I.1 Introduction

La Recherche d'Information (RI) peut être définie comme une activité dont la finalité est de localiser et de délivrer un ensemble de documents à un utilisateur en fonction de son besoin en informations. Le défi est de pouvoir, parmi le volume important de documents disponibles, trouver ceux qui correspondent au mieux à l'attente de l'utilisateur.

L'opération de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI), ces systèmes ont pour but de mettre en correspondance une représentation du besoin de l'utilisateur (requête) avec une représentation du contenu des documents (fiche ou enregistrement) au moyen d'une fonction de comparaison (ou de correspondance). L'essor du web a remis la RI face à de nouveaux défis d'accès à l'information, il s'agit cette fois de retrouver une information pertinente dans un espace diversifié et de taille considérable. Ces difficultés ont donné naissance à une nouvelle discipline appelée Recherche d'Information sur le Web.

I.2 La recherche d'information

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information.

I.2.1 Définitions

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années, nous citons dans ce contexte les trois définitions suivantes :

- ***Définition 1*** *:* La recherche d'information est une activité dont la finalité est de localiser et de délivrer des granules documentaires à un utilisateur en fonction de son besoin en informations [1].

- ***Définition 2*** : La recherche d'information est une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information [2].

- ***Définition 3*** *:* La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information [3].

SRI *:* système de recherche d'information

Toutes ces définition partagent l'idée que la RI a pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes qui reflètent un besoin d'information.

I.2.2 Concepts de base de la recherche d'information

La recherche d'information est considérée comme l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptibles de répondre aux besoins de l'utilisateur. La gestion de ces informations implique le stockage, la recherche et l'exploration des documents pertinents. De ce contexte plusieurs concepts clés peuvent être définis, nous avons donc trouvé utile de les clarifier. A cet effet une synthèse des travaux de [4] et [5] nous a permis de dégager les concepts suivants:

- ***Collection de documents :*** la collection de documents (ou fond documentaire) constitue l'ensemble des informations exploitables et accessibles. Elle est constituée d'un ensemble de documents. Dans le cas général et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. Ces représentations sont étudiées de telles sortes que la gestion (ajout suppression d'un document) ou l'interrogation (recherche) de la base se font dans les meilleures conditions de coût.

- ***Document :*** le document constitue l'information élémentaire d'une collection de documents. L'information élémentaire, appelée aussi granule de document, peut représenter tout ou une partie d'un document.

- ***Besoin d'information :*** la notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis par [6]:

  - Besoin vérificatif *:* l'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit stable, c'est-à-dire qu'il ne change pas au cours de la recherche.
  - Besoin thématique connu *:* l'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connus. Un besoin de ce type peut être stable ou variable : il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label.

RI *:* Recherche d'information

- Besoin thématique inconnu *:* cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations en dehors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

- *Requête :* la requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature. Une requête est un ensemble de mots clés, mais elle peut être exprimée en langage naturel, booléen ou graphique.

- *Modèle de représentation :* un modèle de représentation est un processus permettant d'extraire d'un document ou d'une requête, une représentation paramétrée qui couvre au mieux son contenu sémantique. Ce processus de conversion est appelé indexation. Le résultat de l'indexation constitue le descripteur du document ou de la requête, qui est une liste de termes ou groupes de termes (concepts), significatifs pour l'unité textuelle correspondante, auxquels sont associés généralement des poids, pour différencier leurs degrés de représentativité du contenu sémantique de l'unité en question. L'ensemble des termes reconnus par le SRI est rangé dans une structure appelée dictionnaire constituant le langage d'indexation. Ce type de langage garantit le rappel de documents lorsque la requête utilise dans une large mesure les termes du dictionnaire. En revanche, il y a risque important de perte d'informations lorsque la requête s'éloigne de ce vocabulaire.

- *Modèle de recherche :* il représente le modèle du noyau d'un SRI. Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer. Il est utilisé pour la recherche d'informations proprement dite et est étroitement lié au modèle de représentation des documents et des requêtes.

I.2.3 Les modèles de recherche d'information

Un modèle de RI a pour rôle de fournir une formalisation du processus de RI et un cadre théorique pour la modélisation de la mesure de pertinence. Il existe un grand nombre de modèles de RI textuelle développé dans la littérature. Ces modèles ont en commun le vocabulaire d'indexation basé sur le formalisme mots clés et diffèrent principalement par le modèle d'appariement requête-document. Le vocabulaire d'indexation $V = \{ti\}$, $i \in \{1, ..., n\}$ est constitué de n mots ou racines de mots qui apparaissent dans les documents.

Un modèle de RI est défini par un quadruplet (D, Q, F, R (q, d)) : où

– D est l'ensemble de documents

– Q est l'ensemble de requêtes

– F est le schéma du modèle théorique de représentation des documents et des requêtes

– R (q, d) est la fonction de pertinence du document d à la requête q Nous présentons dans la suite les principaux modèles de RI : le modèle booléen, le modèle vectoriel et le modèle probabiliste.

