for Information Science, 1996.

[Eic94] David Eichmann. The "rbse" spider - balancing effective search against web load. In *Proceedings of First WWW*, MAI 1994.

[ES99] D Eichmann and P Srinivasan. Filters, webs and answers : The university of iowa trec-8 results. In *TREC-8 Proceedings*, pages 259–267. The Eighth Text REtrieval Conference (TREC 8), NIST Special Publication, 1999.

[EW94] Oren Etzioni and Daniel Weld. A softbot-based interface to the internet. *CACM*, 37(7) :72–76, JUL 1994.

[FFBL^{+}99] R Fielding, Frystyk.H, Tom Berners-Lee, J Gettys, and Jeffrey Mogul. *Hypertext Transfer Protocol HTTP/1.1.* HTTP Working Group, internet draft edition, AVR 1999.

[Fie94] T Roy Fielding. Maintaining distributed hypertext infostructures : Welcome to momspider's web. In *Proceedings of First WWW*, JUN 1994.

[FPC98] James French, Allison L Powell, and Walter R Creighton. Efficient searching in distributed digital librairies. In *DL'98*, pages 283–284, June 1998.

[Gaf95] Adam Gaffin. *EFF's Guide to the Internet w.3.1.* EFF, san Francisco CA 94117 USA, FEB 1995.

[GCGMP97] Luis Gravano, Kevin Chang, Hector Garcia-Molina, and Andreas Paepcke. *START Stanford Protocol Proposal for Internet Retrieval and Search*, 1997, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs7-1580/postscript.*

[GKR98] David Gibson, Jon Kleinberg, and Prabhakar Raghavan. *Inferring Web Communities from Link Topology*, 1998, *http ://simon.cs.cornell.edu/home/kleinber/ht98.ps.*

[Gra93] M Gray. *Growth of the World-Wide Web*, 1993, *http ://www.mit.edu :8001/aft/sipb/user/mkgray/ht/web-growth.html.*

[Gre00] D Green. The evolution of web searching. *Online Information Review (United States)*, 24(2) :124–137, AVR 2000.

[Gri00] Luc Grivel. *L'hypertexte comme mode d'exploitation des resultats d'outils et methodes d'analyse de l'information scientifique et technique.* PhD thesis, L'Universite de Droit, d'Economie et des Sciences d'Aix-Marseille, Marseille, JAN 2000.

[Ham96] K Hammond. *ECHO - Even Chaos Holds Order Experimental Meta Search Engine*, 1996, *http ://www.infolab.nwu.edu/echo/.*

[HH97] Stephen P. Harter and Carol A. Hert. Evaluation of information retrieval systems : Approaches, issues and methods. *ARIST*, 32 :1–94, 1997.

[HHMN99] Monika R. Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. Measuring index quality using random walks on the web. In Elsevier Science, editor, *Proceedings of the 8th International World Wide Web Conference*, pages 213–225, MAI 1999.

[HHN00] L Huang, M Hemmje, and E J Neuhold. Admire : an adaptive data model for meta search engines. *Computer Networks*, 33 :p.431–448, Jun 2000.

[Hof00] P Hoffman. *RFC2781, UTF-16, an encoding of ISO 10646*, 2000, *http ://www.landfield.com/rfcs/rfc2781.html.*

[Hu994] *Le thésaurus : Conception, élaboration, gestion.* ASTED, 1994.

[Huy96a] Nam Huyn. *A More Aggressive Use Of Views To Extract Information*, 1996, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs6-1577/postscript.*

[Huy96b] Nam Huyn. *Query Reformulation under Incomplete Mappings*, 1996, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs6-1576/postscript.*

[HVQM99] P Hipola, B Vargas-Quesada, and A Montes. Descripcion y evaluacion de agentes multibuscadores. description and evaluation of multi-searcher agents. *Profesional de la Informacion*, 8(11) :15–24, NOV 1999.

[iC98] School Libraries in Canada. Search tools chart. *School Libraries in Canada*, 18(1) :21–22, 1998.

[ISO86] ISO. Documentation - guidelines for the establisment and developement of monolingual thesauri. Technical report, ISO, 1986.

[ISO00] ISOC. *Internet Society*, 2000, *http ://www.isoc.org.*

[Jon98] Steve Jones. Dynamic query results previews for a digital library. In *DL'98*, pages 291–292, June 1998.

[JP00] Cahier Jean-Pierre. Comment fonctionnent les moteurs de recherche sémantique. *Le Monde Informatique*, 875 S(875 S) :32 –33, DEC 2000.

[JRP00] Lobna Jéribi, Béatrice Rumpler, and Jean Marie Pinon. *Système d'aide à la recherche et l'interrogation de bases documentaires, basé sur la réutilisation d'expériences.*, 2000.

[Key96] John Keyes. Using conceptual categories of questions to measure differences in retrieval performance. In *Proceedings of ASIS*. American Society for Information Science, OCT 1996.

[Kla98] Judith Klavans. Data bases in digital libraries : Where computer science and information management meet. In ACM PRESS, editor, *PODS98*, pages 224–226, JUN 1998.

[Kle98] Jon Kleinberg. Authoritative source in a hyperlinked environment. In *Proc.SIAM*, 1998.

[Koe99] W Koehler. An analysis of web page and web site constancy and permanence. *JASIS*, 50(2) :162–181, FEB 1999.

[Kor97] Robert R. Korfhage. *Information Storage and Retrieval.* 1997.

[Kos93] Martijn Koster. *Guidelines for Robot Writers*, 1993, *http ://info.webcrawler.com/mak/projects/robots/guidelines.html.*

[Kos94a] Martijn Koster. *A Standard for Robot Exclusion*, 1994, *http ://info.webcrawler.com/mak/projects/robots/robots.html.*

[Kos94b] Martijn Koster. *The Web Robots FAQ...*, 1994, *http ://info.webcrawler.com/mak/projects/robots/faq.html.*

[Kos95] Martijn Koster. *Robots in the Web : threat or treat ?*, 1995, *http ://info.webcrawler.com/mak/projects/robots/threat-or-treat.html.*

[Kos96] Martijn Koster. *Evaluation of the Standard for Robots Exclusion*, 1996, *http ://info.webcrawler.com/mak/projects/robots/eval.*

[KT00] Mei Kobayashi and Koichi Takeda. Information retrieval on the web. In *ACM Computing Surveys*, pages 144–173. ACM, JUN 2000.

[Laa97] H v d Laan. Meta web woelers. meta search programs. *Informatie Professional*, 1(1) :28–29, 1997.

[Lar97] Jean-Pierre Lardy. *Recherche d'information sur l'internet, Outils et méthodes.* ADBS EDITIONS, 1997.

[LCK$^+$97] Joseph Lee, David Cheung, Ben Kao, Jax Law, and Thomas Lee. Inteligent agent for matching information providers and consumers on the word-wide-web. *IEEE*, pages 189–199, 1997.

[Leb99] Marie Lebert. *Le multilinguisme sur le WEB*, 1999, *http ://www.ceveil.qc.ca/multifr5.htm.*

[Lel98] Catherine Leloup. *Moteur d'indexation et de recherche.* Eyrolles, 61, Bld Saint-Germain, 75240 Paris, 1998.

[Les96] Lessard. *Introduction à la linguistique français*, 1996, *http ://qsilver.queensu.ca/french/Cours/215.*

[Lèf00] Philippe Lèfevre. *La recherche d'information du texte intégral au thésaurus.* Hermes, 2000.

[LG98a] S Lawrence and CL Giles. Inquirus, the neci meta search engine. *Computer Networks and ISDN Systems*, 30(1-7) :1995, AVR 1998.

[LG98b] Steve Lawrence and Lee Giles. Context and page analysis for improved web search. *IEEE*, 2(4) :38–46, 1998.

[LG98c] Steve Lawrence and Lee Giles. Searching the world wide web. *Science*, 280(5360) :1998, 1998.

[LG99] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740) :107–109, JUL 1999.

[Liu98] Jian Liu. Guide to meta-search engines. *BF Bulletin-Special Libraries Association-Business and Finance Division*, (107) :17–20, JAN 1998.

[LL99] Robert Losee and Anne Lee. Measuring search engine quality and query difficulty : Ranking with target and freestyle. *Journal of the American Society for Information Science*, 50(10) :882–889, 1999.

[Los97] Robert Losee. Comparing boolean and probabilistic information retrieval systems across disciplines and queries. *Journal of the American Society for Information Science0*, 48(2) :143–156, 1997.

[Los00] Robert Losee. When information retrieval measures agree about the relative quality of document rankings. *Journal of the American Society for Information Science0*, 51(9) :834–840, 2000.

[LTLN98] Ee-Peng Lim, Cheng-Hai Tan, Boon-Wan Lim, and Wee-Keong Ng. Querying structured web resources. In *DL'98*, pages 297–298, June 1998.

[Mar] M Marjanovic. *Le monde des thésaurus documentaires*, *http ://www.ceveil.qc.ca/traitem/marjanom/intro.htm.*

[Mar97] Massimo Marchiori. The quest for correct information on the web : hyper search engines. *CNISDN*, 2(29) :1225–1235, 1997.

[MAR99] Pierre-Marie MARTIN. La pertinence de l'information sur internet, internet ou la concierge planétaire. *Kafkaiens Magazine*, (8) :6, 1999.

[McB94] Olivier McBryan. Genvl and wwww : Tools for taming the web. In *Proceedings of First WWW*, May 1994.

[Mer] Marie-Thérèse Mercier. Analyse, indexation documentaire dans un centre de documentation.

[MH98] Joel D Martin and Robert Holte. Searching for content-bases addresses on the world-wide web. In *DL'98*, pages 299–300, June 1198.

[Mik97] Thomas Mike. A web crawler in perl. *Linux Journal*, pages 80–87, AUG 1997.

[Mme99] Mmedium. *Lexique de Michel Cartier*, 1999, *http ://www.mmedium.com/.*

[Mot98] Motrech. *La recherche des documents*, 1998, *http ://www-dist.cea.fr/ext/neuf/moteur/problemes-recherche.html.*

[MyS99] MySQL. *MySQL Reference Manual.* TcX AB, Detron HB and Monty program KB, OCT 1999.

