وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR –ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR –ANNABA

جامعة باجي مختار – عنابة –
كلية العلوم و الهندسة

Faculté des Sciences de l'Ingénieur **Année : 2005**

Département d'Informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme de **MAGISTER**

Une Indexation A Base d'Ontologies Pour Le Filtrage d'Informations Sur Le Web

Option

Intelligence Artificielle Distribuée (IAD)

Par

Mr. ZIANI Radouane

DIRECTEUR DE MEMOIRE : Mohamed Tayeb LASKRI Professeur Université Annaba

DEVANT LE JURY

PRESIDENT :	Pr. Mokhtar SELLAMI	Professeur	Université de Annaba
RAPPORTEUR :	Pr. Med Tayeb LASKRI	Professeur	Université de Annaba
EXAMINATEUR :	Dr. Tahar BENSEBAA	Maître de conférences	Université de Annaba
EXAMINATEUR :	Dr. Khiereddine KHOLLADI	Maître de conférences	Université de Constantine

Remerciements

C'est avec un immense plaisir que j'exprime ma profonde gratitude aux personnes qui ont contribué, directement ou indirectement, à la réussite de ce travail : je n'aurais jamais pu achever ce travail sans le soutien dont j'ai bénéficié.

Un très grand Merci à Mr LASKRI Mohamed Tayeb Professeur des universités et Recteur de l'université Badji Mokhtar–Annaba- pour m'avoir fait l'honneur d'encadrer et diriger ce travail, pour sa patience, malgré mon caractère particulier, et son encouragement prodigué.

Un très grand Merci aussi à mon enseignant Mr SELLAMI Mokhtar professeur à l'université de Annaba, chef de LRI et président du comité scientifique du département d'informatique de m'avoir fait l'honneur de présider le jury de mon mémoire. Je remercier également mon enseignant Mr BENSEBAA Tahar Maître de conférences à l'université de Annaba et Mr KHOLLADI Khiereddine Maître de conférences à l'université de Constantine pour avoir accepté de faire partie du jury autant qu'examinateurs. Je remercie, une autre fois, Mr LASKRI d'avoir rapporté sur mon travail.

Je remercie infiniment ma famille, qui compte beaucoup pour moi : ma mère, mes parents, mes sœurs & frères.

Je remercie également tous mes amis : Tayeb, Hamid, Salih, Reda, Salah, Sofiane, Rochdi, Karim, Billel, Sabri, Toufik ….

Je terminerai ces remerciements par mes collègues : Djahid, Wahid, Hafidi, Mourad, Mohamed, Douadi … Tous les membres du Groupe de Recherche en Intelligence Artificielle (LRI/GRIA) et tous les enseignants de département d'informatique.

Résumé

Le but du filtrage est de recueillir un maximum de documents sémantiquement homogènes, répondant à une problématique claire du point de vue de la compréhension humaine. Dans ce travail, nous proposons un prototype du moteur de recherche intégrant un système de filtrage d'informations. Il restitue à l'usager un véritable pouvoir d'interprétation de son besoin d'informations et des réponses qui lui sont données en lui offrant une boite à outils pour organiser les informations (filtrer les informations puis les proposer aux usagers).

Comme tout système dont le but est la recherche d'information, notre système vise à minimiser à la fois le "bruit" et le "silence" en fonction des requêtes utilisateurs. Pour atteindre ces objectifs, nous avons utilisé un concept issu du domaine d'intelligence artificielle à savoir les ontologies comme formalisme de représentation des connaissances. Les ontologies ont un rôle central dans le filtrage des documents web.

Mots Clés : Filtrage d'informations, Indexation, Ontologie, Catégorisation des documents, Recherche d'informations, Web, Moteur de recherche, Modèle de requête.

Abstract

The aim of the filtering is to collect as many semantically homogeneous documents as possible, thus answering a clear problem statement as for as human comprehension is concerned. In this paper, we propose a prototype of search engine which integrates an information screening system. It restitutes to the user a real interpreting power for his information needs as well as answers supplied by an information organizing tool box (to screen information then to offer it to the user).

Like any system which goal is to seek information, this system aims to minimizing both "noise" and "silence" in relation to users requests. To reach such objectives, we have used a concept derived from artificial intelligence namely ontology like knowledge representation formalism. Ontology plays a central role in web document filtering.

Key Words: Information Filtering, Indexing, Ontology, Document Categorization, Information Retrieval, Wb, Search Engine, Request Model.

ملخص

الهدف من التصفية هو الحصول على أكبر عدد ممكن من الملفات المتجانسة، الرادة على مشكلة واضحة من حيث المفهوم الانساني. في هذه المذكرة، نقترح نموذج لمحرك بحث و الحاوي على نظام تصفية المعلومات. هذا المحرك يشكل بالنسبة للمستعمل سلطته المكينة لترجمة حاجته للمعلومات و الأجوبة المعطاة له على شكل لوحة وسائل من أجل تنظيم المعولمات (تصفية المعلومات ثم عرضها على المستعملين).

كأي نظام هذفه البحث عن المعلومات، نظامنا يهذف الى الحد من الاجابات الزائدة (الفوضى) و عدم العثور على أية اجابة ناجعة بدلالة مطالب المستعملين. من أجل التوصل الى هذه الغاية استعملنا مفهوما منبثقا من مجال الذكاء الاصطناعي ألا وهو الأنطولوجيات كوسيلة لتمثيل المعارف. الأنطولوجيات تلعب دورا مهما في تصفية ملفات الويب.

الكلمات المفتاحية : تصفية المعلومات، فهرسة، أنطولوجية، تقسيم الملفات، البحث عن المعلومات، ويب، محرك البحث، نموذج السؤال.

Liste des figures

Partie 1

Partie 2

Table des matières

Deuxième partie

Le Système de Filtrage Proposé

Chapitre 1 : Les Ontologies

Chapitre 2 : Le Modèle de Filtrage

Introduction

La quantité d'information disponible sur le Web est importante et elle ne cesse de croître. La recherche d'information demeure problématique : il est en effet difficile de trouver ce que l'on recherche malgré l'existence de sites et de moteurs de recherche. Les moteurs de recherche par mots clés renvoient généralement comme réponse à une requête un grand nombre de pages à consulter, ce qui demande à l'utilisateur de faire lui-même le tri et le filtrage dans cette masse d'information. Les résultats ne sont pas tout pertinents et l'information retrouvée n'est pas complète.

La recherche pleine texte n'est pas toujours efficace : La page recherchée peut utiliser un terme sémantiquement proche mais syntaxiquement différent de celui de la requête; les fautes de frappe et les variantes lexicales sont généralement considérées comme étant des termes différents.

Dans les annuaires comme Yahoo les ressources sont indexées et classées manuellement en fonction de catégories qui sont d'ordre trop général pour répondre à des requêtes spécifiques : en effet il y a souvent un recoupement entre les différentes catégories et une certaine ambiguïté quant à leur étendue. Ceci peut dérouter l'utilisateur qui ne sait pas trop dans quelle thématique rechercher son information.

Aujourd'hui, la problématique qui se pose est celle d'une **recherche d'information intelligente** sur le Web. Le Web Sémantique [TIM99] est un espace d'échange qui reste à construire. Un de ses intérêts est d'une part d'apporter suffisamment de renseignements sur les ressources, en ajoutant des annotations sous la forme de *méta-données* et d'autre part, de décrire leur contenu de manière à la fois formelle et signifiante à l'aide d'une *ontologie* pour être interprétables aussi bien par les humains que par les machines.

Nous faisons l'hypothèse que les performances des systèmes de recherche d'informations peuvent être améliorées si on considère l'information circulant sur le Web, non plus comme de simples données numériques (à la manière des bases de données traditionnelles), mais du point de vue de sa sémantique. Il en résulte la nécessité de recourir aux différentes techniques de représentation des connaissances.

Nous avons conçu le prototype d'un système de recherche d'informations qui combine à la fois les diverses fonctionnalités propres au domaine de la recherche d'informations et des fonctionnalités faisant appel aux technologies issues des travaux en ingénieries des connaissances (les ontologies).

C’est dans cette optique que s’inscrit notre mémoire. Dans une partie d’état de l’art, nous nous proposons, dans un premier temps, de passer en revue les notions générales relatives au processus de repérage de l’information, puis, en second lieu, d’introduire les caractéristiques plus spécifiques de la recherche d’information sur Internet, notamment en résumant la typologie des outils actuellement disponibles dédiés à cette fin. Suite à cela nous présentons les déférentes méthodologies d’indexation et en fin mettant l’accent sur quelques travaux sur le filtrage de l’information. Suite à cette mise en contexte théorique, dans la deuxième partie, nous présenterons des définitions et des notions de bases relatives aux ontologies puis nous présenterons et discuterons notre prototype de moteur de recherche qui encapsule un mécanisme de filtrage basé sur l’utilisation des ontologies. Des suggestions d’investigations futures seront également formulées.

Problématique & Objectif

La qualité d'une recherche d'information peut s'exprimer en termes de silence et de bruit. Le bruit correspond aux résultats non pertinents trouvés par l'utilisateur tandis que le silence correspond aux résultats pertinents non trouvés. Les utilisateurs des différents outils de recherche disponibles sur Internet se trouvent confrontés au quotidien à ces deux paramètres. Qui ne s'est jamais retrouvé face à un nombre faramineux de réponses (évoquant un bruit important) ou un trop faible nombre de réponses à la suite de sa requête (signe d'un silence important)? Cette surabondance ou l'absence de réponses dépend principalement de deux facteurs :

- La qualité de la requête formulée par l'utilisateur. En effet, l'utilisation d'un terme non approprié dans une requête peut aboutir à la présence de bruit si le terme utilisé est trop commun, ou à du silence si le concept recherché peut être décrit par plusieurs synonymes et que celui utilisé dans la requête un synonyme peu usité.

- La qualité de l'index interrogé (nombre de documents, qualité intrinsèque et spécificité des documents présents dans cet index). Si l'on suppose que la requête est parfaitement formulée, le nombre de réponses dépend du nombre de documents présents dans l'index de l'outil de recherche interrogé, et donc directement de sa nature (moteur, méta-moteur ou annuaire).

C'est dans ce contexte que s'inscrit l'objectif de notre travail, il s'agit de solutionner le problème de minimisation de bruit évoqué par les moteurs de recherche.

Partie 1

Etat de l'art

Chapitre 1

Les systèmes de repérage de l'information

1. Introduction

Ce chapitre constitue un état de l'art sur les principaux systèmes et modes de recherches d'informations en générale, que ce soit dans les moteurs de recherche dans des bases documentaires ou bien dans les moteurs de recherche sur le web. Nous expliquons, à travers ce chapitre, les principales techniques employées pour le repérage de l'information dans les bases documentaires et dans les bases universelles telle que les moteurs de recherche, ainsi que le principe de fonctionnement global des SRI et les critères d'évaluation des SRI.

2. Les principaux types de systèmes de repérage de l'information (SRI)

2.1. Les systèmes booléens ou traditionnels

Comme leur nom l'indique, ces systèmes se basent sur la logique développée par le mathématicien britannique George Boole [DEL99]. Ils utilisent des opérateurs pour combiner des termes de recherche entre eux, comme s'il s'agissait d'énoncés mathématiques.

Ces systèmes appréhendent un texte comme une suite aléatoire de mots délimités entre eux par des signes de ponctuation, des espaces typographiques ou d'autres caractères tels $%&-/#_~. Ils apparient requêtes et documents via le principe de concordance de modèle (*pattern matching*), et plus particulièrement la recherche de concordances exactes (*exact matches*). Lorsque la requête de l'usager est confrontée au contenu de la base de données, les entrées qui apparaissent sur la liste de résultats sont celles qui contiennent la ou les chaîne(s) recherchée(s), soit dans le texte même du document, soit dans d'autres champs de l'enregistrement (par exemple, les balises META d'un fichier HTML(1)). Les résultats ne font l'objet d'aucun tri [DEL99].

Les principaux opérateurs utilisés sont les suivants :

2.1.3. Les opérateurs dits booléens

- **L'opérateur ET**

Il permet de rendre la présence de mots obligatoire. Il est également symbolisé par son équivalent anglais AND ou par l'espace lorsqu'il est pris par défaut.

Exemple : *commerce ET électronique* repérera tous les documents où ces deux mots figurent.

(1) Les balises META, comme leur nom le suggère, sont des «informations sur l'information» : elles fournissent aux outils de recherche des renseignements spécifiques, par exemple un résumé ou une suite de mots clés relatifs au contenu d'une page Web. Ces codes appartiennent au langage HTML et ne sont pas visibles pour l'utilisateur. Ils s'inspirent du travail effectué pour les documents en sciences humaines dans le cadre de la TEI (*Text* Encoding Initiative), qui visait à spécifier des descripteurs de contenu à l'usage des auteurs et des éditeurs pour différents types de documents.

- **L'opérateur OU**

Il permet de rendre la présence de mots optionnelle. Il est également symbolisé par son équivalent anglais OR ou par l'espace lorsqu'il est pris par défaut.

Exemple : *commerce OU électronique* repérera tous les documents qui comprennent au minimum un de ces deux mots.

- **L'opérateur SAUF**

Il permet d'exclure la présence de mots. Il est également symbolisé par ses équivalents anglais NOT, BUT NOT ou AND NOT.

Exemple : *commerce SAUF électronique* repérera tous les documents où figure le mot *commerce* mais sans qu'y apparaisse le terme *électronique*.

- **Les parenthèses ( )**

Elles permettent de limiter la portée des opérateurs booléens et/ou d'introduire un ordre de priorité entre les différentes parties d'une requête.

Exemple : *(commerce OU paiement) ET électronique* repérera les documents qui contiennent à la fois *électronique* et soit *commerce* soit *paiement* soit ces deux termes.

- **La troncature**

Elle consiste à recourir à l'emploi de masques (*jokers* ou *wild cards*). Généralement symbolisée par * ou ? ou $, la troncature permet d'effectuer des recherches sur des parties de mots. Notons qu'elle est moins flexible dans le contexte de la recherche d'information sur le Web qu'en ce qui a trait aux logiciels documentaires traditionnels (impossibilité de l'appliquer en début de mot[5], nécessité fréquente de saisir un nombre minimum de caractères, etc.). Elle est toutefois intéressante en ce qu'elle permet de faire des recherches sur des mots de même famille et sur les variations de genre et de nombre.

Exemples : *biblio** repérera *bibliothèque*, *bibliothèques*, *bibliothécaire*, *bibliophile*, etc. La troncature peut aussi s'utiliser à l'intérieur d'un mot, pour remplacer un ou plusieurs caractère(s) : *coll$sion* repérera *collision* et *collusion*.

- **La recherche de locutions**

Elle fonctionne habituellement à l'aide des guillemets " " et permet la recherche exacte d'une séquence ordonnée de mots adjacents.

Exemple : *"commerce électronique"* repérera tous les documents où ces deux mots figurent l'un à côté de l'autre et dans cet ordre.

2.1.4. L'opérateur de proximité

La recherche sur la proximité est considérée comme une extension du modèle booléen. L'opérateur de proximité permet de rechercher des entrées où les mots désirés apparaissent à l'intérieur d'une «fenêtre» de voisinage dont l'ampleur varie selon les outils (généralement entre 10 et 100 mots, parfois beaucoup plus). Les formulations les plus habituelles sont anglophones : NEAR ou FOLLOWED BY (dans ce dernier cas, on tient également compte de la linéarité, c'est-à-dire de l'ordre d'apparition des termes). Pour rechercher des termes côte à côte (un peu comme une recherche de locution, mais sans souci de linéarité), on emploie parfois également un opérateur de proximité spécifique, dit *opérateur d'adjacence*. Il est généralement symbolisé par ADJ.

Exemples : *commerce NEAR électronique* repérera les entrées où ces deux termes figurent près l'un de l'autre. *Commerce FOLLOWED BY électronique* exigera, de plus, que l'ordre de saisie des mots soit respecté. *Commerce ADJ électronique*, pour sa part, recherchera les entrées où ces deux termes apparaissent immédiatement l'un à côté de l'autre, peu importe l'ordre d'apparition.

2.2. Les systèmes statistiques ou probabilistes

Les systèmes statistiques ou probabilistes sont une application des recherches menées aux Etats-Unis par G. Salton à partir du milieu des années 1960. Ils vont au-delà des approches booléennes par mots clés, dont ils tentent d'améliorer les performances. Leur but est de permettre le repérage des documents qui s'avèrent similaires à un ensemble de mots. Grâce à des technologies algorithmiques qui exploitent probabilités et statistiques inférentielles, ils repèrent et trient les réponses selon leur degré de correspondance avec la requête de l'usager, c'est-à-dire selon leur chance d'être jugées pertinentes par ce dernier. Ce type de recherche fournit donc non seulement les concordances exactes (*exact matches*) d'une requête, mais aussi celles qui s'en rapprochent (*close matches*). La plupart des outils de recherche sur Internet, soit relèvent de cette catégorie, soit sont des systèmes booléens augmentés de ce type de capacités statistiques, en particulier de fonctions d'évaluation de pertinence (*relevancy ranking*).

De manière très schématique, les systèmes statistiques basent leur fonctionnement sur le dénombrement des occurrences totales de chaque terme (sauf, éventuellement, les mots vides) dans un document, de même que dans l'ensemble de la base de données de l'outil. Toutefois, ceci ne veut pas dire nécessairement que les outils statistiques se bornent à compter les mots de la requête présents dans la base de données et que le document avec le plus d'occurrences

«gagne», car une tactique aussi simpliste tendrait à favoriser exagérément les documents de taille importante.

Un mécanisme supplémentaire d'assignation de poids différenciés aux divers mots existe donc généralement, la formule la plus fréquente consistant à affecter à ces derniers un poids inversement proportionnel à leur fréquence totale d'apparition dans la base de données : un mot relativement «rare» est ainsi doté d'un poids plus considérable qu'un mot très commun. Le principe sous-jacent est que le contenu informationnel d'un terme est inversement proportionnel à sa fréquence d'apparition : autrement dit, plus un mot figure souvent dans un texte ou un ensemble de textes, moins il est discriminant et véhicule en soi d'information.

D'autres facteurs peuvent être considérés dans le procédé de pondération des résultats, par exemple la *densité*, qui tient compte de la fréquence d'apparition d'un mot dans un document et de la taille de ce dernier. Une méthode reliée consiste à appliquer une courbe de pondération déclinante où la première occurrence d'un mot dans un document reçoit plus de poids que la seconde, elle-même supérieure à la troisième, etc. En ce qui concerne l'évaluation des documents, les critères suivants sont également susceptibles d'être utilisés :

- La proximité des mots clés entre eux ;
- L'emplacement des mots clés dans le document.

Depuis peu, dans le cas particulier des SRI sur Internet, on recourt en outre aux indicateurs suivants :

- Le nombre de liens dans la base de données pointant vers une page ou la présence d'un lien en provenance d'un site «important»;
- Le nombre de fois qu'une page est visitée à partir d'une liste de résultats;
- Pour les outils qui incorporent un annuaire, la présence dans l'annuaire de la page concernée.

Par ailleurs, pour tous les systèmes statistiques, la présence de l'ensemble des mots clés de la requête dans un document assure toujours à ce dernier l'émergence en tête de liste des résultats : ainsi, pour une requête comportant à la fois *bananes* et *pommes*, un document avec une occurrence de *bananes* et une occurrence de *pommes* précédera immanquablement un document avec seulement trois occurrences de *bananes*.

2.3. Les systèmes de traitement du langage naturel (T.L.N.)

Le T.L.N. peut être considéré comme un sous-champ du secteur de l'intelligence artificielle. Les recherches qui y sont menées s'appuient sur des disciplines comme la linguistique, l'informatique et les sciences cognitives.

La recherche sur le langage naturel vise la compréhension et la modélisation de la façon dont l'être humain construit le sens d'une phrase ou d'un document, notamment via l'identification des indices exploités pour bâtir cette signification. Puisque l'acquisition du langage chez l'être humain se fait par le biais de l'assimilation progressive des règles et modèles (*patterns and templates*) qui le structurent – les enfants apprenant ainsi, par exemple, à exprimer l'opposition singulier/pluriel ou à construire une phrase, une question ou un ordre –, le T.L.N. pose comme principe que, si nous arrivons à définir ces patrons et à les décrire à un ordinateur, alors nous pourrons enseigner à la machine une partie de la manière dont nous parlons et nous comprenons entre nous. L'experte américaine E. Liddy définit ainsi le T.L.N. :

> *"Natural language processing is a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of particular tasks or applications"* [LID98].

Les systèmes de T.L.N. sont, dans les faits, des systèmes statistiques auxquels l'on adjoint des bases conceptuelles, des bases de connaissances ou des thesaurus, et que l'on dote d'une interface en langue naturelle. La tâche centrale du T.L.N. en ce qui a trait aux SRI concerne la traduction de requêtes et de documents en langage naturel, donc potentiellement ambigus, en représentations internes non ambiguës pouvant être utilisées pour la mise en correspondance et le repérage. Idéalement, ce type de SRI pourrait permettre aux usagers de faire part de leurs requêtes de manière naturelle et avec tous les détails requis (exactement comme ils le feraient avec un bibliothécaire de référence...) et «comprendrait» le sens sous-jacent de la requête dans toute sa subtilité et sa complexité. Ce système permettant une analyse identique des documents de la base de données – peu importe leur nature –, il serait dès lors possible d'effectuer une mise en correspondance conceptuelle à part entière des requêtes et des documents.

La recherche sur le T.L.N. est actuellement en plein essor, car l'interrogation en langue naturelle de bases de données en texte intégral est depuis longtemps considérée comme l'une des clés possibles au «problème de l'usager final» dans le domaine de l'information électronique. (Comme nous le verrons dans le chapitre suivant, les SRI actuels conviennent

surtout à leurs concepteurs et aux spécialistes formés aux procédures d'interrogation...). Les systèmes de T.L.N., s'ils peuvent éventuellement manier les requêtes de type booléen ou statistique, fonctionnent en effet particulièrement bien sur des demandes en «langage ordinaire».

Il existe sept niveaux linguistiques (au moins) à partir desquels les humains extraient le sens d'un texte oral ou écrit et qui sont donc susceptibles d'être incorporés à un système de T.L.N. [SYL00] :

2.3.1. Niveau phonétique/phonologique

Ce niveau réfère à la façon dont les mots sont prononcés. Il n'est pas important en ce qui concerne le repérage de textes écrits, mais s'avère crucial pour la compréhension du langage oral et dans les systèmes de reconnaissance vocale.

2.3.2. Niveau morphologique

En linguistique, le *morphème* désigne la plus petite partie d'un mot porteuse de sens. Ce niveau concerne donc l'analyse componentielle des mots, par exemple l'étude des racines (*chanson* pour *chansonnier*, *chansonnette* ; en anglais *child* pour *childlike*, *childish*, *children*) ou des préfixes et suffixes (*poly-*, *in-*, *-ation*, *-s*). Sous forme de troncature automatique (*stemming*), c'est le niveau le plus communément incorporé dans les SRI, et depuis le plus longtemps. Il est à noter que, plus les langues ont une morphologie riche (ce qui n'est pas le cas de l'anglais), plus l'attention portée dans un SRI à ce niveau linguistique s'avère payante.

2.3.3. Niveau lexical

Le niveau lexical concerne l'analyse du sens des mots (uniquement le sens «du dictionnaire», hors de tout contexte). C'est à ce niveau qu'un SRI peut opérer un étiquetage grammatical des parties du discours (requête/réponse).

2.3.4. Niveau syntaxique

Ce niveau identifie le rôle joué par chacun des mots à l'intérieur d'une phrase et les relations des termes entre eux (le marquage des parties du discours réalisé à l'étape précédente est exploité à cette fin). La structure d'une phrase véhicule, en effet, ce genre d'informations, y compris dans les cas où le sens des mots eux-mêmes demeure inconnu. A titre d'exemple, *Paul frappe Jean* et *Jean frappe Paul* sont des énoncés formés des mêmes mots, mais dont les sens sont bien différents. La position des mots permet ici de déterminer qui est le sujet et qui est l'objet de l'action.

Les systèmes avancés de T.L.N. arrivent à exploiter cette information structurelle, notamment en emmagasinant des représentations de chaque phrase ou en caractérisant les

genres de relations (par exemple, en identifiant comme des définitions les énoncés où des mots sont joints par des expressions comme *est un*).

2.3.5. Niveau sémantique

Ce niveau concerne l'analyse des sens possibles d'une phrase. Les mots à sens multiples y sont désambiguïsés. Puisque, vue de l'extérieur, une chaîne de caractères utilisée dans différents contextes demeure identique, la prise en compte des mots qui l'entourent se révèle nécessaire afin d'identifier le sens en jeu. Dans les SRI, il peut également y avoir expansion des requêtes (*query expansion*) par ajout de synonymes et développement des lieux géographiques (par exemple, *New England* se développera en *Maine, Massachusetts, New Hampshire, Vermont, Rhode Island* et *Connecticut*).

2.3.6. Niveau discursif

Le niveau discursif exploite la structure documentaire des différents genres de textes et de requêtes en vue d'une extraction additionnelle de sens. On peut ainsi tirer parti, par exemple, des traits structurels caractéristiques d'un article de journal, d'un article scientifique, etc. En profitant de cette structure prévisible, le T.L.N. peut déterminer le rôle d'une pièce d'information spécifique dans un document (opinion, fait, prédiction, conclusion, etc.). La résolution des anaphores se fait également à ce niveau.

2.3.7. Niveau pragmatique

Ce niveau réfère au substrat sémantique formé par l'ensemble des connaissances du locuteur sur le monde, connaissances extérieures aux documents ou aux requêtes eux-mêmes mais nécessaires à leur bonne compréhension.

Pour inclure ce niveau dans les systèmes de T.L.N., il s'avère nécessaire de leur adjoindre de gigantesques bases de connaissances où des chercheurs ont recensé patiemment «tout» leur savoir sur le monde. Cette technique est longue et coûteuse; elle présente, en outre, le désavantage de ne pas toujours refléter rapidement les dernières évolutions des connaissances humaines.

La taille de l'objet d'analyse augmente donc au fur et à mesure que l'on avance vers les niveaux supérieurs de compréhension, de même que les difficultés rencontrées par le traitement automatique [LID98].

Bien sûr, tous les systèmes de T.L.N. n'opèrent pas sur l'ensemble de ces niveaux. Les produits qui prennent en charge les niveaux linguistiques élevés sont rares, surtout quand on s'intéresse à la fois au traitement des documents et à celui des requêtes. En réalité, la plupart des systèmes actuels dits de T.L.N. se limitent aux plus bas niveaux de compréhension, et ce, uniquement du côté des requêtes.

En ce qui concerne les SRI sur Internet, la majorité d'entre eux sont actuellement capables de tronquer sur le pluriel/singulier les termes de la requête, ou même d'ajouter/soustraire certaines autres formes d'un mot – essentiellement grâce à la manipulation de suffixes. Certains (INFOSEEK, ASKJEEVES) peuvent, en outre, interpréter quelque peu la syntaxe en «parsant» les éléments de la requête, mais ils s'appliquent pas cette technique au traitement des documents. On commence également à voir apparaître des procédés comme l'identification automatique des noms propres (basée sur la reconnaissance des majuscules et non sur une méthode plus motivée linguistiquement) et celle des locutions (qui semble s'appuyer surtout sur la proximité des mots entre eux) [DEL99].

3. Comment fonctionne un SRI

Le fonctionnement d'un SRI peut être divisé en quatre grandes étapes [DEL99] :

3.1. Le traitement des documents (L'indexation)

C'est l'étape de l'ajout des documents au système et de la construction du *fichier d'indexe*, soit la liste alphabétique de tous les mots présents dans la base de données (les mots vides étant laissés de côté) avec les adresses de chacune de leurs occurrences. Pour les systèmes statistiques, il y a aussi établissement de poids différenciés pour les mots présents dans les documents.

D'autres opérations peuvent éventuellement avoir lieu à cette étape :

1. L'ajout ou la création de bases de connaissances avec des lexiques internes ; des réseaux sémantiques ; de synonymes, …. ;
2. L'extraction additionnelle d'information ou la réalisation d'opérations diverses sur les mots lors du stockage : lemmatisation ; identification des noms propres et/ou communs ; identification du rôle des mots et de leurs relations avec les autres mots de la phrase, du paragraphe, du document ;
3. L'assignation automatique de termes d'indexation ou de larges catégories thématiques ;
4. Le stockage de représentations formelles de chacune des phrases.

3.2. Le traitement des requêtes

Cette étape concerne surtout les systèmes statistiques et de T.L.N., qui doivent accomplir en aval un travail qui est partiellement accompli en amont par le chercheur en ce qui concerne les systèmes booléens : rendre les requêtes compréhensibles par la machine.

Les systèmes statistiques peuvent éventuellement procéder à :

- l'identification des termes importants de la requête ;
- l'identification des racines et des variations de genre et de nombre ;
- l'assignation d'une pondération à chacun des termes de la requête.

Dans leur forme la plus achevée, les systèmes de T.L.N. peuvent mener à bien :

- l'étiquetage de toutes les parties du discours ;
- l'identification des sujets, objets, agents, verbes ;
- l'ajout de synonymes et de formes alternatives pour les noms propres.

Les systèmes de T.L.N. moins développés, pour leur part, se contentent habituellement d'effectuer l'identification des racines et une analyse syntaxique de base.

3.3. L'appariement des requêtes (*query matching*)

Cette étape concerne la mise en correspondance des requêtes avec le fichier d'indexe et, le cas échéant, la base de connaissances.

3.4. La présentation des résultats

Elle peut se faire par date, par champ ou par pertinence présumée par rapport à la requête.