I.2.3.1 Modèle booléen

Le modèle booléen [7] est basé sur la théorie des ensembles. Dans ce modèle, les documents et les requêtes sont représentés par des ensembles de mots clés. Chaque document est représenté par une conjonction logique des termes non pondérés qui constitue l'index du document. Un exemple de représentation d'un document est comme suit : d = t1 ∧t2∧t3...∧ tn. Une requête est une expression booléenne dont les termes sont reliés par des opérateurs logiques (OR, AND, NOT) permettant d'effectuer des opérations d'union, d'intersection et de différence entre les ensembles de résultats associés à chaque terme. Un exemple de représentation d'une requête est comme suit : " q = (t1∧t2) ∨ (t3∧t4) " . La fonction de correspondance est basée sur l'hypothèse de présence/absence des termes de la requête dans le document et vérifie si l'index de chaque document dj implique l'expression logique de la requête *q*. Le résultat de cette fonction est donc binaire est décrit comme suit :

$$RSV(q, d) = \{1,0\}.$$

I.2.3.2 Modèle vectoriel

Dans ces modèles [8], la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance dans un espace vectoriel. Le modèle vectoriel représente les documents et les requêtes par des vecteurs d'un espace à *n* dimensions, les dimensions étant constituées par les termes du vocabulaire d'indexation.

L'index d'un document *dj* est le vecteur = (w1,j ,w2,j ,w3,j , ..,wn,j), où wk,j ∈ [0 1] dénote le poids du terme *tk* dans le document *dj* . Une requête est également représentée par un vecteur = (w1, q, w2, q, w3, q,.., wn, q), où *wk, q* est le poids du terme tk dans la requête q.

La fonction de correspondance mesure la similarité entre le vecteur requête et les vecteurs documents. Une mesure classique utilisée dans le modèle vectoriel est le cosinus de l'angle formé par les deux vecteurs : RSV (q, d) = cos q d.

Plus les vecteurs sont similaires, plus l'angle formé est petit, et plus le cosinus de cet angle est grand. A l'inverse du modèle booléen, la fonction de correspondance évalue une correspondance partielle entre un document et une requête, ce qui permet de retrouver des documents qui ne reflètent pas la requête qu'approximativement. Les résultats peuvent donc être ordonnés par ordre de pertinence décroissante.

I.2.3.3 Modèle probabiliste

Ce modèle est fondé sur le calcul de la probabilité de pertinence d'un document pour une requête [9], [10], [11]. . Le principe de base consiste à retrouver des documents qui ont en même temps une forte probabilité d'être pertinents, et une faible probabilité d'être non pertinents. Etant donné une requête utilisateur *Q* et un document *D*, il s'agit de calculer la probabilité de pertinence du document pour cette requête.

Deux possibilités se présentent : *R, D* est pertinent pour *q* et D, n'est pas pertinent pour q. Les documents et les requêtes sont représentés par des vecteurs booléens dans un espace à n dimensions. Un exemple de représentation d'un document dj et une requête q est le suivant :

dj = (w1, j, w2, j, w3, j, .., wn, j),

q = (w1, q, w2, q, w3, q, .., wn, q). Avec wk, j ∈ [0 1] et wk, q ∈ [0 1].

La valeur de wk, j (resp. wk, q) indique si le terme t_k apparaît ou non dans le document *dj* (resp. q).

Le modèle probabiliste évalue la pertinence du document dj pour la requête q. Un document est sélectionné si la probabilité que le document d soit pertinent, notée *p(R/D)*, est supérieure à la probabilité que d soit non pertinent pour *q*, notée *p(R/D)* où *R* est l'événement de pertinence et est l'événement de non pertinence.

Le score d'appariement entre le document d et la requête *Q*, noté RSV (Q, D) est donné par : $(,) = \frac{(/)}{(/)}$

Ces probabilités sont estimées par de probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent. Cette mesure de similarité entre la requête et les documents peut se calculer par différentes formules. Ce modèle a donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI. Le modèle Okapi BM25 a été développé par Robertson en 1994 dans lequel le calcul du poids d'un terme dans un document intègre des aspects relatifs à la fréquence locale des termes, leur rareté et la longueur des documents

I.3 Les systèmes de recherche d'informations

I.3.1 Définition

*Un **Système de Recherche d'Informations (SRI)** est un système informatique qui permet de retourner à partir d'un ensemble de **documents**, ceux dont le contenu **correspond** le mieux à un **besoin** en informations d'un utilisateur, exprimé à l'aide d'une requête* [3].

Un SRI inclut un ensemble de procédures et d'opérations qui permettent la gestion, le stockage, l'interrogation, la recherche, la sélection et la représentation de cette masse d'informations.

I.3.2 Le processus de recherche d'information

Les différentes étapes du processus de RI, sont représentées schématiquement par le processus en U (**voir *Figure. 1.1***) [3]. La figure illustre particulièrement :

- les notions de documents et de requêtes qui sont des conteneurs d'informations,
- les opérations d'analyse, d'indexation et d'appariement qui permettent globalement de traiter la requête dans le but de sélectionner des documents à présenter à l'utilisateur.