[Nak] Didier Nakache. *Data Mining sur Internet (Glossaire) CNAM Conservatoire National des Arts et Métiers*, *http ://home.nordnet.fr/ dnakache/probatoire.*

[Nic00] S Nicholson. A proposal for categorization and nomenclature for web search tools. *Journal of Internet Cataloging*, 2(3/4) :9–28, MAR 2000.

[NOR] NORMOS. *Internet Engineering Standards Repository*, *http ://www.normos.org/.*

[NoT] Groupe NoTIAL. *Les instances de normalisation*, *http ://www.ceveil.qc.ca/notial/instance.html.*

[Not98] Greg R. Notess. *Search Engines Comparison*, 1998, *http ://www.imt.net/ notess/compeng.html.*

[Not01] Greg R Notess. *Search Engine Statistics*, 2001, *http ://www.searchengineshowdown.com/stats/.*

[NV96] Christian Neuss and Johan Vromans. *Application CGI en Perl pour les Webmasters.* International Computer Press, AUG 1996.

[Per99] Jean Perrot. *La linguistique.* 15e édition, 1999.

[Pyt98] C Pythoud. Utilisation des instruments de recherche disponibles sur internet. use of search tools available on the internet. *ARBIDO*, 13(5) :7–11, MAI 1998.

[Rac99] Jean Racicot. *Multimédium-Lexique : Les inforoutes*, 1999, *http ://www.mmedium.com/cgi-bin/veille/lexique.cgi ?zone=2.*

[Rag95] Dave Ragget. *Hypertext Markup language Specification version 3.0.* W3C, dsr@w3.org, internet draft edition, MAR 1995.

[Raj99] T B Rajashekar. Internet and web search engines. In S.G. Kumar, Information C. P. Vashishth. Ahmedabad, and Library Network Centre, editors, *Proceedings of the Sixth National Convention for Automation of Libraries in Education and Research, Nagpur, India, 18-20 February 1999*, pages Supplement p.1–16), 1999.

[RC99] J Repman and R D Carlson. Surviving the storm : using metasearch engines effectively. *Computers in Libraries*, 19(5) :50–55, MAI 1999.

[RMGMP98] Gerard Rodriguez-Mula, Hector Garcia-Molina, and Andreas Paepcke. *Collaborative value filtering on the Web*, 1998, *http ://elib.stanford.edu :80/Dienst/Repository/2.0/Body/stanford.cs8-63/postscript.*

[Sai99] A P Sainul. Meta search engines : effective tool for information retrieval. In P. S. G. Kumar, India : Information C. P. Vashishth. Ahmedabad, and Library Network Centre, editors, *Sixth National Convention for Automation of Libraries in Education and Research (CALIBER 99), Nagpur, India, 18-20 February 1999*, pages 362–369, 1999.

[SB96] Linda Schamber and Judy Bateman. User criteria in relevance evaluation : Toward development of a measurement scale. In *Proceedings of ASIS*. American Society for Information Science, 1996.

[SC] T Selva and T Charnier. *Utilisation de dictionnaire : apport de l'informatique dans les dictionnaires pour apprenants, projet Alexia*, *http ://comte.univ-fcomte.fr/RECHERCHE/P7/pub/Euralex98/Euralex.html.*

[SE95] Erik Selberg and Oren Etzioni. *Multi-Engine Search and Comparison Using the MetaCrawler*, 1995, *http ://www.w3j.com/1/selberg.169/paper/169.html.*

[SJYD98] Robert G. Sumner, Jr., Kiduk Yang, and Bert J. Dempsey. An interactive WWW search engine for user-defined collections. In *DL'98*, pages 307–308, 1998.

[SM83] Gerard Salton and Michael J McGill. *Information retrieval*, volume 1. McGraw-Hill Book Company, 1983.

[SPCK$^+$99] Tomek Strzalkowski, Jose Perez-Carballo, Jussi Karlgen, Anette Hulth, Pasi Tapanainen, and Timo Lahtinen. Natural language information retieval : Trec-8 report. *Kluwer Academic Publishers*, 1999.

[Sta98] T Stanley. Meta-search engines : where are the limits ? In Learned Information Europe Ltd, editor, *Proceedings of the 2Second International Online Information Meeting, London, 8-10 December 1998*, pages 297–300, 1998.

[Sta99] T Stanley. To meta-search or not to meta-search ? *Assignation ;*, 16(2) :8–10, JAN 1999.

[Sug99] C Sugunavathy. Web search engines. In P. S. G. Kumar, Information C. P. Vashishth. Ahmedabad, and Library Network Centre, editors, *Proceedings of the Sixth National Convention for Automation of Libraries in Education and Research*, pages 316–23. Academic libraries in the Internet era, 1999.

[Sul00] Danny Sullivan. *Invisible Web Gets Deeper*, 2000, *http ://www.searchenginewatch.com/sereport/00/08-deepweb.html.*

[Sun] Sun release. *Perl 5.0 Reference Pages.*

[Swa96] Deborah E. Swain. Information search process model : How freshmen begin research. In *Proceedings of ASIS*. American Society for Information Science, OCT 1996.

[Tan98] W Tan. Subject access on internet : highlights of the metasearch engines. *Journal of Educational Media and Library Sciences ;*, 36(1) :20–29, SEP 1998.

[Tom97] Anastasios Tombros. *Reflecting user information needs through query biased summaries.* PhD thesis, GLASGOW, SEP 1997.

[Tom99] N Tomaiuolo. Are metasearches better searches ? *Searcher*, 7(1) :30–34, 1999 1999.

[TS98] Anastasios Tombros and Mark Sanderson. The advantages of query-biased summaries in information retrieval. In Bruce Croft, Alistair Moffat, C.J van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of ACM SIGIR*, pages 2–10. ACMPRESS, AUG 1998.

[Vos98] A Voss. *Dictionnaire de l'informatique et de l'Internet.* 1998.

[vR79] C.J van Rijsbergen. *Information retrieval.* 1979.

[WAB$^+$99] Allison Woodruff, Paul M. Aoki, Eric Brewer, Paul Gauthier, and Lawrence A. Rowe. An investigation of documents from the world wide web. In *Proceedings of Fifth International World Wide Web Conference*, 1999.

[WCSP96] Larry Wall, Tom Christiansen, Randal.L Schwartz, and Stephen Potter. *Programming Perl.* O'Reilly Associates, second edition, SEP 1996.

[Weg96] Katarzyna Wegrzyn. Etude et realisation d'un robot pour la recherche d'information sur le web. rapport DEA E/193/CRI, Université d'Evry-val-d'Essonne, AUG 1996.

[Wei98] A Weiss. The evolution of world wide web search tools. In *Proceedings of the 2Second International Online Information Meeting, London, 8-10 December 1998*, pages 289–295, DEC 1998.

[ZE99] O Zamir and O Etzioni. Grouper : a dynamic clustering interface to web search results. *Computer Networks*, 31(11-16) :31–44, MAI 1999.

[Ziz95] Mountaz Zizi. *Cartes Dynamiques Intractives : une métaphore spatiale pour l'exploration des espaces informationnels.* PhD thesis, UNIVERSITÉ de Paris XI ORSAY, 91 405 ORSAY Cedex, 1995.

[ZK99] Jin Zhang and Robert R. Korfhage. A distance and angle similarity measure method. *JASIS*, 50(2) :772–779, 1999.

Liste des tableaux

1.1 Exemple de fichier d'indexation direct. 28
1.2 Exemple de fichier d'indexation inversé. 29

2.1 La croissance du site Web par mois en milliers. 52
2.2 Nombre des sites et d'utilisateurs par pays. 53
2.3 Nombre des sites et d'utilisateurs par continent. 54
2.4 La nombre des Sites d'Internet. 56

3.1 Le catalogue de recherche. 62
3.2 Comparaison de différentes moteurs de recherche. 65
3.3 Comparaison de quelques méta-moteurs sur serveur distant. 67
3.4 La comparaison des certains méta-moteurs client. 69
3.5 La comparaison de certains outils avec accès à des bases de données. 70
3.6 Exemple de portails. 71
3.7 Les robots pour rapatrier des sites. 72
3.8 Les liens non valides. Le comparaison de moteurs de recherche 80

5.1 Les moteurs généraux interrogés par FIM-Métaindexeur. 109
5.2 Les portails administratifs interrogés par FIM-Métaindexeur. 109
5.3 Les moteurs de recherche sur les sites de l'administration française interrogés par FIM-Métaindexeur. 110
5.5 Les moteurs utilisés comme moteur de recherche sur les sites interrogés. 112

5.6 La méthode du protocole HTTP utilisée pour l'interrogation des moteurs de recherche. 113

5.7 La table de la correspondance pour les fonctions booléennes. 117

5.8 Recherche des phrases. Acceptation de guillemets. 118

5.9 Acceptation des accents. 120

5.10 Les paramètres optionnels utilisés par le Fim-Métaindexeur. 122

7.1 Serveurs analysés pour les statistiques. 158

7.2 Nombre de mots dans la question. 160

7.3 Longueur des mots dans les requêtes. 161

7.4 Nombre de mots dans la question (TEST1). 162

7.5 Nombre de mots dans la question (TEST2). 162

7.6 Longueur des questions : minimale, maximale et moyenne pour les fichiers TEST1,TEST2 et pour les questions d'Adminet. 162

