Typologie des moteurs de recherche sur le WEB, Rapport tâche T1.1 du projet SIIRI

Jean-Pierre Chevallet

Projet SIIRI

Système Intelligent et Interactif de Recherche d'Information

Rapport Tâche T1.1

Typologie des moteurs de recherche sur le WEB

Auteur

Jean-Pierre Chevallet
Equipe MRIM
(Modélisation et Recherche d'Information Multimédia)

Laboratoire CLIPS-IMAG
Université Joseph Fourier
B.P. 53 Grenoble cedex 9

Table des matières

A. INTRODUCTION 3

B. MOYENS D'ACCES AU WEB 4

C. MODELES DE RECHERCHE D'INFORMATION 6

1 CLASSIFICATION DE BLAIR 6
2 MODELES THEORIQUES ET OPERATIONNELS 10

D. ETUDE DES MOTEURS ACTUELS 11

1 LISTE COMPARATIVE 11
2 LE SYSTEME ALTAVISTA 12
2.1 Historique et caractéristiques 12
2.2 Traitement des documents à l'indexation 13
2.3 La pondération 15
2.4 Présentation des résultats 15
2.5 Sémantique des requêtes 16
2.6 Conclusion 18
3 LE SYSTEME HOTBOT 18
3.1 Historique et caractéristiques 18
3.2 Traitements des documents à l'indexation 18
3.3 Indexation du contenu 18
3.4 Indexation de la forme 19
3.5 La pondération 19
3.6 Sémantique des requêtes 20
3.7 Conclusion 20
4 LE SYSTEME EXCITE 20
4.1 Historique et caractéristiques 20
4.2 Traitement des documents à l'indexation 21
4.3 Sémantique des requêtes 21
4.4 Conclusion 21
5 LE SYSTEME INFOSEEK 21
5.1 Historique et caractéristiques 21
5.2 La pondération 21
6 LE SYSTEME LYCOS 22
6.1 Historique et caractéristiques 22
6.2 La pondération 22
7 LE SYSTEME VOILA 22
7.1 Historique et caractéristiques 22
8 LE SYSTEME GOOGLE 23
8.1 Indexation du contenu 23
9 LE SYSTEME EUROFERRET 24

E. SYNTHESE ET BILAN 25

1 INDEXATION. 25
1.1 Indexation du contenu 25
1.2 Indexation de la forme 26
2 RECHERCHE 27
2.1 Le langage de requête 27
2.2 La présentation des résultats 29

F. CONCLUSION 31

G. REFERENCES 32

A. Introduction

Le développement du réseau Internet, par l'intermédiaire du WEB, conduit à la mise à disposition de quantités importantes d'informations grâce à la standardisation du codage des documents hypertextes (norme HTML). Cette information, si elle est disponible et quasi gratuite, n'en est pas moins difficilement accessible. En effet, si la navigation était le seul moyen de recherche pour atteindre un document, une partie de l'information serait à jamais inaccessible soit par son éloignement (en terme de liens) soit par l'absence de chemin : le web ne forme pas forcément à priori un graphe connexe.

Les moteurs de recherche ont précisément pour objectif de palier à la déficience de la navigation en proposant un accès direct aux pages par leur contenu.

Dans ce document, nous allons analyser les principaux moteurs de recherche, en examinant leurs capacités du point de vue des modèles classiques de recherche d'information. Cette étude est relativement difficile par le peu d'informations qui sont rendues publiques à propos du fonctionnement interne de ces moteurs, ceci pour des raisons commerciales. Néanmoins, nous sommes parvenus de manière indirecte à obtenir le maximum de renseignements sur les possibilités et les limites des moteurs les plus courants du WEB.

L'objectif de ce document est de fournir une typologie des moteurs en fonction de la puissance de leur langage de requête. Il est organisé de la manière suivante : la partie suivante dresse un état des différents moyens d'accéder à de l'information sur le WEB. Parmi ces moyens, nous nous concentrons dans les parties suivantes sur les moteurs de recherche. La partie C rappelle les modèles de recherche d'information sur lesquels ils se basent et la partie D présente les caractéristiques des principaux moteurs ainsi que quelles moteurs qui se distinguent par la mise en œuvre d'un modèle de RI différent des autres.

B. Moyens d'accès au WEB

Il existe actuellement plusieurs moyens d'accéder à de l'information sur le WEB. Ils sont par leur nature très différents : la manière de les utiliser et les résultats que l'on peut en attendre sont également très variés. Ces outils de recherche peuvent se classifier en deux grandes catégories :

1) **Les outils de recherche ponctuels :** ils permettent de répondre à un besoin ponctuel d'information. Parmi ces outils on peut citer :

   - Les outils de recherche thématique : appelés aussi annuaires, ils permettent de localiser un site en parcourant une classification hiérarchique de thèmes. Ces outils recensent un certain nombre de sites au travers de fiches descriptives comprenant en règle générale, le titre, l'adresse (URL) accompagné par un bref descriptif de quelques mots. Cette fiche est placée dans une catégorie permettent de retrouver de l'information en parcourant une structure hiérarchique de thèmes. Cette structure est établie manuellement et l'indexation des pages et des sites web (cf. glossaire), est également réalisée manuellement. On les confond à tort avec les moteurs de recherche car ils permettent souvent de faire une recherche sur leurs catégories. Ils permettent même parfois d'interroger des moteurs de recherche
   - Les moteurs de recherches : issus des systèmes documentaires, et des systèmes de Recherche d'Informations (SRI), ils indexent de manière automatique le contenu de tous les documents accessibles sur Internet (pages HTML, email, etc.). Ils sont caractérisés par leur couverture du Web. Cette couverture va influencer leur rappel, c'est à dire leur capacité a trouver tous les documents pertinents. Ils sont aussi caractérisé par la puissance du langage des requêtes et les choix qui guident l'ordonnancement des réponses : ces éléments influencent la précision des réponses, c'est à dire la capacité à ne retrouver que des documents pertinents. Ils sont aussi caractérisés par la « fraîcheur » de leur indexation. Cette valeur est liée à la vitesse avec laquelle ils parcourent le Web en entier. Finalement leur vitesse de réponse à une requête, et le nombre de requêtes qu'ils sont capables de traiter par jour, est un élément important pour leur popularité auprès de tous les publics du Web.
   - Les méta chercheurs : ces outils permettent d'interroger simultanément plusieurs moteurs de recherche et de fusionner leurs résultats. Certains proposent également une aide à la formulation des requêtes. Ils se basent sur les différences de couverture des moteurs du Web pour espérer augmenter le rappel. Ils se basent aussi sur les différences entre les critères de classification des moteurs pour espérer augmenter la précision des réponses. Ils restent tributaires de la qualité des moteurs qu'ils utilisent.

2) **Les outils de filtrage :** ils permettent de répondre à un besoin d'information qui s'inscrit dans une durée. Leur rôle est donc de filtrer les nouvelles informations pour un besoin précis et stable.

   - Les agents de recherche : appelés aussi « agents intelligents », ils permettent d'automatiser une recherche sur le Web en interrogeant des moteurs de recherche en différé. Ils sont adaptés pour des tâches de veille technologique.
   - Les chaînes : elles permettent de recevoir de l 'information thématique. Elles s'apparentent aux chaînes de télévision ou radio spécialisées. L'utilisateur ne peut que recevoir l'information ou changer de chaîne.

Dans le reste de ce document, nous allons développer uniquement les moteurs de recherche. Ils s'apparentent aux Systèmes de Recherche d'Informations (SRI). A ce titre, il se fondent sur des modèles. Nous présentons donc une classification des modèles de recherche d'information et nous organiserons les moteurs disponibles sur le Web selon cette classification.

C. Modèles de recherche d'information

Cette partie rappelle les modèles de recherche d'information sur lesquels s'appuient les moteurs de recherche.

Un modèle a pour objectif de représenter une partie de la réalité dans un univers formalisé. Un modèle est en adéquation avec la réalité lorsque que la partie décrite est suffisante pour que les distorsions liées au modèle puissent être négligées. Dans le cadre de la recherche d'information, un modèle formalise la manière dont le document est indexé, le langage de requête et la fonction de correspondance chargée d'évaluer la mesure de pertinence d'un document par rapport à une requête. Plusieurs catégorisations sont possible pour détailler ces modèles. Dans [1] les modèles sont classés en 11 catégories selon les critères suivants : expression de la requête, index du document, ordonnancement des réponses, fonction de correspondance. Dans la partie suivante nous présentons succinctement cette classification.