4. Les critères d'évaluation d'un SRI : la précision et le rappel

La pertinence des résultats obtenus suite à une requête est le critère que l'on utilise habituellement lorsque l'on désire jauger l'efficacité et la qualité d'un SRI. Cette pertinence fait appel au jugement de l'usager final (ce dernier ayant toujours raison...) et on la mesure à l'aide de deux grands indicateurs : la *précision* et le *rappel*.

La précision se rapporte au pourcentage des documents repérés qui sont jugés pertinents par l'utilisateur final. Le rappel, quant à lui, concerne le pourcentage de documents, parmi tous ceux de la base de données qui seraient jugés pertinents par l'utilisateur final *s'ils étaient repérés*, qui sont effectivement rapatriés dans les faits.

$$\textbf{La précision} = \frac{\text{Nombre des documents pertinents}}{\text{Nombre total des documents repérés}}$$

$$\textbf{Le rappel} = \frac{\text{Nombre des documents pertinents repérés}}{\text{Nombre de tous les documents pertinents existant dans la base de données}}$$

Les expressions *taux de bruit* et *taux de silence* sont également utilisées pour désigner ces phénomènes respectifs. Le SRI accompli serait donc celui qui parviendrait à retrouver tout ce qui intéresse l'utilisateur tout en ne repêchant rien de ce qui ne l'intéresse pas – en d'autres termes, à atteindre à la fois 100% de rappel et 100% de précision. Il s'agit actuellement d'un

idéal purement théorique puisque, dans les faits, ces deux taux ont plutôt tendance à être inversement proportionnels et à atteindre ensemble un total de 100% au lieu des 200% de la recherche idéale : un système qui favorise la précision voit d'ordinaire son taux de rappel baisser et vice-versa (le plus souvent, c'est la précision qui est privilégiée) [DEL99]. La figure ci-dessous résume cette situation.

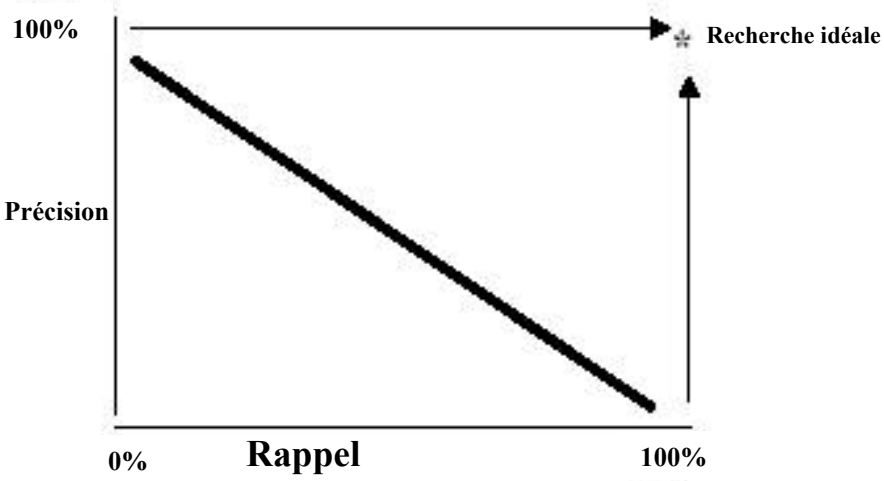


Figure 1 : la précision et le rappel

5. Conclusion

Les performances des SRI en termes de précision et de rappel sont fort variables. Les systèmes booléens et statistiques peuvent ainsi être disposés aux deux extrémités d'un même spectre. Implacables, les systèmes booléens repêchent exactement ce qu'on leur a demandé. Si l'on a bien formulé sa requête, on obtient ce que l'on cherchait ; sinon, on risque de ne rien repérer qui soit utile.

Les systèmes booléens présentent généralement des interfaces peu conviviales, ce qui, conjugué à leur mode d'interrogation à base de mots clés et d'opérateurs logiques et de proximité, contribue à les rendre difficiles à maîtriser pour les usagers non spécialistes. En fait, ils s'avèrent souvent frustrants même pour les experts, qui doivent mémoriser les différences subtiles entre les diverses interfaces booléennes.

Les systèmes statistiques, au contraire, misent sur le rappel (ils se soucient toutefois également de la précision, puisque leur classement présente les résultats les plus pertinents en premier). A ce niveau, ils font un peu mieux que les systèmes booléens, atteignant un taux de rappel de près de 50% [DEL99]. Avec ces systèmes, on obtient non seulement ce que l'on a demandé, mais aussi, éventuellement, ce que l'on *aurait dû* demander... de même que, bien souvent, des documents qui renferment les termes de la requête, mais pas l'information recherchée.

Le T.L.N., se superposant aux systèmes booléens ou statistiques, engendre pour sa part un accroissement à la fois du rappel et de la précision. Utilisé au niveau du traitement du document, il permet une extraction et un stockage plus riche de l'information ; utilisé au niveau du traitement de la requête, il facilite l'expression des besoins d'information grâce à la puissance du langage réel ; utilisé au niveau de l'évaluation des réponses, il simplifie la mise en correspondance avec le sens et l'intention de la requête, améliorant du même coup l'évaluation de pertinence. On peut prévoir que, à l'avenir, les interfaces en langue naturelle s'imposeront comme les préférées de la plupart des utilisateurs.

Les approches booléennes, statistiques et de T.L.N. doivent donc être vues comme étant complémentaires plutôt que concurrentes ou mutuellement exclusives. L'avenir est sans aucun doute à la combinaison d'éléments en provenance de ces diverses méthodes.

Le chapitre suivant sera consacré aux SRI sur Internet, où nous verrons le comportement des systèmes de repérage d'informations sur internet.

Chapitre 2

Les SRI sur Internet

1. Introduction

Le repérage de l'information sur Internet est actuellement une tâche ardue, dont le succès (ou l'insuccès) est tributaire en bonne partie de l'efficacité des outils de recherche. Nous présentons, dans ce chapitre, les caractéristiques des diverses catégories d'outils de recherche d'informations sur intrenet.

2. La recherche d'information sur Internet

Pour mieux comprendre l'intérêt des SRI de plus en plus sophistiqués pour la recherche d'informations sur Internet, il convient de rappeler quelques aspects de ce réseau qui permettront de nous éclairer sur les insuffisances des outils de recherche existants à ce jour.

Sur le réseau Internet chacun peut devenir producteur d'informations et créer son propre site Web. Chaque université, chaque centre de recherche, chaque entreprise, chaque individu peut devenir son propre éditeur. Cette liberté éditoriale explique le grand dynamisme du réseau Internet. Les informations scientifiques et techniques circulent de manière très rapide.

Les échanges et les collectes d'informations se font facilement. La diffusion des connaissances dépend ainsi dans une moindre mesure qu'auparavant du rythme des revues et de ses contraintes (coûts, pagination…). D'une telle situation, résulte également une certaine anarchie éditoriale. Les sources présentes sur le Web sont d'une grande variété mais la validation de l'information par un comité éditorial n'est pas toujours assurée. Il y a donc un besoin d'identifier des ressources pertinentes.

Les ressources apparaissent et disparaissent sans que la plupart des utilisateurs en soient informés. Trouver non seulement les ressources existantes mais aussi, et surtout, les plus récentes et les plus fiables est une véritable gageure. Le Web est affranchi des rythmes de périodicité de parution. Cela permet, si nécessaire, une mise à jour quotidienne, à l'inverse un site peut rester inchangé pendant des mois voir des années ou changer de localisation ou d'appellation.

Les informations disponibles sur Internet sont d'une nature hétérogène. Différents types et formats de documents se côtoient, rendant plus ou moins malaisée la recherche d'informations. De même l'accès à l'information est réalisé suivant de multiples protocoles comme HTTP (HyperText Tranfert Protocol, protocole servant pour le Web), FTP (File Transfert Protocol, utilisé pour le transfert de fichiers), SMTP (Send Mail Transfert Protocol, utilisé pour la messagerie), Gopher (l'ancêtre de web), au travers d'API (Application Programming Interface) comme pour les bases de données ou les applications évoluées sous forme d'applet.

Les données proposées sur le Web sont en général peu structurées malgré le développement des champs « méta-data » dans le langage HTML (HyperText Markup Language) qui donnent des indications (nom, sujet, …) sur le document auquel elles se rapportent.

Les différents mécanismes mis en œuvre dans la sélection des sites par les moteurs de recherche sont le plus souvent opaques à l'utilisateur, que ce soit au niveau de la constitution de l'index, des choix réalisés pour les critères de pertinence, des techniques développées pour interpréter la requête de l'internaute ou encore pour ordonner les résultats. L'indexation des sites trouve rapidement ses limites dans la mesure où elle est souvent générée automatiquement par des logiciels. Ceci explique la prédominance des moteurs de recherche généralistes, indexant le texte intégral des documents HTML. Une réponse à cette problématique est donnée sous la forme d'annuaires dont Yahoo est un exemple. Même si cette approche est une réponse aux questions des utilisateurs elle ne répond qu'en partie au problème de la recherche d'information [SYL00].

L'orientation générale des nouveaux outils de recherche sur Internet, quelle que soit leur nature, est d'offrir aujourd'hui des spécialisations dans certains domaines, comme celui de la finance ou de la santé …etc. Cette approche permet d'utiliser des bases de connaissances spécialisées et fiables dans le domaine considéré.

3. Les difficultés au repérage de l'information sur Internet

Avant de présenter plus en détail les différents types de SRI sur Internet, il peut être utile de rappeler succinctement le contexte général dans lequel évolue la recherche d'information sur le Web. La liste qui suit résume les principales difficultés qu'elle doit gérer:

3.1. Le manque d'habileté et de formation à la recherche des usagers

Une étude de 1996 portant sur le comportement d'usagers inexpérimentés face à un SRI a démontré que le quart d'entre eux n'atteignaient même pas le seuil défini comme minimal d'habileté de recherche. On peut raisonnablement supposer qu'une situation semblable prévaut actuellement sur Internet où les usagers spécialistes des systèmes d'information ne constituent plus qu'une minorité, appelée sans doute à devenir encore plus infime dans les prochaines années [DEL99].

3.2. La couverture limitée des SRI

Sur Internet, il est évident que, plus la base de données d'un outil est imposante et complète, plus ce dernier est susceptible de trouver des réponses à une requête, en particulier pour les sujets obscurs ou très précis. Toutefois, selon un article publié dans la revue *Science,*

le meilleur outil au niveau de la couverture du Web, HOTBOT, n'indexait que 34% des 320 millions de pages estimées disponibles au moment de l'étude. Le pire, LYCOS, ne dépasserait pas les 3%. Cette situation inquiétante ne semble pas destinée à s'améliorer – bien au contraire – en regard de la croissance incontrôlée du Web, et également du fait que l'augmentation de la taille de leur base de données ne semble guère être une priorité chez la plupart des concepteurs d'outils.

Un autre problème considérable, à ce niveau, concerne l'augmentation du «Web invisible», c'est-à-dire des pages dont le traitement pose d'importantes difficultés aux outils de recherche. Les cadres (1) (*frames*) en sont un exemple typique : il n'est pas rare de voir des sites de ce genre de 100 pages ou plus uniquement représentés dans l'index d'un outil par leur page d'accueil (où, en outre, la seule information indexée est souvent l'inscription *This site requires frames*). Les *pages dynamiques* sont également problématiques pour la plupart des outils de recherche, de même que celles qui emploient la technologie XML.

3.3. L'instabilité des ressources

Internet est d'une mouvance intrinsèque : chaque jour, des ressources apparaissent, disparaissent ou déménagent. Comme, par ailleurs, le Web descend en droite ligne du Gopher, il en a hérité plusieurs des caractéristiques, notamment le fait que les liens entre documents ne soient pas bidirectionnels : une ressource vers laquelle pointe un lien n'est pas au courant de cet état de fait. Si cette ressource change, est déplacée ou cesse d'exister, les liens URL ne font donc pas l'objet d'une mise à jour automatique, et demeurent bien souvent «pendants» (d'où le fameux code «Erreur 404»). C'est ce qui explique que les bases de données des outils de recherche comportent inévitablement des liens invalides, en quantité plus ou moins importante selon les cas.

3.4. L'ambiguïté linguistique

3.4.1. La surabondance de synonymes

Cette situation s'explique par la valorisation de la paraphrase dans les textes autres que purement techniques, pour des raisons d'élégance et de style. Elle est également tributaire des différences linguistiques diachroniques, régionales ou professionnelles.

3.4.2. La polysémie

Les données suivantes concernent l'anglais, mais sont intéressantes à titre indicatif. Le *Webster's Seventh Dictionary* recense quelque 60 000 entrées ; or, de celles-ci, 21 488 (soit presque 40%) ont deux sens ou plus [WAC94]. En fait, dans la langue de Shakespeare, un mot aurait en moyenne sept acceptions différentes... La situation est d'autant plus

(1) Les cadres permettent de disposer de plusieurs fenêtres sur une page Web.

préoccupante que ce sont les mots les plus courants qui ont le plus de sens distincts : à titre d'exemple, *run* a 29 sens, qui se subdivisent en près de 125 sous-sens.

A cette polysémie fondamentale des langues naturelles s'ajoutent, en outre, les usages métaphoriques et les comparaisons.

Beaucoup des pièges du traitement automatique sont occasionnés par cette ambiguïté du langage. Les problèmes d'ambiguïté lors du repérage d'information sur Internet sont d'autant plus critiques que les SRI qu'on y trouve actuellement ne parviennent pas à extraire l'information contextuelle contenue dans les documents et les requêtes ; ils ne disposent pas non plus de la masse d'informations sur le monde emmagasinée dans le cerveau des usagers [WAC94].

3.4.3. Les variations orthographiques et les erreurs d'orthographe et de frappe

Un certain nombre de noms communs sont d'orthographe fluctuante, par exemple *clé/clef* ou *fantasme/phantasme* (en anglais, on pourrait citer *gray/grey*, *theatre/theater, aluminium/aluminum*, etc.). Les noms propres présentent eux aussi des variantes : ainsi, dans un texte, ils peuvent figurer en version abrégée dans le titre (pour sauver de l'espace), apparaître en version complète dans le premier paragraphe afin d'établir clairement la référence, puis revenir par la suite sous des formes plus courtes, l'entité ayant déjà été introduite. Outre les synonymes, le chercheur doit donc penser aux diverses variantes orthographiques possibles quand vient le moment d'imaginer les différentes manières dont un concept peut être exprimé (l'emploi de la troncature peut éventuellement lui faciliter un peu la tâche).

Le problème des fautes, pour sa part, est aggravé par l'incorporation de plus en plus fréquente, dans les bases de données, de textes numérisés à l'aide de techniques de reconnaissance optique de caractères (ROC). Selon certaines études, en effet, ces textes, sans une relecture attentive des épreuves, peuvent facilement comporter jusqu'à 30 erreurs par page [DEL99].

3.4.4. Les pertes d'information lors du traitement

Certains phénomènes comme l'ordre des mots, la distinction minuscules/majuscules ou la présence de signes diacritiques et de caractères spéciaux ne sont pas toujours gérés de manière cohérente et efficace par les outils de recherche. Des subtilités comme les distinctions entre *AIDS* et *aids* (*SIDA* et *assistants*), *school library* et *library school* (*bibliothèque scolaire* et *école de bibliothéconomie*), *tache* et *tâche* leur échappent donc souvent, de même que la nécessité de mettre en correspondance des formes comme *online* et *on-line*. Il y a là une perte

d'information importante, puisque la prise en compte des majuscules, par exemple, peut favoriser le traitement des abréviations, des acronymes et des noms propres.

3.4.5. La difficulté de formulation de certains concepts

Sur Internet, les SRI ne permettent pas toujours la formulation de requêtes en langue naturelle, comme le souligne E. Liddy : *«The engines expect minimal one-word or two-word queries and are optimized for them, rather than for sentences, which would enable the user to fully present their information need.»* [LID98]. Cela augmente la difficulté éprouvée à définir des concepts importants mais vagues : «*Speaking in code is difficult, and it leaves out important aspects of thought.*» [DEL99]. Actuellement, la seule solution consiste bien souvent à administrer aux SRI de longues chaînes de synonymes et d'adjectifs.

3.4.6. Les «false drops»

Il faut entendre par là les documents qui sont repêchés suite à une requête, mais qui sont sans rapport aucun avec le sujet. Ce phénomène de bruit est dû en bonne partie au couple bon mot/mauvais sens ; il concerne en particulier les systèmes booléens, qui ne vérifient pas automatiquement la proximité et la fréquence des mots. A un moindre niveau, il affecte également les systèmes statistiques. Le problème des «false drops» illustre avec acuité les limites de ces deux modes de repérage [DEL99].

4. Les outils de recherche généralistes

Ensemble de plusieurs dizaines de millions de documents, le Web connaît plusieurs modes d'indexation au travers d'outils de recherche, mais aucun n'est complet et totalement pertinent. Leurs caractéristiques sont généralement de couvrir divers domaines d'intérêt avec un degré de précision faible, ce qui fait dire d'eux qu'ils sont généralistes.

4.1. Les annuaires

Nous retiendrons comme premier type d'outils de recherche sur Internet les *annuaires* – que l'on appelle également *guides*, *répertoires* ou *catalogues*. Le prototype en fut la WORLD WIDE WEB VIRTUAL LIBRARY, localisée au CERN. Les annuaires sont des regroupements par sujet des ressources d'Internet. Ils consistent en des classements arborescents où l'accès au thème souhaité s'effectue en parcourant une série de rubriques et de sous-rubriques. Comme on peut le lire dans l'aide en ligne de l'annuaire YAHOO!, *«l'analogie avec un arbre s'impose clairement : chaque catégorie du guide, ou branche de l'arbre, abrite plusieurs sous catégories, d'autres branches qui, elles-mêmes, vous donnent le choix entre plusieurs chemins possibles au fur et à mesure de votre balade, etc.»*. En fait, les annuaires, dont les

ramifications successives conduisent à des sujets de plus en plus pointus, pratiquent ce que l'on pourrait appeler le «principe de l'entonnoir».

D'ordinaire, ils incorporent également un moteur de recherche par mot clé, ce qui permet d'effectuer directement une requête sur le sujet souhaité.

Le consultant Internet français Olivier Andrieu propose la définition suivante des annuaires :

> *« Un annuaire est un outil de recherche qui recense un certain nombre de sites (et non de pages) Web au travers de fiches descriptives comprenant, en règle générale, le titre, l'adresse (l'URL) et un bref commentaire d'une longueur allant le plus souvent de 15 à 25 mots au maximum. Chaque site est inscrit dans une ou plusieurs catégorie(s) – on parle également de rubrique(s) –. Ces outils peuvent ainsi être considérés comme les pages jaunes du Web. Lorsqu'un mot clé est saisi dans le formulaire proposé, l'annuaire effectue une recherche sur les occurrences de ce terme dans ses fiches descriptives de site, et non pas dans le contenu des pages du site en question. Il s'agit là de la différence la plus notable avec les moteurs de recherche ».* [OLI04]

On peut résumer ainsi les principales caractéristiques des annuaires :

- ils recensent des *sites* et non des *pages* individuelles ;
- ils structurent leur inventaire selon une classification en général propre à l'outil (certains ont recours aux classifications documentaires traditionnelles comme celle de la Bibliothèque du Congrès de Washington – utilisée notamment par la WORLD WIDE WEB VIRTUAL LIBRARY – ou celle de Dewey, mais le cas demeure rare) ;
- le repérage et la catégorisation des ressources s'effectuent souvent manuellement, au moins en partie. Les annuaires recourent, à cette fin, soit à des professionnels de la documentation (bibliothécaires, documentalistes), soit à des spécialistes des diverses thématiques concernées (par exemple, des médecins pour la rubrique *Santé*), soit encore à des volontaires (rémunérés ou non) ;
- les annuaires incorporent parfois directement des sites Web dans leur base de données (suite à une décision de l'équipe éditoriale ou à une suggestion en provenance des usagers du service) ; toutefois, il est généralement nécessaire d'entreprendre une démarche délibérée d'inscription : le responsable du site à enregistrer doit soumettre ce dernier, qui est alors visité, évalué et – si accepté – inclus dans l'arborescence de l'outil.

Le principe des annuaires présente plusieurs avantages. Tout d'abord, ces instruments permettent de guider l'utilisateur dans ses investigations ; ils s'avèrent donc moins

«intimidants» que la ligne de commande vide des autres outils de recherche. Grâce à la catégorisation effectuée sur l'information, il s'avère aisé pour l'usager de «butiner» entre sites traitant d'un même sujet, un peu comme l'on bouquine devant les rayons d'une bibliothèque. La philosophie des annuaires permet également de limiter le taux de bruit, et s'accompagne d'une substantielle valeur ajoutée due à l'activité humaine de sélection, d'évaluation et d'hiérarchisation des ressources. On note également, bien sûr, certains inconvénients : augmentation du taux de silence (en supposant qu'un document soit classé dans une seule catégorie), couverture relativement restreinte d'un bassin potentiel de millions de sites Web, mise à jour moins rapide que pour les autres outils, dépendance par rapport aux choix éditoriaux des réalisateurs (il n'y a souvent qu'un pas entre l'évaluation des ressources et la censure...). En outre, même si les requêtes de recherche sont possibles, elles offrent en général moins de souplesse et de précision que celles permises dans les outils de type moteur.

De manière globale, on peut donc dire que les annuaires, favorisant le repérage de sites généraux sur un sujet donné, s'avèrent surtout utiles pour des fouilles vastes et thématiques ou pour débuter une recherche d'information encore mal définie. Leur convivialité en faisant par ailleurs les outils de recherche les plus simples d'utilisation, ils sont également tout indiqués pour les débutants [DEL99].

Il convient de souligner que les annuaires disponibles en plusieurs versions linguistiques ne constituent pas autant de copies d'une même base de données simplement coiffées d'interfaces différentes. Il s'agit bien, dans les faits, de bases totalement dissociées ; il importe donc de les interroger successivement et d'effectuer les requêtes dans la langue de l'interface (par exemple, en anglais dans YAHOO! INTERNATIONAL et en français dans YAHOO! FRANCE).

Quelques annuaires en langue anglaise :

Nom	**URL**
Galaxy	http://galaxy.einet.net/
Jassan	ht://www.jassan.com/
Looksmart	http://www.looksmart.com/
Magellan	http://magellan.excite.com/
Open Directory Project	http://dmoz.org/
Snap	htp://www.snap.com/
Yahoo! International	http://www.yahoo.com/

Quelques annuaires en langue française :

Nom	**URL**
Carrefour	http://www.carrefour.net/
CTrouvé	http://www.ctrouve.com/
Francité	http://www.i3d.qc.ca/
Nomade	http://www.nomade.fr/
Yahoo! France	http://www.yahoo.fr/

4.2. Les moteurs

Le second type d'outils de recherche sur Internet est constitué par ce que l'on appelle des *moteurs*. WEBCRAWLER fut le premier instrument de ce genre, en ligne depuis avril 1994 [DEL99]. Si les annuaires évoquent le plan de classification des bibliothèques traditionnelles, les moteurs, pour leur part, ressemblent un peu à ces programmes qui produisent automatiquement des index primitifs en associant, à chaque mot d'un document, la ou les page(s) où il figure – du reste, on les appelle aussi parfois des *index*.

Les moteurs permettent à l'usager de repérer l'information non suite à une navigation thématique, mais via l'interrogation à l'aide de mots clés et de commandes logiques d'une base de données indexée ; leur fonctionnement rejoint ainsi celui des logiciels de gestion documentaire usuels. En général, deux modes de recherche sont disponibles : *recherche simple* (proposée par défaut à partir de la page d'accueil de l'outil, avec plus ou moins de possibilités de recherche) et *recherche avancée* (accessible en option et où des possibilités de recherche variées et approfondies, souvent paramétrables, sont offertes).

Le fonctionnement des moteurs s'appuie sur la collecte de données par des *robots*, lesquelles sont ensuite indexées directement à l'aide des mots qui les constituent. De gigantesques bases de données – autrement plus imposantes que celles des annuaires – sont ainsi élaborées ; elles opèrent *grosso modo* sur le mode des *fichiers inversés (les indexes)* en établissant des correspondances entre des mots et des URL. Les utilisateurs sondent la base à l'aide d'un module d'interrogation qui recourt à un langage de requête plus ou moins standard ; des interfaces conviviales sont généralement mises en place afin de faciliter l'interaction. L'activité des moteurs de recherche, contrairement à celle des annuaires, est entièrement automatisée.

Les robots – qui connaissent diverses autres appellations évocatrices, notamment *spider* («araignée»), *ant* («fourmi»), *worm* («ver de terre» ou «se faufiler»), *wanderer* («vagabond»), *crawler* («nageur»), etc. – sont tout simplement des programmes informatiques qui tournent

sur un ordinateur relié au Réseau et qui explorent systématiquement celui-ci de manière à collecter l'information présente.

Les robots procèdent en repérant les liens hypertextuels d'un document pour ensuite aller visiter les pages vers lesquelles pointe ce dernier. Ils parcourent ainsi rapidement un site, puis d'autres sites qui lui sont liés, et ainsi de suite. Une fois le site indexé, le robot revient régulièrement «capturer» une version plus récente des différentes pages. Il n'est pas rare que le même robot soit utilisé par plusieurs moteurs différents, avec seulement quelques différences de paramétrage. [DEL99]

Généralement, seuls les fichiers ASCII et HTML sont indexés (et non, par exemple, les fichiers compressés ou de type *.pdf*). Le fonctionnement mécanique des robots fait en sorte qu'il est fort difficile de contrôler quelles pages sont récupérées pour être indexées. Le contenu de la base d'un moteur demeure donc essentiellement tributaire des sites utilisés comme points de départ et de la stratégie privilégiée pour la visite des liens (ce peut être une stratégie en largeur, où tous les liens immédiats dans l'ensemble des pages rapatriées sont visités, ou une stratégie en profondeur où, pour une sélection de documents, le robot descend de page en page jusqu'au dernier lien existant) [SYL00]. La Figure suivante décrit le fonctionnement des Moteurs de recherche :

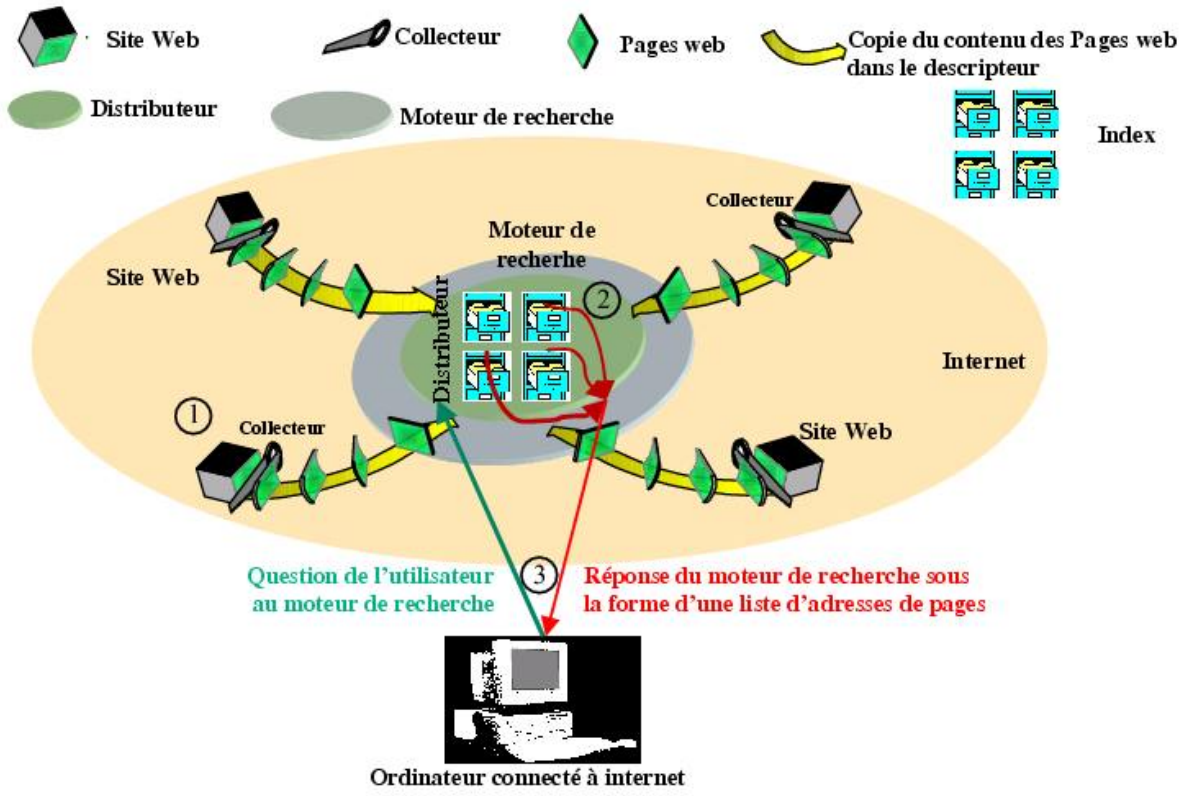


Figure 1. Principe de fonctionnement d'un moteur de recherche

Mentionnons que, contrairement aux annuaires, les moteurs qui se déclinent en plusieurs versions linguistiques ne proposent, en général, que des versions localisées d'une même base de données (EXCITE FRANCE, LYCOS FRANCE). Beaucoup des grands moteurs internationaux ne se donnent, d'ailleurs, pas cette peine et se contentent de doter leur interface anglophone d'une option de recherche de restriction linguistique (ALTAVISTA, HOTBOT, NORTHERN LIGHT).