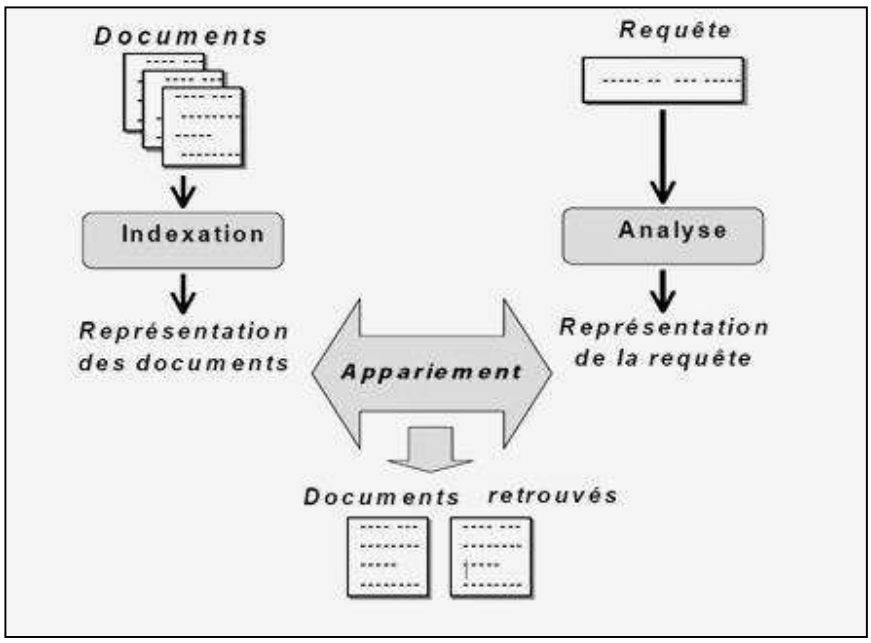

Figure 1.1 : *Processus en U de recherche d'informations*

I.3.3 Notions de document et de requête

I.3.3.1 Document : Le document représente le conteneur élémentaire d'information, exploitable et accessible par le SRI. Un document peut être un texte, une page WEB, une image, une bande vidéo, etc. Dans notre contexte, nous appelons document toute unité qui peut constituer une réponse à un besoin en information exprimé par un utilisateur.

I.3.3.2 Requête : Une requête constitue l'**expression** du **besoin** en informations de l'utilisateur. Plusieurs systèmes utilisent des langages différents pour décrire la requête :

- par une liste de mots clés : cas des systèmes SMART [7] et Okapi [9],
- en langage naturel : cas des systèmes SMART [7] et SPIRIT [12],
- en langage booléen : cas du système DIALOG [13],
- en langage graphique : cas du système NEURODOC [14].

I.3.4 Principales phases du processus de recherche d'information

L'objectif fondamental d'un processus de RI est de sélectionner les documents ''les plus proches'' du besoin en information de l'utilisateur décrit par une requête. Ceci induit deux principales phases dans le déroulement du processus : indexation et appariement requête/documents.

I.3.4.1 L'indexation

Un SRI gère les différentes collections de documents en les organisant sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leur contenu sémantique. L'interrogation de ce fond documentaire à l'aide d'une requête nécessite également la représentation de cette dernière sous une forme compatible avec celle des documents. Ce processus de conversion est appelé **indexation** (également appelé analyse pour la requête).

L'indexation est une étape très importante dans le processus de RI. Elle consiste à déterminer et extraire les termes représentatifs du contenu d'un document ou d'une requête, qui couvrent au mieux leur contenu sémantique. La qualité de la recherche dépend en grande partie de la qualité de l'indexation.

Le résultat de l'indexation constitue, ce que l'on nomme le **descripteur** du document ou de requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leur degré de représentativité du contenu sémantique de l'unité qu'ils décrivent.

Les descripteurs des documents (mots, groupe de mots) sont rangés dans une structure appelée dictionnaire constituant le **langage d'indexation**.

Un groupe de mots est à priori sémantiquement plus riche que les mots qui le composent pris séparément. Cet argument conduit à ne pas considérer simplement les mots simples comme unités de base dans le langage d'indexation mais également des groupes de mots. Ce groupe de mots forme ce que l'on appelle un *thesaurus*. Ce dernier inclut des relations de type linguistiques (équivalence, association, hiérarchisation) et statistiques (pondération) [15].

L'indexation peut être caractérisée par son mode et fonction de pondération.

- **Les Mode d'indexation :** l'indexation peut être manuelle, automatique ou semi-automatique :

  a) indexation manuelle : chaque document est analysé par un spécialiste du domaine correspondant ou par un documentaliste,

  b) indexation automatique : chaque document est analysé à l'aide d'un processus entièrement automatisé,

  c) indexation semi-automatique : le choix final reste au spécialiste du domaine correspondant ou documentaliste, qui intervient souvent pour établir des relations sémantiques entre mots-clés et choisir les termes significatifs.

I.3.4.2 Fonction de pondération

La pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît.

Le pouvoir de discrimination des termes pour décrire le contenu des documents n'est pas identique pour tous les termes. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, [9] a défini la **fonction de pondération** d'un terme dans un document connue sous la forme de ***Tf.Idf***, qui est reprise dans différentes versions par la majorité des SRI [9], [16] et [17]. On y distingue :

- *Tf* (*term frequency*) : cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document. Le *Tf* est souvent exprimé selon l'une des déclinaisons suivantes :

  a. *Tf* : utilisation brute

  b. $0.5 + 0.5 \frac{Tf}{Max(Tf)}$

- *Idf* (*Inverse of Document Frequency*) : mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection. Cette mesure est exprimée selon l'une des déclinaisons suivantes :

  a. $Idf = \log\left(\frac{N}{df}\right)$

  b. $\text{Idf} = \log\left(\frac{N-df}{df}\right)$.

En effet, lors des compagnes d'évaluation internationales, la mesure a eu des performances très limitées dans des corpus de taille très variable. Le problème posé est que les termes appartenant aux documents longs apparaissent très fréquemment et emportent le poids sur les termes appartenant à des documents moins longs. Les documents longs auront alors plus de chance d'être sélectionnés [19].

I.3.4.3 L'appariement requête-document :

Les SRI intègrent un processus de recherche/décision qui permet de sélectionner l'information jugée pertinente pour l'utilisateur. A cet effet, une mesure de similitude (correspondance) entre la requête indexée et les descripteurs des documents de la collection est calculée. Seuls les documents dont la similitude dépasse un seuil prédéfini sont sélectionnés par le SRI.