1 Classification de Blair

Dans cette classification, la notion de descripteur correspond en pratique au termes d'indexations ou aux mots clés des index manuels. Cependant, si l'on envisage une indexation plus complexe comme par exemple à base de graphes conceptuels (voir [2]), ce modèle bien qu'étant le plus simple est tout de même utilisable.

Modèle 1 : Modèle ensembliste d'appartenance

- La requête est un seul descripteur
- Les documents sont indexés par un ensemble de descripteurs
- Pas de pondération à la recherche, pas d'ordre
- Correspondance : le descripteur est dans l'ensemble de l'index.

Le second modèle élargit la requête à un ensemble.

Modèle 2 : Modèle ensembliste d'inclusion

- La requête est un ensemble de descripteurs
- Les documents sont indexés par un ensemble de descripteurs
- Pas de pondération à la recherche, pas d'ordre
- Correspondance : les descripteurs sont inclus dans l'ensemble de l'index.

Le modèle suivant relaxe la fonction de correspondance pour retrouver plus de documents. Il n'y a toujours pas d'ordre dans les réponses.

Modèle 3 : Modèle ensembliste d'intersection

- La requête est un ensemble de descripteurs avec une valeur entière S de seuil
- Les documents sont indexés par un ensemble de descripteurs
- Pas de pondération à la recherche, pas d'ordre
- Correspondance : au moins S descripteurs sont dans l'ensemble de l'index.

Le modèle suivant est toujours à base d'intersection mais introduit un ordre dans les réponses basé sur la taille de cette intersection.

Modèle 4 : Modèle ensembliste d'intersection ordonnées

- La requête est un ensemble de descripteurs avec une valeur entière S de seuil
- Les documents sont indexés par un ensemble de descripteurs
- Les documents sont classés par le cardinal de l'intersection
- Correspondance : au moins S descripteurs sont dans l'ensemble de l'index

Dans le modèle suivant, la notion de seuil disparaît au profit d'un ordre basé sur une comptabilisation pondérée des descripteurs de la requête présents dans le document. Le poids correspond à l'importance relative des termes que désire donner l'utilisateur. Cela demande cependant un effort supplémentaire au moment de l'interrogation.

Modèle 5 : Modèle ensembliste à requêtes pondérées

- La requête est un ensemble de descripteurs ayant un poids
- Les documents sont indexés par un ensemble de descripteurs
- L'ordre est donné par le calcul de la correspondance
- Correspondance : calcul de la somme des descripteurs de la requête présents dans l'index du document

Le modèle suivant est l'inverse du précédent : une pondération est appliquée aux termes des documents. L'utilisateur n'a plus à s'occuper de pondérer les termes. Par contre, il faut décider d'une pondération des descripteurs. De plus cette pondération est établie une fois pour toutes avant l'examen des requêtes.

Modèle 6 : Modèle ensembliste à index pondérés

- La requête est un ensemble de descripteurs
- Les documents sont indexés par un ensemble de descripteurs pondérés
- L'ordre est donné par le calcul de la correspondance
- Correspondance : calcule la somme des poids des descripteurs communs entre le document et la requête.

Finalement le modèle suivant combine la pondération des documents et des requêtes. Il s'agit en fait d'une vision simplifiée du modèle vectoriel de G. Salton. En effet, le calcul de la correspondance dans ce modèle est obtenu par le produit scalaire des vecteurs documents et requêtes obtenu de la manière suivante : la dimension des vecteurs correspond au nombre maximum de descripteurs; le poids dans chaque dimension est égal au poids du descripteur s'il est présent, ou à zéro sinon. Le modèle a les avantages et inconvénients des modèles 5 et 6 précédents. Si les poids subissent une normalisation par division de la norme du vecteur où il apparaissent, on obtient le modèle suivant.

Modèle 7 : Modèle vectoriel généralisé : indexation et requêtes pondérées

- La requête est un ensemble de descripteurs pondérés
- Les documents sont indexés par un ensemble de descripteurs pondérés

- L'ordre est donné par le calcul de la correspondance
- Correspondance : calcule sur les descripteurs en commun de la requête et du document, la somme du produit des poids.

Le modèle suivant est plus connu sous la dénomination de « modèle vectoriel ». Comme on l'a vu, il est une extension du modèle 7 par normalisation des vecteurs. Il est important dans la mesure où il conceptualise la recherche d'information : c'est en fait le premier modèle en tant que tel. Ce modèle est très utilisé en recherche car il favorise les requêtes longues ce qui est le cas dans beaucoup de collections de tests. En pratique il est moins intéressant car les requêtes sont beaucoup trop courtes pour permettre l'assignation automatique de poids. Il a aussi l'avantage de faciliter les calculs de retour de pertinence (relevance feedback).

Modèle 8 : Modèle vectoriel classique (règle du cosinus)

- La requête est un ensemble de descripteurs pondérés
- Les documents sont indexés par un ensemble de descripteurs pondérés
- L'ordre est donné par le calcul de la correspondance
- Correspondance : calcule le cosinus de l'angle formé par les deux vecteurs.

Le modèle 9 est très connu et très utilisé. Il a l'avantage de pouvoir indiquer le refus d'un terme (opérateur NON), mais un inconvénient majeur qui réside dans la difficulté pour l'utilisateur de manier des expressions logiques complexes : les théorèmes de la logique des propositions (ex: la loi de Morgan) ne sont pas forcément connus des utilisateurs. L'autre inconvénient est l'absence de pondération. On peut pallier à cet inconvénient de deux manières : soit en élargissant la correspondance (voir modèle 9bis), soit en proposant une classification basée sur d'autres critères que ceux de la correspondance, comme le nombre de fois où le mot est présent dans le titre ou le reste du document.

Modèle 9 : Modèle booléen

- La requête est une formule logique de descripteur avec les opérateurs ET, OU, NON
- Les documents sont indexés par un ensemble de descripteurs
- Pas de pondération à la recherche, pas d'ordre
- Correspondance : l'opérateur ET est interprété comme la présence des deux termes dans le document, le OU comme la présence de l'un ou de l'autre (non exclusif) et le NON comme l'absence du terme dans le document.

Le modèle suivant n'est pas explicitement dans la classification de Blair. Il correspond à une extension classique du modèle booléen. Plusieurs fonctions de calcul de correspondance pondérée sont possibles. Ces mesures sont parfois issues des recherches sur les ensembles flous ou sur les logiques non classiques. Nous présentons une mesure couramment utilisée.

Modèle 9 bis : Modèle booléen pondéré

- La requête est une formule logique de descripteur avec les opérateurs ET, OU, NON
- Les documents sont indexés par un ensemble de descripteurs pondérés entre 0 et 1
- L'ordre est donné par le calcul de la correspondance

- Correspondance : l'opérateur ET est interprété comme le minimum des poids dans le document des deux descripteurs, le OU correspond au calcul du maximum et le NON comme 1 moins la valeur du poids du descripteur.

Le modèle suivant est actuellement le plus utilisé pour les moteurs du Web (cf partie D). Il introduit des opérateurs d'adjacence qui ne font pas partie de la logique car ce sont des prédicats qui rendent la valeur "vrai" si les deux termes arguments du prédicat sont présents à une certaine distance dans le document. Cette distance peut s'exprimer en nombre de mots, ou bien selon qu'ils sont dans une même phrase ou un même paragraphe. Il est alors possible de formuler les requêtes suivantes : (A AND ((A NEAR B) OR (A NEAR C))) qui signifie que l'on recherche des documents ayant le terme A et dont ce terme est soit proche de B ou bien proche de C. De la même manière que le modèle booléen, il est envisageable d'étendre ce modèle vers un modèle pondéré.

Modèle 10 : Modèle booléen avec recherche plein texte

- La requête est une formule logique de descripteur avec les opérateurs ET, OU, NON avec des opérateurs de proximité.
- Les documents sont indexés par un ensemble de descripteurs avec la conservation de la distance entre les termes.
- Pas de pondération à la recherche, pas d'ordre
- Correspondance : les opérateurs logiques sont interprétés de la même manière que le modèle booléen (9) avec en plus une restriction quand à l'adjacence de termes. Selon l'opérateur utilisé, un document n'est sélectionné que si les deux termes sont à une certaine distance dans le document.

Les deux derniers modèles de la classification de Blair introduisent l'utilisation de thésaurus. Cette utilisation peut être mise en œuvre dans n'importe quel autre modèle précédent. Il s'agit plus d'une dimension supplémentaire à prendre en compte dans les modèles que de véritables modèles à part entière.