Un des avantages de la démarche de type moteur réside dans le fait que l'utilisateur n'a pas à connaître la catégorie (et la structure hiérarchique) dans laquelle pourrait se trouver l'information recherchée, puisque cette dernière n'est pas compartimentée de la sorte et que la recherche s'opère principalement par concordance avec un modèle (*pattern matching*). Par ailleurs, comme l'absence d'intervention humaine équivaut souvent à une absence de déontologie, les moteurs sont en principe plus performants que les annuaires pour repérer des documents à contenu sensible (violence, pornographie) ou carrément sujets à controverse (sites haineux, terroristes, pédophiles, etc.), une caractéristique que l'on peut ou non applaudir mais qui est conforme à l'esprit libertaire et anarchiste du Net [DEL99].

Le taux de rappel obtenu par les moteurs est souvent bon, mais il s'accompagne malheureusement d'une grande quantité de bruit, c'est-à-dire d'une baisse du taux de précision : les moteurs suscitent des réponses très hétérogènes, où les doublons abondent parfois. La (non-) mise à jour des index constitue souvent également une source de problèmes. Autre inconvénient : contrairement aux annuaires, les moteurs abandonnent l'usager à lui-même (rien ne guide ni ne balise la recherche) et ne fonctionnent habituellement pas sur le mode d'un ensemble de réponses qu'il est possible de restreindre et d'affiner successivement : la recherche se fait en un coup et un seul. Enfin, leur utilisation demeure délicate et les recherches peuvent prendre beaucoup de temps.

Généralement plus appréciés des internautes aguerris que des débutants, les moteurs, en un certain sens, sont plus «puissants» que les annuaires. Ils sont donc tout indiqués pour des recherches qui portent sur des sujets fins et précis ou sur un objet dont l'existence est connue d'avance, mais ils risquent de générer des milliers de réponses d'intérêt inégal si la requête s'avère trop vague ou trop commune [DEL99].

Comme on le voit, les moteurs se différencient des annuaires à de nombreux points de vue. On peut résumer ainsi leurs principales caractéristiques :

- Ils recensent des pages individuelles et non des sites en tant qu'entités ;
- Aucune structuration, classification ou hiérarchisation de l'information n'est effectuée
- Leur fonctionnement ne comporte aucune intervention humaine ;
- Il n'est pas absolument nécessaire d'inscrire les pages d'un site auprès des divers moteurs : on peut tout simplement choisir d'attendre que les robots débusquent le site concerné au détour d'un lien, le visitent et en indexent les différentes pages. Cette méthode demeure néanmoins aléatoire et requiert habituellement l'écoulement d'un certain laps de temps. Il est donc nettement préférable d'opter pour la soumission manuelle des URL que l'on désire faire connaître. Pratiquement tous les moteurs offrent, en effet, une fonction de type ***Add a site*** ou ***Add URL***, qui sert à signaler au robot l'adresse de pages à visiter.

Enfin, si les annuaires et les moteurs sont des outils bien distincts, il convient de signaler que de plus en plus de sites de recherche combinent l'accès aux deux genres d'instruments, selon des formules qui privilégient l'un ou l'autre type : moteur agrémenté d'un annuaire (par exemple, VOILA) ou annuaire complété d'un moteur de recherche externe (par exemple, FRANCITE). Une autre tactique consiste à conclure des accords de partenariat avec des sociétés concurrentes : l'annuaire YAHOO!, par exemple, dirige l'internaute sur le moteur INKTOMI en

cas de recherche infructueuse. Les moteurs INFOSEEK FRANCE et EXCITE FRANCE, pour leur part, affichent les catégories et les descriptions de sites de l'annuaire NOMADE.

Quelques moteurs en langue anglaise :

Nom	**URL**
ALTAVISTA	http://www.altavista.com/, http://www.av.com/, http://altavista.digital.com/
EXCITE	http://www.excite.com/
EXCITE version française	http://www.fr.excite.com
HOTBOT	http://www.hotbot.com/
INFOSEEK	http://infoseek.go.com/
LYCOS	http://www-english.lycos.com/
LYCOS version française	http://www.lycos.fr/
WEBCRAWLER	http://www.webcrawler.com/
YAHOO!	http://www.yahoo.com/
YAHOO! version française	http://www.yahoo.fr/
GOOGLE	http://www.google.com
GOOGLE version française	http://www.google.fr/

Quelques moteurs en langue française :

Nom	**URL**
ECILA	http://www.ecila.fr/
LOKACE	http://www.lokace.com/
VOILA	http://www.voila.fr/
VOILA version mondiale	http://www.voila.com/

4.3. Les métamoteurs

Le troisième grand groupe d'outils de recherche est celui des *métamoteurs*. Ce sont des instruments qui visent à faciliter la transmission d'une même requête vers différents moteurs et annuaires.

Les métamoteurs se subdivisent en deux catégories. La première rassemble les *Configurable Unified Search Interfaces* (*CUSI*), que l'on appelle également – de manière plus prosaïque – les *bibliothèques de moteurs* ou les *All in One*. Ce genre d'instrument recense habituellement un grand nombre d'outils de recherche en fournissant un accès direct, sur une même page, à la ligne de commande de chacun d'eux. Utiles dans la mesure où ils permettent la consultation de plusieurs services à partir d'un même site et disposent souvent d'une

interface astucieuse qui évite à l'usager d'avoir à retaper continuellement sa requête, ces métamoteurs de première génération demeurent, toutefois, assez primitifs et ne rendent que peu de services supplémentaires. Ils se chargent tout simplement de communiquer la requête concernée aux différents outils de recherche, généralement de façon séquentielle [DEL99].

Quelques CUSI, parmi bien d'autres :

Nom	**URL**
ALL-IN-ONE SEARCH PAGE	http://www.allonesearch.com/
Easy Search (japonais)	http://www.aist.go.jp/NIBH/~honda/EasySEARCH/index.cgi
GOLDENBRICK (francophone)	http://www.goldenbrick.fr/goldensearch/recherche.html
THE SEARCH PLACE	http://users.isaac.net/duane/search/

Ceci dit, le terme *métamoteur* renvoie la plupart du temps à une seconde catégorie d'outils, à valeur ajoutée ceux-là : les *Simultaneous Unified Search Interfaces* (*SUSI*). Ces métamoteurs fonctionnent en transmettant simultanément la requête de l'usager à plusieurs outils de recherche, principalement des moteurs. La quantité d'outils ainsi «interpellés» est très variable ; elle se situe d'ordinaire entre 5 et 150 [DEL99].

Les métamoteurs récupèrent par la suite les différentes listes de résultats et les façonnent en un document unique. Certains procèdent, en outre, à un classement de pertinence supplémentaire et à l'élimination des doublons. Plusieurs d'entre eux permettent également de configurer la liste des sources à interroger.

Les principaux avantages liés à l'emploi des SUSI ont trait au gain de temps (il n'est plus requis de visiter les outils un à un) et au fait qu'ils dispensent l'usager de la nécessité de s'initier aux modalités d'utilisation de chacun des sites à interroger –entreprise qui s'avère parfois laborieuse en ce qui concerne les modes de recherche avancée. L'utilisation de ces métamoteurs fait face, toutefois, à certains problèmes pratiques. Tout d'abord, il s'avère impossible pour ces logiciels d'exploiter les fonctionnalités avancées des outils de recherche, précisément parce que la syntaxe en est très variable. Leurs requêtes doivent demeurer suffisamment «basiques» pour être acceptées par tous les outils auxquelles elles sont envoyées, ce qui en diminue la puissance. Ensuite, comme le fait remarquer O. Andrieu :

> *[...] les métamoteurs font la synthèse de résultats fournis par plusieurs moteurs différents, classant chacun leurs résultats de façons différentes, sans utiliser les mêmes critères de pertinence. Une synthèse de documents classés de façons*

ainsi disparates est-elle si simple que cela à effectuer, et surtout, est-elle plus pertinente ? La question reste posée... [OLI04]

Ajoutons que les métamoteurs anglophones font généralement preuve de ce que l'on pourrait qualifier de «myopie anglo-saxonne» en ce qui concerne la liste des outils à sonder... Le concept de métamoteur, tout en étant intéressant en soi, demeure donc l'objet d'un certain nombre de réserves. Pris pour ce qu'il est, toutefois, et utilisé un peu à la manière d'un annuaire (pour des recherches larges et thématiques), ce type d'outil peut tout de même s'avérer d'une utilité non négligeable [DEL99].

4.4. Les agents intelligents

Le concept d'agent intelligent recouvre des réalités nombreuses et diverses. Au sens large, les agents intelligents peuvent être définis comme des outils *«permettant d'automatiser, périodiquement ou à la demande, des tâches de façon transparente pour l'utilisateur qui bénéficie des résultats»*. Dans le contexte plus spécifique de la recherche d'information, ces logiciels sont généralement dotés, à des degrés divers, des caractéristiques de base suivantes :

- L'automatisation et l'autonomie du fonctionnement ;
- La mobilité, c'est-à-dire l'aptitude à voyager sur les réseaux ;
- La capacité d'interaction avec des interlocuteurs humains ou mécaniques ;
- La capacité dynamique d'apprentissage.

Contrairement aux annuaires, aux moteurs et aux métamoteurs, les agents intelligents ne forment pas une classe clairement délimitée de SRI sur Internet. D'une part, ils sont souvent incorporés aux outils des autres groupes : les robots que nous avons évoqués précédemment constituent, en fait, un type élémentaire d'agent intelligent, tout comme les métamoteurs sont une application de cette technologie.

D'autre part, les agents varient énormément entre eux au niveau de leurs caractéristiques spécifiques [DEL99]:

4.4.1. Agents de recherche d'informations

Sous le terme d'agents de recherche d'informations, se cache toute une gamme de logiciels allant des moteurs de recherche aux "agents intelligents". Ces agents de recherche d'informations sont d'une grande diversité, remplissant rarement les même tâches [SYL00]. On distinguera quatre fonctions principales :

- Recherche d'information : celle-ci peut se faire de manière "intelligente" par l'utilisation de méta-moteurs perfectionnés, d'outils d'analyse linguistique des requêtes ou par exploration de liens hypertextes à partir d'une URL (adresse Internet) donnée.

- Analyse des informations récupérées : indexation sémantique des résultats, résumé automatique.
- Filtrage, édition, archivage, mise à jour de résultats.
- Navigation off-line (hors ligne ou en local) parmi des pages, ou des sites Web, ou des pages de sites Web téléchargés.

4.4.2. Agents de navigation en local (off-line)

Les agents de navigation en local permettent de télécharger sur un disque dur des sites Web préférés et de les consulter en local. Ils sont à l'origine d'une économie sur le prix de connexion puisqu'il ne faut pas attendre le chargement des pages Web. Ils permettent d'autre part de surveiller la mise à jour régulière de ces sites, c'est à dire que périodiquement le logiciel va vérifier si le site a été modifié. Certains logiciels permettent de plus de faire des recherches au sein des pages téléchargées sur le disque dur. A terme, on peut penser que ces produits vont être peu à peu intégrés dans les navigateurs. Déjà Microsoft Internet Explorer 3.01 pour Macintosh permet de surveiller les modifications d'une page [SYL00].

4.4.3. Agents guides

Comme les agents de navigation « off-line », on peut penser que les agents guides seront intégrés peu à peu dans les navigateurs. Ce sont des applications qui ont pour objectif d'accompagner et d'assister les utilisateurs dans leur navigation par une série de fonctionnalités variées. Les agents guides préparent des réponses à certaines questions de l'utilisateur sur des points comme : où suis-je ?, que se passe t-il sur les pages ?, où devrais-je aller ?, quels sont mes centres d'intérêt ?, etc... Même si tous les agents n'obéissent pas aux même technologies et aux mêmes fonctionnalités ils ont des objectifs communs et sont des outils utiles aux internautes pour améliorer la recherche d'information [SYL00].

Quelques agents intelligents :

Nom	**URL**
AURESYS	http://ms161u06.u-3mrs.fr/hom.html
DIGOUT4U	http://www.arisem.com/index_fr.html
INFORIAN QUEST 98	http://www.inforian.com
MATA HARI	http://www.thewebtools.com
NEARSITE	http://www.nearsite.com
PRICELINE	http://www.priceline.com
SELECTCAST	http://www.aptex.com/productsselectcast.htm
SHOPPING EXPLORER	http://www.shoppingexplorer.com

WEBWHACKER
http://www.bluesquirrel.com/products/whacker/whacker.html
WEBZINGER http://www.webzinger.com

5. Les outils de recherche spécialisés

5.1. Description

Les outils spécialisés constituent une autre classe des SRI sur internet, comme leur nom indique il s'agit des outils spécialisés dans un domaine particulier (Médicine, Biologie, Finance…etc.). L'homogénéité du domaine considéré permet de mettre en oeuvre des techniques de recherche d'information plus sophistiquées que pour dans le cadre de recherches généralistes sur le web et sa spécialisation rend le besoin d'information précise plus pressant [NED02].

5.2 Comparaison des outils spécialisés face aux moteurs de recherche généralistes

Une comparaison en 6 points permet de mieux évaluer les avantages et inconvénients que l'on peut trouver dans ces deux approches [SYL00] :

a) outils ciblés, spécialisés, limitation du bruit par rapport aux robots,

b) données présentées plus homogènes,

c) valeur ajoutée humaine pour la description des ressources, la sélection, la validation et la catégorisation des ressources,

d) suivi du répertoire et actualisation (liens obsolètes) ,

e) les répertoires ont une mise à jour moins rapide,

f) les répertoires offrent une expression de la requête utilisateur plus évoluée que par des équations booléennes qui sont moins proches du domaine d'intérêt.

5.3 Quelque exemples d'outils spécialisés dans le domaine médical

Cliniweb

Il s'agit d'un index des informations cliniques disponibles sur le Web. Il met à la disposition des étudiants et professionnels de la santé une interface de navigation et de recherche de données cliniques. Sa base de données contient une liste de sources d'informations indexées grâce à la nomenclature MeSH, et retrouvées par l'intermédiaire du système SAPHIRE [HER95]. L'objectif de CliniWeb est de structurer les ressources cliniques de haute qualité disponibles sur le Web pour les enseignants, les praticiens et les chercheurs dans le domaine médical. Un robot recherche et indexe les sites qui sont censés intéresser le public visé. CliniWeb permet un accès rapide aux diverses ressources qui ont été repérées et

organisées selon des thèmes spécifiques. De plus CliniWeb se veut un banc d'essai pour la définition de méthodes optimales destinées à l'évaluation de ressources cliniques disponibles sur le Web. L'indexation des sites avec des codes MeSH est réalisée à la main par un spécialiste du domaine [HER96]. Cet outil est composé d'une:

- base d'URLs de données cliniques : le seul critère de sélection de pages Web est la présence d'informations cliniques utiles. Ces pages proviennent essentiellement des sites affiliés aux agences gouvernementales ou aux écoles de médecine.
- indexation des URLs grâce à l'arborescence MeSH. Bien que l'assignation des termes MeSH soit effectuée manuellement, le processus est assisté par le système SAPHIRE. Ce système prend en entrée du texte libre et identifie les termes MeSH correspondants.
- interface de navigation et de recherche d'URLs : un outil permettant d'insérer une liste d'URLs sous chaque terme MeSH.

La difficulté à maintenir la base de données témoigne de l'une des limites de l'outil. De plus l'indexation manuelle par un étudiant en médecine dépourvu d'expérience dans ce domaine représente une autre difficulté. Il faut citer également les limites concernant les recherches. Puisque l'utilisateur n'est censé s'intéresser qu'à la maladie, non pas au diagnostic ou au traitement. La découverte de nouvelles ressources et leurs indexations se révèle être une tâche très difficile. CliniWeb est disponible sur le site de "the Oregon Health Sciences University" (OHSU) : http://www.ohsu.edu.cliniweb.

CISMeF

CISMeF [DAR98] a pour objectif de cataloguer et d'indexer les principales ressources médicales francophones. CISMeF utilise pour structurer l'information d'une part le thésaurus MeSH « Voir glossaire » et d'autre part le format RDF du Dublin Core. CISMeF s'adresse autant aux professionnels de la santé médecins qu'aux infirmières, sages-femmes, vétérinaires, soigneurs et nutritionnistes. Même le grand public y a accès. Les ressources enregistrées dans CISMeF sont décrites suivant une hiérarchie à 5 niveaux : 1) meta term, 2) catégorie, 3) mots clés, 4) subheadings, 5) type de source. Les "types" dans CISMeF sont une généralisation des "types" de publications de Medline. Des types spécifiques des ressources disponibles sur Internet ont été rajoutés, par exemple : associations, informations patients, etc...

CISMeF contient un index thématique qui inclut les spécialités médicales ainsi qu'un index alphabétique. Il respecte le Net-Scoring constitué d'un ensemble de 48 critères établis

dans le but d'améliorer la qualité de l'information de santé diffusée sur l'Internet. Ces critères s'inspirent d'une réflexion d'un groupe nord américain [SOU03].

6. Conclusion

Sur Internet, le processus de repérage de l'information est confronté à de multiples difficultés, certaines spécifiques (comme l'instabilité des ressources), d'autres communes à tous les systèmes d'information (par exemple, les problèmes découlant des ambiguïtés langagières). Les outils de recherche développés pour tenter de gérer cette situation sont très nombreux à l'heure actuelle ; ils continuent de se multiplier à un rythme effréné et il n'est sans doute pas exagéré de prétendre qu'il en apparaît de nouveaux presque tous les jours [DEL99]. Face à un tel foisonnement, l'internaute moyen est souvent tenté de s'en tenir à la consultation d'un service ou deux parmi les plus connus, tels ALTAVISTA, YAHOO! ou GOOGLE. Pourtant, il est au contraire impératif, lorsque l'on mène des recherches d'information sur Internet, de ne pas se cantonner à un seul outil ni même à un seul genre d'outils. Ici aussi, la complémentarité est le maître mot : aucun outil de recherche n'offre de couverture parfaitement exhaustive [DEL99]; en outre, il semble que les recoupements entre les portions d'Internet couvertes par les différentes bases d'outils de même type demeurent assez minimes, bien qu'il soit fort difficile d'évaluer la situation à ce niveau. Comme, par ailleurs, les différents types d'outils ont été conçus pour répondre à des besoins distincts (recherches simples, générales ou thématiques pour les annuaires et métamoteurs ; recherches complexes ou pointues pour les moteurs ; recherche spécifique à un domaine donné pour les outils spécialisés), il s'avère beaucoup plus judicieux d'employer en parallèle plusieurs outils.

Chapitre 3

L'indexation dans les SRI

1. Introduction

Tout outil de recherche d'information fonctionne à partir de fichiers d'index, cœur de la technologie utilisée, représentant l'information textuelle stockée dans des documents électroniques.

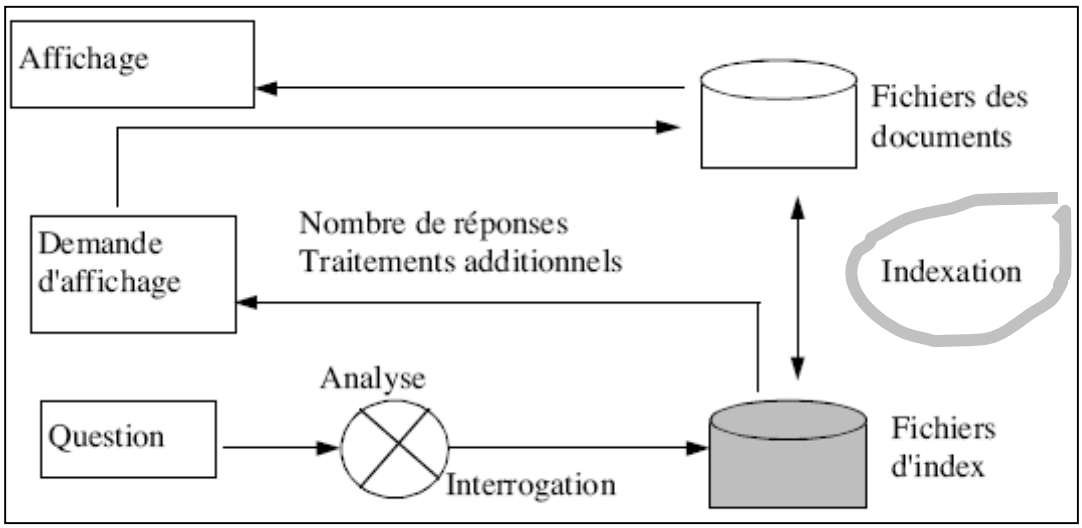


Figure 1. L'indexation dans les outils de recherche

La recherche documentaire classique utilise des index de type mots-clés, où les entrées d'index sont des mots ou expressions normalisés d'un langage contrôlé – comme un thésaurus par exemple [CAT98]. C'est une technique éprouvée, qui permet effectivement de retrouver de manière certaine des informations indexées. La recherche est exacte et confronte une équation de recherche booléenne au contenu des fichiers d'index : elle est parfaitement efficace et ne laisse place, au plan informatique, à aucune incertitude quant au fait que les documents trouvés correspondent bien à l'équation de recherche. En revanche, elle demande des ressources importantes pour indexer les documents et la qualité des résultats obtenus en recherche dépend directement de la qualité de l'indexation.

Si les documents électroniques sont des informations textuelles non structurées, les entrées d'index peuvent être beaucoup plus nombreuses et peu ou pas normalisées. Pour être efficace, la recherche ne peut pas se borner à une simple analyse de présence/absence de mots, au risque de fournir des taux de silence et de bruit trop importants.

C'est effectivement sur ces questions que les technologies ont très sensiblement évolué depuis quelques années. Dans ce chapitre nous donnerons un survole sur les différentes techniques d'indexation utilisées par les outils de recherche.

2. L'indexation

2.1. Définition et concepts

Les spécialistes de la documentation assurent le stockage et la diffusion de l'information (fixée sur différents supports). Or ces opérations exigent au préalable un traitement intellectuel des documents : l'indexation. L'indexation documentaire va dégager dans ceux-ci l'information nécessaire à son repérage en l'établissant ensuite dans des listes, tables ou index.

Dans une phase préalable dite d'analyse, on soumet des textes ou documents à l'indexation, ceci en utilisant des mots tirés du langage naturel, c'est-à-dire des mots clés contenus dans les textes/documents. L'indexation consiste à extraire tous les éléments d'information (concepts) éventuellement utiles à leur usage ultérieur et, dans une deuxième phase, à transposer ces données brutes dans un langage tiré d'un lexique documentaire normalisé ou codé (langage documentaire) permettant le stockage et la communication de l'information. La transcription en langage documentaire se fait au moyen d'outils d'indexation tels les thesaurus, les répertoires de sujets ou les classifications [MEY01].

L'indexation des documents collectés est une étape très importante dans le processus de recherche. En effet, de la qualité de l'indexation dépend la qualité de la recherche. Un index est une liste des mots retenus avec pour chacun d'eux les documents dans lesquels ils apparaissent.

2.2. Le processus d'indexation

2.2.1. L'extraction des mots clés du document

C'est une étape qui peut sembler triviale au premier abord, et qui pourtant constituera la base de tout le reste du processus d'indexation. Il faut donc que cette phase soit d'une qualité maximale.

L'ensemble des développeurs des moteurs de recherche considèrent un mot comme étant une chaîne de caractères. Cette chaîne est constituée d'au moins un caractère et pouvant contenir par exemple:

- N'importe lesquelles des 26 lettres de l'alphabet en majuscules (A-Z).
- N'importe lesquelles des 26 lettres de l'alphabet en minuscules (a-z).
- Le tiret (-).
- N'importe lesquels des caractères accentués du jeu de caractères ISO-8859-1.

2.2.2. La normalisation des mots clés

Ce traitement consiste à retrouver pour un mot sa forme normalisée (généralement le masculin pour les noms, l'infinitif pour les verbes, le masculin singulier pour les adjectifs, etc.). Ainsi, dans l'index ne sont conservées que les formes normalisées, ce qui offre un gain de place appréciable, mais surtout, si le même traitement est effectué sur la question, cela permet d'être beaucoup plus souple et rapide dans la recherche : par exemple, si l'utilisateur effectue une recherche avec un verbe, les documents comportant ce verbe dans toutes ses formes conjuguées seront pris en compte, et pas seulement les documents contenant le verbe dans la forme entrée par l'utilisateur. Pas tous les moteurs de recherche mettent en place actuellement un tel traitement.

2.2.3. L'élimination des mots vides

Cette étape revêt une importance certaine dans la mesure où elle constitue un facteur d'une grande influence dans la précision de la recherche. Le fait de ne pas éliminer les mots vides provoque inévitablement du bruit. L'élimination des mots vides (le, la, et, de, des...) doit se faire aussi bien à l'indexation qu'à l'interrogation (élimination des mots vides de la question) [MEY01].

3. Méthodes d'indexation

3.1 Méthode manuelle

L'indexation manuelle est ajustée par des corrections humaines et est principalement adoptée par les outils de recherche de type thématique (les annuaires). Il s'agit d'un contrôle manuel des informations récoltées soit par un robot de recherche soit par soumission des utilisateurs. Les informations sont ensuite classées et cataloguées par grandes catégories. Les catalogues thématiques ainsi créés sont construits à la main par des personnes filtrant les sites en fonction de leur qualité, pertinence et fiabilité. Cette méthode manuelle apporte une valeur ajoutée certaine mais la mise à jour est moins rapide et la couverture beaucoup moins large. Par ailleurs, l'intervention humaine peut biaiser la couverture géographique, linguistique ou thématique de la base [MEY01].

Le rôle de l'indexeur est d'attribuer au document un certain nombre de descripteurs. On distingue les descripteurs suivants [MEY01] :

3.1.1. Mots-clés unitermes

Ces descripteurs sont formés d'un seul mot. Par conjonction de plusieurs unitermes, on obtient des expressions composées.

3.1.2. Descripteurs composés

Ils sont constitués d'expressions de deux ou trois termes. On peut utiliser des expressions de différents types:

- Nom-adjectif (ex: Droit social).
- Nom-préposition-nom (ex: Histoire de la musique).
- Des termes possédant un trait d'union (ex: Libre-échange) .
- Des termes avec rejet (ex Boole, algèbre de) bien adapté pour les catalogues manuels.
- Des termes avec un qualificatif (ex: Mercure (métal), Mercure (planète))

3.1.3. Descripteurs structurés

Un descripteur structuré contient plusieurs informations sous une même entrée dite "vedette". On fait se succéder les descripteurs dans l'ordre suivant :

- Tête de vedette, significative du sujet.
- Sous-vedette de point de vue.
- Sous-vedette de localisation géographique.
- Sous-vedette de localisation chronologique.
- Sous-vedette de forme (dictionnaire, bibliographie, congrès).

3.1.4. Codes numériques

Les descripteurs peuvent être codés pour représenter simplement des notions difficiles à réduire en quelques mots.

3.1.5. Indices de classification

3.1.5.1. Classification hiérarchiques ou non

On considère que la connaissance est constituée d'éléments emboîtés à différents niveaux. La maintenance de la classification est délicate car elle nécessite de trouver à quel niveau de connaissance correspond le document à insérer. Elle entraîne l'utilisation d'indice très grand.

Exemple : en classification Dewey on a :

5 Sciences

51 Mathématiques

512 Algèbre

513 Arithmétique

3.1.5.2. Classification à facettes

On essaye de regrouper les connaissances sans que chacune soit dépendante d'une autre. Par exemple on utilise le principe de Raganathan que l'on peut décomposer chaque sujet en éléments inclus dans des domaines sémantiques pré-définis :

- Personnalité
- Matière
- Energie
- Espace

3.2. Méthodes automatiques

Pour réduire les coûts et uniformiser l'indexation, on fait appel aux méthodes d'indexation automatiques. Elles sont basées sur l'analyse de la forme de surface, sur les similitudes et sur le sens du document. Les méthodes automatiques sont utilisés en général par les moteurs de recherche généralistes [MEY01].

3.2.1. Indexation en texte intégral

Elle utilise pour l'indexation tous les mots lexicaux (noms, verbes, adjectifs) présents dans le document. Un anti-dictionnaire élimine les mots ayant peu de valeur documentaire (mots vides).

3.2.2. Indexation partielle du document

Elle consiste à n'utiliser que quelques termes jugés significatifs. On s'intéresse à quelques renseignements sur le document : son titre et/ou les premières lignes du texte par exemple.

La qualité de la recherche est évidemment directement proportionnelle à la qualité des renseignements partiels.

3.2.3. Indexation par méthodes statistiques

Les méthodes statistiques partent de deux principes :

- Il existe une relation entre la fréquence d'un terme à l'intérieur d'un document et son importance pour la représentation du document.
- Pour l'indexation du document, il existe une relation inversement proportionnelle entre l'importance d'un terme, et le nombre total de documents de la base de donnée contenant ce terme. Les termes rares seront de ce fait privilégiés.

Le document est indexé par ses ***n*** descripteurs, suivi de leur pondération qui obtenue par produit du facteur d'importance interne au document et du facteur d'importance externe :

- Le facteur d'importance interne c'est la fréquence du descripteur dans le document.
- Le facteur d'importance externe c'est l'inverse du nombre de documents de la base qui comportent ce descripteur.

On définit une qualité de discrimination d'un terme dans la base en distinguant :

- Les termes très fréquents, peu spécifiques qui "attirent" les documents autour d'eux.
- Les termes très rares, utiles pour la précision de la recherche mais pas pour la recherche elle-même, vu qu'ils "rejettent" les documents.
- Les termes de fréquence moyenne qui ont tendance à regrouper des documents relativement semblables.

Ainsi, en ajoutant à la liste des termes pondérés, le regroupement des termes rares et des descripteurs composés constitués des termes peu discriminants, on améliore la couverture de recherche. Ceci se fait au détriment des temps de calculs qui sont à répéter à chaque indexation d'un nouveau document.