7.7 Longueur des mots dans les requêtes (TEST1). 162

7.8 Longueur des mots dans les requêtes (TEST2). 162

7.9 Longueurs des mots : minimale, maximale et moyenne. 164

7.10 La structure de base de données. Table STAT_REPONSE. 166

7.11 La structure de base de données. Table REPONSE. 167

7.12 La structure de base de données. Table DOCUMENT. 167

7.13 La structure de base de données. Table PAR_MOT. 168

7.14 Le degré de la dépendance selon le coefficient de corrélation. 177

7.15 La liste de serveurs analysés. 177

7.16 Liste de serveurs regroupés. 180

7.17 Tableau des paires de serveurs avec une corrélation > 0.8 182

7.18 Tableau des paires de serveurs avec une corrélation > 0.7 182

7.19 Tableau des paires serveurs avec une corrélation > 0.6 183

7.20 Tableau des paires serveurs avec une corrélation > 0.5 184

Table des figures

1.1 Exemple de trois relations sémantiques entre les descripteurs dans le thésaurus.[Lèf00] 34

1.2 Exemple de représentation du bruit et du silence en recherche de l'information 37

1.3 Méthode des hyperliens. Liens simples, niveau n. 41

1.4 Méthode des hyperliens. Liens multiples, niveau 2. 42

2.1 Exemple de structure hypertexte 47

2.2 Structure de document HTML 48

2.3 Un code de page HTML simple 49

2.4 Interprétation de code HTML dans le navigateur Netscape. 49

2.5 La croissance du Web. 51

3.1 Le modèle de moteur de recherche. 64

3.2 Comparaison de couverture des différents moteurs. 66

3.3 Comparaison de couverture de certains moteurs. 68

3.4 Statistiques de changement de taille pour les bases de données [Not01] 77

3.5 Statistiques du changement de taille pour les bases de données (pour les partenaires de Inktomi)[Not01] 78

4.1 Schéma du méta-moteur. 90

4.2 Les modules principaux de méta-moteur. 95

4.3 Analyse de la question posée. Le schéma général. 97

5.1 Traduction de question posée vers la requête spécifique de chaque moteur. 116
5.2 Couverture de moteurs qui donnent des résultats corrects pour les différentes méthodes automatiques. 126

6.1 Schéma d'interfonctionnement des modules et des données de FIM-Métaindexeur. 130
6.2 Module d'entrée des données. 131
6.3 L'algorithme de création d'une requête. 133
6.4 Algorithme de récupération de réponses. 135
6.5 Algorithme d'analyse des réponses récupérées. 136
6.6 Formulaire pour insertion de paramètres d'un moteur. 138
6.7 FIM-Métaindexeur. La page de formulaire. 139
6.8 FIM-Méta indexeur. Sélection du serveur. 143
6.9 Tri des réponses. 144
6.10 FIM-Métaindexeur. La page des réponses. 146
6.11 FIM-Métaindexeur. La page de résultats, les statistiques. 147
6.12 Page d'aide pour FIM-Métaindexeur 149
6.13 Les statistiques globales. 150
6.14 FIM Métaindexeur. Statistiques des réponses. 151
6.15 Nombres de réponses en fonction de temps de réponse. 152
6.16 Statistiques sur les temps de réponse. 153
6.17 Statistique de temps de réponse en fonction de l'heure 154

7.1 Statistiques sur la longueur de la question pour les questions de fichiers-tests (TEST1 et TEST2) et les questions d'Adminet. 163
7.2 Statistiques sur la longueur des mots dans les questions pour les fichiers-tests (TEST1 et TEST2) et les questions d'Adminet. 165

7.3 Nombre de réponses selon la requête, pour chacun des moteurs. . . . 169

7.4 Le temps de réponse minimal. 170

7.5 Le temps de réponse maximal. 171

7.6 Le temps de réponse moyen. 172

7.7 Les temps de réponses : minimal, maximal, moyen. 173

7.8 Corrélation entre le temps de réponse et nombre de réponses calculée pour chacun des moteurs. 175

7.9 La corrélation thématique de serveurs d'administration. 178

7.10 Le graphe de dépendance pour les paires de serveurs avec la corrélation ≥ 0.5. 179

7.11 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.9 181

7.12 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.8 182

7.13 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.7 182

7.14 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.6 183

7.15 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.5 184

7.16 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.4 185

7.17 Corrélation entre serveurs regroupés selon la nouvelle numérotation. Coefficient de corrélation > 0.3 185

7.18 Histogramme des corrélations de paires de serveurs. 186

7.19 Corrélation pour le serveur d'EDF. 187

7.20 Corrélation pour le serveur de Ministère du la Justice. 188

7.21 Corrélation pour le serveur de l'Université de Caen. 189

7.22 Proportions de différents codes de retour (code HTTP). 190

7.23 Taille (en nombre de mots) de documents-réponses pour tous les documents. 192
7.24 Taille des documents-réponses. Compartiment [0 - 5000] mots. . . . 192
7.25 Taille des documents-réponses. Compartiment [5000 - 50 000] mots. . 193
7.26 Taille des documents-réponse. Compartiments [50 000 - 350 000] . . 193
7.27 Taille de documents-réponses pour le serveur de Sénat. 194
7.28 Taille de documents-réponses pour le serveur de Assemble Nationale. 195
7.29 Rendement maximal de liens valides. 196
7.30 Rendement minimal de liens valides. 197
7.31 Rendement moyen de liens valides. 198
7.32 Rendement de liens valides. 199
7.33 Existence de mots de la requête dans le document complet. 200
7.34 Existence de mots de la requête dans le corps *(body)* du document . 201
7.35 Existence de mots de la requête dans le *title* du document 202

8.1 La page de sites favoris de Vincent Ramette. 206
8.2 La page des références dans le domaine des sciences politiques du département de Droit de Université de Lille. 207
8.3 La page de CoCoon, le méta-moteur qui référence FIM-Métaindexeur pour la recherche dans le domaine du droit. 208

9.1 Fonctionnement de FIM Métaindexeur. 246
9.2 Taille de documents-réponses pour le serveur de l'Association Agropolis 249
9.3 Taille de documents-réponses pour le serveur de Cour de Cassation . 250
9.4 Taille de documents-réponses pour le serveur de l' Assemblée Nationale. 250
9.5 Taille de documents-réponses pour le serveur de Service-Public. . . . 251
9.6 Taille de documents-réponses pour le serveur de Ministère de la Fonction Publique 251
9.7 Taille de documents-réponses pour le serveur de Ministère de la Justice. 252

9.8 Taille de documents-réponses pour le serveur de Ministère des Affaires Étrangères. 252

9.9 Taille de documents-réponses pour le serveur de Bibliothèque Publique d'Information. 253

9.10 Taille de documents-réponses pour le serveur d'Adminet-JO. 253

9.11 Taille de documents-réponses pour le serveur de Ministère de la Recherche. 254

9.12 Taille de documents-réponses pour le serveur de Commissariat général du Plan. 254

9.13 Taille de documents-réponses pour le serveur de Secrétariat d'Etat au Commerce Extérieur. 255

9.14 Taille de documents-réponses pour le serveur de Université de Caen Basse-Normandie. 255

9.15 Taille de documents-réponses pour le serveur de Institut National de l'Audiovisuel. 256

9.16 Taille de documents-réponses pour le serveur de Ministère de l'Economie, des Finances et de l'Industrie. 256

9.17 Taille de documents-réponses pour le serveur de Ministère de l'Education Nationale. 257

9.18 Taille de documents-réponses pour le serveur de Sénat. 257

9.19 Taille de documents-réponses pour le serveur de Institut National de Recherche en Informatique et en Automatique. 258

9.20 Taille de documents-réponses pour le serveur de l'EDF. 258

9.21 Taille de documents-réponses pour le serveur de Institut National de la Recherche Agronomique. 259

9.22 Taille de documents-réponses pour le serveur de l'Institut Curie. . . . 259

9.23 Taille de documents-réponses pour le serveur de Laboratoire Central des Ponts et Chaussées. 260

9.24 Taille de documents-réponses pour le serveur de l'Université de Valenciennes. 260

Annexe A
Technologies utilisées

FIM-Métaindexeur

1. Informations générales

Type : méta-moteur distant [Section 3.1.3.1],
URL : <http ://veneux.ensmp.fr/metaindexeur/meta.cgi>,
Mise en service : 1998.

2. Environnement technique

Exécution : machines du CRI de l'École de Mines à Fontainebleau,
Système : UNIX Solaris 8 SPARC,
Accès Internet : liaison spécialisée à 512 kbits/s,
Serveur Web : Apache 1.3

3. Langage utilisé

PERL : (*Practical Extraction and Reporting Language* écrit par Larry Wall est distribué sous Licence Publique GNU [WCSP96, Chr96])

Le code de tous les modules de FIM-Métaindexeur (interface utilisateur, agent de récupération, procédure d'analyse, procédure d'évaluation, procédure de vérification, etc. [Figure 6.1, section 6.1]) décrits dans le chapitre 6 est écrit en Perl.

4. Fonctionnement

Le fonctionnement de FIM-Métaindexeur est illustré sur la figure 9.1.

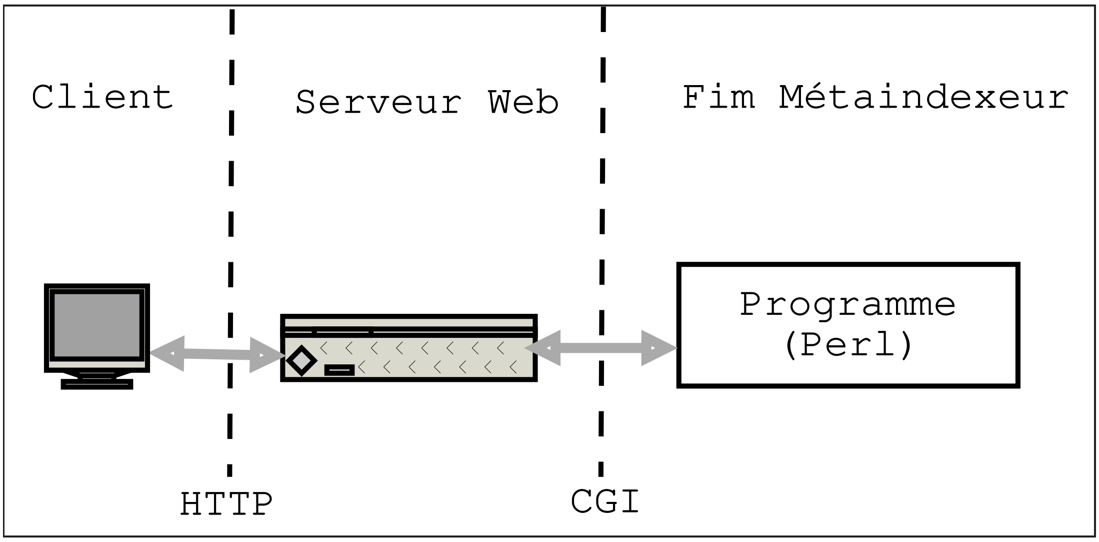


Fig. 9.1 – Fonctionnement de FIM Métaindexeur.