Extension modèle 11 : Thésaurus : expansion simple

- La requête est un descripteur
- Les documents sont indexés par un ensemble de descripteurs
- Expansion de la requête : le terme est étendu par tous les termes sémantiquement proches selon le thésaurus. Dans le cas du modèle booléen l'ajout se fait par une disjonction (OU).

Cette variante introduit des poids dans le thésaurus qui expriment des variations de degré de ressemblance sémantique. Cela permet à l'utilisateur de contrôler l'introduction des nouveaux termes dans sa requête. Ce contrôle peut être fait par la nature des relations dans le thésaurus lorsqu'elles sont connues (ex: généricité, spécificité, synonymie). Le choix des termes peut aussi être laissé à l'appréciation de l'utilisateur.

Extension modèle 12 : Thésaurus : expansion pondérée

- La requête est un descripteur
- Les documents sont indexés par un ensemble de descripteurs
- Expansion de la requête : le terme est étendu par tous les termes sémantiquement proches selon le thésaurus si son poids dépasse un seuil donné par l'utilisateur. Dans le cas du modèle booléen l'ajout se fait par une disjonction (OU).

2 Modèles théoriques et opérationnels

La classification précédente permet un de donner une partition et un guide pour caractériser les systèmes de recherche d'information. Nous proposons dans cette partie un regroupement de ces modèles en catégories plus larges. Depuis quelques années, les modèles de recherche d'information sont étudiés de manière théorique. Cela implique qu'il existe maintenant deux catégories de modèles [1 pp37]:

- Le modèle opérationnel : ces modèles conduisent directement vers une concrétisation pratique sous la forme d'un programme. Ils sont la formalisation de la réalisation d'informatique d'un SRI.
- Les modèles théoriques : à l'inverse des modèles opérationnels, ces modèles ont pour objectif de décrire le cadre dans lequel doit s'inscrire le système et son modèle opérationnel associé. Ces modèles ont une plus grande généralité et un caractère fédérateur qui leur permet de structurer les modèles opérationnels.

Les modèles théoriques se scindent en deux familles :

1) Les modèles logiques : ces modèles se fondent sur l'utilisation d'une logique mathématique. L'index d'un document et la requête sont vus comme des formules dans cette logique, la fonction de correspondance est une déduction logique partant de la formule du document pour arriver à la requête. Selon la logique choisie (classique, modale, floue, non monotone, etc.) , la puissance de description de ces modèles est très variable.
2) Les modèles probabilistes : ces modèles identifient la notion de pertinence comme un événement ayant une valeur de probabilité en fonction d'une requête pour un document donné.

La classification de Blair dans la partie précédente, propose une organisation des modèles opérationnels. Ils peuvent tous être décrits par une théorie logique ou bien une théorie probabiliste, sauf les modèles explicitement booléens (modèles 9, 9bis et 10), qui bien évidemment relèvent du modèle théorique logique.

Dans la suite de ce document, nous utiliserons cette classification pour catégoriser les moteurs de recherches du Web

D. Etude des moteurs actuels

Dans cette partie nous détaillons les systèmes de recherche d'informations disponibles sur le Web, les plus significatifs. Il existe actuellement une grande quantité de systèmes accessibles. Les critères que nous choisissons sont en premier la taille du corpus couvert par le système. En effet, la taille du Web est un défi à l'usage de technologies sophistiquées car le système doit toujours pouvoir indexer le plus de pages en un minimum de temps pour assurer une cohérence entre les index et le contenu réel des pages. Il doit en plus pouvoir résoudre des requêtes en quelques dizaines de secondes pour limiter l'attente de l'utilisateur, et finalement, si le système a du succès, pouvoir assurer un grand nombre de connections simultanément.

Dans l'analyse que nous faisons de ces systèmes, nous présentons le modèle d'indexation car c'est à partir des contraintes de cette modélisation que découlent les possibilités d'interrogation. Comme très peu d'informations techniques sont disponibles à propos de ces systèmes, nous avons en fait travaillé à l'inverse, c'est à dire partir des langages de requêtes.

Nous présentons en premier le système Altavista. Il nous servira de référence comparative pour la description des autres systèmes. Nous évitons ainsi une énumération redondante et fastidieuse de caractéristiques car ces systèmes sont en fait très semblables : ils sont quasiment tous basés sur le même modèle théorique, le modèle logique booléen. Seul le système EurroFerret se distingue en se basant sur le modèle vectoriel.

1 Liste comparative

Pour aider l'utilisateur à s'y retrouver dans les moteurs de recherche, il existe des pages ou des sites qui regroupe de l'information au sujet de ces moteurs. Par exemple, le site http://www.abondance.com décrit les principaux moteurs. Nous les avons regroupé dans le tableau ci-dessous classé suivant la couverture du Web, c'est à dire le nombre de pages indexées. Ces tableaux représentent donc les 10 moteurs qui couvrent le plus des pages.

Nous indiquons dans ces deux tableaux (cf. Tableau 1 et Tableau 2), la date de lancement du système, c'est à dire la date où le système a été mis en accès libre sur le réseau. Le nombre de page (en millions) donne une idée de la couverture. Le rafraîchissement est le temps annoncé par chaque système pour parcourir et ré-indexer toutes ses références. Finalement, la portée indique la zone de couverture. Par exemple, le système Voilà se limite aux pages françaises alors que le système EurroFerret se limite aux pages européennes.

	AltaVista	HOTBOT	voila	Northern Light	excite
Date lancement	12/1995	5/1996	7/1998	8/1997	10/1995
Nb de pages	140 M	110 M	100 M	80M	55 M
Rafraîchissement	2 sem	4 sem	2 sem	4 sem	6 sem
Portée	Monde	Monde	Monde	Monde	Monde

Tableau 1

	Euroferret	LYCOS	infoseek	Google!	WEBCRAWLER
Date lancement	1996	6/1995	1/1994	1998	4/1994
Nb de pages	36 M	30 M	30 M	25 M	2 M
Rafraîchissement	4 sem	3 sem	3 sem		1 sem
Portée	Europe	Monde	Monde	USA	

Tableau 2

Les parties suivantes font l'analyse de ces dix moteurs. La conclusion donnera un aperçu comparatif de ces moteurs.

2 Le système Altavista

2.1 Historique et caractéristiques

Ce système (http://www.altavista.com/) est selon leurs auteurs, un des moteurs de recherche ayant la plus vaste couverture d'indexation. Ils annoncent indexer 140 millions de pages Webs et 16000 newgroupes. Ce système semble très visité puisqu'ils annoncent un nombre d'accès de 21 millions d'utilisateurs différents par mois avec 32 millions de requêtes traitées par jour. Du point de vue matériel, le système tourne sur 23 machines au total.

D'un point de vue historique, ce système est le résultat d'une recherche amorcée durant l'été 1995 dans les laboratoires de la société Digital à Palo Alto en Californie. Il etait prévu au départ pour tester des nouvelles machine de la marque, mais il a rapidement pris un place majeur dans les moteurs de recherche à indexation automatique. Ce système n'est pas le premier qui ait présenté ses services sur le web, mais il a réussi par la couverture qu'il propose, à devenir leader dans de type de service.

Le modèle de recherche d'information utilisé est le modèle booléen étendu aux opérateurs de proximité (modèle No 10) avec un classement des résultats selon des critères de fréquence de termes.

Il propose deux types de recherches disponibles : basique et avancée. Cela correspond en fait à deux langages de requêtes différents mais qui recouvrent les mêmes opérateurs. C'est uniquement la syntaxe qui change ainsi que les valeurs par défaut comme le choix de l'ordonnancement des réponses.

2.2 Traitement des documents à l'indexation

Nous séparons dans cette partie ce qui concerne le contenu même des documents indexés c'est à dire les informations les plus proches du sens, des caractéristiques plus externes de ces documents. Nous entendons par caractéristiques externes, ou caractéristiques de formes, toutes les informations non liées à la signification (sémantique) mais plus proche de la présentation. De manière simplifiée, le contenu est ce dont parle le document, alors que la forme est comment il en parle.

a Indexation du contenu

Les documents sont analysés pour en extraire le contenu. Dans le cas de ce système, le contenu consiste en une séquence de termes dans une langue donnée. Un terme est défini de manière apparemment très simple comme une séquence de caractères n'appartenant pas à des séparateurs. Il semble qu'aucun traitement ne soit particulier à la langue du document. Les sigles par exemple ne sont pas traités correctement. Par exemple, le symbole + est considéré comme un séparateur de caractères. Il n'y a donc pas de moyen de rechercher des documents sur le langage « C++ ».