La fonction de correspondance est un élément clé d'un SRI, car la qualité des résultats dépend de l'aptitude du système à calculer une pertinence des documents la plus proche possible du jugement de pertinence de l'utilisateur [20].

Il existe deux types d'appariement :

→ ***Appariement exact***

Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés.

→ ***Appariement approché***

Le résultat est une liste de documents censés être pertinents pour la requête. Les documents retournés sont triés selon un ordre de mesure. Cet ordre reflète le degré de pertinence document/requête

I.4 Recherche d'information sur le Web

I.4.1 Outils de recherche d'information

Il existe de nombreux outils de recherche d'information sur le Web, ces outils qui se spécialisent en fonction des services utilisés et du type d'information qu'ils recensent.

On qualifie d'ailleurs très souvent aujourd'hui tout interface de recherche et d'interrogation de moteur de recherche et ce quelle que soit la source interrogée et le système informatique utilisés [21]. Il convient en effet de distinguer différents types d'outils de recherche sur l'Internet.

Un premier critère de classification des outils de RI repose sur le mode de recherche proposé. Il distingue entre les outils par navigation arborescente (comme les annuaires) ou hypertexte (comme les listes de signets), et les outils par requête (comme les moteurs, fondés sur l'utilisation de mots-clés). Cette distinction n'est plus pertinente aujourd'hui, tant l'imbrication est forte entre les mêmes outils [22].

Un deuxième critère reste toujours valable, en dépit des apparences : celui du mode d'indexation des ressources. Selon ce critère, on distingue les annuaires thématiques, qui procèdent à un référencement des sites Web et les moteurs de recherche, qui fonctionnent par collecte et indexation automatisées des pages Web (et non des sites). Cette distinction, 'historique', est moins nette aujourd'hui, à cause de l'imbrication des annuaires et des moteurs : Google utilise l'annuaire de l'Open Directory, Yahoo a son propre moteur, etc. Dans le cadre de notre thèse, nous distinguons trois catégories d'outils pour la recherche d'information sur le web: les moteurs de recherche, les annuaires et les méta-moteurs. Cette distinction qui repose également sur le mode d'indexation reste essentielle, car elle induit des usages et des technologies très différentes. Ainsi un annuaire thématique va-t-il référencer des sites Web, là où un moteur indexera toutes les pages d'un site ? En effet, l'annuaire facilitera le défrichage, le premier repérage des ressources dans un domaine ou un secteur défini par l'organisation arborescente proposée, alors qu'un moteur de recherche permettra de trouver un document très précis. Enfin les méta-moteurs permettent d'interroger en une seule fois différents outils de recherche, qu'ils soient de type annuaire ou de type moteur. Nous présentons dans ce qui suit ces trois catégorie d'outils pour la RI, et nous mettons l'accent sur les moteurs de recherche du fait qu'ils seront utilisés comme support de validation dans la partie contribution de notre thèse.

I.4.1.1 Moteur de recherche

Un moteur de recherche est une application permettant de retrouver des ressources (pages web, images, vidéo, fichiers, etc.) associées à des mots quelconques. Certains sites Web offrent un moteur de recherche comme principale fonctionnalité ; on appelle alors moteur de recherche le site lui-même (Google Vidéo par exemple est un moteur de recherche vidéo). Ces outils de recherche sur le web sont constitué de « robots », encore appelés bots, spiders, crawlers ou agents qui parcourent les sites à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des annuaires) pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés.

Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Ce sont des moteurs dits desktop qui combinent la recherche parmi les fichiers stockés sur le PC et la recherche parmi les sites Web — on peut citer par exemple Exalead Desktop, Google Desktop et Copernic Desktop Search, etc. Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

1- Le correcteur orthographique : il permet de corriger les erreurs introduites dans les mots de la requête, et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique.

2- Le lemmatiseur : il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche.

3- L'anti dictionnaire : utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots "vides" (tels que "de", "le", "la") qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

En ce qui concerne les caractéristiques, les moteurs de recherche ont un fonctionnement commun, mais diffèrent par un certain nombre de critères [23]. Pour ce qui est de commun, rappelons simplement qu'ils procèdent tous des même étapes :

- D'abord l'exploration du web, durant laquelle ils vont collecter les informations sur chaque page rencontrée.

- Puis l'indexation, durant laquelle ils vont enregistrer dans une base de données les informations collectées.

- Enfin la recherche, durant laquelle ils vont rechercher les données collectées en fonction des mots clés.

Si tous les moteurs passent par ces étapes communes, ils ont tous leurs différences. Voici quelques éléments sur lesquels ils se différencient [20] :

- D'abord la manière d'explorer le web.

- Ensuite, le choix des informations qu'ils vont récupérer des sites visités. Certains moteurs vont conserver le titre de la page, la description qu'en a fait le créateur de la page (Meta tag), parfois une partie du contenu de la page, …

- La manière de construire l'index, et sa taille. C'est d'ailleurs un des critères de performance le plus souvent mis en avant. En effet, plus un moteur indexe de pages, plus il a de chance de vous fournir un résultat correspondant à votre requête. Google passe pour avoir l'index le plus important. Les chiffres sont néanmoins difficiles à obtenir.

En septembre 2005, Google indiquait avoir indexé 24 milliards de pages. Actuellement, si l'on demande à Google de présenter toutes les pages contenant simplement le chiffre '1' (c'est une astuce pour estimer le nombre de pages indexées), il propose 21 milliards de pages. Yahoo! en propose 40 milliards, il existe deux manières :

a. La manière de rechercher dans l'index.
b. La manière de présenter les résultats (on parle alors d'interface).