3.2.4. Indexation par les citations

Dans cette méthode, on considère qu'un document répond à une question de deux manières :

- Par un apport personnel de la part de l'auteur.
- Par un ensemble d'autres documents cités en relation avec la question.

De même, on indexe le document de la façon suivante :

- Un ensemble de termes d'indexation.
- Un ensemble de documents indexant.

On peut calculer la similitude de deux documents à partir des termes d'indexation et des relations de co-citation qui induisent des agrégations de documents. On distingue trois types de co-citations :

- Un document A cite un document B.
- Les documents A et B sont cités simultanément par un document C.
- Deux documents A et B citent le même document C.

3.2.5. Indexation par des méthodes sémantiques

L'indexation est composée d'une liste de descripteurs adaptés et une représentation du document dans un langage de description tenant compte des relations sémantiques entre termes. Pour qualifier l'environnement sémantique d'un terme, on utilise un thesaurus qui comporte les relations de type hiérarchique, d’équivalence ou d’association.

Les réseaux sémantiques sont en fait, une extension de la notion de thesaurus. Dans ces réseaux, les termes d'indexation sont des nœuds d'un graphe, et les arcs sont porteurs d'une relation sémantique.

L'indexation est donc composée de la liste des descripteurs qui comportent leurs propriétés :

- "est un"
- "sorte de"
- "partie de"
- "synonyme de"

Les descripteurs héritent des propriétés de la classe à laquelle ils appartiennent, en fonction de la relation sémantique.

3.3. Mise à jour des index

Afin de garantir les performances de réponses d'un moteur de recherche, les index doivent être régulièrement mis à jour. L'évolution des outils de recherche permet actuellement une mise à jour en temps réel ou en léger différé. Le procédé de mise à jour diffère entre un moteur de recherche sur Internet et un moteur d'une base de données.

Dans le cas premier, la mise à jour consiste simplement à parcourir la base de données du moteur Internet, d'indexer les nouveaux documents et de supprimer les entrées d'index correspondant à des documents qui n'existent plus. Dans le deuxième cas, les documents nouveaux ou modifiés peuvent être mis à jour à la volée [MEY01].

Au besoin, les index ne doivent pas seulement être mis à jour ou entièrement reconstruits mais également optimisés afin de garantir des taux de réponses fiables et rapides.

4. Structure des fichiers d'index

La structure logique et physique des fichiers d'index dépend de la structure et de la richesse de la base de données adoptée par chaque outil de recherche. La tendance générale est plutôt le recours à l'utilisation des fichiers séquentiels indexés reliés entre eux par des pointeurs telles des tables d'une base de données relationnelles reliées par des clés étrangères [MEY01].

Pour illustrer le principe, on peut simplifier en imaginant trois fichiers :

- le premier contient les mots,
- le second contient une liste ordonnée de mots avec les références aux documents qui les contiennent

- le troisième répertorie, selon l'outil, les positions des mots clés dans la phrase, le paragraphe, la section, le chapitre etc.

Le fait de répertorier les positions des mots dans les paragraphes et les phrases permet à l'outil de recherche de recalculer ces positions lors de la requête de l'utilisateur et de retourner les documents contenant les mots ayant le même degré de proximité que les positions contenues dans l'index.

Exemple : Cas d'un moteur de recherche référençant les mots par numéro de document, numéro de paragraphe, numéro de phrase et la position du mot dans la phrase.

Enregistrement du document dans la base de données
Numéro: 10
Titre: Informatique de gestion
URL: http://www.is-net.ch
Texte: SQL , langage d'interrogation de bases de données.

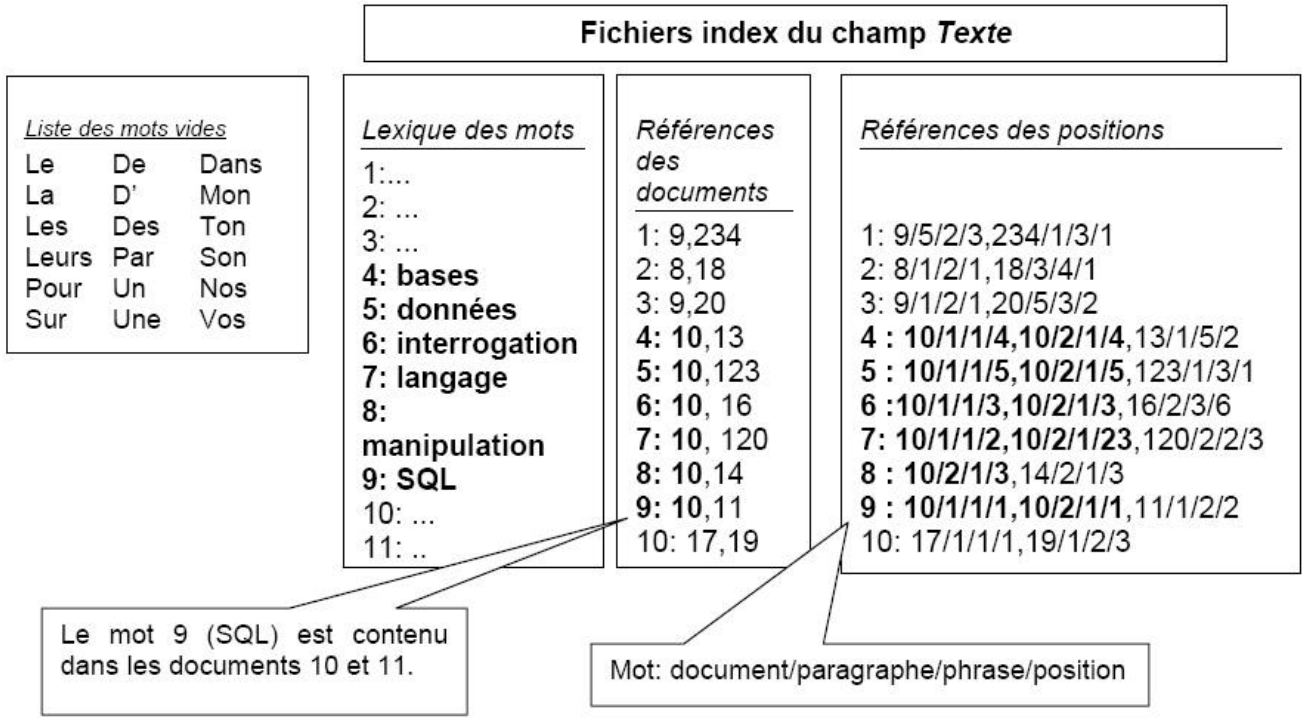


Figure 2. Structure des fichiers d'index

Au cas où l'index des positions serait mis en oeuvre, si l'utilisateur fait une requête sur «langage d'interrogation », le moteur de recherche retournera uniquement les documents contenant les mots «langage » et «interrogation » l'un à côté de l'autre. Un outil de recherche ne gérant pas les positions des mots et se contentant d'utiliser l'index du champ texte seulement retournera tous les documents contenant «langage » d'une part, ceux contenant

«interrogation » d'autre part et ceux contenant «langage » et «interrogation » à la fois qu'ils soient l'un à côté de l'autre ou pas.

5. Le web sémantique

D'après la définition de Tim Berners-Lee « le web sémantique permettra (contrairement au web actuel qui est vu comme un web syntaxique) de rendre le contenu sémantique des ressources web interprétables non seulement par l'homme mais aussi par machine » [TIM99].

L'objectif du web sémantique est donc de tendre vers un web dont la sémantique des données serait à la fois compréhensible par des utilisateurs humains et appréhensible par des entités informatiques (moteurs de recherche, serveurs d'informations, agents de recherche) [BID01].

En ce qui a trait avec l'indexation documentaire, le web sémantique comporte de nouvelles technologies d'indexation et de représentation de l'informations.

5.1. L 'annotation sémantique des documents

L'analyse « full-text » réalisée par les moteurs de recherche a vite montré ses limites et encouragé l'utilisation de méta-données au sein des documents accessibles via le web. C'est ainsi que se sont succédées de nombreuses propositions pour encourager l'ajout de méta-données, certaines juste syntaxique (balise <META> de HTML), d'autres plus précise mais évolutives (exemple : le Dublin Core) et plus récemment des modèles extrêmement souples tels que le RDF (Ressource Description Framework). En ajoutant des méta-données aux documents, on souhaite rajouter au web la sémantique qui lui manquait [TAN00].

L'annotation (le marquage ou le balisage) sémantique des données sur le web ouvre de nombreuses perspectives d'amélioration de la qualité des moteurs de recherche. L'approche la plus populaire consiste, donc, à décrire dans des méta-données le contenu sémantique des pages web. Pour que différentes applications puissent utiliser et échanger ces méta-données, il faut spécifier un modèle pour les représenter. Comme extension au langage XML, le W3C (World Wide Web Constrium) a recommandé le langage RDF pour décrire de telles méta-données et le vocabulaire conceptuel sur le quel reposent ces méta-données peut s'exprimer dans un schéma RDF (RDFS) [ZEB04].

5.2. L'annotation RDF des documents

Le langage XML apporte la structure (syntaxe) comme un arbre de syntaxe abstraite, mais, il n'apporte rien sur la signification, le sens ou la sémantique. La sémantique permet de définir la signification des balises, donc des contenus [TUA01].

Exemple : (en XML)

```
<Livre>
        <Auteur> John Maynard Keynes</Auteur>
        <Titre>General Theory of Employment …</Titre>
….
</Livre>
```

Ou bien

```
<aaaa>
        <bbbb> John Maynard Keynes</bbbb>
        <cccc>General Theory of  Employment …</cccc>
….
</aaaa>
```

Le sens apporté par cette structure syntaxique peut être le suivant :

<Livre> : un livre est un genre de document.

<Auteur>, <Titre> : Les documents ont un auteur, qui est une personne, un titre qui est un littéral.

Le RDF est utilisé pour décrire des annotations sémantiques décrivant le contenu des documents, afin de permettre des recherches d'information en utilisant ces annotations.

Cette description du contenu repose sur un modèle partagé, c'est à dire une ontologie ; cette ontologie est définit par RDF Schema.

Le RDF repose sur un modèle de triplet : [Sujet, Prédicat, Objet] qui représente une description (ou méta-donnée) sur la ressource.

- Sujet est la ressource que l'on veut décrire.
- Prédicat est une propriété de la ressource.
- Objet est la valeur pour la propriété.

Le triplet est présenté sous la forme de graphe orienté. Sujet et objet sont des nœuds, prédicat est un arc de sujet à objet.

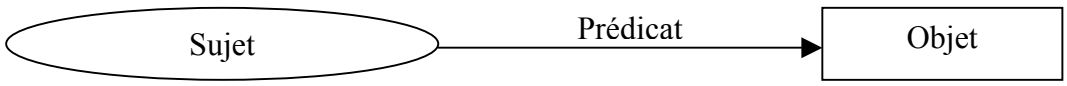


Figure 3. Graphe d'un modèle RDF.

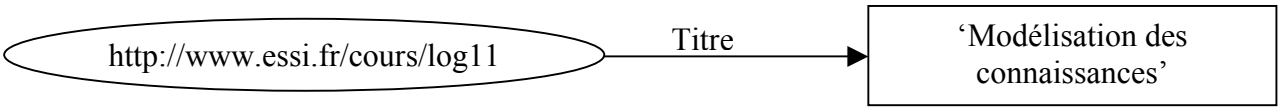
Alors, on dit que *sujet* a une propriété nommée *prédicat* avec la valeur *objet.*

Exemple :

Une description « modélisation des connaissances » est un titre de document de la page web http://www.essi.fr/cours/log11

Titre(http://www.essi.fr/cours/log11, 'modélisation des connaissances')

Peut être modélisée par le paragraphe suivant :



La description RDF en syntaxe XML est la suivante :

<rdf : Description about= http://www.essi.fr/cours/log11>
<titre> modélisation des connaissances </titre>
</rdf : Description>

En ajoutant du sens à cette description, par exemple

- La ressource est cours,
- Un cours a des enseignants
- L'enseignant est un chercheur de l'INRIA, etc.

L'annotation sémantique devient donc :

```
<rdf : Description about= http://www.essi.fr/cours/log11>
<rdf : type rdf :resource='#Cours'/>
<titre> modélisation des connaissances </titre>
<num>Log11</num>
<enseignant>
   <chercheur>
        <rdf :about=http://www.inria.fr/Olivier.Corby'>
        <nom> Olivier Corby</nom>
        <institut>INRIA</institut>
  </chercheur>
</enseignant>
</rdf : Description>
```

La modélisation RDF ne donne que la capacité d'échange de modèle. Elle ne permet pas à l'utilisateur de définir des vocabulaires qui portent une sémantique de description de ressources. Un schéma RDF avec des concepts de bases peut offrir cette capacité. Le schéma est lui même défini sous RDF et on l'appelle RDFS. On peut distinguer trois sortes de concepts :

- Concepts élémentaires
- Concepts pour la définition de schéma
- Concepts utilitaires.

Ces concepts sont définis par l'utilisateur sous forme de classes & sous-classes pour les sujets, les objets et les prédicats.

Plus récemment le W3C a recommandé une extension pour RDF & RDFS pour l 'annotation sémantique des documents, c'est le langage DAML avec une base de connaissances ontologique OIL.

5.3. Le Web Sémantique et la notion d'ontologie

5.3.1. Les ontologies

Une ontologie est définie comme la conceptualisation des objets reconnus comme existant dans un domaine, de leurs propriétés et de connaissances heuristiques associées. Elle permet de représenter la connaissance d'un domaine sous un format informatique en principe utilisable pour différentes applications. La conceptualisation consiste en un ensemble d'objets, les concepts, et en un ensemble de relation entre les concepts. Une ontologie peut prendre la forme d'une hiérarchie de termes atomiques ou bien accompagnés de définitions, de schémas conceptuels spécifiant la structure de certaines connaissances, d'une théorie, d'un ensemble de règles logiques définissants un ensemble de termes et de contraintes sur leur utilisation. [SYL00] (un état de l'art sur les ontologies sera mis en œuvre dans ce qui suit de ce mémoire).

5.3.2. L'exploitation des ontologies dans le web sémantique

Dans le cadre du web sémantique, beaucoup de travaux cherchent à localiser la connaissance en effectuant des annotations de documents. Ces annotations sont actuellement ajoutées manuellement aux documents. Cette approche est difficilement applicable à grande échelle sur une grande quantité de documents dont le contenu peut varier au cours du temps. D'où la nécessité d'utilisation des méthodes automatiques pour la localisation de connaissances dans les documents web.

La localisation des informations s'effectue sous la forme d'une indexation automatique des documents. Cette indexation est relative à une connaissance dans un domaine particulier. Cette connaissance sera, donc, représentée par une ontologie du domaine (dite « orientée terminologie »).

Dans telle ontologie, chaque étiquette de concepts est étendue par tous ses synonymes possibles. Les étiquettes sont désambiguïsées en utilisant un thésaurus (WordNet « Voir glossaire »). Les concepts des pages web sont repérés en utilisant des outils issus du T.L.N, en

utilisant un thésaurus général et en exploitant une mesure permettant de déterminer, pour un concept donné, son potentiel de représentativité du contenu de la page.

Ensuite, les concepts de l'ontologie permettent d'indexer le document en fonction des concepts qu'il contient. Cette méthode d'indexation permet aussi d'évaluer les documents par rapport à un domaine donné.

L'intérêt principal de cette méthode est son utilisation dans la construction d'une réponse à une requête. En effet, la recherche d'un concept du domaine est immédiate et les opérateurs booléens peuvent être modifiés pour prendre en compte la structure de l'index et les relations définies dans l'ontologie [IRI01].

Une application importante de cette méthode se situe dans le cadre du recrutement sur internet, dont l'objectif est de repérer dans une grande masse de CV en ligne, ceux répondant à un certain nombre de critères de compétence dans un domaine particulier. (Projet COMMONCV) [ZEB04].

6. Conclusion

A travers ce chapitre nous avons vu les différents modes d'indexations des documents électroniques dans les différents outils de recherche d'informations. Malgré l'évolution technologique employée dans ces outils le problème de dépistage de l'information pertinente se pose toujours.

Les méthodes d'indexation manuelle sont bonnes pour un lot très restreint d'informations (tel que les annuaires ou les systèmes spécialisés), elles ne peuvent être généraliser pour une base de données aussi grande tel que celle des moteurs de recherche généralistes.

Les méthodes automatiques ont montré leurs limites sur le plan pratique tel que l'on connaît au jour d'huit.

Les nouvelles technologies d'organisation, d'annotation et d'indexation des ressources sur internet dans le cadre du web sémantique permettent de rendre le contenu sémantique des ressources web interprétables non seulement par l'homme mais aussi par les machines. Le web sémantique est en cours de développement et reste le rêve du W3C.

Chapitre 4

Quelques travaux sur le filtrage de l'information

1. Introduction

La quantité d'information disponible sur le Web est importante et elle ne cesse de croître. La recherche d'information demeure problématique : il est en effet difficile de trouver ce que l'on recherche malgré l'existence de sites et de moteurs de recherche. Les moteurs de recherche par mots clés renvoient généralement comme réponse à une requête un grand nombre de pages à consulter, ce qui demande à l'utilisateur de faire lui-même le tri et le filtrage dans cette masse d'information

Dans ce chapitre nous exposons quelques approches et techniques pour remédier au problème de filtrage.

2. La sous-utilisation des SRI et la surcharge d'information

Comme nous avons motionné dans le premier chapitre, la qualité d'une recherche d'information peut s'exprimer en termes de silence (le rappel) et de bruit (la précision). Le bruit correspond aux résultats non pertinents trouvés par l'utilisateur tandis que le silence correspond aux résultats pertinents non trouvés. Les utilisateurs des différents outils de recherche disponibles sur Internet se trouvent confrontés au quotidien à ces deux paramètres. Qui ne s'est jamais retrouvé face à un nombre faramineux de réponses (évoquant un bruit important) ou un trop faible nombre de réponses à la suite de sa requête (signe d'un silence important)? [IHA00] Cette surabondance ou l'absence de réponses dépend principalement de deux facteurs :

- La qualité de la requête formulée par l'utilisateur. En effet, l'utilisation d'un terme non approprié dans une requête peut aboutir à la présence de bruit si le terme utilisé est trop commun ou à du silence si le concept recherché peut être décrit par plusieurs synonymes et qui est utilisé dans la requête un synonyme peu usité. Les opérateurs booléens permettent, sous certaines conditions, de lutter efficacement à la fois contre le bruit (opérateurs ET et SAUF) et le silence (opérateur OU). L'utilisation de vocabulaire contrôlé de type Thésaurus, hélas très rarement disponible sur les outils de recherche sur internet, contribue également à augmenter la qualité des requêtes.
- La qualité de l'index interrogé (nombre de documents, qualité intrinsèque et spécificité des documents présents dans cet index). Si l'on suppose que la requête est parfaitement formulée, le nombre de réponses dépend du nombre de documents présents dans l'index de l'outil de recherche interrogé, et donc directement de sa nature (moteur, méta-moteur ou annuaire). Ainsi, dans le cas des moteurs et des méta-moteurs généralistes, le nombre de réponses à une requête peu spécialisée est généralement extrêmement élevé (parfois

jusqu'à plusieurs millions!), avec habituellement un bruit important sur l'ensemble des réponses. En revanche, l'avantage de ces outils est de fournir toujours des réponses, même si le sujet de la requête s'avère extrêmement spécialisé. Dans le cas des annuaires généralistes, le nombre de réponses est toujours plus faible par rapport aux moteurs et aux méta-moteurs du même type. Si cela s'avère être un avantage indéniable dans le cas de requêtes peu spécialisées (car le nombre de réponses proposé est non seulement plus faible, mais aussi limité aux seuls sites répondant à certains critères de qualité pour lesquels ils ont été retenus), cela s'avère être un inconvénient pour des requêtes spécialisées qui produisent souvent très peu ou pas de réponses.

Les travaux de [JON99], de [SPI99] et de [BUR97] sur l'usage des moteurs de recherches ont montré que les ressources du système sont sous-utilisées et que les outils mis à la disposition de l'usager final pour explorer le nombre élevé de réponses sont insuffisants et inadaptées [IHA00]. Les requêtes des usagers sont pauvrement formulées (moins de deux termes et absence d'opérateurs booléens) et ceux-ci ne visualisent pas plus de deux pages web. La richesse des réponses obtenue est ignorée. Dans le cas du moteur de recherche ALTAVISTA [SIL98], 85% des usagers se contentent des dix premiers résultats fournis sur la première page et 78% des requêtes ne sont pas modifiées dans le but de les améliorer. Les tactiques élaborées pour réduire ce problème de surinformation sont rudimentaires. Seulement un usager sur deux tente de réduire le nombre de réponses en ajoutant souvent un terme à l'équation d'origine. Les usagers préfèrent changer le contenu de la requête plutôt que de modifier sa structuration logique, ce qui pose le problème de la pertinence des outils logico-analytiques mis à leur disposition et suggère d'autres modalités d'exploration.

3. Les approches du problème

La tendance de l'informatique documentaire aujourd'hui est de répondre à ce problème en se centrant sur le rôle de l'usager pour filtrer, adapter, personnaliser sa recherche. Placer l'usager final au centre des études est devenu l'une des évolutions les plus marquantes ces dernières années en informatique documentaire.

Essentiellement, trois directions se font jour à ce sujet :

1. Soit l'on offre à l'usager un ensemble d'outils pour qu'il construise lui-même son parcours de recherche. L'usager reste maître de sa recherche : l'ambition du système est de lui proposer des filtres afin qu'il puisse construire un parcours au fur et à mesure de l'évolution de son besoin d'information. L'information est organisée à la sortie et c'est

l'usager qui a la tâche de trier cette masse d'information : c'est l'approche "filtrage d'information".

2. Soit l'information est organisée au moment où on la saisit dans le système de recherche : c'est l'approche "métadonnées" et catalogage qui permettent de simplifier la récupération ultérieure de l'information.
3. Soit on pré-calibre la recherche de l'usager en fonction d'une connaissance de celui-ci obtenue directement à partir de questions qui lui sont posées (diffusion sélective d'information traditionnelle par l'intermédiaire de déclarations de profils). Il reçoit des réponses sans qu'il ait la maîtrise de l'évolutivité de son besoin d'information [IHA00].

4. Expérimentations de techniques de filtrage

4.1 Proposition de filtrage par les langages documentaires : Le système Cathie.

Une des solutions au problème de surcharge d'information, consiste à construire des interfaces utilisateur qui regroupent automatiquement les résultats en catégories. Cette méthode (clustérisation) est apparue avec l'introduction du modèle vectoriel. Elle a été constamment améliorée. Plusieurs travaux récents ont permis de donner un support visuel à cette catégorisation des documents. Le professeur Khorflag (1998) a effectué, à l'Université de Pittsburg, un ensemble de recherches pour mettre au point des interfaces facilitant la visualisation de l'information, et par conséquent, le repérage. Il a conçu trois prototypes : VIBE (*Visual Information Browsing Environment*), GUIDO (*Graphical User Interface for Document Organization*) et BIRD (*Browsing Interface for Retrieving Documents*). On peut aussi citer les prototypes suivants :

- TileBars [HEA95]
- Scatter/Gather [CUT92]
- InfoCrystal [SPO93]

Dans cette approche les chercheurs ont utilisés la classification de Dewey. La classification hiérarchique de Dewey exemplifie deux fonctions : la collacation (inclusion) et la partition (exclusion). L'inclusion rapproche les objets et les idées semblables. Mais dans un domaine d'information très vaste, il est tout aussi important d'exclure l'information non désirée qu'inclure ce qui est recherché. La partition peut être opérée en divisant une grande quantité d'information en parties plus petites comme moyen d'isoler la partie qui a la plus grande probabilité d'être pertinente. [CHA95].

Les récents travaux de [IHA00] ont tenté d'apporter des contribution à ces problèmes de surcharge en utilisant un certain nombre d'approches permettant d'assister l'usager dans le

choix des termes et dans l'élaboration de stratégies de recherches. Le prototype CATHIE (CATalogue Hypertexte Interactif et Enrichi) proposé par [IHA00] tend à remédier en partie à divers déficits que les analyses d'usage ont mis en évidence. Il associe la richesse des vocabulaires contrôlés, les possibilités de visualisation et de navigation de l'hypertexte et la puissance du modèle probabiliste. Après une requête, l'usager peut exploiter ces quatre stratégies :

- Effectuer une recherche dans le lot de documents trouvés
- Reformuler la question à travers les vedettes matières proposées
- Établir un filtrage thématique des termes et/ou documents
- Afficher la notice et voir les documents similaires.

[IHA00] ont décidé que les vedettes matières issues de la liste RAMEAU doivent être regroupées selon le champs sémantique. Deux types de classification sont effectuée. La première concerne un calcul de fréquence des vedettes matières construites (VMC) sur un lot de documents. La seconde consiste à structurer ces VMC selon les domaines. De ce fait, plus de sémantique ont été intervenu en établissant une classification des vedettes par domaine.

Exemple d'utilisation de système CATHIE

Après une recherche sur la question "access" et si l'usager décide de spécifier les termes et les réponses relatifs au domaine informatique, CATHIE affiche ces nouveaux dossiers (figure 1).

Indiquer le domaine (Ici informatique)

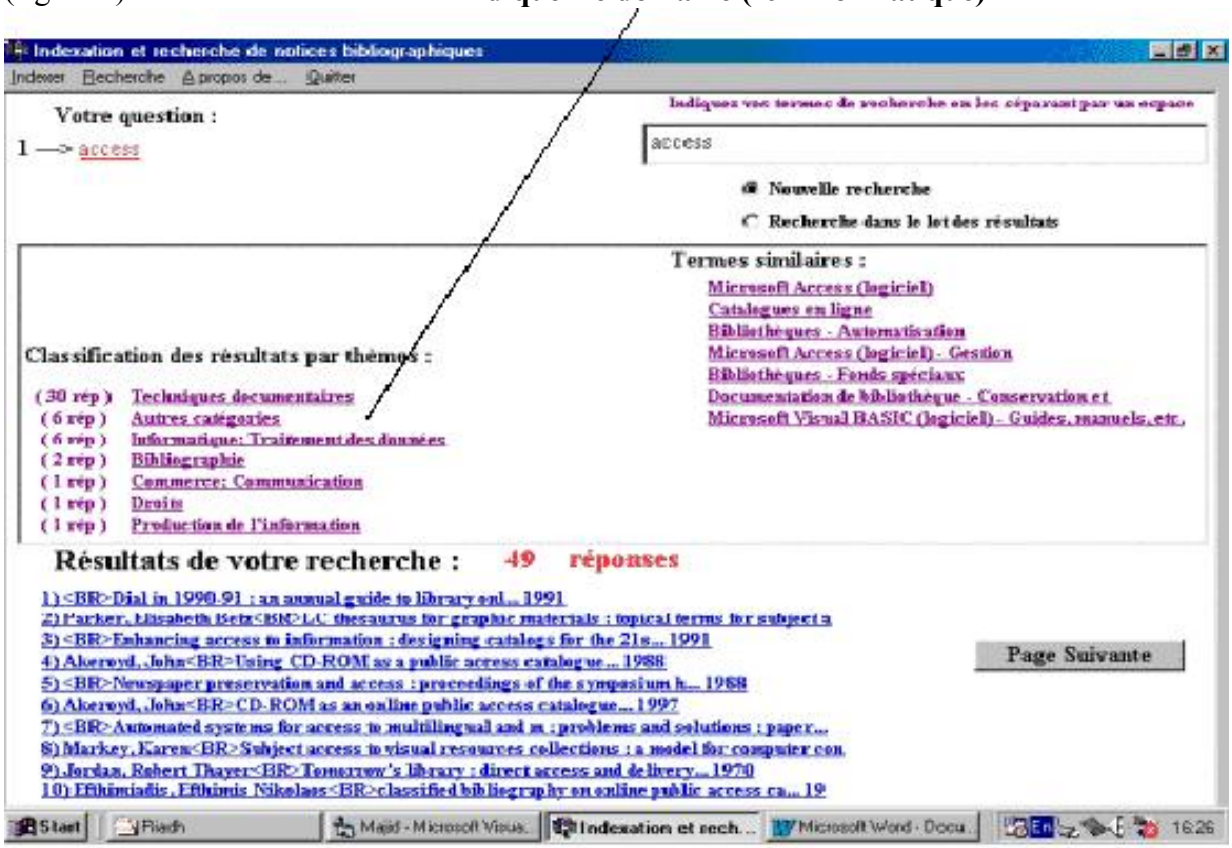


Figure 1. Filtrage d'information dans CATHIE (exemple 1)

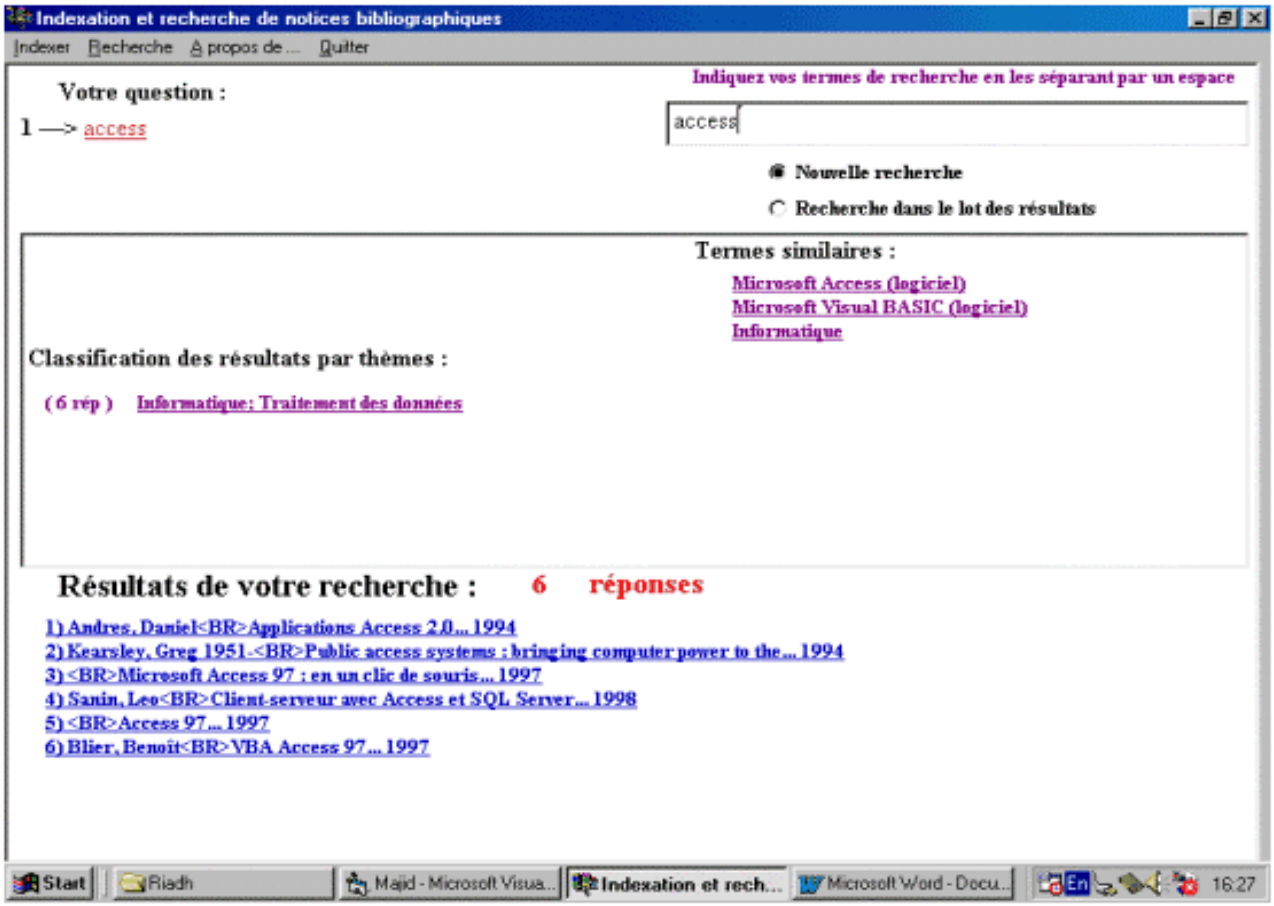


Figure 2. Filtrage d'information dans CATHIE (exemple 2)

- Le prototype CATHIE est encore expérimental et de nombreuses améliorations d'ordre techniques pourraient lui être apportées.