Requête (a) : *C++*
Le résultat ne prend en compte que la lettre C.

Tout le texte d'une page est indexé jusqu'à 100 Ko. Au-delà seul les liens sont indexés et au dela de 4 Mo plus rien n'est indexé (référence [Abondance])

Il n'y a pas dans ce système d'anti-dictionnaire comme on peut le constater avec la requête suivante :

Requête (b) : *le la les un une des*
*Altavista annonce : 7.218.130 pages répondant à cette requête. La page en tête est : **LES PONTS, UNE HISTOIRE D'HOMMES... - ACTUALITE 10 - LIEGE AU FIL DES PONTS** nbsp; ACTUALITE 10 - LIEGE AU FIL DES PONTS. Photo 1: panorama des ponts de Liège (vu des hauteurs de la Citadelle) Photo 2: des ouvriers au travail sur..*
***URL:** met.wallonie.org/publications/src/actu10/p02.html*

En effet, tous les mots de cette requête sont des mots outils de la langue, sauf éventuellement le terme « la » qui est soit un article défini soit un substantif (la note de musique). On remarque pourtant que ce système élimine parfois certains termes de la requête. Il semble que ce soit des mots clés qui indexent trop de documents. Cela ressemble à la notion de fréquence inverse documentaire, mais en fait, il semble plutôt que ce soit la fréquence globale sur tout le corpus qui soit utilisée pour éliminer un terme. En effet, lorsqu'un terme d'une requête est éliminé, ce système renvoie un chiffre qui dépasse parfois les 140 millions de pages qui est le nombre total de pages qu'il est censé indexer. Ce chiffre représente donc le nombre d'occurrence du terme dans tout le corpus. Par exemple la lettre « a » dans une requête est éliminée et le chiffre proposé en justification est 654.856.130, qui est supérieur au nombre de pages.

Dans la table suivante nous avons la liste des fréquences retournées pour des requêtes d'une seule lettre. On voit apparaître un seuil entre 24 et 51 millions d'occurrence.

Lettre	Fréquence globale en million	Eliminé
a	654	Oui
i	237	Oui
e	103	Oui
d	68	Oui
c	54	Oui
b	51	Oui
f	24	Non
g	23	Non
j	23	Non
h	22	Non
k	16	Non

Tableau 3

Par contre, lorsque l'on soumet une combinaison de deux termes, le résultat est soit à nouveau l'élimination des termes (requête « a b »), soit les termes sont acceptés (requête « b c ») avec un comptage qui semble être la fréquence de cooccurrence. Mais lorsque l'on indique la conjonction des termes (requête « +b +c ») ces termes sont à nouveau refusés. On constate un comportement voisin avec les requêtes avancées lorsqu'on demande un classement sur le terme. La requête sans classement « a AND f » permet d'obtenir une réponse avec 6 millions de pages. Cette même requête avec un classement sur la lettre f donne 16 millions de réponses, alors qu'un classement sur la lettre a ne donne plus de réponses ! On peut en déduire que la limite ne semble pas être d'ordre théorique, mais d'ordre pragmatique : à partir d'une certaine fréquence globale, le système ne peut plus ordonner les documents.

Les documents que nous retrouvons en tête avec ce type de requêtes à une seule lettre, sont ceux dont les titres contiennent des lettres isolées. En fait, il s'agit d'une sorte de « mise en page » erronée de la part des auteurs qui ont cru bon de rendre leurs titres plus lisibles en séparant chaque lettre par un espace. Le système prend alors chaque lettre comme un mot isolé (ex : « J u s t . F l u t e s . O n l i n e »)

Le nombre exact de documents répondant à une requête doit être demandé par une requête explicite. En dehors de cette demande, le nombre indiqué est une approximation. La requête « a AND f » avec un classement sur f indique un nombre de page égal à 16 millions, ce qui est supérieur aux 6 millions qu'annonce ce système pour cette requête.

b Indexation de la forme

Dans ce système l'analyse à l'indexation permet de repérer différents aspects des documents ayant rapport aux attributs hypertextes (les liens) ou multimédias (images, sons).

La reconnaissance des liens hypertextes lors de l'indexation, est le premier élément vers une indexation de la structure des documents. Dans ce système cependant, seul la présence d'un lien avec le nom de l'adresse HTTP est indexée. La signification de ce lien : composition (spécifique ou générique), référence, sens de lecture, etc., n'est pas prise en compte. Nous énumérons ci dessous les parties de la forme des documents HTML indexés séparément du reste du contenu de la page. Cela permet à l'interrogation qu'une partie de la requête porte

exclusivement sur cet aspect du document. Nous indiquons entre parenthèses la syntaxe à utiliser dans une requête sous la forme d'un exemple.

- Le titre : ce système sépare l'indexation du titre d'une page du reste de la page. Cela lui sert dans le calcul de la pondération (title: »Laboratoire CLIPS »).
- L'ancre : elle correspond au texte visible d'un lien hypermédia. Ce texte peut être considéré comme significatif du contenu de la page cible sauf dans le cas où son information est vide comme « cliquer ici » (anchor : »click here »).
- Le lien hypermédia : ce lien est l'identificateur unique de la page
- Le nom d'une "appel Java"

2.3 La pondération

Ce ne sont pas directement les termes des documents ou des requêtes qui sont pondérés, mais une pondération est appliquée pour classer les documents. Il n'y a pas de pondération au niveau des requêtes. Cela signifie qu'on ne peut pas indiquer qu'un terme est plus important qu'un autre.

Le poids d'un document qui va déterminer son rang est calculé à partir de :

- la fréquence du terme dans le document
- La position du terme dans le document : si le terme apparaît dans le titre ou dans le début du document, il sera jugé plus pertinent
- La position relative des termes entre eux : si les termes sont une même fenêtre de mots (une dizaine de mots)

Requête (c) : *le*
Altavista annonce : 17.701.286 pages.
Page en tête :
1. Le site officiel de l'Olympique de Marseille
index
URL: *www.olympiquedemarseille.com/*

Dans cette réponse, la fréquence dans le titre du mot est de 1, alors qu'elle est nulle dans le reste du texte.

Cette requête montre que l'apparition du terme de la recherche dans le champ titre est un critère prépondérant à son apparition dans le reste du texte.

2.4 Présentation des résultats

Ce qu'affiche Altavista est soit le titre et les premiers mots de la page, soit le contenu de méta tags comme la description de sa page par l'auteur. Ce système annonce le nombre d'apparition du terme dans tout le corpus. Exemple, si l'on prend un nom propre pour interroger, nous obtenons un faible nombre de réponses et la valeur du nombre de mots …

Requête (d) : *Chaudiron*
Il y a 41 pages qui répondent à cette requête, et le nombre d'occurrence de ce mot est également 41. On en déduit qu'il n'y a qu'une seule occurrence de ce terme par page.

2.5 Sémantique des requêtes

Ce système comme d'autres (Hotbot par exemple) , a la particularité de disposer de deux langages d'interrogation. Cela correspond en fait à deux syntaxes. Une des syntaxes correspond à un sous langage du modèle booléen en terme de puissance d'expression, c'est à dire, en terme de sémantique. Les requêtes de ce sous langage sont alors transformables dans le langage booléen général.

La syntaxe des opérateurs de l'algèbre de Boole est très mathématique et a de quoi rebuter un utilisateur non averti. C'est probablement pour cette raison que le sous langage est proposé par défaut dans ces systèmes. Malheureusement, la méconnaissance de la sémantique de ce sous langage peut vite engendrer une confusion dans son utilisation. Par exemple, l'opérateur binaire AND entre deux termes apparaît dans le sous langage comme un opérateur unaire, le + devant le terme.