4.1 Interface utilisateur

Les pages de formulaire et de résultats sont créées dynamiquement sous forme de programmes CGI.

4.2 Paramètres de Fim-Metaindexeur

Toutes les paramètres concernant les moteurs interrogés [6.2.1] sont sauvegardés dans un fichier texte structuré.

4.3 Intégration avec le serveur Web

L'application est réalisée sous forme de programmes CGI.

CGI (*Common Gateway Interface*) définit les protocoles de communication entre un serveur HTTP et des programmes externes dans un environnement standard.

4.4 Communications externes (interrogation des moteurs)

La structure fonctionnelle de communication avec les moteurs interrogés est décrite au chapitre 5.

Protocole HTTP : Méthode GET ou POST selon le moteur [Section 5.6],

Agents de récupération : agents indépendants lancés en parallèle (les agents respectent le *Standard for Robot Exclusion* [Kos94a]),

Parallélisme des requêtes : 50 connections simultanées au maximum,

Temps de réponse : réglable par l'utilisateur dans les paramètres optionnels [Section 6.2.3.1], fortement dépendant du temps de réponse des moteurs interrogés,

Les statistiques

L'analyse des résultats décrits au chapitre 7 a été effectuée avec les outils suivants :

1. Base de données

Base de données : MySQL version 3.23,

Interface d'administration : KMySQLadmin version 0.4.4 (sous environnement KDE),

Interface utilisateur : kMySQL version 1.21 (sous environnement KDE)

2. Analyse des tests

Insertion des données dans la base : programme écrit en Perl,

Recherche et extraction des données : procédures utilisant le langage SQL,

Calculs : plusieurs programmes écrits en Perl et SQL

3. Présentation des résultats

Data Explorer de IBM, version 4.1.0 (*Open Visualisation Data Prompter*)

Annexe B : Les statistiques sur la taille des documents-réponses

Les résultats d'analyse et les statistiques réalisées sont présentés dans les figures 9.3 à 9.24.

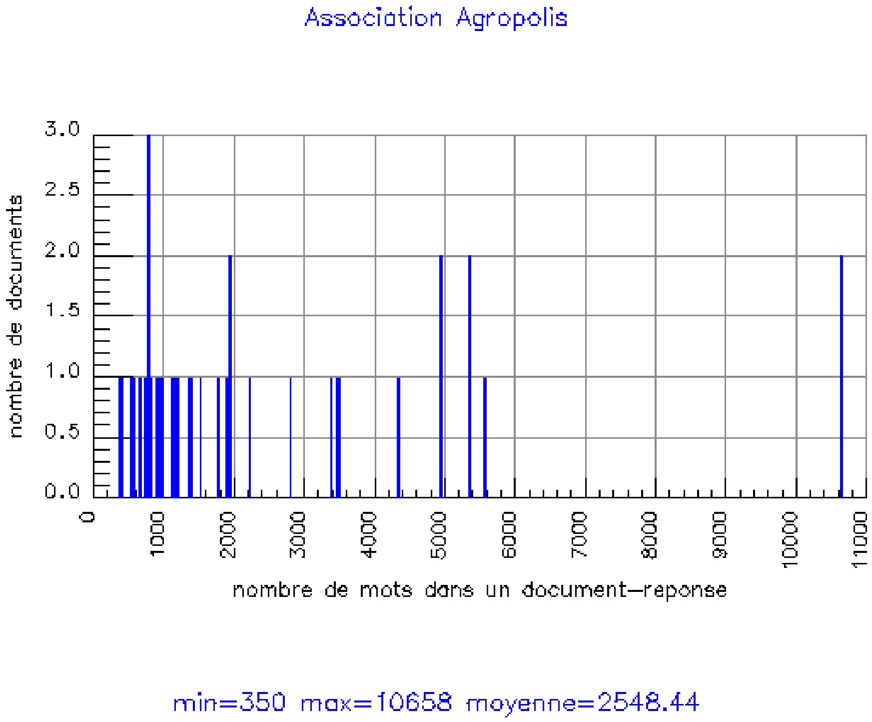


FIG. 9.2 – Taille de documents-réponses pour le serveur de l'Association Agropolis

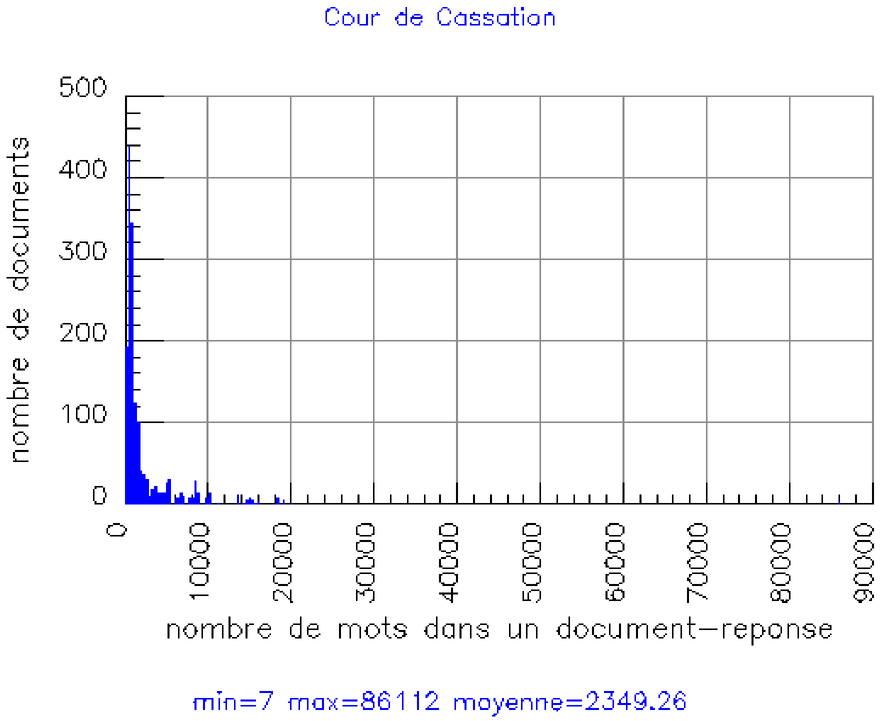


FIG. 9.3 – Taille de documents-réponses pour le serveur de Cour de Cassation

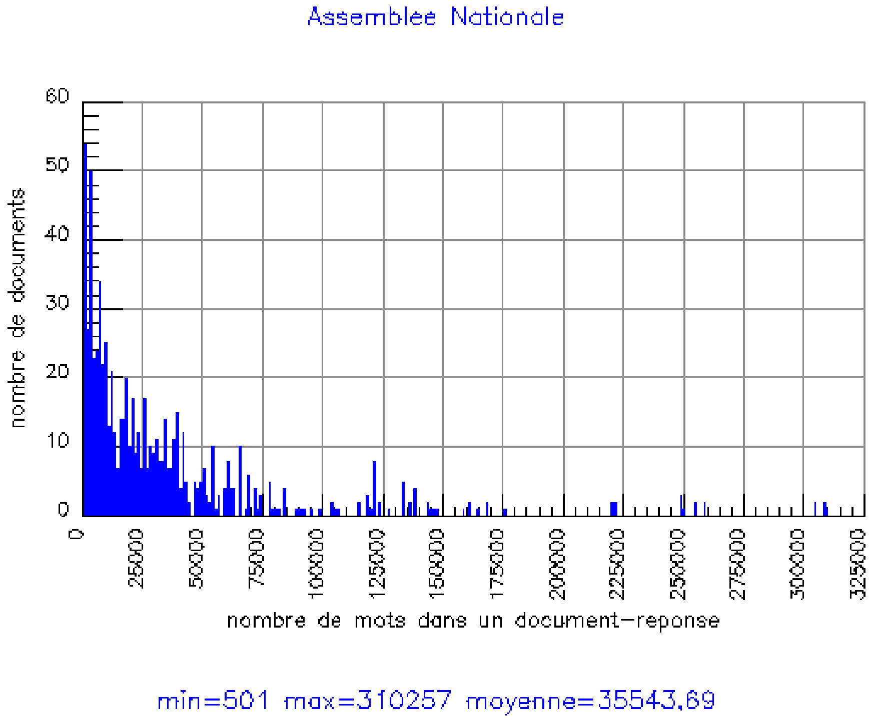


FIG. 9.4 – Taille de documents-réponses pour le serveur de l' Assemblée Nationale.

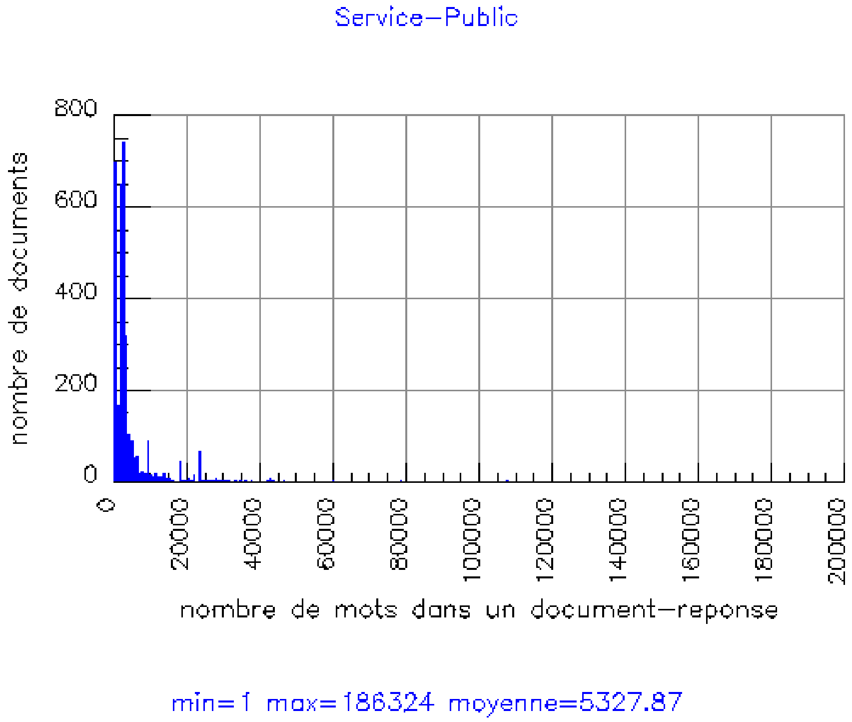


FIG. 9.5 – Taille de documents-réponses pour le serveur de Service-Public.