Si Google présente sobrement le titre, un extrait, et quelques autres éléments, d'autres moteurs ont une interface beaucoup plus riche avec présentation d'images, de critères de pertinence, des graphiques, de mots clés, …

Enfin la popularité des moteurs de recherche n'est pas absolue, généralement, nous ne considérons souvent que le moteur de recherche Google. Toutefois, son hégémonie n'est pas aussi importante dans le reste du monde. Nous parlons également souvent du trio Google, Yahoo, Bing. Auxquels nous rajoutons le moteur de recherche chinois 'Baidu' qui représente selon une étude du cabinet Comscore qui réalisée en janvier 2011 : 3,3 milliards de requêtes, soit 5,4 % du total, mais également 73 % du marché chinois (450 millions d'internautes). Baidu se classe dans le 'Top 3' des moteurs de recherche les plus utilisés dans le monde. Le tableau 1.1, présente des statistiques sur le classement mondial des moteurs de recherche pour les années 2009 et 2010.

Parts de marché des moteurs dans le monde		
Moteur	*Janvier 2009*	*Janvier 2010*
Google	63,1%	62,8%
YAHOO!	12,2%	11,9%
Baidu 百度	04,6%	04,5%
bing	03,1%	03,1%

Tableau 1.1 : *Classement mondial des moteurs de recherche (2009/2010)*

I.4.1.2 Les annuaires

L'annuaire (ou directory en anglais) est une liste de liens subdivisés en catégories suivant une structure en arbre, accompagnée d'une brève description. Bien que ce procédé fût pionnier en la matière, il tend à disparaître. En effet, le fait de devoir sélectionner les catégories dans lesquelles on recherche suppose que l'on sache exactement où chercher. Et on peut se demander où se positionne le site qui appartient à plusieurs catégories. Mais à cette2question, les moteurs utilisant ce procédé vous répliqueront qu'ils se trouvent dans toutes celles susceptibles de correspondre. Néanmoins, on doit reconnaître aux annuaires un

gros avantage, celui de mettre en quelque sorte dans le contexte, ainsi les recherches dans la base de données sont diminuées, en plus d'obtenir des résultats plus pertinents.

Les annuaires sont donc des outils basés sur le recensement humain de l'information. Ils signalent des sites et des ressources de l'Internet comme un catalogue de bibliothèque signale des livres ou bien encore comme les pages jaunes signalent des entreprises. On distingue dans ce contexte deux catégories d'annuaires [24].

a) Les annuaires commerciaux (Tableaux) :

Ils se financent grâce à la publicité. Ils ont en principe une couverture dit "générale" (ils couvrent toutes les disciplines). Ils peuvent concerner le monde ou une zone régionale, nous citons parmi eux :

- Annuaires généralistes internationaux : le plus connu est sans doute 'Yahoo Directory', mais il existe aussi 'DMIZ' de l'Open Directory Project et l'annuaire de 'Lycos'.

- Annuaires régionaux commerciaux : ce sont les annuaires qui recensent des sites en fonction de leur langue. Dans le cas de des annuaires francophones nous citons la version française de 'Yahoo Directory' ou encore l'annuaire 'Francité'.

- Les annuaires qui recensent d'autre pays ou parties du monde: comme l'annuaire 'Wohaa' pour l'Afrique et l'annuaire russe 'Yandex'.

b) Les annuaires non commerciaux :

Sont des annuaires élaborés par des individus de façon bénévole ou bien par des institutions. Ils sont soit généraux soit spécialisés. Leur préoccupation consiste toujours à identifier les ressources et les sites en tenant comptes de leur qualité :

- Annuaires à couverture (généraliste): comme le 'Vlib' (Virtual Library) et l'annuaire 'Resource Discovery network'.

- Annuaires à couverture thématique ou spécialisée : comme le répertoire en sciences humaines 'Voice of the Shuttle' et le répertoire de ressources juridiques 'Findlaw'.

Enfin, la distinction est importante entre les annuaires et les moteurs de recherche. Le tableau1.2, présente une comparaison entre ces deux outils pour la RI.

I.4.1.3 Comparaison entre les annuaires et les Moteurs de recherche :

Indexation de sites par des documentalistes Indexation de mots par des robots Recherche sur des sites et sur des catégories Recherche en texte intégral sur des pages web Avantages : choix des informations, classement raisonné par catégories et sous-catégories

En a résumé les avantages et les inconvénients de chaque une en tableau suivante

Annuaires	Moteurs
Indexation de sites par des documentalistes	Indexation de mots par des robots
Recherche sur des sites et sur des catégories	Recherche en texte intégral sur des pages web
Avantages : choix des informations, classement raisonné par catégories et sous-catégories	**Avantages** : plus d'exhaustivité et mise à jour plus rapide
Inconvénients : moins d'exhaustivité et mise à jour moins rapide	**Inconvénients** : capture de pages web sans classement raisonné
À retenir : L'exploration des catégories s'avère souvent plus fructueuse que celle des sites	À retenir : La recherche par mots-clés donne de meilleurs résultats sur les moteurs

***Tableau 1.2** : Comparaison entre annuaires et moteurs de recherche*

I.4.1.4 Les méta-moteurs

Ils sont de création plus récente. Ils constituent en fait la première génération des agents dits "intelligents". Ils permettent d'interroger en une seule fois différents outils de recherche, qu'ils soient de type annuaire ou de type moteur, afin de fournir une réponse plus exhaustive. Deux catégories de méta-moteurs: ceux en ligne et ceux consistant en un "logiciel client" à installer sur son ordinateur (le plus connu: COPERNIC) [24]. Le principe de fonctionnement des méta-moteurs est différent, Certains indexent l'information contenue dans différents annuaires et moteurs, d'autres les interrogent simultanément de façon dynamique. Certains de ces méta-moteurs retraitent plus au moins les réponses (tri, dé doublonnage). Ils permettent ainsi de rechercher de façon plus large sur le Web. Toutefois, cela peut également générer du "bruit" (réponses non pertinentes).