L'étude de ces nouveaux modèles d'interaction (visualisation de l'information, catégorisation thématique, reformulation interactive, etc.) est encore récent.

4.2. Un prototype de système de recherche d'information personnalisé selon le profil des utilisateurs : Le système Profil-Doc

La conception du système de recherche d'information Profil-Doc s'appuie sur deux hypothèses:

Un découpage et une caractérisation des documents

Une identification du profil de l'utilisateur.

L'étude a été restreinte à l'**I**nformation **S**cientifique et **T**echnique (IST), le public visé quant à lui, est composé de chercheurs en IST.

Brièvement, les fondements du projet Profil-doc sont basés sur le fait que l'augmentation continuelle de la masse d'information à consulter rend de plus en plus pénible la recherche de l'information pertinente, ceci est d'autant plus vrai lorsque l'on consulte des

bases de données en texte intégral. Le sens d'un texte étant donné, non seulement par son contenu mais aussi par sa structure, l'idée dominante de Profil-doc est que les parties de document auront un usage différencié a priori suivant le besoin de l'utilisateur [CHR00].

Dans ce cadre, un prototype de système, alliant une description du document et une caractérisation de l'usager. Ainsi, en fonction de l'usager et de son profil, le système, pour une requête particulière, n'utilisera que certaines parties du document initial.

Les parties du texte exploitées en fonction de la structuration du document sont appelé les **U**nités **D**ocumentaires (UD).

4.2.1 Hypothèses théoriques

Propriétés de description des UD

Outre les propriétés classiques de description des documents (titre, auteur, pays, année, environnement de production, environnement de diffusion) , les unités documentaire sont repérées et caractérisées selon trois propriétés [CHR00] :

Type	résumé Table des matières introduction description de contexte description de thème environnement expérimentation résultats discussion méthode conclusion bibliographie
Forme discursive	descriptif, narratif, argumentatif, discours rapporté
Style de présentation	Littéraire littéraire avec données numériques données numériques calcul représentation

Profil de l'utilisateur

Les utilisateurs se définissent selon quatre propriétés :

Niveau Educationnel	Maîtrise / DEA / Recherche
Champ Disciplinaire	SIC / Informatique / Agronomie / Pharmacie....
Etapes de recherche	Constitution d'une bibliographie Définition du sujet Faisabilité

	Expérimentation Interprétation des données Rédaction Repérage des approches expérimentales Plan de travail Compréhension de la problématique Etat de l'art Synthèse bibliographique Dégagement des nouveaux axes de recherche Mise à jour des connaissances
Type de recherche	Recherche pointue Recherche généraliste

La définition de ces modalités a été proposée à partir d'un questionnaire effectué auprès de chercheurs en SIC, sciences pharmaceutiques et sciences physiques.

Le filtrage de l'information

L'enquête précédemment notifiée avait aussi pour objectif de faire apparaître certain usages de lecture de l'IST par les chercheurs, en particulier, quel pouvait être les liens entre les caractérisations de parties de discours utiles ou non en fonction des caractérisations de profil d'utilisateurs.

Les résultats ont été stockés dans une matrice, la matrice d'association. Les valeurs 0, 1 ou * contenues dans chacune des cases reflétant respectivement un lien établit, aucun lien, un lien non définit entre une modalité de caractérisation d'UD et une modalité de description de profil d'utilisateur. La stratégie de filtrage proposé dans ce système de recherche est une stratégie à deux temps pour sélectionner les UD à extraire de la base.

1. Il construit un *vecteur résultant* qui renseigne les propriétés valides pour le profil de l'utilisateur.
2. Il extrait les UD de la base selon différentes stratégies documentaires régies par une *fonction d'aiguillage*.

Trois méthodes théoriques sont proposées pour construire le vecteur résultant et 9 fonctions d'aiguillages permettent d'extraire les UD de la base, soit au total **27 stratégies de filtrage différentes**.

4.2.2 Expérimentation

Les hypothèses de Profil-doc étant théoriques, un prototype de système a été construit pour permettre d'en tester l'impact lors d'une recherche d'information.

Le Prototype Profil-doc

Ce prototype est composé de :

- Une base de données d'UD structurées suivant des caractéristiques de description présentées précédemment. Le système utilisée pour indexer et stocker cette base est le système SPIRIT qui a la particularité d'utiliser conjointement une analyse statistique et linguistique. Grâce au module SPIRIT-W3, développé par la branche DIST du CEA de Saclay et commercialisé par TGID, cette base est interrogeable sur Internet.
- Une interface d'interrogation (consultable à l'adresse http://recodoc.univ-lyon1.fr/cgi-bin/nphconex) permettant d'une part de poser une question en langage naturel et d'autre part de filtrer les UD selon trois *stratégies documentaires* privilégiant soit le document (avec l'environnement de production, l'environnement de diffusion) ou l'unité documentaire (avec la caractérisation de l'UD selon le type, la forme discursive et le style).
- Une interface de consultation dans laquelle les résultats se présentent sous forme d'une liste ordonnée d'UD triées suivant une valeur de pertinence calculée par SPIRIT [RAD88], le texte de l'UD étant accessible à partir du titre par lien hypertextuel. Ce type d'interface a été développée de manière a pouvoir utiliser l'application BeFor [CHA96] permettant l'interrogation en rafale de moteurs de recherche sur Internet. En effet, pour tester les hypothèses de filtrage 3 utilisateur ont été simulé : ayant des profils différents, utilisant 8 des 27 stratégies de filtrage et posant chacun 652 questions. Donc au total il y'a plus de **15000 interrogations différentes** [MIC99]. Pour chaque interrogation la distance est calculé, distance entre la réponse du système et la réponse qui aurait été obtenue sans aucun filtrage. Ce type de calcul permet d'évaluer le facteur d'impact des stratégies.

Les résultats

Les profils se distribuent différemment selon les fonctions d'aiguillage, aucune d'elles ne semble privilégier un profil particulier. Moins la fonction d'aiguillage est complexe et moins elle est fédératrice de réponses différentes, c'est à dire plus son facteur d'impact est petit. Une stratégie documentaire fondée sur la caractérisation des unités documentaires produit des résultats plus personnalisés qu'une stratégie fondée sur la caractérisation du document (par exemple le type d'environnement éditorial). Parmi les stratégies documentaires fondées sur la caractérisation des unités documentaire, le type d'unité documentaire est plus discriminant que la forme discursive ou le style [CHR00].

5. Conclusion

Ce chapitre constitue notre point de départ (ou un bout de fil) du chemin de la solution à proposer. Depuis la naissance des SRI les utilisateurs se trouvent confrontés au quotidien au problème de surcharge d'informations (des informations bruitées). Deux grands axes se font jour à ce sujet : soit l'on offre à l'usager un ensemble d'outils pour qu'il construise lui-même son parcours de recherche, soit on pré-calibre sa recherche en fonction d'une connaissance de l'usager obtenue directement par l'intermédiaire de déclaration de son profil.

Dans la première direction l'usager reste maître de sa recherche : l'ambition du système est de lui proposer des filtres afin qu'il puisse construire un parcours au fur et à mesure de l'évolution de son besoin d'information. Dans la seconde, il reçoit des réponses sans qu'il ait la maîtrise de l'évolutivité de son besoin d'information.

Donner un rôle à l'usager dans la recherche d'information (première direction), rendre le système interactif, c'est de lui permettre d'agir sur la sortie des résultats grâce à des mécanismes qui regroupent et ordonnent les informations trouvées.

Partie 2

Le Système de filtrage proposé

Chapitre 1

Les ontologies

1. Introduction

Dans ce chapitre nous donnerons des définitions et des notions fondamentales relatives aux ontologies. Les ontologies ont été introduites en Intelligence Artificielle (IA) il y a 15 ans, le terme d'ontologie est cependant usité en philosophie depuis le XIXième siècle. Dans ce domaine, l'Ontologie désigne l'étude de ce qui existe, c'est à dire l'ensemble des connaissances que l'on a sur le monde [FED02]. En IA, de façon moins ambitieuse, on ne considère que *des* ontologies, relatives aux différents domaines de connaissances. C'est à l'occasion de l'émergence de l'Ingénierie des Connaissances que les ontologies sont apparues en IA, comme réponses aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques.

2. Notion d'ontologie

Parmi toutes les définitions du terme ontologie de données dans la littérature, la plus citée est celle de T. Gruber : « Une ontologie est une spécification explicite d'une conceptualisation, c'est à dire, une description d'une partie du "monde" (un domaine) en termes de concepts et de relations entre ces concepts. » [MAR01].

L'ontologie a un rôle clé dans la représentation et l'utilisation des connaissances. Elle fournit une définition cohérente et non ambiguë du vocabulaire utilisé pour représenter la connaissance, mais elle ne se limite pas à une simple liste de termes ; elle doit aussi fournir l'interprétation sémantique de ces termes.

Les **concepts,** dans une ontologie, représentent les objets, les notions ou les idées du domaine.

Les **relations** représentent les liens conceptuels pouvant exister entre les concepts [FED04].

Exemple : Par exemple, dans le domaine de la géométrie, les concepts de ((*Point*)), de ((*Droite*)) et de ((*Plan*)) doivent être pris en compte, ainsi que la relation ((*appartenir-à*)) entre un *Point* et une *Droite* ou un *Plan*.

3. Que représente-t-on dans une ontologie ?

Quatre grands types de caractéristiques nous permettent de préciser ce qui peut être représenté dans une ontologie ainsi que le processus de modélisation :

3.1. Le type d'ontologie

Les méthodes en Ingénierie des connaissances ont répertorié plusieurs types d'ontologie liés à l'ensemble des objets conceptualisés et manipulés au sein d'un SBC:

(1) L'ontologie du domaine, conçues dans des domaines variés : la médecine, l'automobile, la chimie, l'aéronautique….

(2) L'ontologie générique (ontologies de haut niveau « top-level »), qui repère et organise les concepts les plus abstraits du domaine,

(3) L'ontologie d'une méthode de résolution de problème où le rôle joué par chaque concept dans le raisonnement est rendu explicite (*p. ex. signe* ou *syndrome* dans le cadre du raisonnement médical),

(4) L'ontologie d'application qui se veut une double spécialisation : de l'ontologie du domaine d'une part et d'une ontologie de méthode, d'autre part.

(5) Finalement, l'ontologie de représentation qui repère et organise les primitives de la théorie logique permettant de représenter l'ontologie (*p. ex.* la *frame ontology* d'ONTOLINGUA de Gruber ou l'ontologie de « propriétés » de Guarino & Welty).

3.2. Les propriétés

Une ontologie est non seulement le repérage et la classification des concepts mais c'est aussi des caractéristiques qui leur sont attachées et qu'on appelle ici des propriétés. Ces propriétés peuvent être valuées. En s'intéressant aux taxinomies en sciences naturelles, les vertébrés ont un tégument (la peau) comportant des poils – p. *ex.* pour les mammifères – ou des plumes – *p. ex.* pour les oiseaux. Dans une ontologie sur le monde animal, on pourra ainsi avoir les concepts de « mammifère » ou « d'oiseau » pour lesquels est précisé le type de tégument, respectivement à poil et à plume. En pratique, un attribut « tégument » pourra être attaché aux concepts et sa valeur variera suivant le concept auquel on fait référence [OLF03].

3.3. La relation « is-a »

La relation de subsomption *is-a (est-un)* qui définit un lien de généralisation – *i.e. hyperonymie*– est utilisée pour structurer les ontologies. Cette relation qui permet formellement l'héritage de propriétés est un choix qui s'impose depuis ARISTOTE. Elle doit être complétée par d'autres relations pour exprimer la sémantique du domaine [OLF03].

3.4. Les autres relations.

Les relations unissent les concepts ensemble pour construire des représentations conceptuelles complexes qui vont être autant de connaissances nécessaires au SBC que l'on construit. Si la connaissance construite correspond à un concept dans le monde modélisé, celui-ci est dit *défini*, à l'opposé des concepts insérés dans l'arborescence de l'ontologie qui sont dits *primitifs*. Par exemple, si l'on définit l'*appendicite* comme une *inflammation localisée-sur* l'*appendice*, c'est un concept dit défini. Dans l'exemple précédent, *localisée-sur*

est une relation binaire qui se définit par les concepts qu'elle relie et par le fait qu'elle est, comme les concepts, insérée dans une hiérarchie, ici de relations.

La relation *is-a* qui structure l'ontologie est une relation du même type que les autres. Elle a cela de spécifique que c'est elle qu'on a justement choisi comme relation de structuration de l'arborescence ontologique. Elle est donc implicite dans cette ontologie. Au niveau des choix, il faut aussi remarquer que les concepts et relations de l'ontologie sont duals l'un par rapport à l'autre. Un concept primitif pourrait être un concept défini, une relation pourrait se retrouver implicitement définie au sein d'un concept primitif. Ce sont les choix assumés du concepteur de l'ontologie qui auront permis de décider de ce qui est essentiel – et donc primitif – ou non. Ainsi, on peut décider que le fait, pour un être humain, d'être un étudiant est temporaire donc non définitoire. On caractérise alors les êtres humains avec une relation de rôle social qui permettra de préciser une fonction d'étudiant ou de professeur.

Un autre choix de conception qui doit être fait durant la conception d'une ontologie est de décider si une connaissance doit être modélisée dans une propriété ou à l'aide d'une relation pointant sur un autre concept. Un critère peut être de dire que c'est une propriété dès lors que les valeurs possibles sont d'un type dit primitif (entier, chaîne de caractères), et c'est une relation dès lors que les valeurs possibles sont d'un type dit complexe c'est-à-dire un autre concept de l'ontologie. Mais cette frontière peut aussi être remise en question.

Enfin, dans certains cas, il peut être nécessaire de compléter la structuration de l'ontologie par la relation *is-a* avec une relation de partie-tout ou *méronymie*. Ce type de relation est, par exemple, indispensable en anatomie médicale où il est nécessaire de décrire des organes ou des systèmes et ce qui les compose. Cette relation n'est pas sans poser des problèmes de modélisation dans la mesure où elle est, selon les situations, transitive ou intransitive [OLF03].

Les réflexions sur les ontologies dans le contexte du Web sémantique s'appuient sur ces différents acquis en notant que dans le cas où l'ontologie est utilisée comme repérage et structuration de méta-données, le fait qu'elle permette de faire des inférences est moins mis en avant et donc moins recherché que dans le cas où l'ontologie est utilisée au sein d'un module logiciel type SBC, nécessitant justement d'effectuer des inférences.

4. La construction des ontologies

La construction d'une ontologie formelle se révèle être un exercice délicat. La difficulté dépend, bien sûr, de la taille de l'ontologie à construire. Mais les problèmes ne sont pas forcément là où on les attend a priori ! L'expérience montre en effet que les problèmes de

modélisation des connaissances ontologiques (comment « mettre en primitives » les connaissances, c'est-à-dire décider quels sont les concepts et les relations, et quelle est leur notion ?) sont généralement plus difficiles à résoudre que les problèmes de **représentation** (comment coder les connaissances dans les constructions du langage opérationnel ?).

Différentes méthodologies ont été élaborées pour la création d'ontologies. Parmi les projets qui ont proposé des méthodes et des outils de construction d'ontologies, citons : Methontology, (KA)2, Terminae, Ontolingua, …. Chaque méthode définie un processus de développement. Une généralisation de ces processus de mise au point d'une ontologie repose sur l'enchaînement des phases suivantes [TIX01] : 1) l'**acquisition** des connaissances ; 2) la **modélisation** des connaissances ; 3) la **représentation** des connaissances.

Ces étapes sont sous le contrôle de l'application pour laquelle l'ontologie est construite. L'application détermine les connaissances à acquérir et guide ensuite les choix de modélisation, puis de représentation.

Pour chacune de ces étapes, l'Ingénierie Ontologique – sous discipline de l'Ingénierie des Connaissances concernée par la construction et la maintenance des ontologies – offre des méthodes et des outils pour assister le « développeur » (nom que nous donnons dans cette session à la personne chargée de la construction de l'ontologie) [SEB00].

4.1. Acquisition des connaissances

Acquérir les connaissances ontologiques pour une application suppose de répondre aux deux questions suivantes :

- Quels concepts **existent** dans le domaine concerné par le développement de l'ontologie ?
- Quels concepts sont **pertinents** vis-à-vis de l'application ?

Un domaine correspond à un ensemble de pratiques, réalisées par des personnes que nous appellerons « experts » (il peut s'agir de techniciens, d'ingénieurs, de gestionnaires, *etc*.). Identifier les concepts d'un domaine revient à étudier la façon dont ces experts conceptualisent leurs pratiques. À défaut de pouvoir accéder directement à leurs représentations mentales, on accède à la façon dont ils **parlent** et **rendent compte** de leurs pratiques. Ainsi, les concepts que l'on cherche à identifier sont exprimés en langue et c'est l'**analyse d'expressions linguistiques** qui doit permettre de révéler les concepts [SEB00].

4.2. Modélisation des connaissances

La modélisation des connaissances consiste à « mettre en primitives » les entités conceptuelles identifiées à l'étape précédente, c'est-à-dire à décider quels sont les concepts et les relations, et à déterminer leur notion. À cette étape, le développeur ne s'embarrasse pas

des contraintes liées à l'utilisation d'un quelconque langage formel : ne mélangeons pas tous les problèmes ! D'autant que la modélisation des connaissances conduit, en général, à effectuer un choix entre plusieurs alternatives [SEB00].

4.3. Représentation formelle des ontologies

Les ontologies sont représentées au moyen de langages formels dédiés, offrant des structures de données adaptées à la représentation de concepts. Parmi ces langages, on distingue :

- Les **langages d'échange** d'ontologies sur le Web, dont la syntaxe est basée sur le langage XML.
- Les **langages opérationnels** qui implémentent les ontologies à des fins d'inférences, pour constituer un composant d'un système d'information.

Les langages opérationnels se distinguent, à leur tour, par les services inférentiels qu'ils apportent. De manière générale, ces services permettent de raisonner sur :

- Le contenu de l'ontologie elle-même, pour en vérifier la cohérence et aider à sa construction.
- Des données exprimées au moyen des notions de l'ontologie, pour déduire de nouvelles connaissances.

4.4. Quelques bons principes

T. Gruber propose ainsi un certain nombre de principes respecter pour construire une ontologie [CHA02]:

Clarté. Les ambiguïtés doivent être réduites, quand une définition peut être axiomatisée, elle doit l'être. Dans tous les cas, des définitions en langage naturel doivent être fournies.

Cohérence. Une ontologie doit être cohérente. Les axiomes doivent être consistants. La cohérence des définitions en langage naturelle doit être vérifiée autant que faire se peut.

Extensibilité. L'ontologie doit être construite de telle manière que l'on puisse l'étendre facilement, sans remettre en cause ce qui a déjà été fait.

Biais d'encodage minimal. L'ontologie doit être conceptualisée indépendamment de tout langage d'implémentation. Le but étant de permettre le partage des connaissances (de l'ontologie) entre différentes applications utilisant des langages de représentation différents.

Engagement ontologique minimal. Une ontologie doit faire un minimum d'hypothèses sur le monde : elle doit contenir un vocabulaire partagé mais ne doit pas être une base de connaissances comportant des connaissances supplémentaires sur le monde à modéliser.

5. La réutilisation d'ontologies

Une aide à la modélisation des connaissances consiste à réutiliser des ontologies. Les enjeux de la réutilisation sont de permettre d'élaborer des ontologies de plus grande taille et de meilleure qualité, tout en réduisant les coûts de développement. Toutefois, la réutilisation a elle-même un coût et, avant de pouvoir importer des parties d'ontologies existantes dans une nouvelle ontologie, le développeur se doit de réaliser les tâches suivantes :

- **Localiser** des ontologies candidates à la réutilisation.
- **Abstraire** de ces ontologies la conceptualisation sous-jacente, ce qui revient à faire abstraction de la syntaxe du langage (toujours particulier !) dans lequel ces ontologies sont spécifiées.
- **Évaluer** ces conceptualisations en termes, notamment, de choix de modélisation effectués et de couverture de domaine.

Dans la pratique, on observe que les ontologies de domaine sont difficilement réutilisables car, même si elles portent sur un même domaine ou un domaine proche, elles n'ont pas forcément été développées pour réaliser une même tâche. Or, les choix de modélisation sont fonction de la tâche réalisée par l'application ayant motivé la construction de l'ontologie. En revanche, les ontologies de haut niveau, parce qu'elles sont indépendantes de tout domaine, se prêtent plus volontiers à la réutilisation [SEB00].

6. Langages de représentations d'ontologies

Il existe plusieurs modèles et langages de représentation pour la modélisation d'ontologie. Nous allons, dans ce qui suit, classer les différents langages formels en fonction des modèles sur lesquels ils reposent.

Parmi les langages de représentation développés au niveau conceptuel, trois grands modèles sont distingués: les logiques de description, les graphes conceptuels et les langages de frames [OLF03].

6.1. Les logiques de description

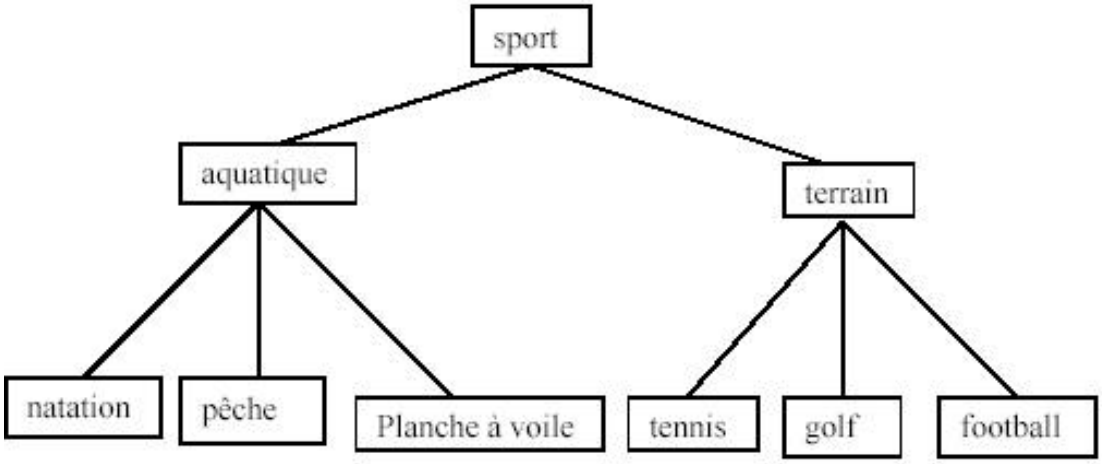
Les logiques de description permettent de représenter les connaissances sous forme de concepts, de rôles et d'individus. Les rôles sont des relations binaires entre concepts et les individus sont les instances des concepts. Les propriétés des concepts, rôles et individus sont exprimées en logique des prédicats, en particulier les propriétés de subsomption. Au niveau terminologique sont définis les concepts, les rôles et leurs propriétés.

Les faits portant sur des individus (types des individus et relations entre individus) sont exprimés au niveau factuel.

LOM et KL-ONE sont des exemples de systèmes implémentant ce modèle. KL-ONE est un langage de représentation dont la sémantique est bien fondée, il a une sémantique externe qui ne dépend ni de la représentation ni des algorithmes qui opèrent sur celle-ci. D'autres langages appartenant à la même famille que celle de KL-ONE sont: BACK, CANDIDE, CLASSIC, DAML+OIL.

Tous les langages de la famille KL-ONE sont des langages hybrides: L'idée des langages hybrides remonte à KL-ONE. Les buts des langages hybrides est de donner une généralisation des réseaux sémantiques et langage de frames, donner une sémantique précise aux langages (autre que celle qui dérive de l'implantation) et permettre à un système de raisonner sur son état de connaissance. Un langage hybride comporte deux composantes:
un langage terminologique qui consiste en une description des connaissances générales comme les classes d'objets (les concepts) et les relations entre classes (les attributs et les rôles) et un langage assertionnel qui est une description des connaissances spécifiques comme les objets ou individus appartenant à ces concepts [OLF03].

Exemple:



En logique de description:

Aquatique $\subseteq$ sport
Terrain $\subseteq$ sport
Natation $\subseteq$ Aquatique
Pêche $\subseteq$ Aquatique
Planche à voile $\subseteq$ Aquatique
Tennis $\subseteq$ Terrain
Golf $\subseteq$ Terrain
Football $\subseteq$ Terrain

Contrainte:

Tennis = (and terrain (atleast 2 personnes))

6.2. Les graphes conceptuels

Les graphes conceptuels ont été initialement conçus pour l'analyse et la compréhension du langage naturel. De par leur similarité avec les réseaux sémantiques, leur pouvoir d'expression est tel qu'ils sont directement applicables à la représentation des connaissances. Les graphes conceptuels sont décomposés en deux niveaux: le niveau terminologique où sont décrit les concept, les relations et les instances de concepts, ainsi que les liens de subsomption entre concepts et entre relations et le niveau assertionnel où sont représentés les faits, les règles et les contraintes sous forme de graphes où les sommets sont des instances de concepts et les arcs des relations [OLF03].

Ce formalisme est implémenté, entre autres, dans COGITANT, une plateforme de développement de SBC utilisant les graphes conceptuels et PROLOG+CG, une extension de PROLOG basée sur les graphes conceptuels [Kab00].

6.3. Langages de frame (frame-based languages)

Introduit dés les années 70 par Minsky [MIN75] comme une modélisation de base pour la représentation de connaissances dans le domaine d'Intelligence Artificielle (IA), le modèle des frames a depuis été adapté à d'autre problématiques puisqu'il a donné naissance au modèle objet, qui envahit peu à peu les différentes branches de l'informatique.

L'idée de frame est très simple. Un « frame » est dans ce contexte un objet nommé, qui est utilisé pour représenter un certain concept dans un domaine. Une frame représente n'importe quelle primitive conceptuelle et est dotée d'attributs (slots), qui peuvent porter différentes valeurs (facets), et d'instances.

Il y a une correspondance entre les systèmes de frame et ceux orientés objet où les classes et les instances correspondent avec les frames, les attributs et les associations de classes avec les slots. Entre les frames, il y a aussi la spécialisation qui donne l'héritage dans les concepts de frame. une frame F1 est plus spécifique qu'une frame F2 si toute instance de F1 est instance de F2.

Le formalisme F-logic était proposé comme le fondement logique pour les langages de frame et orientés objet. Il permet de comprendre le modèle sémantique dans tous les langages de frame et d'aider à la construction d'une base de connaissances.

Parmi les langages de frame on peut citer YAFOOL (Yet Another Frame Based Object-Oriented Language), OML (Ontology Markup Language), Ontolingua, KRL (Knowledge Representation Language), SHIRKA et bien d'autres encore.

SHIRKA fondé sur un modèle classe/instance. C'est un modèle totalement uniforme dans lequel tout schéma, attribut ou facette est une instance d'un schéma de plus haut niveau.