Requête simple	Requête avancée équivalente	Nombre de documents
Chevallet	Chevallet	126
Bruandet	Bruandet	155
Chevallet Bruandet	Chevallet OR Bruandet	257
Chevallet +Bruandet	Bruandet	155
+Chevallet Bruandet	Chevallet	126
+Chevallet +Bruandet	Chevallet A ND Bruandet	24
-Chevallet Bruandet	Bruandet AND NOT Chevallet	131
Chevallet -Bruandet	Chevallet AND NOT Bruandet	102
+Chevallet -Bruandet	Chevallet AND NOT Bruandet	102
Chevalet Chevallet -Bruandet	(Chevalet OR Chevallet) AND NOT Bruandet	501
Chevalet +Chevallet -Bruandet	(Chevallet) AND NOT Bruandet	102
+Chevalet Chevallet -Bruandet	(Chevalet) AND NOT Bruandet	400
+Chevalet +Chevallet -Bruandet	(Chevalet AND Chevallet) AND NOT Bruandet	1

Tableau 4

Nous présentons dans cette partie l'algorithme de transformation d'une requête simple en requête avancée que l'on peut déduire au vu des résultats que fournissent ces deux modes d'interrogation. La syntaxe des requêtes simples autorise deux types d'opérateurs unaires : le +et le -. Le + binaire existe dans les requêtes simples mais n'est pas documenté. Nous avons constaté qu'une requête simple est sémantiquement équivalente à une requête avancée construite de la manière suivante :

- Tous les termes participent systématiquement au tri des documents
- L'ordre sur les termes n'a pas d'importance
- Une requête sans aucun opérateur est traduite par un opérateur OR entre chaque terme.
- Une requête avec au moins un terme ayant l'opérateur + est traduite par une expression avec des AND entre tous et uniquement les termes ayant cet opérateur +. Les autres termes participent simplement au tri.

- Un terme précédé du symbole – est ajouté à l'expression avec les opérateurs AND NOT, avec les autres termes de la requête non précédés de l'opérateur -. Les autres termes constituent une requête construite de la même manière qu'en l'absence des termes négatifs.

Nous avons confirmation de cet algorithme en examinant le nombre de documents retourné dans la requête avec la syntaxe simple, et la requête équivalente dans la syntaxe avancée (cf. Tableau 4). On peut remarquer dans ce tableau que les chiffres avancés par Altavista sont cohérents. En effet, lorsque l'on calcul les cardinaux des ensembles de documents on a :

Card(Chevallet) = 126

Card(Bruandet) = 155

Card(Chevallet AND Bruandet) = Card(Chevallet ∩ Bruandet) = 24

Card(Chevalet AND NOT Bruandet) = Card(Chevallet) - Card(Chevallet ∩ Bruandet) = 102

Card (Bruandet AND NOT Chevallet) = Card(Bruandet) - Card(Chevallet ∩ Bruandet) = 131

Card(Chevallet OR Bruandet) = Card(Chevallet) + Card(Bruandet) - Card(Chevallet ∩ Bruandet) = 257

Il faut noter qu'un espace doit obligatoirement être présent pour séparer les mots avec l'opérateur "+". Si les mots sont joints, alors cet opérateur est celui de l'adjacence. Cette fonctionnalité (ou bug ?) n'est pas documentée dans l'aide en ligne d'Altavista ni dans aucun document décrivant ce système (Cf. Tableau 5)

Requête simple	Requête avancée équivalente	Nombre de réponses
Jean+Pierre+Chevallet	"Jean Pierre Chevallet"	45
Jean +Pierre +Chevallet	Pierre AND Chevallet	59

Tableau 5

Altavista propose un chiffre qui représente le nombre d'occurrence d'un terme. Pour des nombre important de réponses, il semble que ce chiffre soit incorrect comme on le remarque dans le tableau 6 ci-dessous.

Requêtes	Nombre de réponses	Nombre d'occurence
Conceptual	176840	125790
Chevallet	126	173

Tableau 6

En effet, le nombre d'occurrence total pour le mot « Conceptual » est annoncé comme étant 125790. Or le nombre de page annoncé est de 176840 : l'un des deux chiffres sont alors incorrect car pour qu'une page soit sélectionnée, il faut qu'il y ait au moins une occurrence du terme dans la page donc le nombre d'occurrence devrait toujours être supérieur au nombre de pages.

Nous allons finalement étudier le comportement purement logique des opérateurs en examinant le nombre de documents dans le cas de formules logiques toujours vraies (tautologies) ou toujours fausses (absurdités).

a	**51 M**
NOT a	37 M
a AND NOT a	0
a OR (NOT a)	99 M
(NOT a) OR a	100 M

Tableau 7

On constate que les requêtes logiquement fausses fournissent bien aucune réponse alors que les tautologies devraient répondre tout le corpus. Le seule chose que l'on peut constater c'est que probablement pour des raisons techniques liées à la rapidité des réponses, les chiffres indiquant le nombre de réponses sont très approximatifs. Cela explique peut être les résultats du tableau précédent.

2.6 Conclusion

Ce système est une bonne illustration d'un système à base de modèle booléen semi-pondéré sans aucune analyse linguistique des documents et des requêtes. Il a pour avantage la rapidité de réponse et probablement d'indexation et sa grande couverture du corpus du WEB. Son inconvénient majeur est justement l'absence de traitement linguistique. Mais ce traitement pourrait très bien faire l'objet d'un système frontal dont le rôle serait une assistance à la formulation des requêtes à l'aide d'analyse et d'informations thématiques contenues dans des thésaurus.

3 Le système Hotbot

3.1 Historique et caractéristiques

Le moteur Hobot (http://www.hotbot.com) est un concurrent direct d'Altavista en terme de couverture du Web et vitesse de réponse. Ce système a été mis en place par la société Wired (aujoud'hui Wired Digital, racheté en octobre 1998 par Lycos) en mai 1996 soit 6 mois après Altavista.

Le modèle de recherche d'information est de type booléen plein texte, comme Altavista (modèle 10) avec deux interfaces pour les requêtes : une interface simplifiée avec un langage booléen limité, et un langage booléen complet.

3.2 Traitements des documents à l'indexation

3.3 Indexation du contenu

Le respect des majuscules n'est pas systématique comme pour Altavisa et une requête recherchant la société « NeXt » ne permet pas de retrouver cette société. Ce système fait par contre la différence entre un terme accentué et un terme sans accent. De même manière

qu'Altavista, les caractères non alphabétiques ne font pas partie des termes indexés. Il n'est donc pas possible, par exemple de rechercher des pages sur la langage « C++ ».

Requête (e) : *elephant*

1. Elephant Books
Books About Elephants 1. The Story of Babar, the Little Elephant ~ Usually ships in 24 hours Jean De Brunhoff, Jean Debrunhoff / Hardcover / Published 1966 Our Price: $9.80 ~ You Save: $4.20 (30%) Read more about this title.. 2. Horton Hears a Who..
99% 3/21/99 http://www.perpetualpreschool.com/newpage8.htm
permet de retrouver des pages anglaises

Requête (f) : *éléphant*

1. Affichage du volume : éléphant
Affichage du volume = LE livre L'éléphant / Serinam Lefakir Lausanne : EPFL, 1995 123 p. : ill. ; 24 cm (Collection « les animaux » ; vol. 4. Sér. B, Animaux qui marchent) Remarquez : en bas du descriptif du titre du livre « L'éléphant » se trouve le...
99% 3/6/95 http://admpc3.epfl.ch/col_elep.htm
ne retrouve que des pages françaises

3.4 Indexation de la forme

Les requêtes peuvent porter sur :

- Les mots du titre de la page
- Les liens dans la page vers une autre adresse : cette fonction est identique à la fonction Link d'Altavista.
- Le domaine de la page : pour limiter la recherche aux pages dont le nom du serveur se termine par un certain domaine.
- Limite de profondeur sur le serveur (depth :n) : il est possible de limiter les pages ayant une position limite en nombre de dossiers à parcourir sur l'adresse HTTP. Cette fonctionnalité est un embryon d'indexation sur la structure d'un site.
- Type de documents contenus (feature :x) : on peut limiter les pages en imposant qu'elles contiennent un certain type de fichier. Par exemple on peut choisir les types acrobat (fichier .pdf), applet Java, des formulaires HTML, des images, des tables, de la vidéo, des fichiers son etc. Il est même possible d'indiquer le type d'extension que le fichier doit contenir.
- La date du document : soit celle fournie dans la page, soit celle qui correspond à la date d'indexation par le robot.

3.5 La pondération

Contrairement à Altavista, il semble qu'il y ait directement une pondération associée aux termes qui indexent les documents. Cette pondération semble être une mesure combinant plusieurs critères dont le nombre d'occurrences du terme par rapport à la taille du document. Par contre, il n'y a pas de pondération possible des termes des requêtes. La pondération est obtenue par une combinaison des critères suivant :

- Nombre d'occurrences du terme dans le document
- Présence du terme dans le titre
- Présence du terme dans la zone des mots clés
- Taille de la page : c'est une différence par rapport à Altavista.
- La cohérence : ce système se permet d'éliminer les documents qui présentent un contenu qui présente manifestement des aspects pour tromper le moteur et apparaître artificiellement en tête de liste. On peut citer les techniques qui répètent le même mot clé, ou qui contiennent du texte non visible servant uniquement à manipuler la fonction de pondération.