La parade mise en œuvre par certains méta-moteurs consiste à limiter le nombre de réponses de chaque outil interrogé (ce qui est indispensable et permet ainsi d'obtenir les réponses en principe les plus pertinentes) [24].

I.4.2 Algorithme des moteurs de recherche d'information

Différentes études ont suggéré de tenir compte de la popularité des documents afin d'améliorer les performances de la recherche d'information. Le PageRank [25] de Google et le HITS [26] de Kleinberg sont deux algorithmes fondamentaux qui utilisent les liens hypertextes pour classer les résultats d'une requête. Généralement, ces algorithmes fonctionnent en deux temps : Dans une première étape, un moteur de recherche retourne une liste de documents répondant à la requête posée, en fonction des termes de la requête et des termes d'indexation des documents. Dans une seconde étape, ces systèmes tiennent compte des liens hypertextes pour classer ces documents [27].

I.4.2.1 Hyperlink-Induced Topic Search (HITS)

Kleinberg fut un des premiers à s'intéresser aux propriétés de connectivité du graphe représentatif d'Internet et de son apport dans la détection de la pertinence d'une page à une requête [28]. Quelques constatations simples sont à l'origine de ses travaux dans ce domaine.

On retrouve d'une part les pages qui semblent être très importantes et jouent le rôle d'autorité sur un sujet donné et d'autre part les documents possédant un grand nombre de liens vers des pages faisant autorité sur un sujet. On distingue ainsi les pages *autorités* ayant un grand nombre de liens entrants et les pages *hubs* ayant un grand nombre de liens sortants et regroupant les autorités d'un même sujet.

Le but de l'algorithme HITS est de déterminer les hubs et les autorités qui renforcent leurs relations mutuellement sur un sujet donné. Ainsi Kleinberg dénombre les bons hubs comme des pages pointant vers beaucoup de bonnes autorités et les bonnes autorités comme des pages pointées par beaucoup de bons hubs [29].

Supposant W la matrice d'adjacence du sous-graphe orienté G. Notons respectivement par X et Y les deux vecteurs colonnes pivot et autorité de dimension (n * 1) contenant les poids pivots et les autorités correspondant à chaque nœud du sous-graphe G. Klineberg utilise un processus itératif afin de calculer ces poids. Le poids autorité du nœud i, Xi, est égal à la somme des poids pivots de tous les nœuds citant le nœud i et, pareillement, le poids pivot du nœud i, Yi, est égal à la somme des poids autorités de tous les nœuds que cite le nœud i.

En pratique, le résultat de HITS est composé de deux listes ordonnées : une liste de bonnes pages autorités et une autre de bonnes pages pivots qui seront renvoyées à l'utilisateur. L'utilisateur a l'embarras du choix entre les deux listes et il peut être intéressé par une liste au détriment de l'autre selon la recherche demandée [30].

I.4.2.2 PageRank

Quelques moteurs de recherche, dont le plus connu est Google, ont pris le pari d'utiliser un autre mode de classement des résultats. Les pages Web sont ordonnées selon leur popularité, une page qui est la cible d'un très grand nombre de liens est probablement non seulement une page validée (page parcourue par un grand nombre de lecteurs, qui ont jugé bon de la citer en référence) mais aussi une page détenant un contenu utile à un grand nombre d'utilisateurs. L'approche du PageRank qui a fait la spécificité du moteur de recherche Google, repose sur la notion de propagation de popularité. Le principe consiste à évaluer l'importance d'une page en fonction de chacune des pages pointant vers elle. La propagation met en avant les pages qui jouent un rôle particulier dans le graphe des liens, avec l'hypothèse suivante : *"une page est importante quand elle est beaucoup citée ou citée par une page très importante*".

La mesure de PageRank (PR) proposée par [25] est une distribution de probabilité sur les pages. Elle mesure en effet la probabilité PR, pour un utilisateur navigant au hasard, d'atteindre une page donnée. Elle repose sur un concept très simple : un lien émis par une page A vers une page B est assimilé à un vote de A pour B. Plus une page reçoit de votes, plus cette page est considérée comme importante. Le PageRank se calcule de la façon suivante :

- Soient $T1$, $T2$, ..., Tn : n pages citant une page A. Notons $PR(Tk)$ le PageRank de la page Tk, $S(Tk)$ le nombre de liens sortants de la page Tk, et d'un facteur compris entre 0 et 1, fixé en général à 0.85. Ce facteur d représente la probabilité de suivre effectivement les liens pour atteindre la page A, tandis que (1-d) représente la probabilité d'atteindre la page A sans suivre de liens. Le PageRank de la page A se calcule à partir du PageRank de toutes les pages Tk de la manière suivante :

$$PR(A) = (1 - d) + d \quad \frac{PR(T)}{S(T)}$$

Initialement, toutes les pages sont équiprobables, leur valeur de PR est alors égale à $1/n$, n

Étant le nombre de documents de la collection.