Les valeurs apparaissant dans une facette sont toujours des instances ou des références à des instances. L'inférence des valeurs d'attribut indéterminées ne fait pas appel à des règles, mais à des filtres décrits par des schémas. Les mécanismes d'inférence qu'il utilise sont : héritage, attachement procédural, filtrage, valeur par défaut, spécialisation et classification.

Cependant, il faut noter l'arrivée en 1999 de XOL ("XML Ontology Exchange Language) langage d'échange qui utilise la syntaxe XML. Un schéma XML est automatiquement généré.

La phase de génération utilise des règles de passage qui permettent d'exprimer les connaissances existantes dans l'ontologie en terme de Schéma-XML. Un prototype a été réalisé, qui permet de faire l'analyse lexicale et syntaxique du fichier XOL entrant et de générer un Schéma-XML en sortie. Son développement a été inspiré d'Ontolingua et de OML [OLF03].

A l'heure actuelle, on connaît RDF, RDF schéma, DAML.

7. Des outils et langages d'ontologies pour le Web

Si on part de la définition d'une ontologie du domaine, comme un modèle formel décrivant les concepts du domaine et leurs relations, un langage pour les ontologies doit offrir les primitives épistémologiques nécessaires pour décrire des concepts de l'ontologie (classes), leurs propriétés et leurs relations et des restrictions sur ces propriétés [GOL03].

7.1 Modélisation

Différents éditeurs d'ontologie qui supportent ces définitions sont actuellement proposés en particulier, Protégé 2000, OntoEdit et OILEd qui sont deux éditeurs pour le langage DAML+OIL. Protégé-2000 offre un environnement graphique interactif pour la conception d'ontologies. Un arbre permet une navigation rapide et simple dans la hiérarchie.

- Le modèle de Protégé-2000 est basé sur les frames. Il aide à définir les classes et les hiérarchies avec héritage multiple; les attributs, les restrictions de valeurs de ces attributs, leurs facettes, comme les restrictions de cardinalité, valeurs par défaut, ainsi que des attributs inverses, des métaclasses et la hiérarchie de métaclasses.

- OILEd est un éditeur graphique développé par l'Université de Manchester qui permet à l'utilisateur de construire des ontologies représentées en DAML+OIL. Le modèle de OILEd est basé sur DAML+OIL. Tout en offrant une interface de modélisation de type « frame », il supporte toute l'expressivité de la logique de description OIL et DAML+OIL. Les classes sont définies en terme de leurs super-classes, de leurs propriétés avec restrictions de type, et en outre la possibilité de définir des axiomes, par exemple pour définir des classes disjointes. Le modèle permet de définir des descriptions complexes comme valeur des attributs, à

l'opposé de la plupart des éditeurs de frames où les classes doivent être nommées pour pouvoir être utilisées [GOL03].

7.2. Représentation

Il existe de nombreuses présentations des langages standards ou en cours de standardisation du W3C (http://www.w3.org): XML, RDF, RDFS, DAML+OIL, OWL, leur structuration en couche, et leur évolution.

XML [W3C] **:** si HTML est un langage d'annotation pour décrire la présentation du contenu d'un document, XML est un langage pour décrire sa structure. Il est possible d'utiliser ses propres balises. XML et XML Schema sont des langages qui pourraient être suffisants pour publier ou échanger les données.

RDF [W3C] **:** est un langage pour décrire les ressources du Web et leurs méta-données. Elles sont décrites par des triplets [Propriété Prédicat Objet]. RDFS introduit en plus la possibilité de définir des classes et des hiérarchies, des propriétés, de contraindre leur domaine. RDF et RDFS peuvent être éventuellement suffisants pour une exploitation des méta-données (documentation) ou une navigation classiques sur le Web. Mais pour une recherche de documents ou une navigation plus intelligente sur le Web, un langage formel plus expressif est nécessaire.

Le langage DAML+OIL [W3C] **:** est un langage conçu par le groupe du W3C WebOnt dans le but de dépasser la simple « présentation » d'informations sur le Web pour aller vers l'interopérabilité, la compréhension et le raisonnement sur ces informations. DAML+OIL est le résultat de la fusion de OIL résultant du projet européen OntoKnowledge, et de DAML-ONT, issu du projet DARPA DAML (American Agent Markup Language). Il doit ses primitives de modélisation intuitives aux « frames », sa syntaxe aux standards XML et RDF, sa sémantique formelle et ses mécanismes de raisonnement aux logiques de description. D'un point de vue formel DAML+OIL est basé sur la logique de description expressive étendue du constructeur oneOf et de types de données. DAML+OIL permet de définir en outre un ensemble d'axiomes. DAML+OIL est accompagné de différents outils : un éditeur mais surtout un classifieur FaCT (Fast Classification of Terminology) qui permet de détecter automatiquement les incohérences, et classer automatiquement les concepts d'une ontologie.

Le futur standard OWL [W3C] : OWL (Ontology Web Language) est le successeur de DAML+OIL. OWL fournira trois sous langages d'expressivité croissante : OWL Lite, OWL DL et OWL Full. La sémantique formelle des logiques de description est définitivement acquise, et ces différentes couches de langages, de niveau d'expressivité et de complexité différents, sont en phase avancée de standardisation au W3C. OWL Lite aura moins de

constructeurs de base, en particulier pas la disjonction ni la négation (mais qui pourraient être capturés). Les différentes fonctionnalités souhaitées pour chacun de ces langages détermineront les constructeurs finaux retenus pour chacun d'eux, le centre de la question étant l'opposition entre *expressivité* et *complexité–propriétés* attendues des algorithmes (décidabilité, correction, complétude). OWL-DL garantirait la complétude et la décidabilité, tandis que OWL Full offrant un maximum d'expressivité et la liberté de syntaxe de RDF, serait sans garantie computationelle. Il reviendra au développeur de l'ontologie de choisir le langage qui convient à ses besoins.

En bref, XML fournit la syntaxe de la couche transport, RDF/RDF(S) les primitives ontologiques de base (le modèle simple de données de RDF et les schémas de RDF), DAML+OIL la couche logique (la sémantique formelle de OIL et DAML+OIL); enfin d'autres couches (notamment règles) viendront étendre DAML+ OIL [GOL03].

8. Ontologie versus thesaurus

Des modélisations conceptuelles ou terminologiques existent depuis longtemps dans le domaine de la recherche d'information au sein des bibliothèques et dans le domaine de la terminologie, par exemple en médecine où il existe, entre autres, des thesaurus de spécialités répertoriant l'ensemble des termes médicaux à utiliser pour décrire l'activité médicale – *i.e.* un vocabulaire contrôlé. En reprenant un vocable de Bourigault *et al*, les différentes Ressources Terminologiques ou Ontologiques (RTO) élaborées dans différents domaines doivent être précisées et *conceptuellement* caractérisées pour bien comprendre leur signification par rapport à une modélisation conceptuelle et formelle et si elles peuvent être ou servir d'ontologies et à quelles conditions. Pour cela, nous allons reprendre ici, rapidement, trois définitions par rapport à des produits terminologiques existant parallèlement aux ontologies, les *thesaurus*, *classifications* et *terminologies* :

> **Un thesaurus** est un ensemble de termes normalisés fondé sur une structuration hiérarchisée. Les termes y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Organisé alphabétiquement, il forme un répertoire alphabétique de termes normalisés pour l'analyse de contenu, le classement et donc l'indexation de documents d'information (dans de nombreux cas, les thesaurus proposent aussi une définition des termes utilisés).
>
> **Une classification** est l'action de distribuer par classes par catégories (rien n'est dit sur le type d'objets classifiés). C'est aussi le résultat de cette action.

Une terminologie est un ensemble des termes particuliers à une science, à un art, à un domaine. Les termes y sont également définis par un texte en langue naturelle et caractérisés par différentes propriétés linguistiques ou grammaticales suivant l'usage prévu de cette terminologie. Avec leur mise sur support informatique, les terminologies ont beaucoup évolué et sont parfois enrichies de relations entre termes, formant ainsi un réseau terminologique.

À partir de là, on peut s'intéresser à quelques ressources Knowledge Management et autres, par exemple WORDNET ou le MeSH « «Voir glossaire », pour vérifier quelle est leur nature exacte, ce qu'on peut en faire et pourquoi [OLF03].

9. Le champ d'application des ontologies

Le champ d'application des ontologies ne cesse de s'élargir et couvre les systèmes conseillers (systèmes d'aide à la décision, systèmes d'enseignement assisté par ordinateur), les systèmes de résolution de problèmes ou les systèmes de gestion de connaissances. Un des plus grands projets basés sur l'utilisation d'ontologies consiste à ajouter au Web une véritable couche de connaissances permettant, dans un premier temps, des recherches d'informations au niveau sémantique et non plus simplement syntaxique. A terme, il est prévu que des applications internet pourront mener des raisonnements utilisant les connaissances stockées sur la Toile.

Autre que le web sémantique, les ontologies ont été largement utilisées dans le cadre des applications web et des moteurs de recherche. Une de ces applications est l'utilisation d'une ontologie du domaine pour affiner une requête à l'aide d'un treillis de Galois [SAF03]. [SAF03] ont étudie comment aider un utilisateur qui effectue une recherche dans un entrepôt thématique à affiner sa requête quand celle-ci retourne trop de réponses. En utilisant une ontologie de domaine et un ensemble de ressources annotées avec les termes de cette ontologie, [SAF03] ont montrés comment utiliser un treillis de Galois pour élaborer, en interaction avec l'utilisateur, une requête plus précise qui réponde mieux à ses attentes.

Un autre type d'application est la Représentation de méta-connaissances pour le développement de Web Sémantiques d'Organisation [COR03]. [COR03] ont présenté l'apport du langage réflexif d'ontologie DefOnto (défini et implanté dans l'équipe [COR03]) pour représenter les différents modèles (de l'organisation et de l'information) et leur ontologie associée. Plus particulièrement, ils ont montré l'intérêt de la représentation de méta-connaissances pour rendre compte du modèle de l'information. Ils ont montré également l'utilisation des services inférentiels de DefOnto pour le développement d'un moteur de recherche exploitant les différents modèles de connaissances.

Les ontologies ont été combinées, aussi, avec les systèmes multi-agents pour la recherche d'informations. [MAR00] ont utilisé les ontologies comme moyen de partage des connaissances entre agents dans leur architecture multiagent pour la recherche sur internet.

Le champ d'application des ontologies a touché aussi le domaine de classement des documents web dans des classes homogènes. Les travaux de [BEN03] portant sur le classement des documents ont utilisé un ensemble de documents de base (classement de base), dont chacun est caractérisé par un (petit) ensemble de concepts d'une ontologie, et ils ont donné une méthode pour regrouper en classes les documents qui ont une sémantique proche, cette méthode est basée sur un calcul de similarité entre les ontologies et les nouveaux documents à classés, à condition qu'il y'ait des liens hypertextes entre ces nouveaux documents et les documents des classes de base [BEN02].

10. Conclusion

Dans ce chapitre nous avons donné un survole sur les ontologies, ceci, sans entrer dans les détailles des ontologies en tant que branche dans la science de l'ingénierie des connaissances.

Dans ce chapitre nous avons expliqué les différents composants des ontologies, les méthodologies de construction et les langages de modélisation et de représentation des ontologies. C'est l'offre en matière de représentation et de modélisation des connaissances à l'aide des ontologies et l'offre en matière de langage de représentation et de modélisation des ontologies qui nous ont poussés à utiliser les ontologies dans notre système de filtrage d'informations que nous allons détailler dans le chapitre suivant.

Chapitre 2

Le Modèle de Filtrage

1. Introduction

Les SRI sur le web fournit toujours des réponses bruitées, un utilisateur qui cherche des informations sur un sujet donné risque d'avoir des résultats qui sont sans aucun rapport avec son sujet.

Bien que les méthodes d'indexation de pages Web se soient notablement améliorées ces dernières années, la pertinence des réponses fournies est loin d'être au niveau des attentes des internautes et le problème du bruit existe toujours.

L'idée de filtrage est de proposer des filtres aux usagers pour les aider à trouver l'information pertinente.

Dans ce chapitre nous détaillerons notre système (moteur) de recherche d'informations proposé et qui porte comme objectif de solutionner le problème de bruit. Ce système combine à la fois les diverses fonctionnalités propres au domaine de la recherche d'informations et des fonctionnalités faisant appel aux technologies issues des travaux en ingénieries des connaissances (les ontologies).

Dans un premier temps nous donnons notre principe de filtrage, puis nous décrivons les parties structurelles du système, suite à cela nous détaillerons son architecture et nous discuterons l'approche employée (les ontologies) sur la quelle se base notre contribution. En fin nous donnerons certaines discussions relatives à la mise en œuvre de cette solution.

2. Principe général du système

Le système proposé s'appuie sur la notion de regroupement des documents Web trouvés en plusieurs domaines auxquels ils font parties. Dans la littérature le nombre de domaines est limité et dénombrable, on distingue entre autre : le domaine mathématique, domaine biologique, domaine mécanique, domaine informatique, domaine commercial, domaine publicitaire … etc [PIE02]. La figure 1 illustre le principe global de filtrage proposé, il s'agit de regrouper les documents faisant partie à des domaines les plus similaires à leurs contenus.

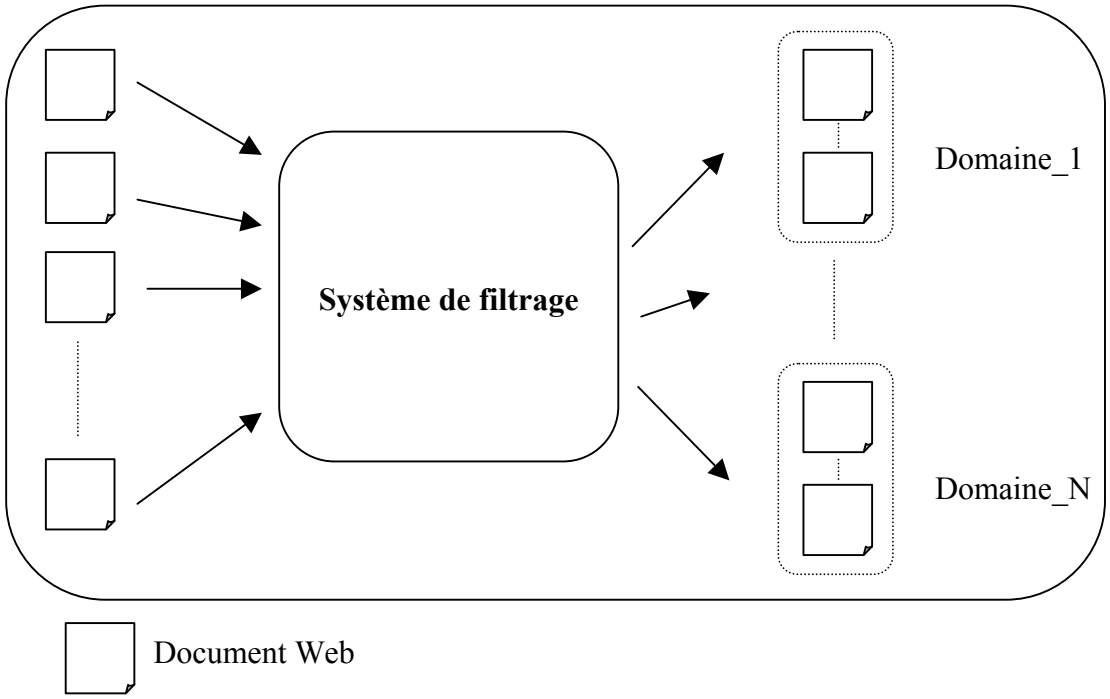


Figure 1. le système de filtrage

Le fait de regrouper les résultats et les affecter à des domaines auxquels ils correspondent, libérera l'utilisateur de faire lui-même le parcours et la consultation des différents résultats, en général un nombre très important des documents renvoyés. Ainsi que l'élimination des résultats non pertinents et qui ne correspondent pas à ses centres d'intérêts.

Donc dans notre approche, le filtrage des résultats revient à créer des sous-ensembles homogènes et à arranger les documents fournis comme réponses dans ces sous-ensembles. L'utilisateur final aura comme réponse une liste des domaines relatifs aux résultats qui répondent à ses besoins en terme d'informations et pourra par la suite naviguer dans le domaine qui lui convient. La figure 2 illustre les différentes interactions entre l'utilisateur et le moteur de recherche proposé:

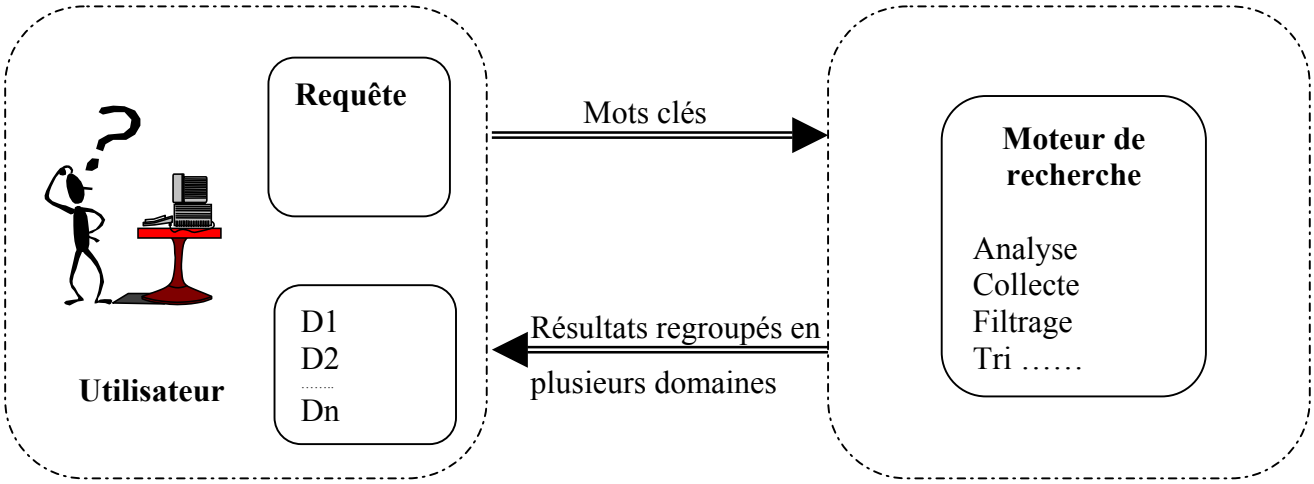


Figure 2. Interaction utilisateur/Moteur de recherche

3. Parties structurelles du système

Le système est divisé en six unités jouant chacune un rôle bien déterminé. Comme le montre la figure 3, ces unités sont :

- Unité de communication avec l'utilisateur ;
- Unité de traitement de la requête ;
- Unité de recherche des documents relatifs aux mots clés formulés par l'utilisateur ;
- Unité d'indexation des ressources web;
- Unité de représentation et hiérarchisation des domaines ;
- Unité de filtrage et de tri des résultats.

Ces unités sont modulaires puisque chacune est spécialisée dans un ensemble de fonctions. Dans ce qui suit nous détaillons chaque unité et nous étudions sa relation avec les autres parties du système.

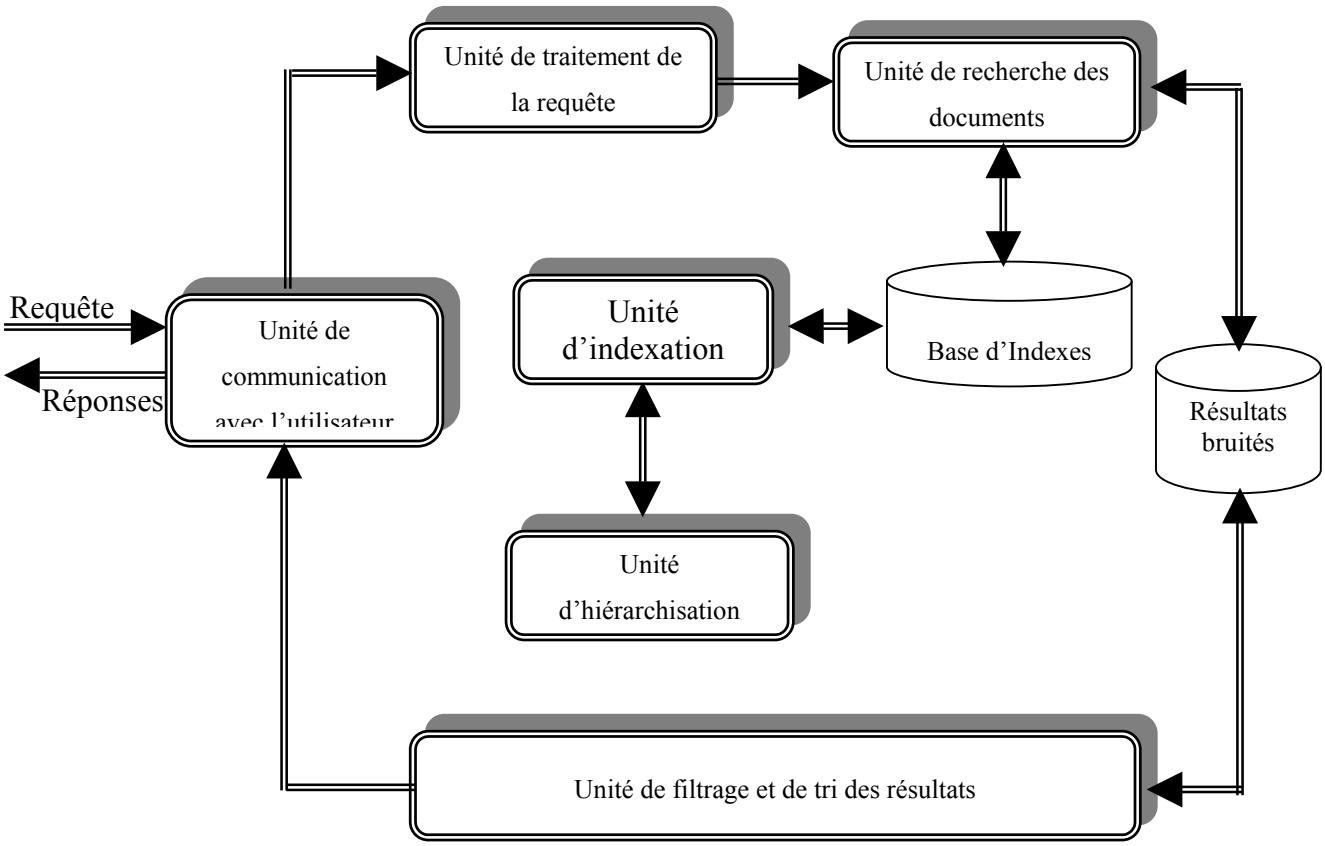


Figure 3. Les parties structurelles du système

3.1. Unité de communication avec l'utilisateur

Cette unité est chargée des communications entre le système et l'usager. Elle assure toutes les interactions entre l'usager et le système. Elle a pour rôles de, premièrement, récupérer la requête à analyser via une interface de recherche simple, cette interface web permet à l'utilisateur de saisir et d'envoyer sa requête. Puis, en deuxième lieu, de transmettre les résultats fournis par le système à l'usager et de les visualiser sous forme de plusieurs catégories.

3.2. Unité de traitement de la requête

L'unité de traitement de la requête reçoit de l'unité de communication la requête à satisfaire et procède aux opérations suivantes :

- La vérification lexicale de la requête : le mode de recherche permet par le système est le mode booléen, c'est pour cette raison que la vérification se limite au niveau lexical parmi tous les niveaux du traitement linguistique connus;
- La correction et le raffinement automatique de la requête : cette correction et ce raffinement concernent les mots mal saisis ou erronés;

- La représentation interne de la requête : pour des raisons d'améliorer le taux de réponses et la pertinence de l'information à fournir, une représentation interne par rapport aux différents composants du système est délivrée par cette unité.

3.3. Unité d'indexation des ressources

Cette unité est constituée d'un robot qui parcourt le Web régulièrement dans un intervalle du temps afin de remplir une grande base de données nommée base d'indexes : BDI. Le robot analyse les documents rencontrés (page HTML, fichiers PDF,...) et les indexe. Les informations qui doivent être extraites de ces documents sont les suivantes :

- L'URL du document.
- Le titre de document.
- Un résumé textuel du document.
- Une liste des mots clés décrivant le document.
- La classe (le domaine) du document.

En ce qui concerne la classe, le classement d'un document n'est pas arbitraire mais il est basé sur la similarité entre ce document et les hiérarchies des domaines. Une fois trouvé, un document doit être affecté à une classe donnée.

Puisque le robot parcourt le web régulièrement, on s'assure que le contenu de la BDI est mis à jour et que les versions des documents indexés sont les plus récentes. Ensuite la BDI sera mise à disposition des autres unités.

3.4. Unité d'hiérarchisation des domaines

Cette unité constitue un stock de données linguistiques des domaines, les données dans ce stock sont bien organisées et bien structurées en hiérarchie des domaines. L'hiérarchie d'un domaine consiste en un réseau linguistique (terminologique) ou une arborescence terminologique de ce domaine. Exemple d'une hiérarchie de domaine informatique:

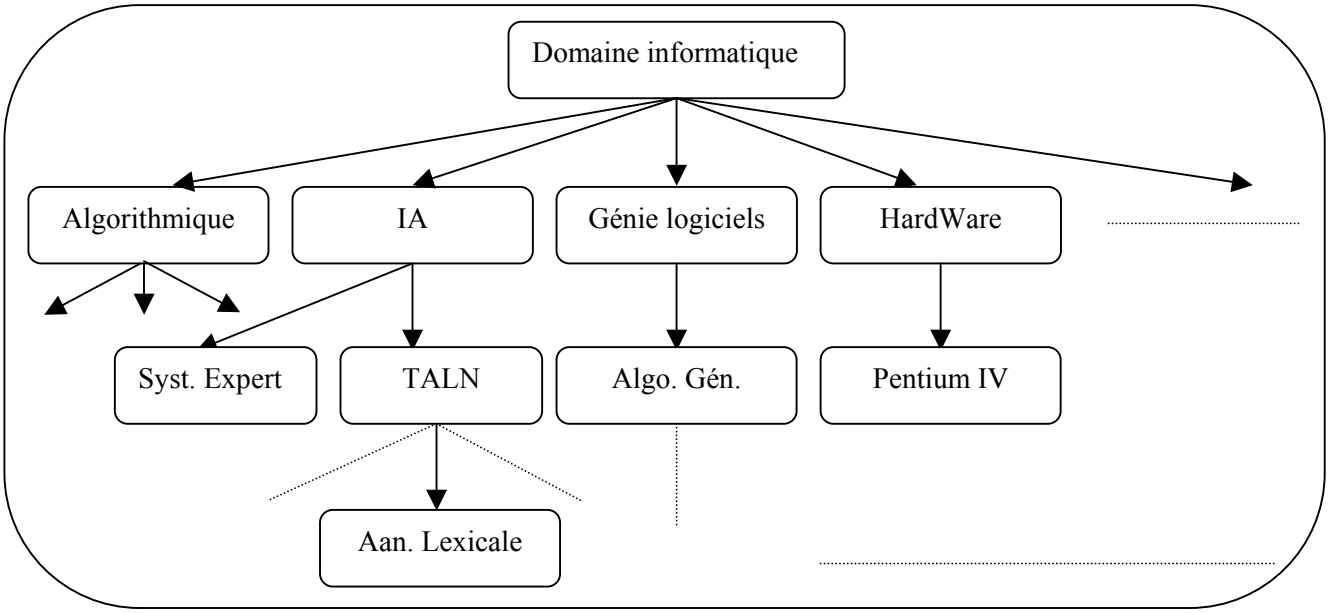


Figure 4. Exemple d'une hiérarchisation du Domaine *Informatique*

3.5. Unité de recherche des documents

Comme son nom indique, cette unité sert à la recherche des documents relatifs aux mots clés composant la requête, ceci est réalisé par l'interrogation de la base des documents (base d'indexes). Après l'interrogation de la base, nous obtenons une nouvelle base plus restreinte par rapport à la BDI et qui contient divers documents syntaxiquement proches de la requête mais sémantiquement bruités, d'où la nécessité d'appliquer la stratégie de filtrage.

3.6. Unité de filtrage et de tri

Après une opération de recherche des documents, l'unité de filtrage et de tri s'occupe de filtrer puis de trier les documents trouvés. L'opération de filtrage est celle déjà mentionnée auparavant, puisque chaque document possède une classe, le filtrage consiste à construire les classes des documents trouvés, c'est à dire que le système regroupe les documents des même classes.

L'opération de tri consiste à trier les documents d'un même domaine et d'ordonner la liste des domaines. Ainsi les questions qui se posent à ce stade sont :

- « Quel est le domaine qui sera visualisé en tête ? Quel est le second ? Le troisième ? …etc. En d'autre terme sur quel critère on va se baser pour ordonner la liste des documents »
- « Comment trier les documents d'un même groupe ? »

Nous avons fixé des critères pour répondre à ces deux questions qui seront détaillés dans (respectivement) les sections 4.9 et 4.8.

4. Architecture détaillée du système

Dans cette section, nous décrivons l'architecture détaillée de système. L'exploitation des connaissances est une technique classique pour favoriser la recherche d'information.

L'originalité de notre approche est d'utiliser un formalisme de représentation des connaissances pour le filtrage d'informations (les ontologies). La particularité de notre approche réside dans le développement d'un système de recherche d'information intelligent.