3.6 Sémantique des requêtes

Comme Altavista, Hotbot a deux syntaxes de requêtes. La recherche simple permet de choisir soit la conjonction entre tous les mots (ET) soit la disjonction (OU). Il est possible aussi de rechercher une séquence de termes.

Une particularité interressante pour ce système est la possibilité à l'interrogation de produire les variations grammaticales d'un terme (féminin, pluriel). C'est une façon simple de palier à l'absence de lemmatisation à l'indexation.

3.7 Conclusion

Ce système est très proche d'Altavista en terme de fonctionnalité. Les différences sont dans quelques possibilités supplémentaires de filtrage comme celle qui traite de l'adresse URL, ou comme la prise en compte dans le classement de la taille des documents

4 Le système Excite

4.1 Historique et caractéristiques

Ce système (http://www.exite.com) est l'un des plus utilisé sur le réseau. Il est né en 1993 à l'initiative d'étudiants de l'Université de Stanford en Californie. Une société a rapidement été créée pour développer cette technologie. Cette société a racheté les moteurs WebCrawler et Magellan.

Sa base est plus faible que les systèmes Altavista et Hotbot. Il a comme caractéristique originale de posséder un thésaurus important qui lui permet de faire de l'expansion de requêtes mais aucune information n'est disponible sur cette particularité. Une autre caractéristique intéressante est la possibilité de faire un semblant de retour de pertinence avec la possibilité de demander un document qui ressemble à un document retourné dans la liste. Ce n'est pas un retour de pertinence complet puisqu'on se limite à un seul document.

Une autre caractéristique originale est la possibilité comme Altavista de faire de l'expansion de requêtes à l'aide d'une série de mots proposés à partir des documents retrouvés.

4.2 Traitement des documents à l'indexation

a Indexation du contenu

La requête « C++ » donne ici des résultats corrects mais il ne semble pas pour autant que l'analyse des documents permette de retrouver des mots contenant des symboles. La requête semble être traitée à part.

b Indexation de la forme

4.3 Sémantique des requêtes

Contrairement aux autres moteurs, Excite ne permet par de recherche par termes tronqués grâce à un caractère joker (comme *). Il permet comme les deux autres des recherches sur des mots consécutifs. Les requêtes sont booléennes avec la sémantique habituelle mais sont limitées et on ne peut pas directement utiliser les opérateurs dans une expression parenthésée mais il faut indiquer dans des champs les mots reliés par un ET, ceux par un OU et ceux précédés pas un NON.

4.4 Conclusion

Il s'agit donc d'un système basé sur un modèle de type booléen avec recherche plein texte (modèle 10)

5 Le système Infoseek

5.1 Historique et caractéristiques

Le moteur Infoseek (http://infoseek.go.com) est un moteur de recherche assez ancien puisqu'il a été mis en service dès 1994. La taille de son index est bien en deçà des chiffres d'Altavista ou Hotbot. Il n'a pas de fonctionnalités supplémentaires par rapport aux autre moteurs étudiés. Par contre, il ne permet pas la recherche par troncature, ni la prise en compte des accents. Sa pondération tenant compte du nombre de documents où un terme apparaît dans la base (frequence documentaire inverse) laisse présager de meilleurs classements que les autres moteurs.

Une particularité de la présentation des résultats est la possibilité de trouver des documents similaires à ceux de la réponse : c'est une sorte de retour de pertinence limité à 1 document.

Un dernier aspect original est que les réponses sont groupées par site, pour ne pas encombrer la page des résultats. On peut ensuite voir toutes les réponses de ce site dans une autre page.

5.2 La pondération

Le classement des documents se fait selon les critères suivants :

- La présence des mots dans le titre ont une pondération plus importante que les mots du contenu.
- Le nombre d'occurrences du mot dans le document
- La fréquence inverse documentaire : c'est une caractéristique rarement utilisée dans les moteurs de recherches. Les mots apparaissant dans beaucoup de documents du le corpus

ont un poids plus faible ; cela correspond au calcul de la fréquence documentaire inverse (inverse document frequency : idf)

6 Le système Lycos

6.1 Historique et caractéristiques

Le moteur Lycos date de 1995 et a le même ordre de grandeur de couverture Web qu'Infoseek. L'adresse générale : http://www.lycos.com renvoie à une page en France : http://www.fr.lycos.de La page anglaise est à l'adresse : http://www-english.lycos.com

Comme tous les autres moteurs, l'indexation ne permet pas la recherche de termes non composés de lettres comme « Canal+ ». Le modèle employé est comme les autres le modèle booléen plein texte avec pondération des documents.

Une originalité (peu courante) de ce système est la possibilité d'influencer le calcul de la correspondance en influençant la pondération des différents éléments qui interviennent.

6.2 La pondération

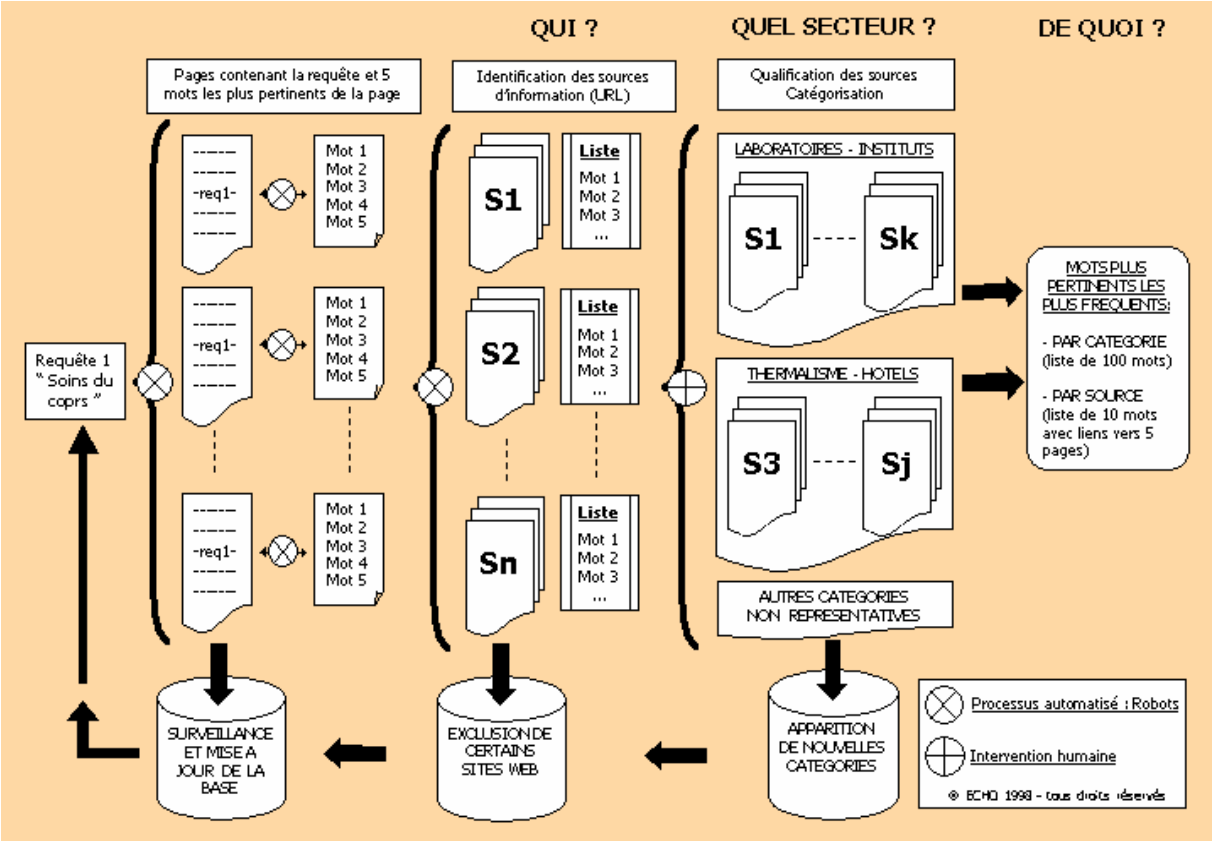
Elle se base sur les critères suivants :

- Le nombre de mots qui correspondent : cela permet de dire que le modèle employé ressemble à un modèle ensembliste avec pondération (modèle No 6).
- Le nombre d'occurrences du terme dans le document
- La présence du terme près du début du document
- La présence des mots les uns près des autres
- L'apparition dans le titre
- Les mots dans l'ordre exact

7 Le système Voilà

7.1 Historique et caractéristiques

C'est un système de recherche (http://www.voila.com) construit par la société Echo (http://www.echo.fr) basée à Sophia Antipolis dans le sud de la France. Selon ses auteurs, ce système indexe 6 millions de pages francophones, 100 millions de pages internationales. Cette société propose aussi un système de veille marketing qui permet par exemple d'analyser un corpus de mots représentatifs d'une thématique. En fait, il doit analyser des sites Web en fonction d'un thème restreint. Je ne pense pas que la technologie utilisée soit très originale, mais c'est plutôt le type d'utilisation de l'information disponible sur le Web qu'il faut noter ici. En effet, ce sont des services qui proposent de nouvelles informations à un utilisateur de type « push » (par mail par exemple) sur un thème donné.



a Figure 1 : Fonctionnement de Internet Word Map **© ECHO 1998 - Tous droits réservés**

8 Le système Google

Ce système [3] est un produit expérimental issu de l'université (http://google.stanford.edu) de Stanford. Ce système va évoluer vers une commercialisation (http://www.google.com).