I.4.3 Architecture des moteurs de recherche

I.4.3.1 Architecture générale des premiers moteurs de recherche

L'architecture originale utilisée par Altavista représente la première catégorie de systèmes. Il s'agit d'une architecture très simple qui se divise en deux parties distinctes. On retrouve d'une part un crawler et d'autre part l'interface d'interrogation du moteur de recherche et le système d'analyse des requêtes proposés par les utilisateurs du système [31].

Le cœur du système repose sur un index inversé permettant d'associer des mots à un ou plusieurs documents. La demande de l'utilisateur est traitée en interrogeant l'index inversé pour connaître les documents dans lesquels apparaissent le plus souvent les mots de la requête [32].

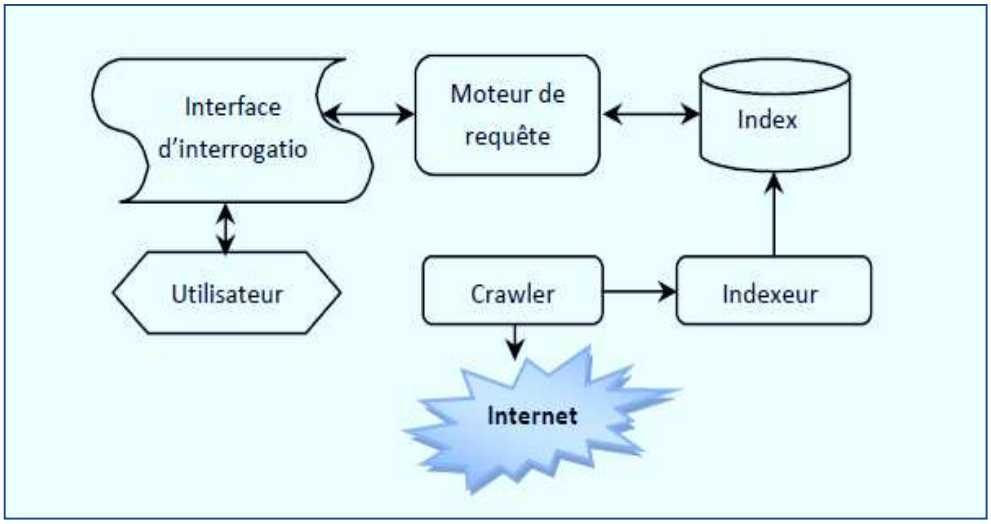

***Figure 1.2** : Architecture originale du moteur de recherche Altavista.*

I.4.3.2 Architecture distribuée et adaptative

Des variantes de l'architecture précédente, basées sur le modèle indexeur-crawler, ont été imaginées pour gommer les défauts inhérents à sa conception. L'une d'entre elle, appelée « Harvest » c'est révélée très innovante en matière de distribution des ressources.

Le récolteur *:* est chargé de collecter et d'extraire périodiquement des informations d'indexation - textes, images - depuis plusieurs sites Web.

Le broker *:* quant à lui, fournit le mécanisme d'indexation et l'interface d'interrogation sur les données amassées par le récolteur. On retrouve ici, le mécanisme indexeur-crawler identifié dans la section précédente. Cependant, plusieurs brokers et plusieurs récolteurs peuvent communiquer ensemble, chacun se spécialisant dans un domaine précis.

Lorsqu'une requête est émise sur un broker dont le domaine traité ne correspond pas à ses capacités, celui-ci transmet la requête à une autre entité capable de la gérer.

C'est un système totalement adaptatif dans lequel il est possible de configurer les brokers et les récolteurs de manière à répartir le besoin en ressources sur un ou plusieurs domaines particuliers. [31].

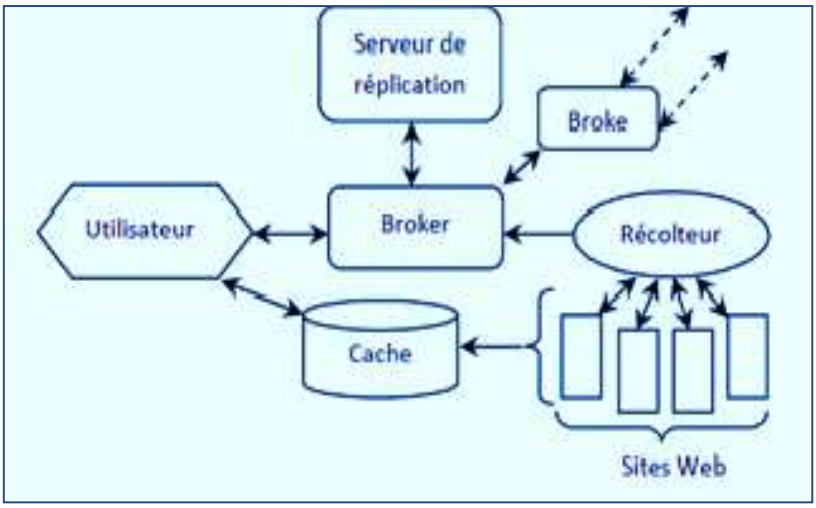

Figure 1.3 *: Architecture du système Harvest [31].*

I.4.5.3 Architecture moderne d'un moteur de recherche

L'architecture du moteur de recherche Google est certainement une des plus efficaces actuellement. Elle ne repose pas sur un système monolithique mais sur un grand nombre de machines classiques coopérant ensemble. Ce système peut se décomposer en plusieurs parties comprenant :

- Un sous-système d'exploration d'Internet
- Un indexeur
- Un analyseur de la topologie d'Internet formée par les liens hypertextes : et un sous-système de présentation et d'exécution de requêtes.
- Un serveur d'URL garde la mémoire des liens des pages à visiter. Des robots chargés d'explorer le Web récupèrent ces liens afin de télécharger les documents correspondant et les stocker dans une base de données recensant la totalité des pages indexées. Cette opération est réalisée continuellement et alimente et met à jour en permanence la base de documents du moteur. Périodiquement, cette base est analysée pour réaliser un index inversé reliant des termes aux documents les contenant. D'autres informations sur les termes sont extraites comme leur position dans le document, la taille de la police utilisée ou sa fonte.