L'architecture suivante donne le détaille des différentes parties étudiées précédemment :

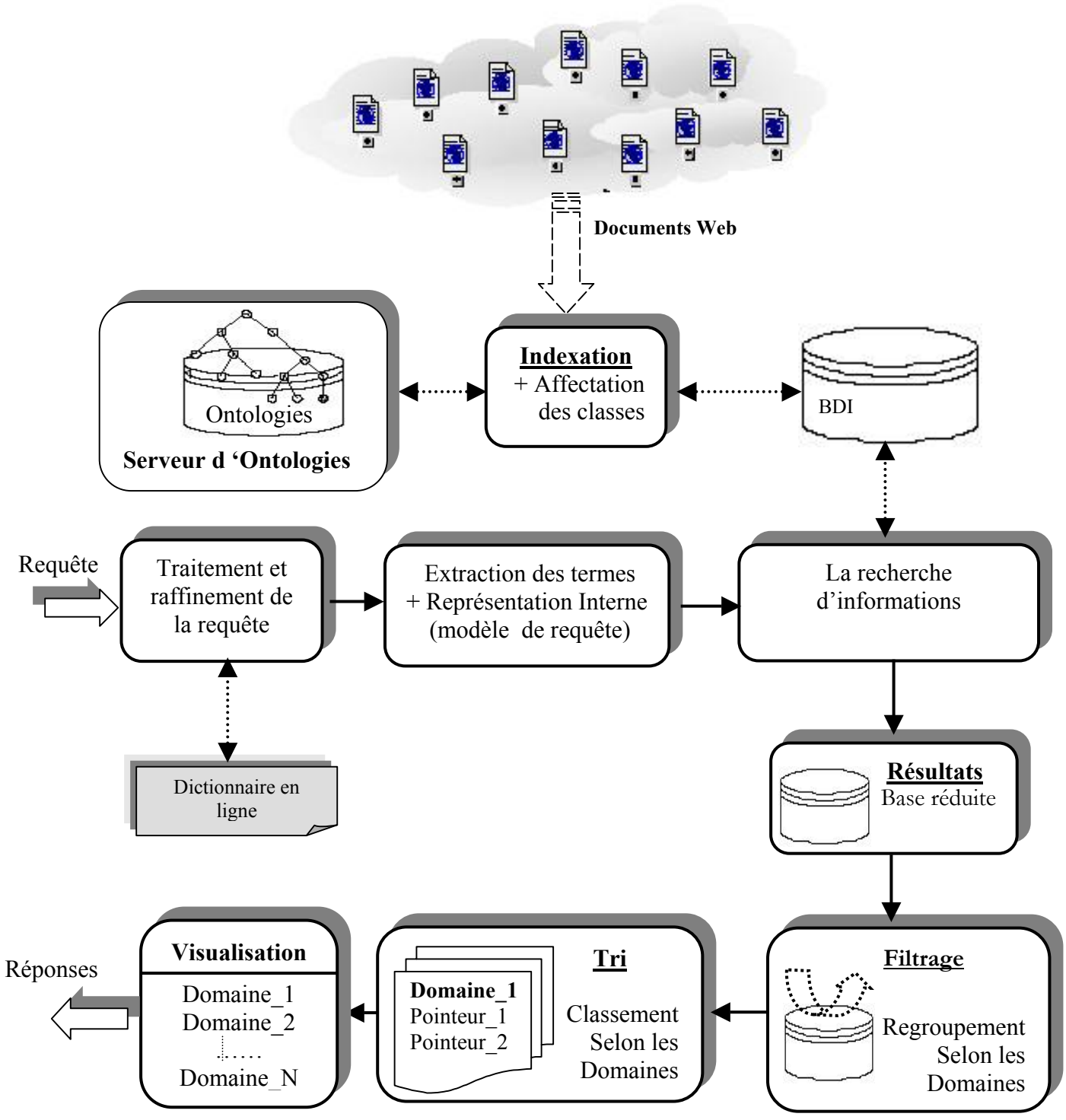


Figure 5. Architecture détaillée du système

4.1. Traitement et raffinement de la requête :

Après la phase de récupération de la requête, on applique un processus de traitement et de raffinement sur cette dernière. Le traitement et le raffinement de la requête comportent ***La vérification lexicale des mots composants la requête*** : s'il y'a une erreur lexicale (mot mal saisi …) le système la corrige automatiquement en proposant à l'usager de remplacer les mots erronés par d'autres mots les plus proches. Cette phase est accompagnée par un dictionnaire électronique en ligne.

4.2. Le dictionnaire en ligne

Le dictionnaire en ligne est un dictionnaire électronique comportant une base des mots. Chaque mot dans ce dictionnaire est accompagné par une liste des synonymes et de toutes ses drivés possibles (pluriel, singulier, toutes les conjugaisons possibles s'il s'agit d'un verbe…etc.).

Du point de vue de sa fonction, il s'agit d'un instrument de contrôle lexical des mots employés dans les requêtes des utilisateurs.

4.3. Le modèle de la requête :

Avant d'établir le modèle de la requête, on doit extraire les termes clés depuis cette dernière, ceci en éliminant les déterminants (le, la, un, une.), les prépositions (dans, de, avec...) et les opérateurs logiques (+, -), c'est à dire qu'on élimine les mots ayant peu de valeur documentaire et en ne gardant que les noms, les verbes, les adjectives.

Afin d'améliorer le nombre de réponses, le système utilisera des synonymes des termes extraits de la requête. A la fin du processus, le système génère une représentation de la requête, c'est ce qu'on appèle le modèle de la requête : *mots clés : listes des synonymes et les opérateurs logiques existants entre ces mots clés.*

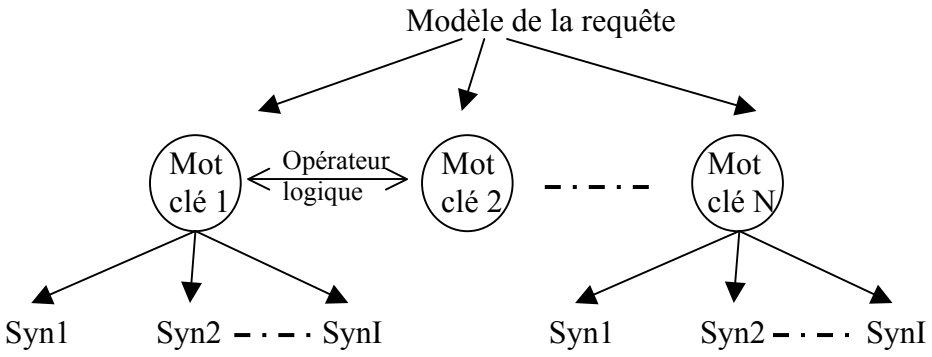


Figure 6. Le modèle de la requête

4.4. L'indexation

L'indexation se débute par un parcours du Web, ce parcours commence par l'introduction manuelle d'une liste des adresses des sites web les plus connus, puis le système aspire tous les documents référencés par des liens hypertextes se trouvant au niveau des pages web initiales et les indexe, ensuite ceux référencés dans ces nouvelles pages et ainsi de suite. En générale la structure hypertextuelle des pages web est représentée comme suit :

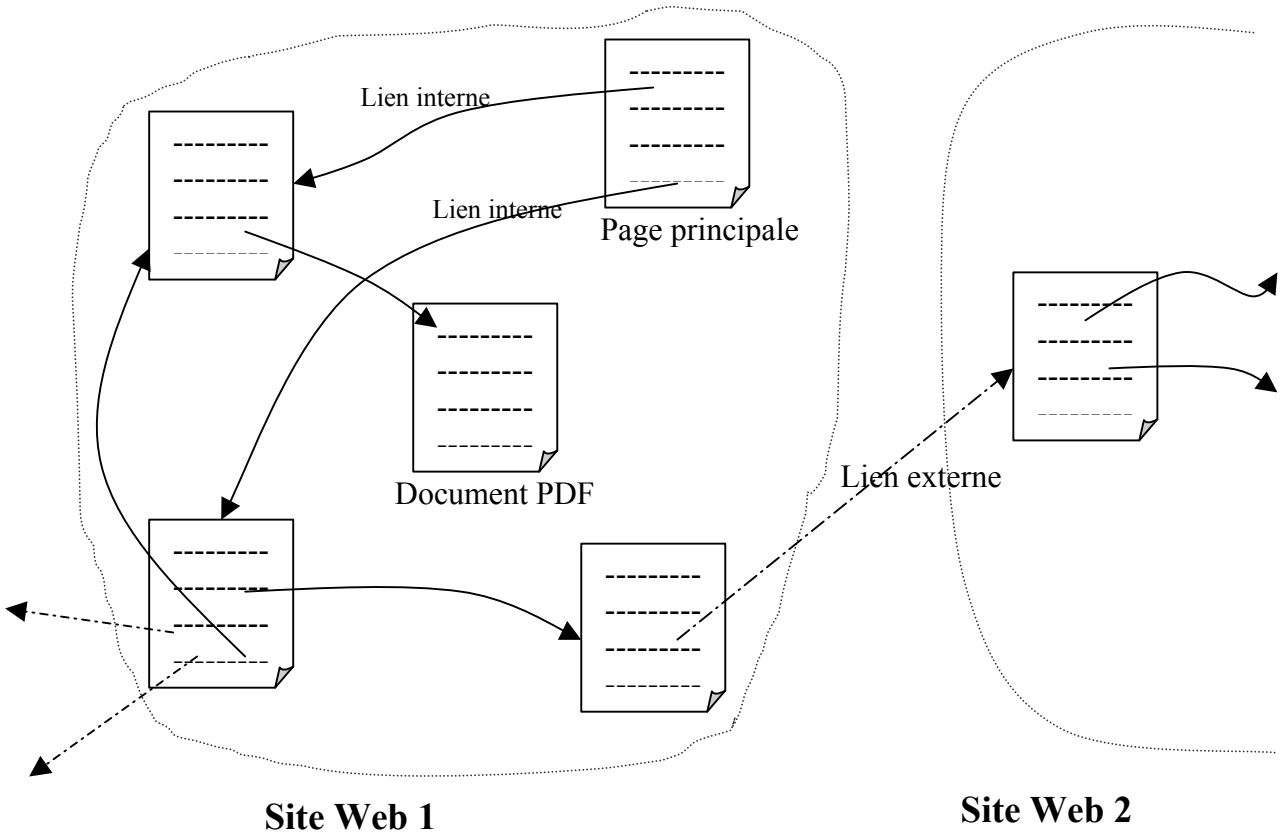


Figure 7. La structure du Web

Chaque document rencontré sera indexé (peut importe son type page HTML, document PDF, Word….). Le problème déclenché par le parcours des sites web est la notion de cycle. Le cycle se produit quand on a un retour vers un site déjà visité ou indexé, c'est à dire qu'une nouvelle page référence un site ou un document déjà visités.

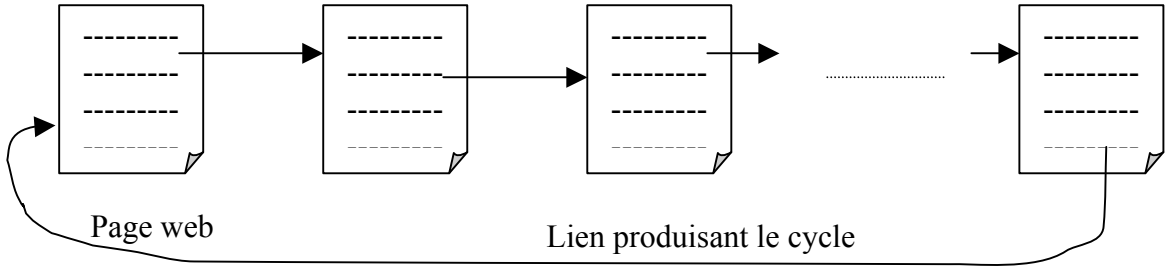


Figure 8. Le problème du cycle

Le cycle est considéré comme un problème, car si le système continu à suivre les documents formés par le cycle on risque de ne jamais terminer l'indexation et d'avoir une boucle infinie d'indexation des mêmes documents concernés par le cycle. Pour remédier à ce phénomène des mécanismes de détection et d'élimination du cycle sont utilisés par le moteur de recherche et qui s'énoncent comme suit :

- Puisque les adresses des documents déjà visités sont stockées dans la base d'indexes, le système vérifie chaque nouvelle référence, si elle coïncide ou non avec la liste des URLs stockée dans la BDI ;
- Si la référence coïncide avec l'un des adresses de la BDI alors le système l'ignore et va chercher une autre référence, sinon il indexe le document référencé.

L'indexation est de type Full-Text (texte intégral), le choix de ce type d'indexation permet de prendre en compte tout le contenu des documents. L'indexation en texte intégrale signifie que tous les mots sont concernés et permet de mieux représenter les documents par leurs contenus au lieu par des informations superficielles. Les informations nécessaires pour indexer ces documents sont les suivantes :

- **Un numéro d'ordre** : Ce numéro est le rang de document dans la base d'indexes, chaque document possède un numéro qui permet de l'identifier d'une manière unique dans la base.
- **L'URL** : l'URL (Uniform Ressource Locator) représente le chemin d'accès (la trace) au document ;
- **Le titre du document** : c'est cette information qui sera visualiser comme titre du document après une réponse du système. Dans les fichiers HTML, le titre de la page est délimité par les deux balises : ***<Title> Titre de la page <\Title>***. Dans les documents formatés (PDF, DOC …) cette information se trouve au niveau de descripteur de lien référençant ce document.
- **Un résumé textuel du document** : une partie de ce résumé sera afficher comme réponse à une requête pour décrire à l'usager le contenu du document fourni comme réponse. Le résumé d'un document est la partie textuelle du document. La structure d'un document web comporte des textes, des images, des liens hypertextes, …etc. Le résumé est formé par tous les textes du document, ceci en éliminant les objets graphiques et les liens hypertextes.
- **Une liste des mots clés décrivant le document** : la liste des mots est obtenue par une analyse Full-Text du document, qui consiste à garder comme mots clés tous les mots

lexicaux (Noms, Verbes, Adjectifs) présents dans le document et d'éliminer les mots ayant peu de valeur documentaire (mots vides : le, la les, de, …etc.). Pour chaque mot clé on fait associer une fréquence, cette fréquence est le nombre d'apparition de ce mot dans le document.

- **La classe du document** : cette information nous permet de déduire à quel groupe (domaine) est affecté ce document s'il sera fourni comme réponse. Nous donnerons dans les paragraphes qui suivent un détaille d'obtention de cette information.

Donc en résumé on peut schématiser la structure logique de la base d'indexes comme suit :

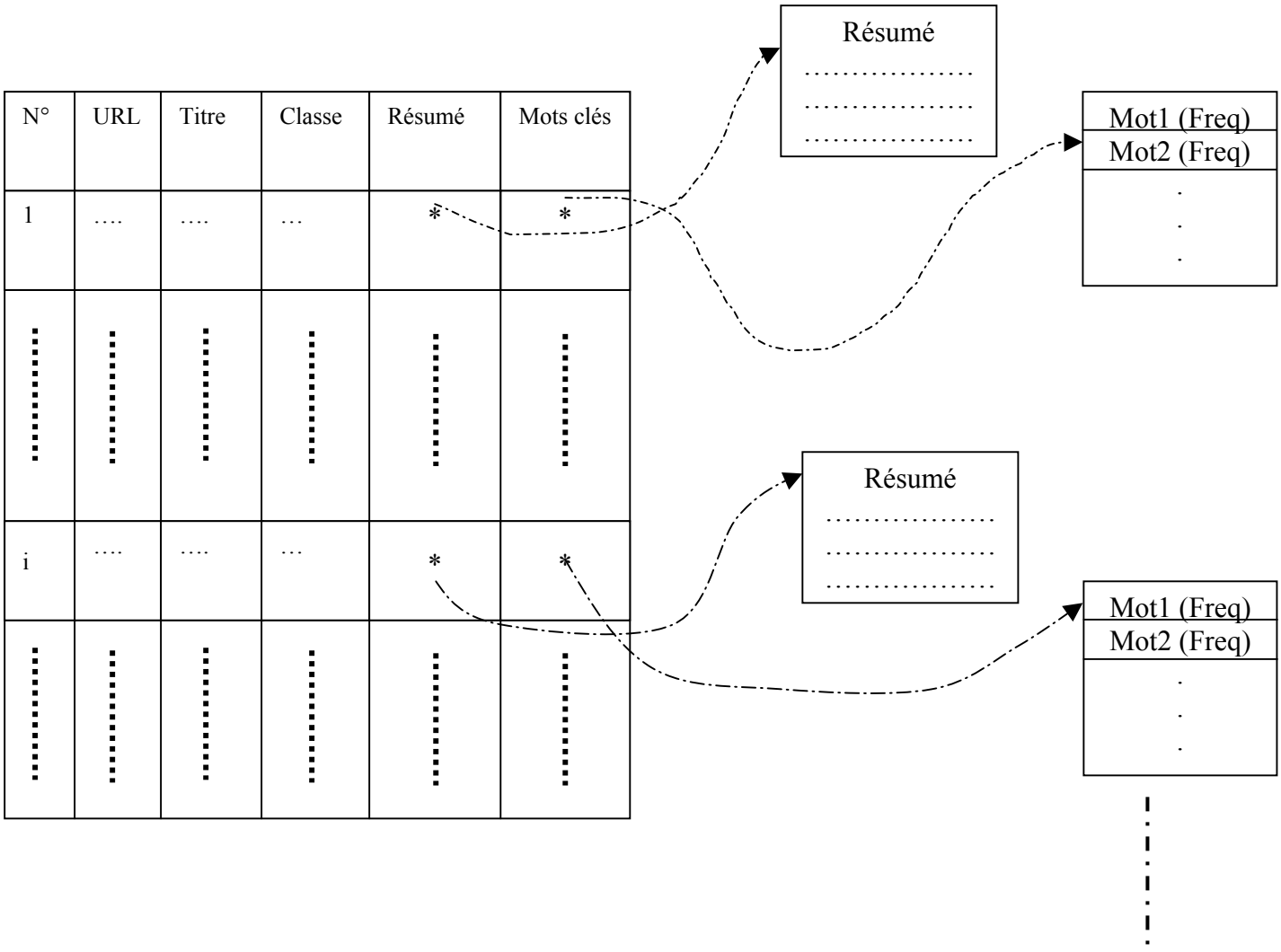


Figure 9. Structure logique de la Base d'Indexes

La phase d'indexation se termine lorsqu'il ne reste plus des liens à visiter ou lors que tous les liens restants produisent des cycles. La taille de la base d'indexe dépend directement de deux facteurs : le nombre de liens introduits manuellement au départ et le nombre de références figurants dans les pages web visitées. En effet, si le nombre de liens d'origine est

trop grand alors la taille de la BDI s'agrandit et réciproquement si ce nombre est faible, de même que si les pages visitées sont riches en terme de liens hypertextes vers d'autres pages, la taille de la BDI s'agrandit et vice-versa.

A chaque intervalle du temps on doit vider la BDI pour relancer l'indexation à nouveau. Ce choix de vidange la BDI est lié en bonne partie à l'évitement de cycle et à la mise à jour des documents circulants sur le Web.

4.5. Les ontologies

Les ontologies que nous avons utilisé sont des ontologies linguistiques (orientées terminologies) qui jouent un rôle central pour la représentation des connaissances des différents domaines dans notre système. Les concepts sont les noms des domaines, les sous domaines et les vocabulaires des domaines. La relation sémantique entre les concepts est définie de la manière suivante :

- *EST_COMPOSE_DE* : un domaine D *EST_COMPOSE_DE* un sous domaine D1, un domaine D *EST_COMPOSE_DE* vocabulaire V1.

Ainsi, par exemple, l'ontologie du domaine informatique peut être illustrée par la figure suivante :

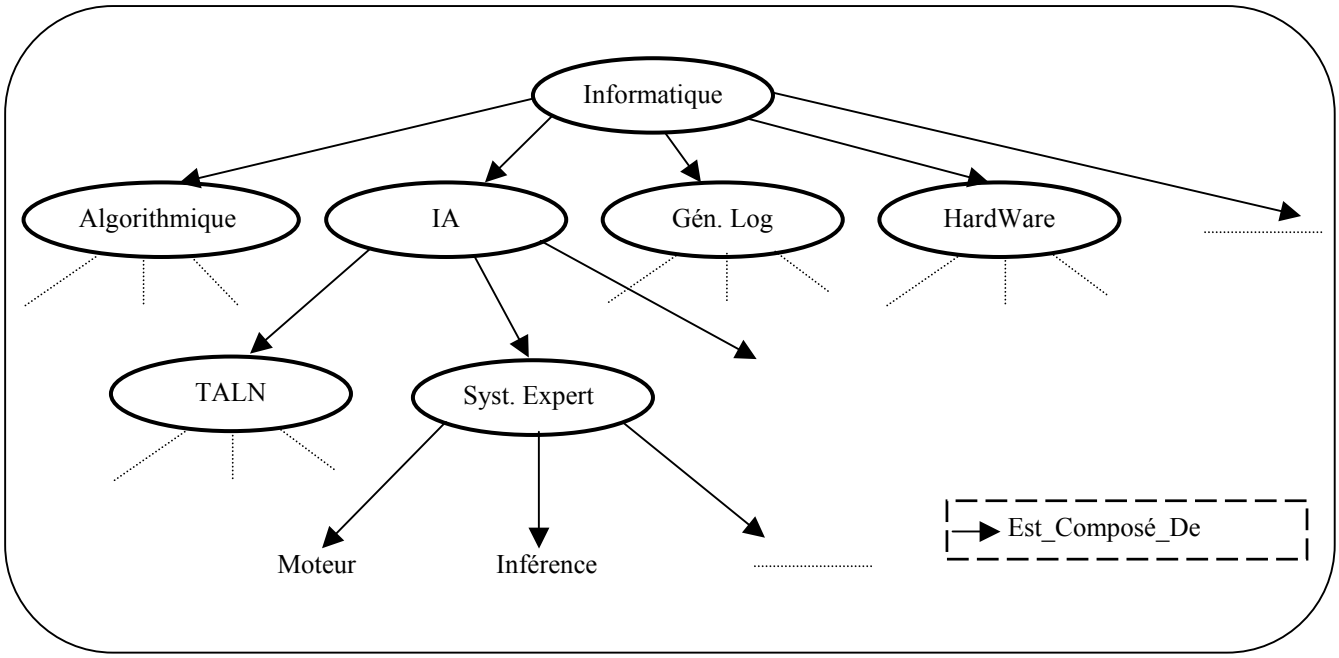


Figure 10. Un fragment d'un arbre ontologique du domaine *INFORMATIQUE*

Les différentes ontologies sont regroupées dans un serveur appelé « *Serveur d'ontologies* » qui s'occupe de la gestion des ontologies et assurant toutes communications avec les autres parties du système.

4.5.1. Exploitation des ontologies pour le classement des documents

La structure des ontologies est exploitée pour l'hiérarchisation des domaines sous forme d'un réseau terminologique, la navigation dans les hiérarchies des concepts des domaines et la classification des documents indexés par le moteur. Chaque domaine possède une ontologie, il est totalement défini à l'aide des termes et relations de l'ontologie.

La navigation dans une ontologie, grâce à un indexe alphabétique, permet d'appréhender les termes du domaine et leurs relations.

L'utilité majeure d'une ontologie est son exploitation par le moteur de recherche **au niveau de l'indexation**, l'affectation d'un document à une classe se fait par le calcul d'une mesure de similarité entre le document et le domaine le plus proche. Ce calcul de similarité consiste à comparer les mots clés extraits du document et les termes de l'ontologie du domaine. L'algorithme semi formel suivant résume la succession des étapes afin de classer les documents :

DEBUT
  POUR « Chaque Document Rencontré » **FAIRE**
    DEBUT
      - Extraire les mots clés décrivant le document ;
      - Interroger le serveur d'ontologies ;
        - le serveur navigue dans les ontologies pour rechercher les mots clés ;
      - Affecter (Document, Classe) ;
    FIN ;
FIN.

Les mots clés sont ceux contenus dans la base d'indexes et la navigation dans une ontologie s'effectue via la relation sémantique EST_COMPOSE_DE. Les ontologies sont parcourues en mode Depth (profondeur d'abord). La similarité entre un document et une ontologie se fait par la recherche des mots du document dans les ontologies une par une. La décision de correspondance entre le document est l'une des ontologies se fait par le biais de taux d'appartenance d'un document à un domaine (nombre maximum de mots clés trouvés dans les ontologies).

4.5.2. La construction des ontologies

Dans cette section nous donnerons les grandes lignes des étapes de construction des ontologies des différents domaines.

Nous proposons d'utiliser le processus de *la fouille du texte* (textmining) pour la construction automatique ou semi-automatique des ontologies.

La construction manuelle des ontologies nécessite le recours à plusieurs experts des différents domaines : des mathématiciens, des biologistes, des informaticiens …etc. qui est une tâche très difficile. Pour cela nous souhaitons rendre cette tâche automatique ou semi-automatique tout en exploitant la base d'indexes BDI.

Nous souhaitons découvrir des connaissances à partir de la BDI (en particulier à partir des résumés textuels) qui seront exploités dans le processus de la recherche d'information. Nous appliquons une technique de textmining appelée *Règle d'Association* dans le but d'extraire des associations intéressantes et non triviales à partir de la base.

Les règles d'associations ont été initialement utilisées en analyse des données puis en fouille de données afin de trouver des régularités, des corrélations dans des bases de données relationnelles de grandes tailles [CHE02]. Nous nous intéressons à la découverte des règles d'associations booléennes pour construire les hiérarchies des domaines (ontologies). Une règle d'association booléenne RA est de la forme :

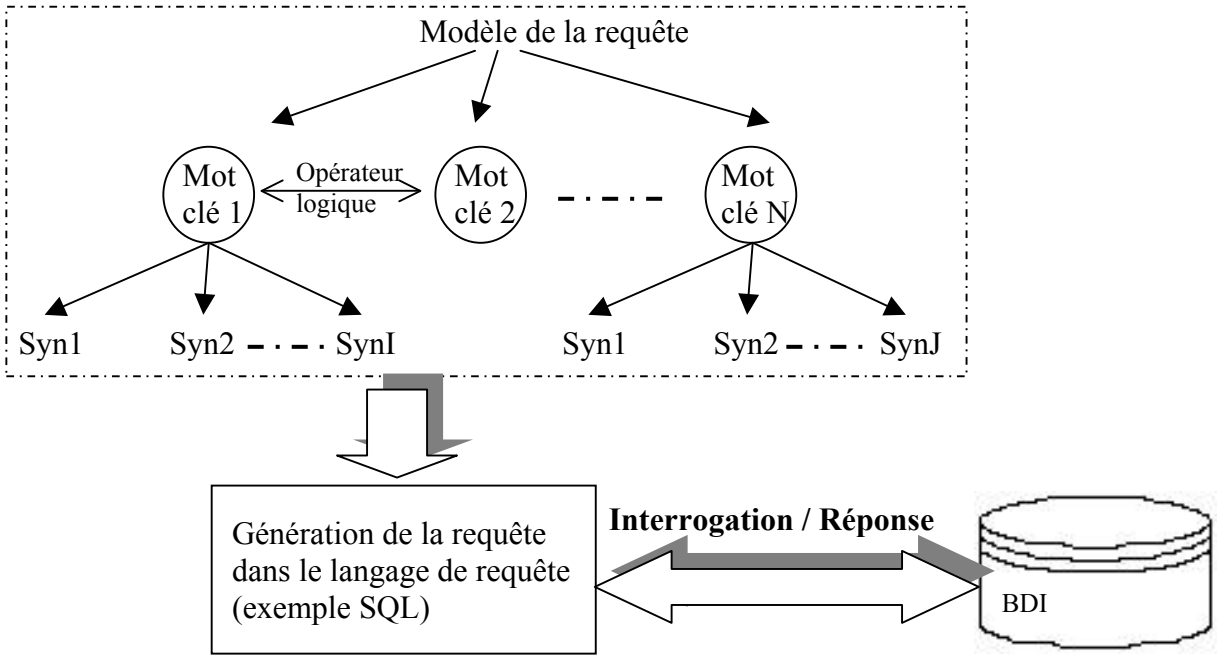
$$\mathbf{RA : a_1 \wedge a_2 \wedge a_3 \wedge \ldots \wedge a_i \rightarrow a_{i+1} \wedge \ldots \wedge a_n}$$

Elle s'interprète intuitivement de la manière suivante : si un objet possède les attributs {a1,…,ai} Alors il a tendance à posséder également les attributs {ai+1,…, an}. Dans notre contexte les objets sont les résumés textuels (corpus textuels) et les attributs sont les termes des domaines. Nous souhaitons à la lumière d'utilisation de textmining l'extraction des termes des domaines et les relations existantes entres eux afin de construire les ontologies.

Après cette étape de construction des ontologies, les ontologies doivent être validées par des experts humains avant leur mise en œuvres dans le système.

4.6. La recherche d'informations :

Le repérage d'informations s'effectue par une interrogation de la base d'indexes à l'aide d'un langage de requête. La structure de la requête dépend étroitement du modèle de requête (en fonction des mots clés, des synonymes et des opérateurs booléens).



4.7. Le regroupement des documents

Après la phase de repérage des descripteurs des documents, on se trouve avec un lot documentaire qui est pour l'instant bruité, la particularité des documents contenus dans ce lot ce qu'ils sont étiquetés par des étiquettes spécifiques : les classes.

Le regroupement est virtuel en fonction des classes trouvées, en faisant l'assemblage de ces documents par toutes les classes, on trouve la liste de tous les domaines renvoyés.