Il a pour but d'expérimenter de nouvelles techniques d'indexation. Par exemple, il implémente la notion de probabilité qu'en naviguant au hasard, on arrive à une page donnée. Cette information est approximée à partir du nombre de pages qui référence une page. Cette information est aussi appelée "l'indice de popularité".

Les possibilités de requêtes sont limitées à une séquence de termes car l'essentiel de ce que propose ce système est réalisé à l'indexation. La proximité des termes par exemple augmente le score du document.

Ce système n'est pas basé sur un modèle booléen. Je pense qu'il faut plutôt le classer parmis les modèles ensemblistes à index pondérés (modèle 6).

8.1 Indexation du contenu

Tous les mots des pages sont indexés à l'exception des termes trop rares. La position des mots dans la page est limitée à 12 bits. Ce système est donc limité à la reconnaissance de la position des 4095 premiers mots d'une page.

Le texte des liens est traité de manière spéciale. Il n'est pas simplement ajouté au reste de la page. Ici le texte d'un lien à ajouté à la page sur laquelle il pointe.

9 Le système Euroferret

Ce système de la société Muscat (http://www.muscat.co.uk/) est un des rare systèmes qui ne soit pas basé sur le modèle booléen, peut être le seul qui indexe plus d'un million de pages. Ses créateurs prétendent même qu'il ont plus de pages européennes que les système Altavista, Hotbot et Infoseek. Le site annonce les chiffres de 36 millions de pages européennes pour EuroFerret, 24 millions pour Altavista et 18 millions pour Hobot. Les chiffres ont été calculés sur la base de requêtes posés aux systèmes. Nous n'avons pas trouvé de Il est basé sur le modèle vectoriel. Il semble que les calculs soient du type probabilistes.

Ce système est également un des rares qui propose une lemmatisation des termes, en fait par troncature sans tenir compte de la langue. Le fait qu'il soit basé sur le modèle vectoriel, lui permet de proposer un retour de pertinence (relevance feedback). Il suffit de sélectionner les documents qui sont jugés pertinents pour la requête et de demander une reformulation automatique.

Le système fonctionne sur deux serveurs Sun Ultra 1 (140 MHz). De ce fait;, le service offert souffre d'une certaine lenteur. L'indexation des documents ne se fait que sur les 60 mots les plus importants et sur 12 expressions clés. Cela leur permet de réduire la taille de l'index puisqu'ils n'ont en moyenne que 1Ko d'index par documents. Leur index total est donc de l'ordre de 36 Go.

Nous n'avons pas trouvé d'informations sur le choix des mots et des expressions. Nous pensons qu'il s'agit très probablement de mesures à base de fréquences des termes et d'inverse de fréquences documentaires. Mais cela reste à vérifier.

E. Synthèse et bilan

Les tableaux ci-après synthétisent la typologie que nous avons étudiée pour les principaux moteurs de recherche du WEB

1 Indexation.

Comme décrit dans les parties précédentes, nous distinguons l'indexation du corps du document que nous appelons contenu, de l'indexation des autres aspects des documents que nous appelons la forme.

1.1 Indexation du contenu

Nous commençons par la phase d'indexation avec le modèle de recherche d'information qui est utilisé selon la classification donnée dans la partie C.1. Les Tableau 8 et Tableau 9 indiquent les traitements du corps du documents. Seul le système Eurroferret utilise un antidictionnaire pour filtrer les mots vides. Malheureusement, comme la langue n'est pas (correctement?) reconnue, cela ne semble fonctionner que pour l'anglais.

	AltaVista	HOTBOT	Northern Light	excite	Euroferret
Modèle	10 et 11	10	10	9 et 11	7 ou 8
Pondération document	tf	tf	tf	tf	tf idf ?
Anti-dictionnaire	–	–	–	–	●
Elimination des mots par la fréquence	●	–	–	–	–
Lemmatisation ou troncature	–	–	–	–	●
Minuscules/Majuscules	●	●	–		–
Accents	●	●	–	●	–
Position relative des mots	●	●	●	–	–

Tableau 8 : indexation du contenu

La lemmatisation qui est réalisée au moment de l'indexation est à différencier de la troncature possible à l'interrogation. Encore une fois, seul le système Euroferret réalise une troncature à l'indexation.

L'indexation de la position relative des mots permet à l'interrogation d'interroger en donnant une contrainte sur la distance maximale où deux termes peuvent apparaître dans un document.

Nous n'avons pas mis dans la table la reconnaissance de la langue du document car cette reconnaissance quand elle existe, n'influence pas l'indexation : tous les systèmes indexent au niveau du mot et ne reconnaissent pas les termes sauf EuroFerret qui sélectionne les mots.

	LYCOS	infoseek	Google!	voila	WEBCRAWLER
Modèle	9	9 et 10	6	10	10
Pondération document	tf	tf et idf	tf	tf	tf
Anti-dictionnaire	–	–	–	–	–
Elimination des mots par la fréquence	–	–	● (trop rare)	–	–
Lemmatisation ou troncature	–	–	–	–	–
Minuscules/Majuscules	–	●	●	–	–
Accents	●	–		–	●
Position relative des mots	● (25 max)	●	● (4095 max)	●	●

Tableau 9 : indexation du contenu

1.2 Indexation de la forme

Parmi les éléments des documents, se trouvent des informations non visibles à l'utilisateur comme les mots clés ajoutés par le créateur de la page à indexer. Ces mots clés sont laissés au libre arbitre du rédacteur de la page, et aucune vérification de la cohérence de cette information avec le contenu effectif des pages n'est possible.

	AltaVista	HOTBOT	Northern Light	excite	Euroferret
Indexation du titre	●	●	●	●	?
Balise keywords	●	●	●		?
URL de la page	●	Uniquement le nom du serveur	●	●	–
Nom fichiers (images,son,etc)	●	●	–	–	–
URL des liens sortant	●	●	●	–	–
Texte des liens sortant	●	●	?	–	–
Indice de popularité	–	●	–	●	–
Date du document	●	●	●	–	–

Tableau 10

L'URL de la page peut être indexée : ce sont alors les différents noms des sous répertoires qui entrent dans l'index. La recherche peut alors se faire sur le nom du serveur, sur la profondeur dans la hiérarchie, ou tous simplement une URL avec des répertoires aux noms explicites sera retrouvée au même titre que le contenu de la page.*

Les noms de fichiers contenus dans une page peuvent être indexés au même titre que les URL qui permettent la navigation à partir d'une page.

	LYCOS	infoseek	Google!	VOILA	WEBCRAWLER
Indexation du titre	●	●	●	●	●
Balise keywords	–	●	?	●	–
URL de la page	–	Uniquement le nom du serveur	●	●	–
Nom fichiers (images,son,etc)	–	●	–	●	–
URL des liens sortant	●	●	●	–	–
Texte des liens sortant	–	●	●	–	–
Indice de popularité	●	●	●	–	●
Date du document	–	–	–	–	–

Tableau 11

L'indice de popularité correspond à un indice relatif au nombre de pages qui pointe sur la page concernée. L'hypothèse qui est faite est qu'une page très référencée par d'autre pages reflète sa popularité dans le WEB et donc très probablement son intérêt de la part des utilisateurs, indépendamment de son contenu véritable.

2 Recherche

Dans cette partie, nous nous intéressons aux possibilités de recherche. En premier nous comparons la richesse des langages de requêtes.