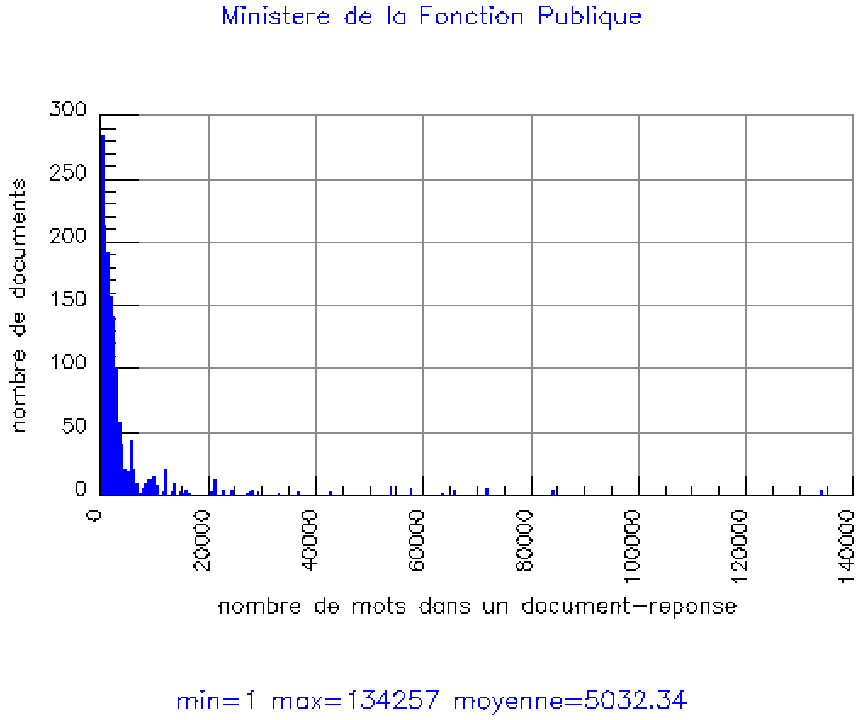


FIG. 9.6 – Taille de documents-réponses pour le serveur de Ministère de la Fonction Publique

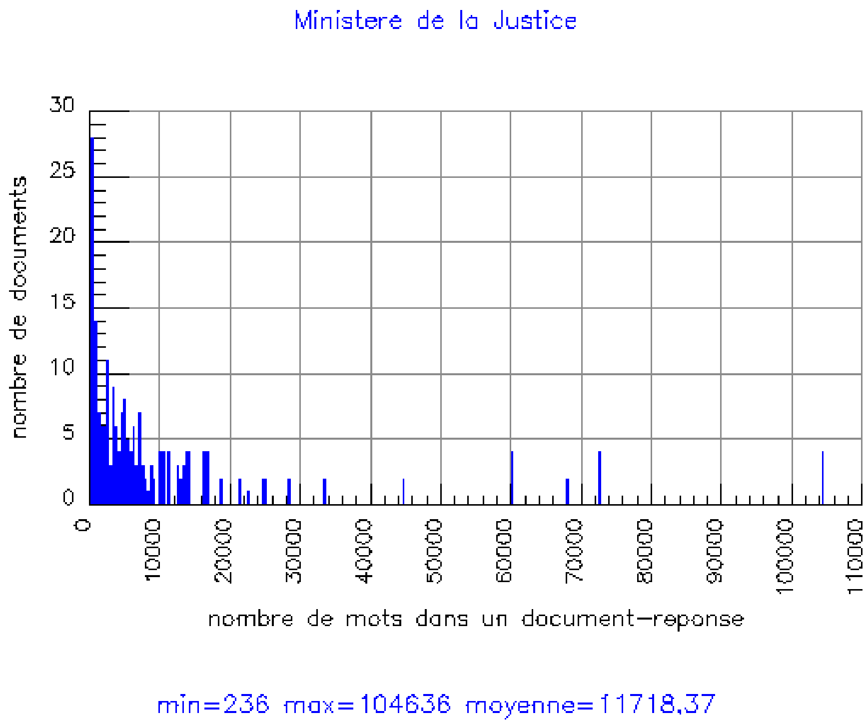


FIG. 9.7 – Taille de documents-réponses pour le serveur de Ministère de la Justice.

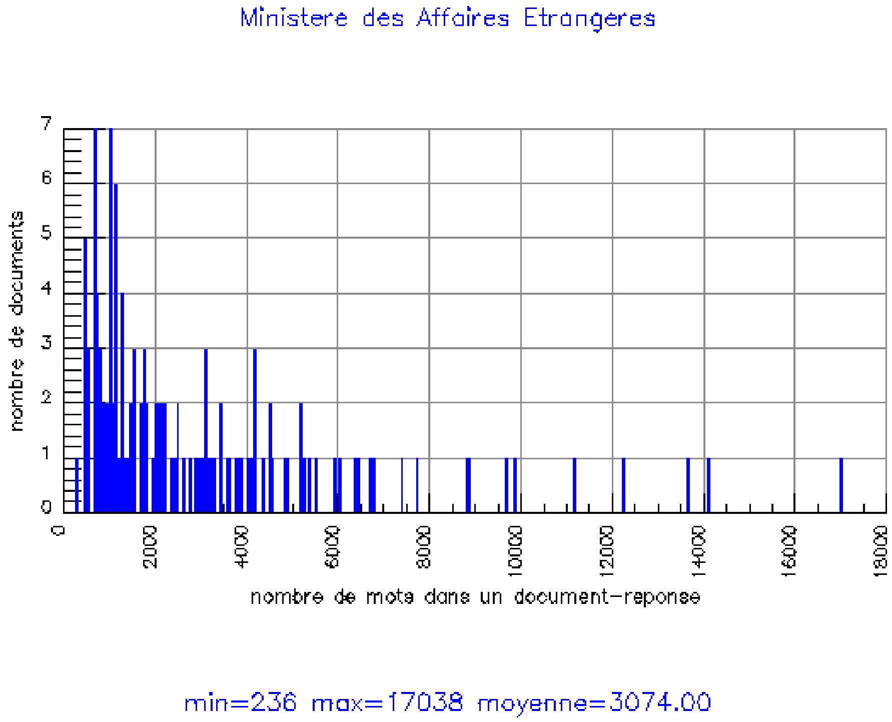


FIG. 9.8 – Taille de documents-réponses pour le serveur de Ministère des Affaires Étrangères.

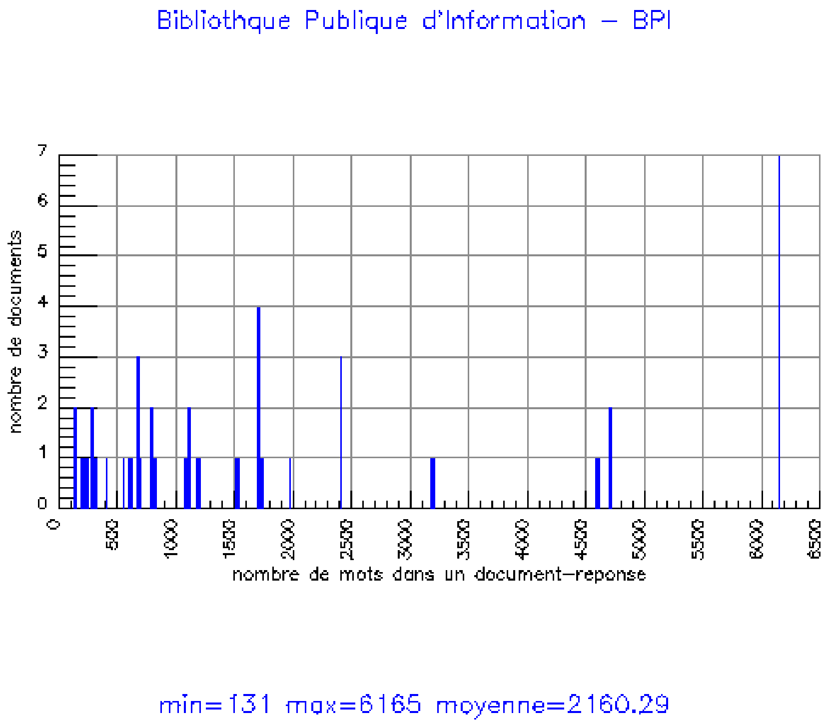


Fig. 9.9 – Taille de documents-réponses pour le serveur de Bibliothèque Publique d'Information.

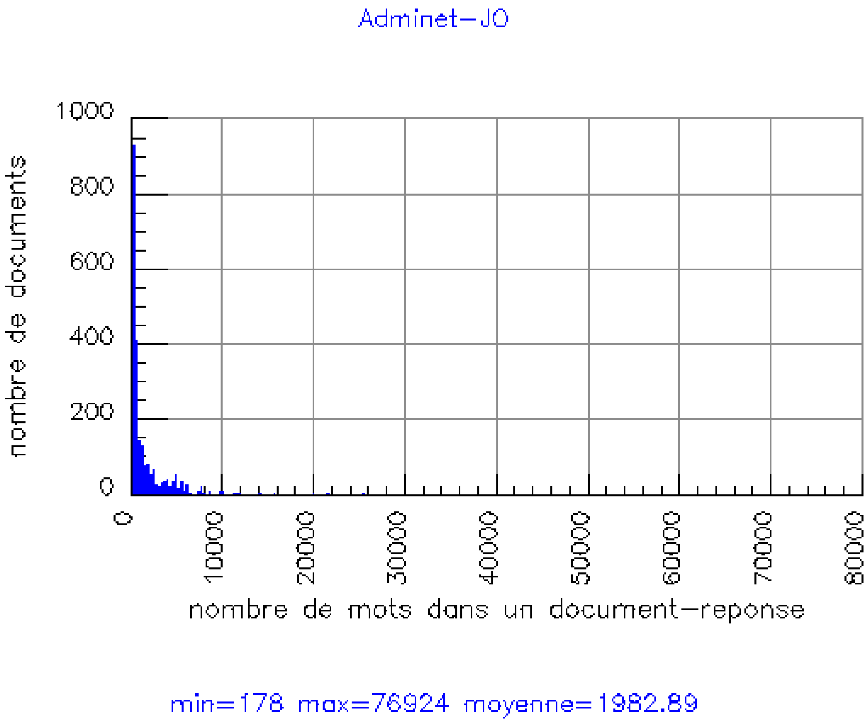


Fig. 9.10 – Taille de documents-réponses pour le serveur d'Adminet-JO.