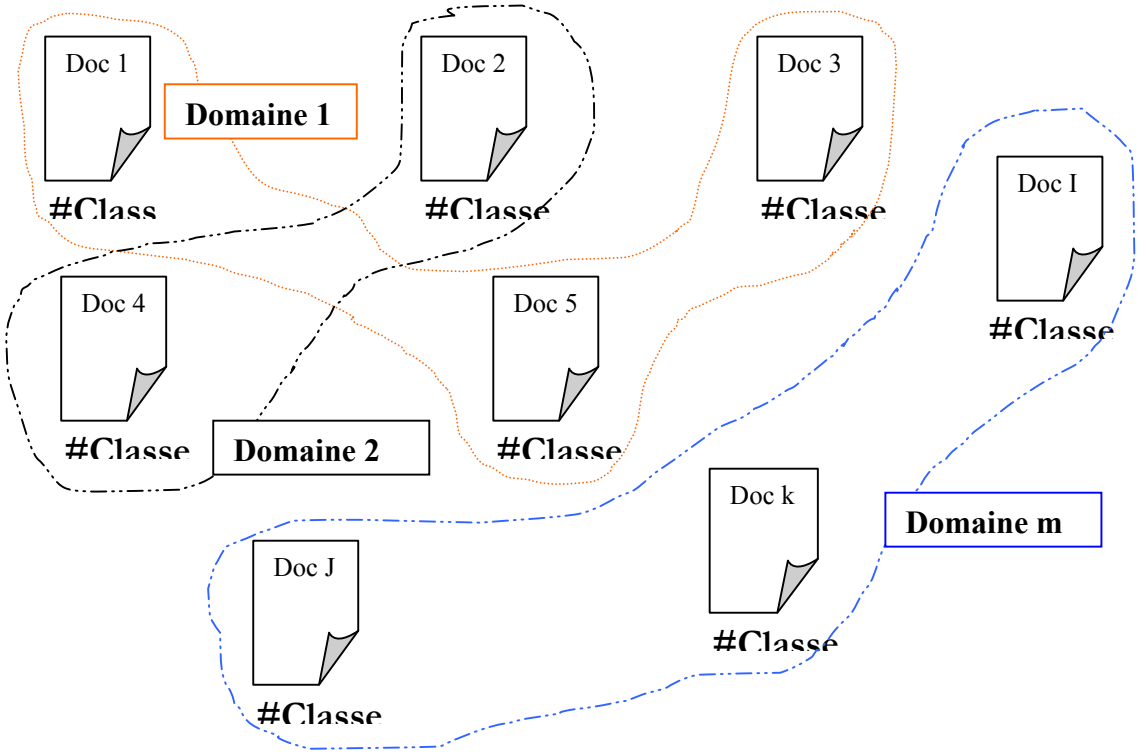


Figure 11. Assemblage des documents par domaines

4.8. Le tri des documents selon les domaines :

On entend par tri selon les domaines, le tri des documents dans un même domaine, pour chaque document on fait associer un poids relatif au nombre de mots clés de la requête, présents dans ce document, par rapport au nombre total des mots du document :

$$\text{Poids(Document X)} = \frac{\text{Le nombre des mots clés présent dans le document}}{\text{Le nombre total des mots composants le document}}$$

Le nombre de mots clés présents dans le document est la somme des fréquences des mots trouvés dans le document.

Chaque document possède un poids, le classement dans le domaine s'effectue selon l'ordre croissant des poids, c'est à dire que le document qui possède le plus grand poids sera classé en tête de liste, le suivant est celui de poids plus faible que le premier et ainsi de suite …etc.

Remarque : Un document possédant le maximum des mots clés de la requête sera privilégié par rapport à un autre qui contient un nombre plus faible de mots clés même si ce dernier a un poids plus grand par rapport au premier.

Exemple : Soit une requête utilisateur composée de 4 mots clés (sans mots vides), après une phase de recherche on a trouvé deux documents A et B de même domaine. A contient, par exemple, 3 mots clés de cette requête avec un poids égale 0.69 et B contient les 4 mots clés de la requête avec un poids de 0.32. Dans ce cas là le document B sera classé avant A par ce qu'il répond au mieux à la requête d'usager.

4.9. Le tri des domaines :

Pareille aux documents, les domaines doivent être aussi ordonnés. L'ordre de visualisation des classes par rapport aux utilisateurs est relatif à la densité documentaire par classe, c'est à dire au nombre des documents regroupés par classe. Le domaine possédant le plus grand nombre de documents sera classé en premier lieu et ainsi de suite.

4.10. La présentation des résultats

C'est par le biais de présentation des résultats que l'usager fait une première évaluation des résultats qui lui semblent intéressants. Les informations affichées doivent permettre à l'usager deux choses essentielles :

- Comprendre globalement comment le système fonctionne (quels ont été les mots considérés comme vides, quels ont été les mots retenus, les mots inférés par le système, …etc.) ;

- Evaluer facilement la liste des domaines renvoyés pour prendre son chemin.

4.10.1. Les informations affichées par le système

Ce Problème doit être vu suivant deux angles : les informations générales concernant la recherche, les informations portant sur chaque classe et chaque document-réponse :

4.10.1.1. Les informations générales

a) Nombre de classes constituées

Cette information va permettre à l'utilisateur d'évaluer la pertinence de sa question. Si le nombre est un peu élevé, il va essayer de préciser sa requête et de relancer la recherche. En plus de ce nombre de classe, à chaque classe est associé le nombre de document contenu dans celle ci, cette information permet à l'utilisateur de savoir quel domaine peut se rapprocher à sa requête.

b) Rappel de la question posée par l'utilisateur

Cette information ne semble pas extrêmement importante, puisque l'utilisateur sait quelle question il a posé. Certes, mais c'est une information importante et même essentielle du contexte de l'interrogation, et elle se doit d'être présente sur la page de résultat. Prenons un exemple concret : L'utilisateur imprime la page des résultats, comment pourra-t-il se souvenir un mois plus tard à quelle question correspond cette liste de documents si la question n'apparaît pas ?

4.10.1.2. Les informations propres à chaque document

a) La localisation du document

C'est nécessaire pour le contexte de l'URL, ceci pour connaître le nom, la localisation, le type de document (Word, HTML, Pdf, etc.), le type de protocole et de quelle organisation il provient.

b) Le titre du document

Cela peut constituer une information vitale pour un minimum de souplesse d'utilisation.

c) L'extrait du document

Il s'agit de deux ou trois lignes du document ce qui nous donne des informations sur le contenu du document.

e) La taille du document

Même s'il ne s'agit pas d'une information vitale, la taille du fichier est fournie par la majorité des moteurs de recherche Internet ou base de données. Elle permet tout de même de se rendre compte si le document a plutôt la taille d'un résumé ou d'un livre entier.

f) La date du dernier mise à jour du document

Cette information donne à l'utilisateur une bonne idée de la "fraîcheur" de l'index et donc de la pertinence de la recherche.

g) La mise en évidence des mots

La mise en évidence des mots de la question dans les documents-réponses est le moyen le plus commode de se rendre compte de la pertinence d'un document. En effet, si le système indique à l'utilisateur que tel document contient quatre des mots de la question, alors qu'un autre n'en contient qu'un, le premier semble beaucoup plus proche de ce qu'il recherche.

5. Discussions

Dans cette section nous donnerons certaines problématiques qui peuvent se déclencher avec notre modèle.

La première problématique concerne le multilinguisme des documents web. On entend par multilinguisme des documents web la diversité des langues des documents se trouvant sur internet (Anglais, Français, ….). Les ontologies que nous avons définies dans notre système sont ontologies linguistiques en langue française, donc les documents en langues différentes de la langue française ne seront pas traités par notre prototype.

Pour résoudre ce problème, on peut enrichir les ontologies par d'autres concepts d'autre langue et d'associer des nouvelles relations d'équivalences entre les concepts déjà définis et ces nouveaux concepts. En suite les documents en langue autre que la langue française peuvent être traité avec notre système. Comme ça nous avons étendu notre moteur de recherche à couvrir plus de sites web indépendamment de la langue utilisée dans ces sites.

La deuxième problématique est celle de page de « frame » (cadre). Un site web est composé généralement d'une page d'accueil qui contient une page centrale ou noyau de la page d'accueil, des cadres, des liens hypertextes pointant vers les restes des documents de site web ou vers d'autres sites, des textes …etc. Les cadres eux même sont des pages web contenant peu d'informations significatives vis à vis les utilisateurs et qui doivent être indexés et classés. Les cadres se trouvent généralement en tête de la page d'accueil, dans la partie gauche de la page, dans la droite, … etc. Une solution à ce problème consiste à considérer les pages composées de plusieurs pages sous forme des cadres comme étant une seule page et les indexer.

6. Conclusion

Dans ce chapitre, nous avons décrit notre prototype de moteur de recherche qui intègre un système de filtrage à base des ontologies. Le filtrage proposé s'appuie sur le regroupement des résultats en plusieurs classes homogènes vis-à-vis leurs contenus.

Le système proposé restitue à l'usager un véritable pouvoir d'interprétation de son besoin d'information et des réponses qui lui sont données en lui offrant une boite à outils pour organiser les informations (filtrer les informations puis les proposer aux usagers). Il faut encore ajouter, que dans le système proposé, nous avons cherché à gérer la multidisciplinarité et la multi-thématicité des termes.

Conclusion générale & perspectives

La nécessité d'associer de la sémantique aux réponses renvoyées par les moteurs de recherche sur le web pour en faciliter le traitement par des utilisateurs est aujourd'hui unanimement reconnue. Structurer les résultats fournis comme réponses et leur associer du sens sont indispensables au «jaillissement» de l'information.

Notre objectif était de considérer les résultats trouvés comme des entités qui peuvent avoir une sémantique, en attribuant des étiquettes non plus seulement aux contenus des ressources mais aussi aux contextes de ces ressources.

Le modèle de filtrage présenté propose un point de vue original sur la catégorisation des documents du web en utilisant des ontologies des domaines pour faire ceci. Ce modèle considère que les ressources se trouvant sur Internet peuvent faire partie à des domaines limités tel que l'on connaît dans le monde réel : un domaine donné peut avoir ses concepts appropries et qui ont entre eux des relations de générique ou spécifique.

La sémantique d'un document est extraite en tenant compte de son contenu et en utilisant une base d'ontologies des domaines. Dans notre modèle l'indexation n'est plus le processus classique utilisé actuellement sur les moteurs de recherche d'information, mais un processus permettant d'ajouter des informations en plus pour décrire les documents existant sur le web.

Le prototype de moteur de recherche d'information proposé permettra à l'utilisateur de retrouver de l'information en considérant les aspects suivants :

Filtrage : Le filtrage permet d'éliminer les bruits relatifs aux résultats fournis comme réponse suite à une requête, l'élimination des bruits signifie tout simplement le regroupement de tous les résultats trouvés en plusieurs **domaines**. Le regroupement touche tous les documents car on ne sait jamais quels sont les centres d'intérêts des utilisateurs, il se peut que la plus part de ces documents constituent des bruits.

Domaine (Classe) : les domaines permettent à l'utilisateur de retrouver un sous-ensemble de pages qui peut être plus proche à ses centres d'intérêt.

Les perspectives, qui nous semblent primordiales, sont l'amélioration de processus de construction des ontologies (textmining) et la réalisation pratique de notre moteur de recherche.

Bibliographies

[BID01] A. Bidault, C. Froideveaux, G. Giraldo, F. Goasdoué, C. Reynaud. « Construction d'outils pour l'integration de données du Web». Contribution de l'equipe IASI du LRI à l'action spécifique « Web Sémantique ». 2001.

[BEN03] Benjamin Nguyen, Iraklis Varlamis, Maria Halkidi et Michalis Vazirgiannis. « Construction de Classes de Documents Web». Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003.

[BEN02] Benjamin Nguyen, Iraklis Varlamis, Maria Halkidi. « Organising Web Documents into Thematic Subsets using an Ontology (THESUS) ». Journees Web Semantique 2002.

[BOU05] M. Boughanem. « Recherche d'information ». Université Paul Sabatier de Toulouse, Laboratoire IRIT. 2005.

[BUR97] Bruza, P.D. and Dennis, S. « Query re-formulation on the Internet: Empirical Data and the Hyperindex Search Engine ». In Proceedings of the RIAO97 Conference - Computer-Assisted Information Searching on Internet, 488-499, Centre de Hautes Etudes Internationales d'Informatique Documentaires. 1997

[CAT98] Catherine Leloup. Principes de fonctionnement des moteurs de recherche. 14 mai 1998.

[CHA02] Charlet Jean « L'INGENIERIE DES CONNAISSANCES DEVELOPPEMENTS, RESULTATS ET PERSPECTIVES POUR LA GESTION DES CONNAISSANCES MEDICALES ». Mémoire d'Habilitation à diriger des recherches Université Pierre et Marie Curie. Soutenu le 10 décembre 2002.

[CHA95] Chan L M. *Classification, present and future*. Cataloging and classification quarterley. 1995.

[CHA96] Charron J. « Méthodes et outils d'exploration multilingue sur Internet en vue d'une veille technologique sur un domaine restreint ». *Thèse à soutenir Université Paris VII. Débutée en 1996*.

[CHE02] Cherfi Hacène et Yankin Toussaint, Adéquation d'indices statistiques à l'interprétation de règles d'association. Article présenté dans la JADT 2002 : 6e journées internationales d'analyse statistique des données textuelles.

[CHR00] Christine Michel - Université Bordeaux III -, Lainé-Cruzel Sylvie - Université Claude Bernard Lyon-. Profil-Doc : Un prototype de système de recherche d'information personnalisé selon le profil des utilisateurs. 2000.

[COR03] C. CORMIER, J.-Y. FORTIER, G. KASSEL, C. BARRY. « Représentation de métaconnaissances pour le développement de Webs Sémantiques d'Organisation »

Laboratoire de Recherche en Informatique d'Amiens- Université de Picardie Jules Verne. Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003.

[CUT93] Cutting, D.R., Karger, D.R., & Pedersen, J.O. *Constant Interaction-Time Scatter/Gather Browsing of Very Large Document Collections*. In Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh. 1993

[DEL99] Delisle Cynthia. « Le filtrage d'information sur internet : convergences et divergences entre outils de recherche » . Mémoire de DEA en Sciences de l'Information et de la Communication, option : Systèmes d'information documentaire. école nationale supérieure des sciences de l'information et des bibliothèques. Université de Bourgogne, IUT de Dijon. Septembre 1999.

[FEL99] Feldman, S.E. «NLP meets the Jabberwocky : natural language processing in information retrieval». *Online*, 1999 Disponible sur le Web : *http://www.onlineinc.com/onlinemag/OL1999/feldman5.html*

[FED02] FREDERIC FURST. « L'ingénierie ontologique ». RAPPORT DE RECHERCHE N° 02-07. Octobre 2002.

[FED04] Frédéric Fürst. « L'opérationalisation des ontologies : une méthodologie et son application au modèle des Graphes Conceptuels ». 2004

[GOL03] C. GOLBREICH, O. DAMERON, B. GIBAUD, A. BURGUN. « Comment représenter les ontologies pour un Web Sémantique Médical? ». Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003.

[HER95] Hersh WR, Hickam D. « Information retrival in medcine :The SAPHIRE experience ». MEDINFO 1995.

[HER96] Hersh WR, Brown K, Donohe LC et al. « CliniWeb : Managing Clincal Information on the World Wide Web ». JAMIA, 1996.

[HEA95] Hearst M . *TileBars: Visualization of Term Distribution Information in Full Text Information Access*, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Denver, CO, 1995.

[IRI01] « Contribution à l'action spécifique 'Web Sémantique' -Thème- 'Ingénierie des connaissances'». Equipe « connaissances, informations et données » IRIN Nantes 2001. Adresse web : http :www.sciences.univ-nantes.fr/irin/Theme_IC

[IHA00] Ihadjadene Madjid et Laurance Favier, Vers des systèmes de découverte et de filtrage d'information documentaire : Quelle stratégie faut-il mettre en place?

ACSI2000, Association Cannadienne des sciences de l'information. Travaux du 28e congrès annuel. 2000.

[KAB00] KABBAJ M. « From Prolog++ to Prolog+CG : a CG objet-oriented logic programming language ». In *Proceedings of the International Conference on Conceptual Structures (ICCS'00)*, Springer LNAI 1867, pages 540-554, 2000.

[JER02] Jérôme Euzenat, Jean-Eric Pin et Remi Ronchaud. « Research challenges and perspectives of the Semantic Web » . Report of the EU-NSF strategic workshop. Held at Sophia-Antipolis, France. 11 January 2002

[JON99] Jones, S. Cunningham SJ. *An analysis of usage of a digital library*. Proceedings of the 2th european conference for digital library, Crete. 1999.

[LID98] Liddy, E.D. «Enhanced text retrieval using natural language processing». *ASIS Bulletin*, 1998, 24 (4). Disponible sur le Web : *http://www.asis.org/Bulletin/Apr-98/liddy.html.*

[MAR00] Marc Côté et Nader Troudi. « NetSA : Une architecture multiagent pour la recherche sur Internet ». Université Laval 2000.

[MAR01] Marie-Sophie Segret, Pierre Pompidor et Danièle Hérin. « Extraction et intégration d'informations semi structurées dans les pages web –Projet Chimère ». Montpellier 2001.

[MEY01] MEYLAN Eddy. « Introduction théorique à la gestion de données textuelles. Bases de données relationnelles objets ». ISNet15_07, Informatique de gestion et systèmes d'information. Juin 2001.

[MIC99] Michel C. « Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et évaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs ». *Thèse soutenue à l'université Lyon II le 6 janvier 1999*.

[MIN75] MINSKY, Marvin. « A Framework for Representing Knowledge ». I*n* WINSTON,P.H., *The Psychology of Computervision*, New York, McGraw-Hill, 1975, 211-277.

[NED02] Claire Nédellec, Adeline Nazarenko. « Application de l'apprentissage à la recherche et à l'extraction d'information - Un exemple, le projet *Caderige* : identification d'interactions géniques ». 2002.

[OLF03] Olfa Jenhani. « Ontologies pour le WEB: relations, construction d'ontologies et méthodes de raisonnement pour la génération de langue naturelle ». INRIA- ARC GeNI. Mai 2003.

[OLI04] Abondance : recherche d'information, référencement et promotion de sites Web http://www.abondance.com/ Maintenu par **Olivier Andrieu.**

[OLF03] Olfa Jenhani. « Ontologies pour le WEB: relations, construction d'ontologies et méthodes de raisonnement pour la génération de langue naturelle ». INRIA- ARC GeNI. Mai 2003.

[PIE02] Pierre-André BUVET et Fabienne MOREAU CYBION, mise en place d'un moteur des recherches intégrant *INTEX.* 2002

[PIE03] Pierre Pompidor, Michel Sala, Danièle Hérin, Une méthode incrémentale d'extraction de connaissances didactiques sur le Web Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003

[RAD88] Radasoa H. « Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles ». *Thèse. Université Paris Sud. Centre d'Orsay - 28 Novembre 1988.*

[SAF03] B. SAFAR, H. KEFI. « Apport d'une ontologie du domaine pour affiner une requête à l'aide d'un treillis de Galois ». *Université Paris-Sud,CNRS (LRI) & INRIA (Futurs)*. Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003.

[SEB00] Sébastien Perpette. « Définition d'une méthode de construction d'ontologies : application à la gestion des connaissances d'une équipe de recherche ». Laboratoire de recherche en informatique d'Amines, équipe d'ingénierie des connaissances, Université de Picardie Jules Verne. Janvier 2000.

[SPI99] Spink, J. Bateman, and B. J. Jansen (1998). *Searching Heterogeneous Collections on the Web: Behavior of EXCITE Users*. Proceedings of the 1998 National Online Meeting, May, New York, 1998

[SPO95] Spoerri, A. *InfoCrystal: A Visual Information Retrieval Interface*. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, 367.

[SOU 03] SOUALMIA LF et DARMONI SJ, « Une terminologie orientée Ontologie pour la recherche d'information sur la toile ». Laboratoire PSI, CNRS, INSA de Rouen. Article présenté dans la Journée Francophones de la Toile - JFT'2003. 30 juin, 1 et 2 juillet 2003.

[SYL00] Sylvain Aymard. « Médiation avec des sources d'informations en santé : contribution au projet ARIANE ». Thèse de doctorat en science de l'information et de la communication. Université d'AIX-Marseille III. 18 décembre 2000.

[TUA01] Tuan Ta Anh. « Principe de RDF & RDFS ». - Laboratoire BD – ENST 2001

[TAN00] Tanguy Larher & Katia Milbeau. « Des outils en évolution- les méta-données. ». Direction des ressources et de l'ingénierie documentaires- Internet au quotidien- CNPD- Octobre 2000.

[TIX01] B. Tixier. La problématique de la gestion des connaissances rapport de recherche. Septembre 2001

[WAC94] Wacholder, N. et R.J. Byrd**.** «Retrieving information from full text using linguistic knowledge». In Martha E. Williams (éd.) : *Proceedings of the 15th National Online Meeting 1994*. Learned Information, Inc., New York, 10-12 May 1994. Medford (New Jersey): Learned Information, Inc., 1994.

[W3C] www.w3.org

[ZEB04] Zebdi Abdelmoumen. « Un modele de Méta-Document pour dépister l'information pertinente sur le web ». Thése de magister présentée et soutenue publiquement à l'université Badji-Mokhtar, Annaba. 2004.

[ZIA04] ZIANI Radouane. « Vers un système de filtrage d'informations sur le web ». Article présenté en tant que communication au sein du SNIB'04 : Séminaire National d'Informatique à Biskra. 04-06 Mai 2004.

[ZIA05] ZIANI Radouane et LASKRI Mohamed Tayeb. « Un système de filtrage d'informations sur le web à base d'Ontologies». Article présenté en tant que communication au sein de la conférence internationale COSI'05 : Colloque sur l'Optimisation et les Systèmes d'Informations (COSI). Béjaia 12-14 Juin 2004.

Glossaire

Altavista : C'est un moteur de recherche très populaire avec la plus grande base de données sur Internet, indexant plus de 140 millions de pages. Son URL principale est http://www.altavista.com. Jusqu'en 1998, ce moteur était utilisé par Yahoo pour la recherche d'informations. Altavista indexe tous les mots d'une page et les nouvelles pages sont rajoutées dans la base de données très rapidement, généralement, dans les deux jours ouvrables. Il demande de soumettre juste la première page de site, le robot d'Altavista explorera le site et indexera les pages.

Applet : C'est un petit programme dans la page web, souvent écrit en Java, qui s'exécute au niveau du navigateur. Il est possible que la présence de ce programme stoppe l'indexation de la page par le robot.

Ask Jeeves : Un méta moteur de recherche a qui il est possible de poser des questions en anglais. Ce service est utilisé par Altavista et trouvable à http://www.askjeeves.com.

Aspirateur : Un « Aspirateur de Site » va aller récupérer toutes les données contenues par un site web et les sauvegarder sur le disque dur. De la sorte, l'utilisateur peut le consulter en local sans être connecté.

Aspirer : Télécharger exhaustivement les fichiers formant un site web, généralement à l'aide de l'outil adéquat, fort logiquement appelé un aspirateur.

Balise : Caractère particulier, ou série de caractères, utilisés pour la mise en forme d'un document (souvent du texte), et qui sera invisible pour l'utilisateur final. Un exemple type est l'insertion des liens d'un document hypertexte. Voir aussi ancre, étiquette, marque, tag en anglais.

CGI : Common Gateway Interface - interface standard entre le serveur web et d'autres programmes fonctionnant sur la même machine. Les programmes CGI sont tous les programmes qui manipulent des données d'entrée et de sortie selon la norme CGI. Dans la pratique, les programmes CGI sont employés pour manipuler des formulaires et des requêtes de base de données et produire ainsi un contenu non-statique de pages web.

DHTML : Dynamic HyperText Markup Language. HTML dynamique, version 4 d'HTML, utilisant les feuilles de style.

Frame : Concept inventé par Netscape, consistant à diviser la fenêtre d'un browser web en plusieurs petites fenêtres, dans chacune desquelles on affiche un document HTML différent (tout comme on peut avoir plusieurs documents sous Word par exemple). Chaque Frame possède son propre URL. Équivalent par ailleurs aux fenêtres MDI. La meilleur traduction semble bien être cadre.

JVM : Java Virtual Machine. interpréteur du code Java qui permet l'exécution du programme, sur une machine en particulier (le code Java restant le même d'un système à un autre).

lien mort : Lien hypertexte pointant vers une ressource n'existant plus, n'ayant jamais existé, ou sur laquelle des restrictions d'accès ont été imposées.

Hotbot : C'est un des plus grands moteurs de recherche avec ces 140 millions de pages référencées. Il utilise la base de données, la puissance de Inktomi. Les nouvelles inscriptions sont prise en compte sous deux semaines voire plus. Son adresse est http://www.hotbot.com.

HTML : HyperText Markup Language - le (principal) language utilisé pour écrire des pages Web.

HTTP : HyperText Transfer Protocol - le (principal) protocole de communication entre les serveurs web et les navigateurs (clients).

Inktomi : Cette base de données est utilisée par certains des plus gros moteurs de recherche, dont HotBot. Inktomi est aussi utilisé par Yahoo quand une requête n'est pas trouvée dans la base de données de Yahoo.

Modèle vectoriel : Proposé par Salton dans le système SMART (1970). L'idée de base consiste à représenter les documents et les requêtes sous forme de vecteurs dans l'espace vectoriel engendré par tous les termes de la collection de documents (Un terme = une dimension)

MeSH (*Medical Subject Heading***)** : C'est un thesaurus médical. C'est le thesaurus d'indexation de la base bibliographique MEDLINE. Il est traduit en français par l'INSERM et sert aussi de thesaurus au site CISMeF. Le MeSH offre une organisation hiérarchique et associative et comprend jusqu'à neuf niveaux de profondeur.

Northern Light : Un moteur de recherche avec la possibilité d'accéder de manière payante à une collection spéciale d'articles sur les affaires, la santé et la consommation. Le premier moteur de recherche a bannir les méta moteur de recherche de sa base de données. L'adresse URL est http://www.northernlight.com.

Servlet : Applet destinée à être exécutée sur le serveur et non pas chez le client.

Serveur : Un ordinateur, un programme ou un processus qui répond aux demandes d'informations d'un client. Sur l'Internet, toutes les pages web sont stockées sur des serveurs y compris les Moteurs et Répertoires de recherche qui sont accessibles de l'Internet.

SHTML : Document sur le web, en général un document HTML, qui sera traité par le serveur avant son envoi.

RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié) : c'est le langage d'indexation élaboré et utilisé par la Bibliothèque nationale de France, les

bibliothèques universitaires, ainsi que de nombreuses autres bibliothèques de lecture publique ou de recherche.

Langage d'indexation précoordonné, RAMEAU est composé d'un vocabulaire de termes reliés entre eux et d'une syntaxe indiquant les règles de construction pour l'indexation.

A la différence d'un thésaurus, la liste d'autorité encyclopédique n'est pas constituée a priori mais au fur et à mesure des besoins d'indexation et évolue sur la base des propositions faites par le réseau de ses utilisateurs.

Robots : Ce sont des programmes automatisés qui se «promènent » sur le World Wide Web et visitent des pages Web. Ils ont lu le texte sur une page et passent par des liens afin de voyager de page en page. Ce que signifie vraiment ceci est qu'elles "lisent" ou rassemblent l'information du code source de chaque page. Selon le moteur de recherche, les robots prennent surtout le titre et la description de méta tags. Les robots continuent ensuite en « lisant » le texte du corps de la page dans le code source. Ils prêtent également attention à certaines étiquettes telles que les titres et le texte.

URL : Uniform Resource Locator. Sur le web, c'est la méthode d'accès à un document distant, créant ainsi un lien hypertexte (On peut aussi désigner de cette manière des serveurs en ftp anonyme ou des sites gopher. En fait, le type de connexion peut être : file, ftp, gopher, http, news ou wais). Un URL est dit « long » quand il contient des données concernant le client et pas seulement le serveur.

W3C : Le World Wide Web Constium (W3C) est une organisation non gouvernementale mise sur pied de Tim Berners Lee en octobre 1994 au Massachusetts Institute of Technology (MIT) en collaboration avec le Centre européen de recherche nucléaire (CERN) et avec le soutien d'organismes de recherche américain et européens. L'Institut National de Recherche en Informatique et en Automatique (INRIA), en 1995, puis l'université Keio, en 1996, sont respectivement devenus les hôtes du W3C en Europe et en Asie.

Web invisible : Le web invisible est l'ensemble des ressources non indexées par les moteurs de recherche. On parle aussi de *deep web* (web profond) pour le désigner contrairement au *surface web*. Le web de surface (visible) est indexé et le web profond est difficile d'accès pour les moteurs de recherche (robots).

Web dynamique : Ce sont des pages web avec des informations qui changent ou sont changées automatiquement en fonction d'une base de données ou d'éléments provenant de l'utilisateur. Il est possible, certaines fois, de se rendre compte que cette technique est utilisée quand l'URL fini avec les extensions suivantes: **.asp, .cfm, .cgi, .shtml, .asp ou .php**. Il est aussi possible d'avoir des pages avec un contenu dynamique et finissant avec les extensions

habituelles à savoir **.html** ou **.htm**. Les moteurs de recherche référencent ces pages dynamiques de la même manière que les pages avec un contenu statique. Attention, les adresses qui contiennent le caractère ? ne sont généralement pas indexées.

WORDNET : C'est une base de données lexicales. Les termes y sont organisés sous formes d'ensembles de synonymes, les *synsets*. Chaque *synset* est un concept lexicalisé. Ces concepts lexicalisés sont reliés par des relations conceptuelles (is-a, *has-a*). Les concepteurs de WORDNET affirment ainsi construire une ontologie linguistique. WORDNET est un énorme dictionnaire hypermédia de l'anglais-américain (plus de 100 000 *synsets*) et sa richesse et sa facilité d'accès en font un intéressant outil pour la recherche d'information ou d'autres tâches comme le traitement du langage naturel.