Cette analyse permet également d'extraire tous les liens hypertextes des documents rencontrés afin d'alimenter le serveur d'URL. Cette base de liens est utilisée afin de calculer le PageRank permettant de trier les documents de l'index par pertinence décroissante [31].

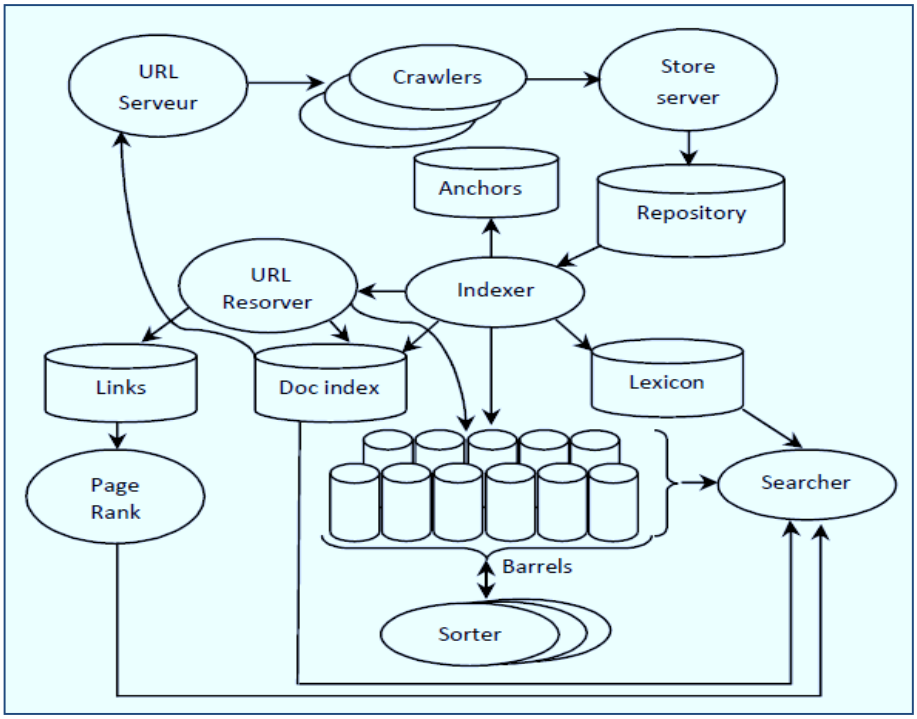

***Figure 1.4** : Architecture du moteur de recherche Google. [31]*

I.4.6 Défis des moteurs de recherche

La double révolution que traversent actuellement les moteurs de recherche selon Eric Boutin concerne :

- Le passage d'un indicateur de pertinence absolu à un indicateur de pertinence centré sur l'utilisateur. L'indicateur de pertinence est l'algorithme qui permet de hiérarchiser les pages renvoyées par le moteur de recherche. Classiquement les pages sont hiérarchisées les unes par rapport aux autres en fonction de critères définis par le moteur de recherche. Il est possible de penser des systèmes dans lesquels l'internaute attribuerait à certains critères des préférences particulières.
- Le passage d'une formulation de la requête intégrant exclusivement le sujet de la recherche (requête de type 'topic related') à une requête multicritères introduisant d'autres dimensions. Il ne s'agit alors plus de hiérarchiser les résultats d'une recherche les uns par rapport aux autres mais de filtrer les résultats obtenus pour n'afficher que ceux qui correspondent au souhait de l'utilisateur.

I.5 Conclusion

Dans ce chapitre nous avons présenté les principales notions et concepts de la recherche d'information, des systèmes de recherche d'information et ceux des outils de recherche sur le web.

A travers les différentes sections que nous avons présentées, nous concluons que la recherche d'information, s'attache à définir des modèles et des systèmes afin de faciliter l'accès à un ensemble de documents se trouvant dans des bases documentaires ou encore sur le web. Le but est de permettre aux utilisateurs de retrouver les documents dont le contenu répond à leur besoin en information, il s'agit donc de retourner l'ensemble de documents pertinents. Cependant, nous constatons que la notion de pertinence dépend de la satisfaction de l'utilisateur d'une part, et des différents sens portés par les termes de la requête d'une autre part. Cette constatation constitue le point faible de la recherche d'information classique, elle représente également le point de départ pour de nouveaux paradigmes de recherche. Nous nous intéressons dans le cadre de cette thèse à deux nouvelles orientations en RI : d'abord, la recherche contextuelle d'information qui modélise le contexte de l'utilisateur, de la requête et celui du système de recherche lui-même, elle utilise à cet effet des mécanismes comme les profils des utilisateurs et les historiques des recherche. Puis, la recherche sémantique d'information, qui utilise des ressources externes, généralement les ontologies, comme support à la modélisation des phases d'indexation et de recherche. Dans le cadre de cette thèse, nous souhaitons apporter des contributions pour améliorer la recherche d'information en prenant en compte le contexte et la sémantique.