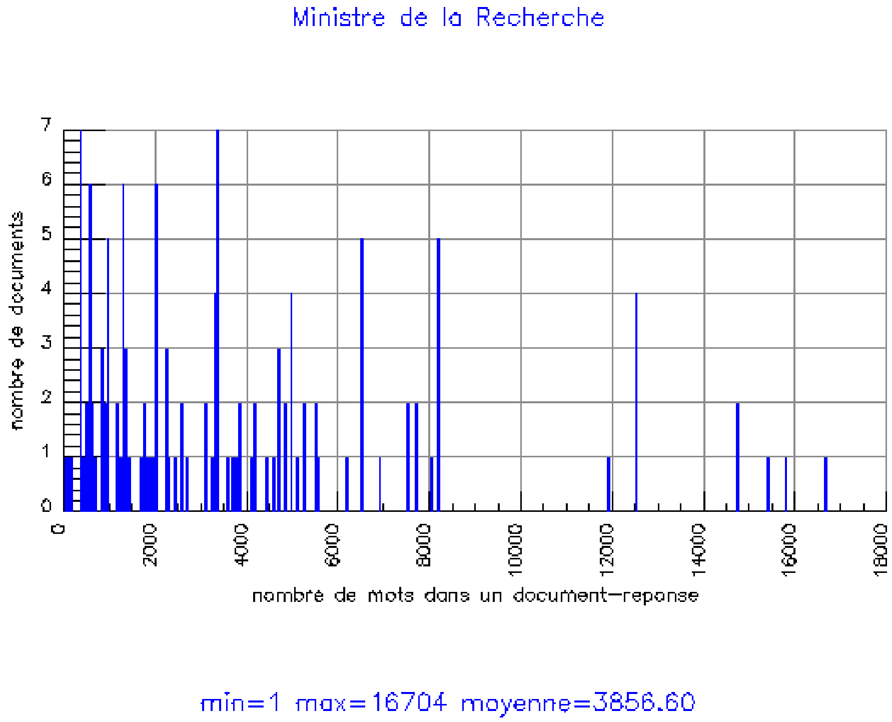


FIG. 9.11 – Taille de documents-réponses pour le serveur de Ministère de la Recherche.

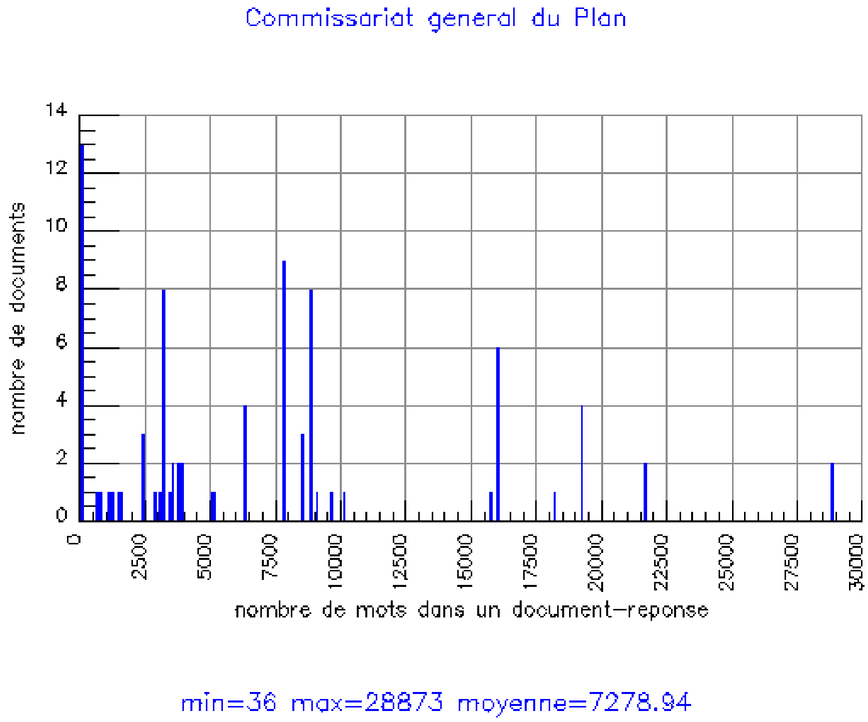


FIG. 9.12 – Taille de documents-réponses pour le serveur de Commissariat général du Plan.

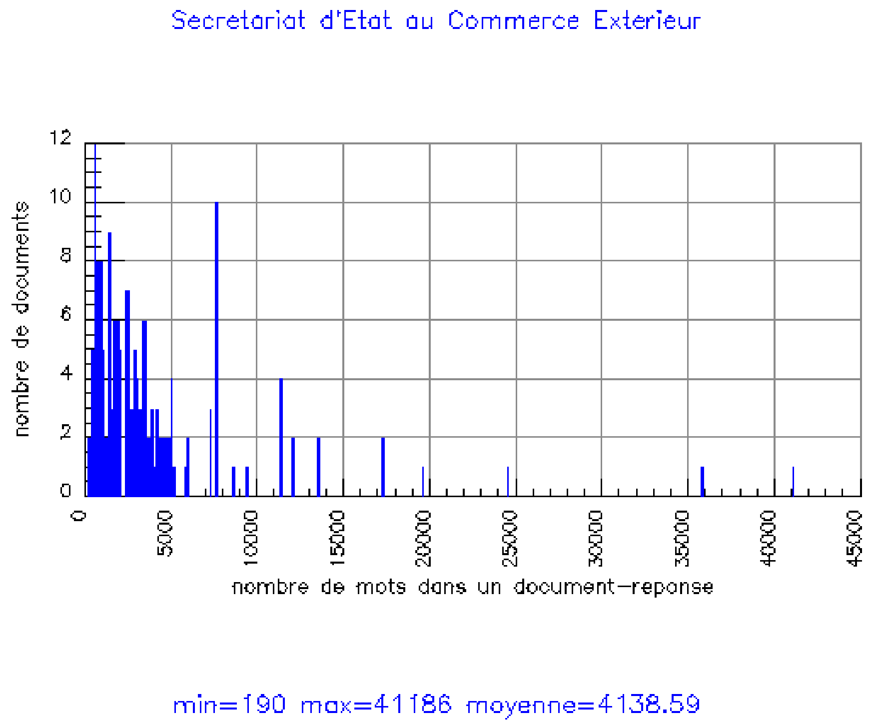


FIG. 9.13 – Taille de documents-réponses pour le serveur de Secrétariat d'Etat au Commerce Extérieur.

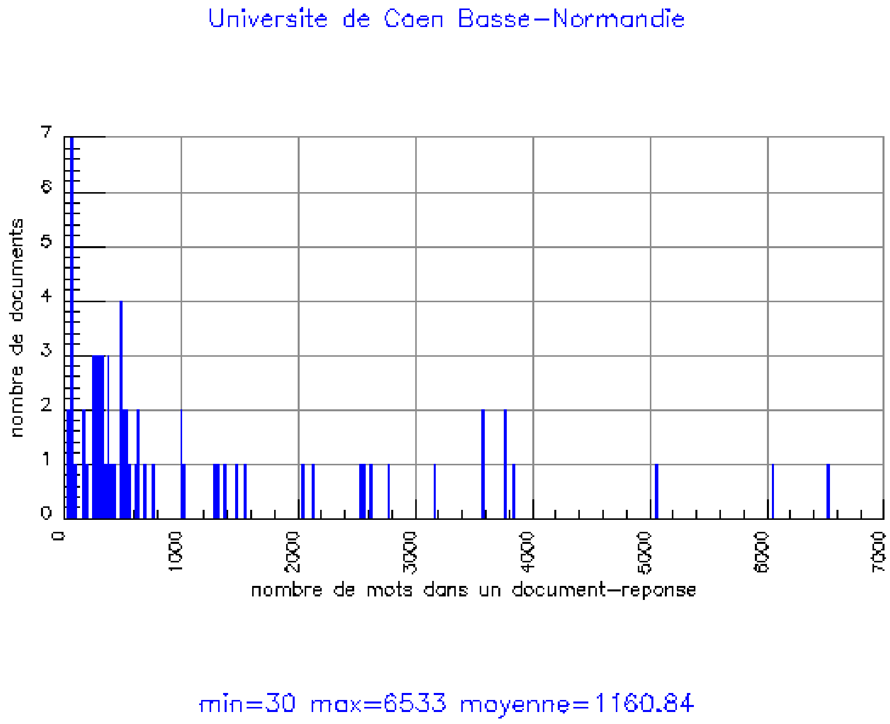


FIG. 9.14 – Taille de documents-réponses pour le serveur de Université de Caen Basse-Normandie.