2.1 Le langage de requête

Comme tous les systèmes étudiés, sauf EuroFerret, sont à base de modèle booléen, ils possèdent tous les opérateurs AND et OR de la logique booléenne. Ce qui les différencie par contre c'est pour certains l'absence de l'opérateur NOT, et pour d'autre l'impossibilité.

	AltaVista	HOTBOT	Northern Light	excite	Euroferret
Opérateur NOT	●	●	●	AND NOT	–
Elimination des mots par la fréquence	●	–	–	–	–
Proximité	10		●	–	–
Adjacence	●	●	●	●	–
Joker de troncature	●	●	●	–	–
Syntaxe logique complête	●	●	●	–	–
Expressions	–	–	–	●	–
Variation morphologique	Vérification orthographe	–	–	–	–
Profondeur dans l'URL	–	●	–	–	–

Tableau 12

L'opérateur de proximité permet de réduire la recherche dans une fenêtre d'une certaine taille. C'est cette taille qui est indiqué dans la table. L'opérateur d'adjacence, la plupart du temps présent avec la syntaxe des guillemets (""), permet de forcer la présence de plusieurs termes en séquence. On peut noter que les systèmes Altavista (et Google?) proposent un opérateur de proximité. Seul le système WebCrawler permet de faire varier dans la requête la taille de cette proximité de 2 à 50 mots. C'est aussi le système qui a la plus faible couverture. Il est vrai que cette fonctionnalité consomme de l'espace de stockage.

Le caractère joker de troncature permet de palier à l'absence de lemmatisation. Seuls les trois premiers systèmes permettent une troncature à droite, c'est à dire en fin de mot.

Pour tous les systèmes à base de modèle logique, les requêtes sont des expressions logiques des opérateurs de bases. Seul quelques systèmes offrent la possibilité d'utiliser la syntaxe complète de l'algèbre de Boole. Les autres n'autorisent pas le parenthésage et limitent donc quelque peu le langage d'expression à :

(t1 AND … ti) AND (ti+1 OR ti+2 OR … ti+n) AND (NOT ti+n+1 AND NOT ti+n+2 AND NOT … ti+n+m)

pour les requêtes ayant i termes obligatoire, n termes optionnels et m termes exclus.

	LYCOS	infoseek	Google!	voila	WEBCRAWLER
Opérateur NOT	●	●	–	AND NOT	●
Elimination des mots par la fréquence	–	Limité a 1 doc	–	–	–
Proximité	●	–	–	–	Variable jusqu'à 50
Adjacence	●	●	–	●	●
Joker de troncature	●	–	–	–	–
Syntaxe logique complète	–	–	–	–	●
Expressions	●	–	–	–	–
Variation morphologique	●	–	–	●	–
Profondeur dans l'URL	–	–	–	–	–

Tableau 13

Seul le système Hotbot offre la possibilité de limiter la profondeur dans l'URL d'une page. Nous ne savons pas vraiment si cette fonctionnalité est utile car la position d'un fichier dans un site ne reflette pas forcément son importance. Néanmoins, conjugé avec le contenu de l'URL, et de la nature de la page recherchée (personnelle, officielle), et compte tenu du besoin (recherche l'information ou un point de départ de navigation), on peut imaginer une utilité à cette contrainte si elle est encapsulée dans un outils d'aide à la formulation de requête comme nous désirons le faire dans le projet SIRII.

Deux systèmes proposent de palier à l'absence de lemmatisation par la possibilité d'interroger les variations morphologiques des termes de la requête. Le système Altavista ne propose pas cette variation, mais propose une vérification de l'orthographe. Par exemple :

Requête (g) : *"pomme de terrre"*
Altavista corrige la faute de frappe des 3 "r" du mot "terre"

2.2 La présentation des résultats

Tous les systèmes présentent leurs résultats dans une page de liens avec le titre de la page et parfois le début du texte du document.

La mesure de pertinence synthétise en un seul chiffre un indice de qualité calculé par le système. C'est une composition de plusieurs mesures. Certains systèmes groupes les documents et permettent d'obtenir une liste de documents proche d'un document fournit en réponse. C'est une sorte de retour de pertinence très simplifié.

Le groupement par site permet de présenter les pages en ne faisant apparaître qu'un seule page par site. L'utilisateur a alors la possibilité d'examiner à part les autres pages du site répondant à la requête. Cette présentation peut faciliter la lecture des résultats.

	AltaVista	HOTBOT	Northern Light	excite	Euroferret
Mesure de pertinence	–	●	●	●	●
Document proche	–	● (les 5 plus visités)	–	●	–
Groupement par site	–	●	–	–	–
Retour de pertinence	–	–	–	–	●
Expansion de requête	●	–	–	●	●

Tableau 14

Le retour de pertinence permet de sélectionner les documents qui correspondent à la notion de pertinence de l'utilisateur, et permet aussi d'exclure les documents qui ne correspondent pas. Le système adapte alors ses paramètre de mesure de correspondance en conséquence. Seul le système Euroferret implante cette fonctionnalité.

	LYCOS	infoseek	Google!	voila	WEBCRAWLER
Mesure de pertinence	–	●	●	–	–
Document proche	–	●	–	–	–
Groupement par site	–	●	–	●	–
Retour de pertinence	–	–	–	–	–
Expansion de requête	–	–	–	●	–

Tableau 15

L'expansion de requêtes permet à l'utilisateur de compléter as requête en choisissant dans une liste de termes proposé par le système en fonction de la requête initiale.

F. Conclusion

L'état des lieux des moteurs de recherche sur le WEB, nous permet de dire que pour l'instant les systèmes proposés sont basés sur des technologie très basiques sans aucun traitement de l'information (langue, termes). Ces systèmes sont plus proches des données brutes que du besoin de l'utilisateur. Ce dernier doit alors faire de gros effort pour trouver une requête capable d'approcher son besoin. Seul les systèmes EuroFerret et Google proposent une approche légèrement plus élevée que les autres systèmes.

Nous avons constaté que lorsqu'une version simplifiée de la syntaxe des requêtes était fournie, sa sémantique n'était pas évidente. Nous en concluons que pour réaliser une application frontale à ces systèmes, il est plus sur, quand c'est possible, d'utiliser la syntaxe complète du langage de l'algèbre de Boole.

Il est frappant également de constater que la technologie mise en œuvre dans des systèmes date en faite des années 50, c'est à dire des tous débuts des recherches dans le domaine de la recherche d'information. La raison est probablement technique car pour couvrir la masse d'information colossale que représente le Web, ces techniques frustres sont faciles à implanter. Mais c'est aussi un choix technologique car l'exemple du système Google montre que l'on peut s'écarter du système basique booléen et fournir out de même des bons résultats avec une syntaxe des requêtes réduite à une séquence de mots.

Notre proposition dans le projet SIRII d'utiliser ces moteurs de recherche comme élément basique à travers une interface d'analyse de la requête prend alors tous son sens : le moteur de recherche est vu comme une simple mais très vaste base de données d'indexation plein texte simple. Le rôle de cette interface sera alors de tirer au mieux partie des spécificités des moteurs que nous avons étudiés dans ce projet.

G. Références

http://searchenginewatch.com : un site très complet, mis à jour régulièrement. Contient des information sur la qualité des moteurs à des fin commerciales.

http://www.abondance.com/outils/comparatif.html : contient les chiffres d'un comparatif entre les moteurs.

http://www.ambrosiasw.com/~fprefect/matrix/matrix.html : une comparaison de quelques moteurs de recherche.

http://www.doubleclick.net/advertisers/altavista/whatis_av.htm : Les chiffres et l'historique sur Altavisat proviennent de cette page.

http://www.lub.lu.se/desire : il s'agit d'un projet Européen sur la recherche d'information. Il contient des information intéressante sur les moteurs mais qui sont un peu anciennes.

http://www-scd-ulp.u-strasbg.fr/urfist/Fiches_Techniques/fiches.html AYMONIN David. (1996). 'Liste des fiches techniques concernant les outils de recherche d'information sur Internet', URFIST Alsace Lorraine Franche-Comté. Mise à jour permanente En une ou deux pages un résumé complet des astuces et techniques d'utilisation des moteurs, des répertoires thématiques, et de bien d'autres outils et services permettant de mener des recherches d'information sur le Web.

---

[1] D.C. Blair, Language and Representation in Information Retrieval, *Elsevier Science Publishers, 1990*

[2] Jean-Pierre Chevallet, Un Modèle Logique de Recherche d'Informations appliqué au formalisme des Graphes Conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels, *Thèse de Université Joseph Fourier, Grenoble, 1992*

[3] . Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of 7th World Wide Web Conference*, 1998