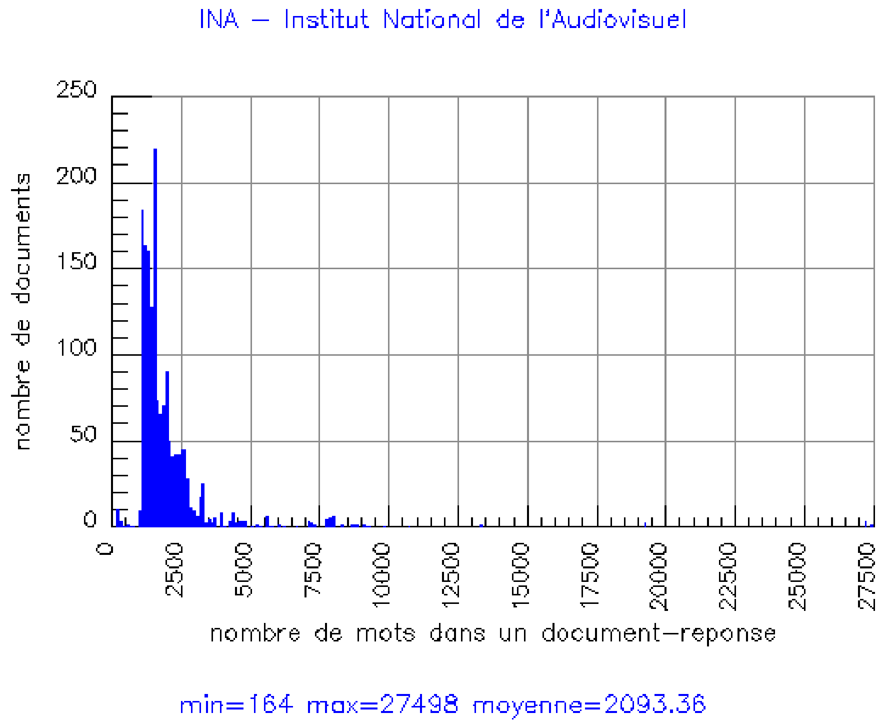


FIG. 9.15 – Taille de documents-réponses pour le serveur de Institut National de l'Audiovisuel.

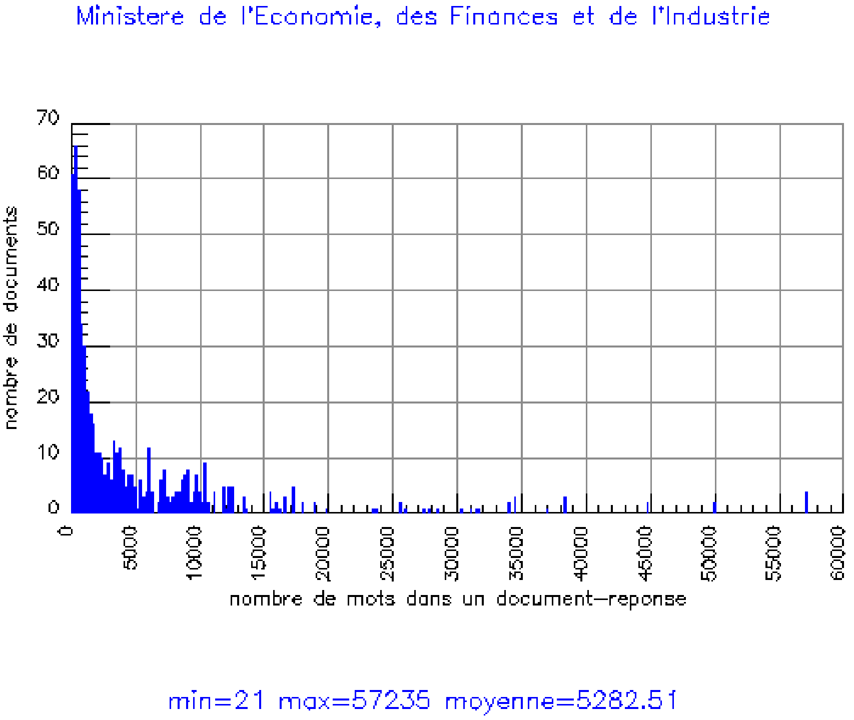


FIG. 9.16 – Taille de documents-réponses pour le serveur de Ministère de l'Economie, des Finances et de l'Industrie.

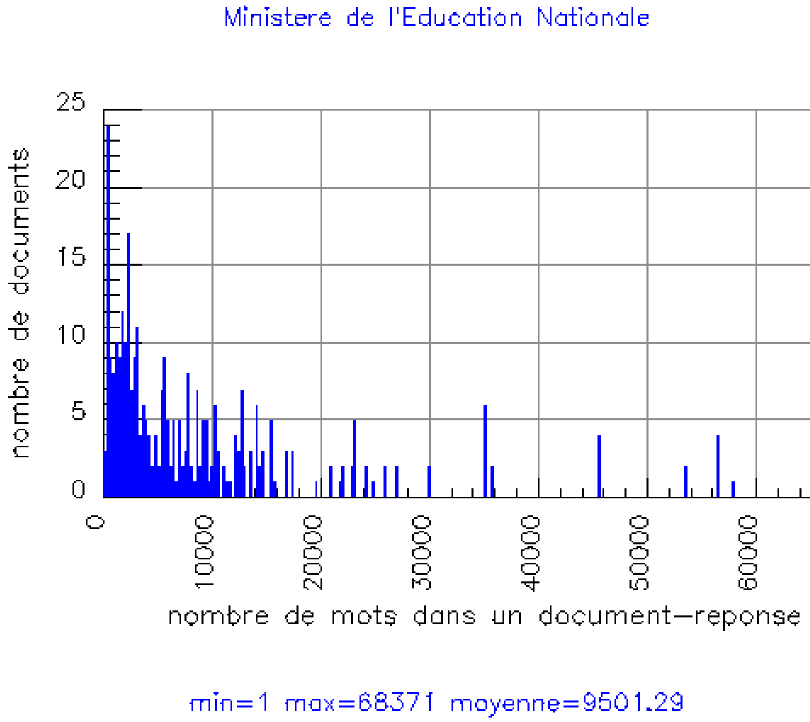


FIG. 9.17 – Taille de documents-réponses pour le serveur de Ministère de l'Education Nationale.

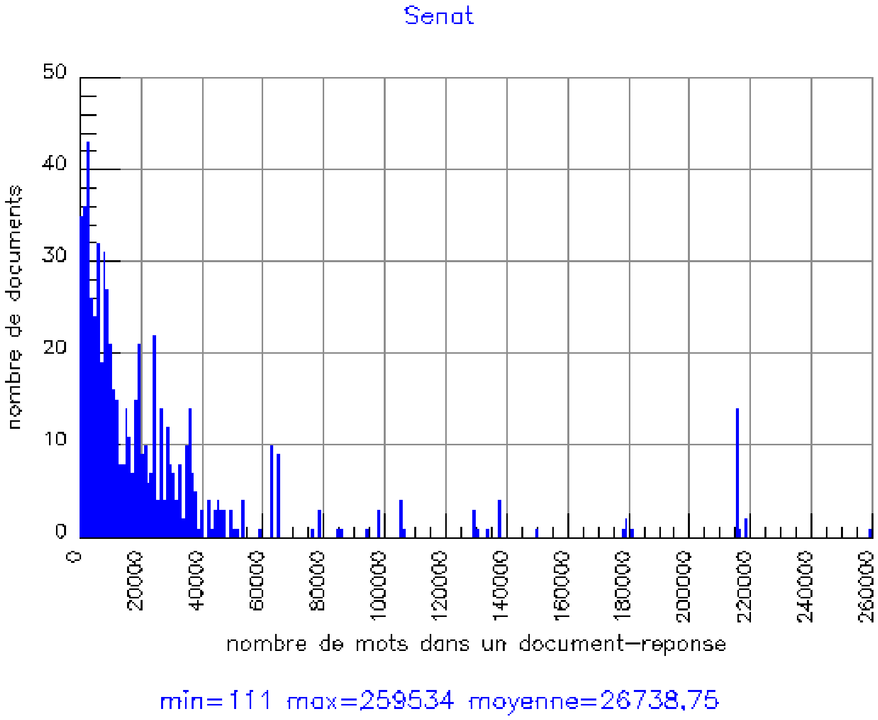


FIG. 9.18 – Taille de documents-réponses pour le serveur de Sénat.

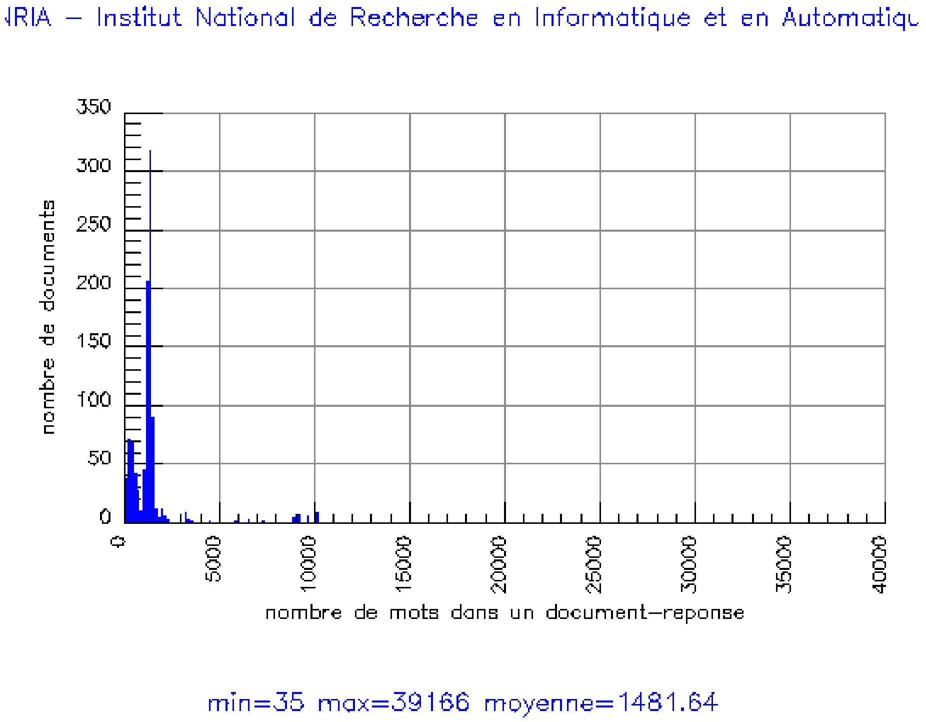


FIG. 9.19 – Taille de documents-réponses pour le serveur de Institut National de Recherche en Informatique et en Automatique.

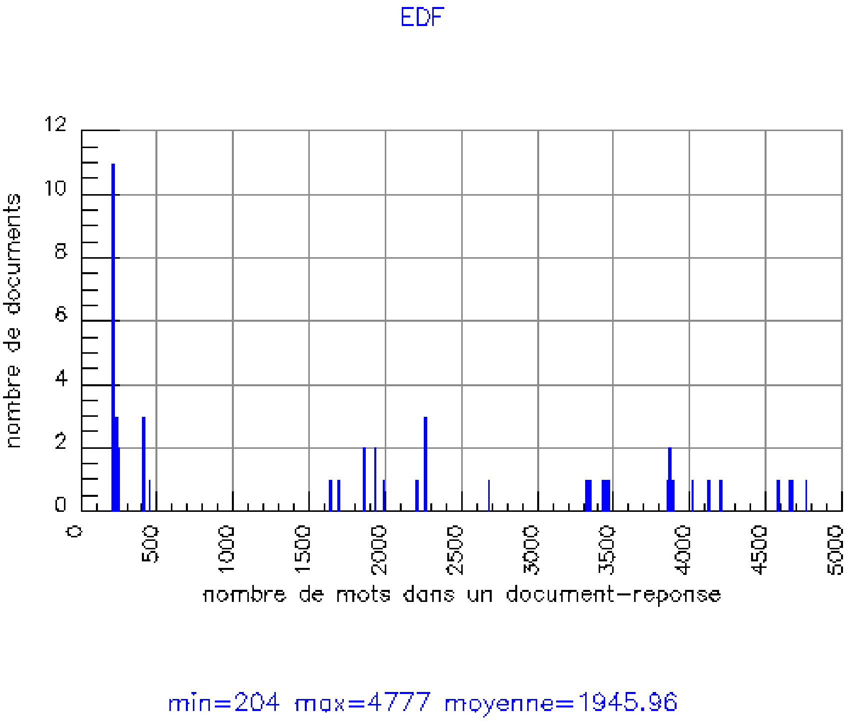


FIG. 9.20 – Taille de documents-réponses pour le serveur de l'EDF.

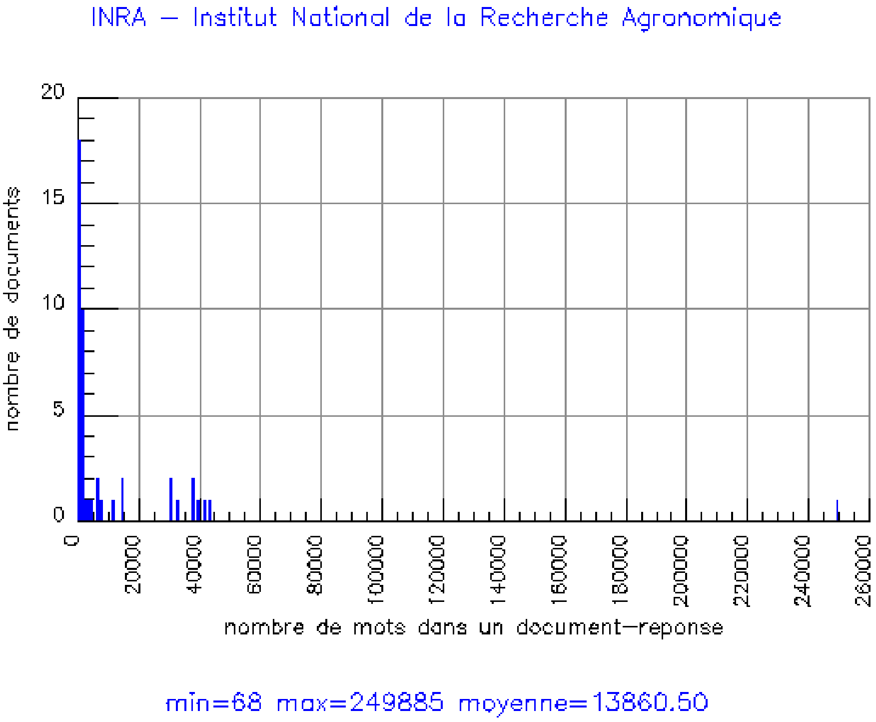


FIG. 9.21 – Taille de documents-réponses pour le serveur de Institut National de la Recherche Agronomique.

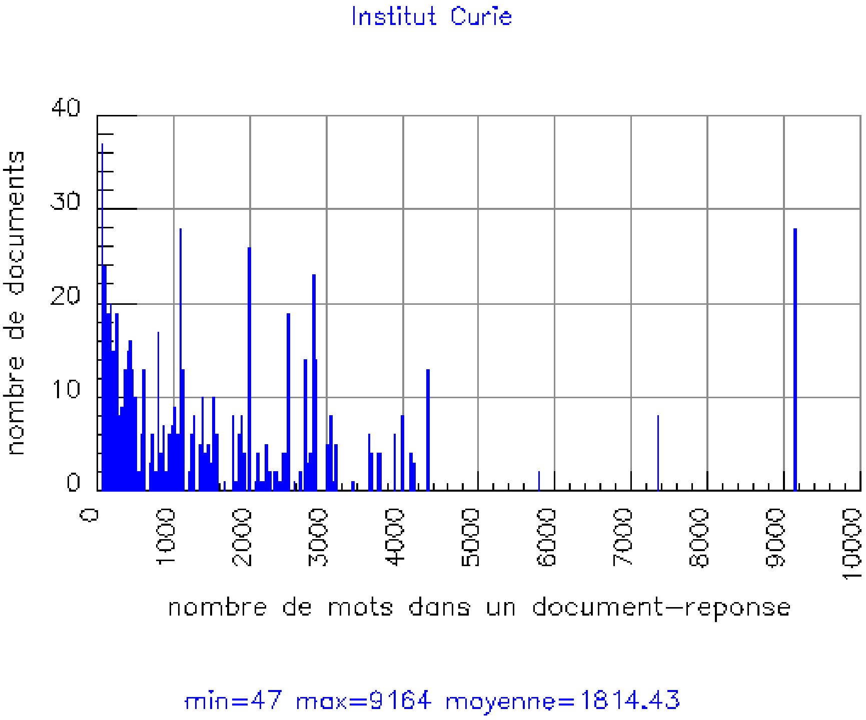


FIG. 9.22 – Taille de documents-réponses pour le serveur de l'Institut Curie.

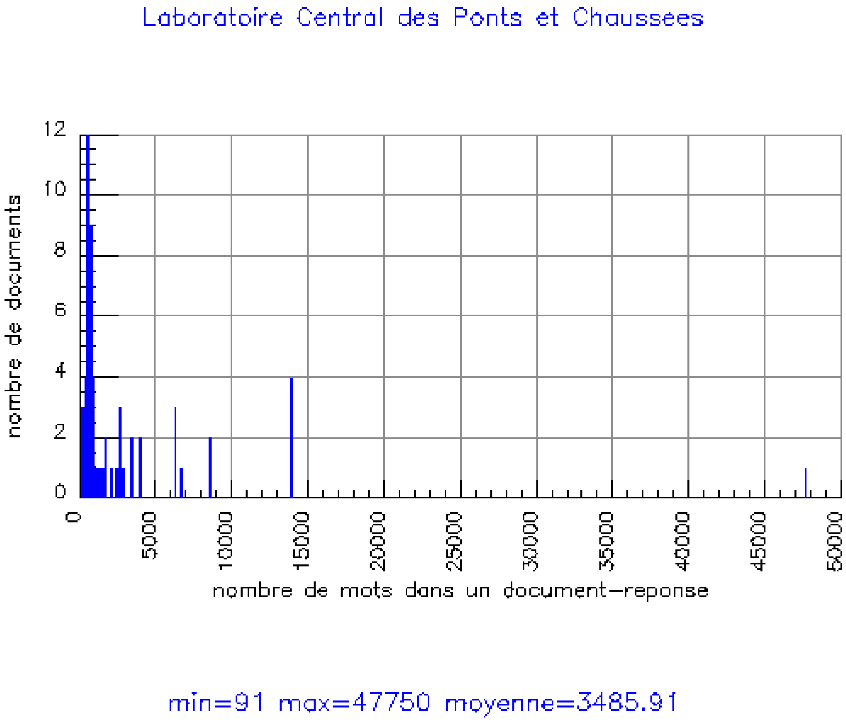


FIG. 9.23 – Taille de documents-réponses pour le serveur de Laboratoire Central des Ponts et Chaussées.

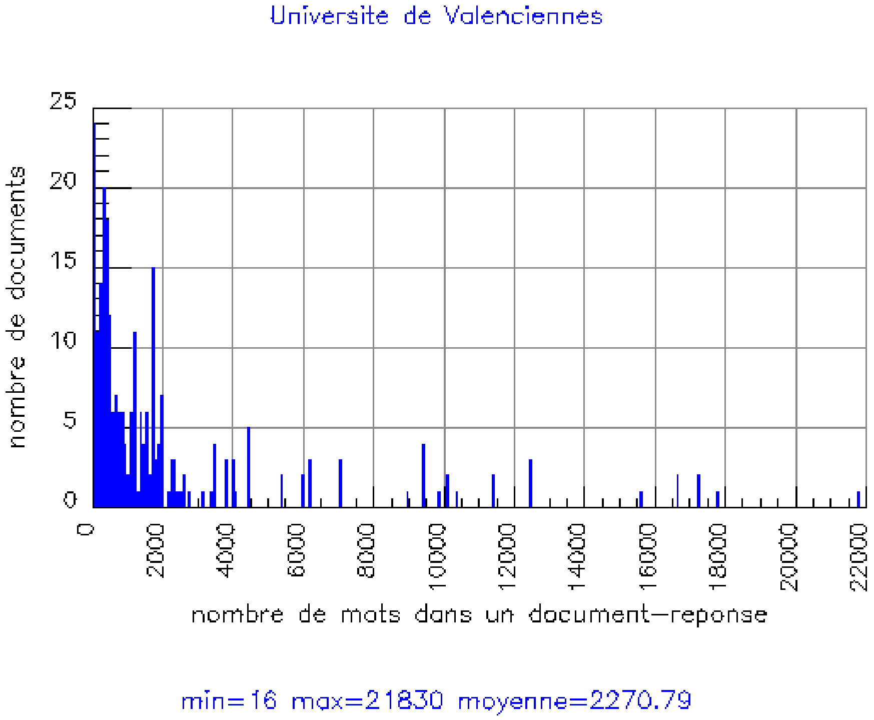


FIG. 9.24 – Taille de documents-réponses pour le serveur de l'Université de Valenciennes.