De la pertinence à l'utilité en recherche d'information : le cas du Web

Brigitte Simonnot

De la pertinence à l'utilité en recherche d'information : le cas du Web

Brigitte SIMONNOT
Maître de conférences, SIC

Centre de Recherche sur les Médias, Université de Metz
IUT de Metz
Île du Saulcy 57045 Metz Cedex 1 France
Brigitte.Simonnot@iut.univ-metz.fr

Résumé
Les paradigmes mis en œuvre par les systèmes de recherche documentaire traditionnels se retrouvent dans la conception des moteurs de recherche sur le Web. Cependant, les concepteurs doivent innover pour s'adapter à ce contexte particulier, qui a un impact important sur le fonctionnement des outils. A travers une analyse de la littérature scientifique et des observations réalisées durant des sessions de formation à la recherche, nous analysons dans quelle mesure l'usager est pris en compte dans la conception même de ces systèmes, en comparant les implantations des paradigmes aux recherches sur le besoin d'information et le processus informationnel. Une analyse des différentes étapes du processus de recherche d'information souligne la complexité et les nombreuses dimensions du concept de pertinence en recherche documentaire.

Mots-clés
Moteur de recherche, Web, Usager, Pertinence/Utilité

Title
From Relevance to utility in Information Research : the case of the Web

Abstract
The different paradigms on which traditional information retrieval systems are built are reused to design Web search engines. However, designers have to innovate in order to adapt to this particular documentary context, which has an important impact on the way tools operate. Through an analysis of French and English scientific literature, and observations made during Web search training sessions, we try to define how far users are taken into account in the design of such systems, by comparing paradigm implementation to recent researches on information needs and the informational process. An analysis of the different steps of an information retrieval process highlights the complexity and the many dimensions of the concept of relevance in document retrieval.

Keywords
Web Search Engines, User, Relevance/Utility

Introduction

Si le développement d'Internet et du Web a relancé l'intérêt pour la recherche documentaire en l'éclairant d'une lumière nouvelle, c'est que, depuis le milieu des années 1990, le Web s'est imposé comme une source d'information de plus en plus cruciale. Publier sur le Web est en effet à la portée du très grand nombre, en évitant les contraintes liées à la chaîne éditoriale et en réduisant les délais de mise à disposition et d'accès à l'information de manière très significative. Face à cette importante masse de documents multiformes, les moteurs de recherches, annuaires et méta-moteurs, sont devenus indispensables. Cependant, leur mode de fonctionnement est le plus souvent opaque pour l'utilisateur. Nous analysons les évolutions que ce nouveau contexte introduit pour les systèmes de recherche

Recherches récentes en Sciences de l'information - convergences et dynamiques, actes du colloque international MICS-LERASS, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410.

d'information (SRI) à travers une grille des formes de pertinence, concept qui est au cœur de toute démarche documentaire.

Par ailleurs, une véritable démarche documentaire n'est pas un acte isolé dans le temps : il convient de parler de sessions de recherche d'information car la recherche d'information peut être assimilée à une situation d'apprentissage, à un processus en spirale qui permet peu à peu d'accéder à de nouvelles connaissances. Comment cette dimension temporelle du processus de recherche d'information est-elle prise en considération par les systèmes de recherche d'information actuels sur le Web ? Ces outils aident-ils à un affinement de la recherche ? Apportent-ils une assistance à la reformulation du besoin d'information ? Quelle continuité est rendue possible entre les requêtes successives ?

Une analyse de la littérature scientifique – essentiellement française et anglo-saxonne – et des observations réalisées durant des sessions de formation à l'usage des moteurs de recherche nourrissent notre analyse. Ces observations ont été réalisées sur deux groupes d'étudiants (80 de niveau BAC+1 et 40 de niveau BAC+3), effectuant des recherches sur un sujet librement choisi puis sur dix sujets imposés. Cette analyse se limite ici aux moteurs de recherche qui présentent leurs résultats sous forme de listes de références classées par pertinence décroissante.

Enfin, nous confrontons les concepts de pertinence et d'utilité. Les critères privilégiés par les systèmes de recherche d'information actuels correspondent-ils aux critères traditionnels d'évaluation de la qualité d'une recherche ?

1. DÉMARCHE DOCUMENTAIRE ET PROCESSUS DE RECHERCHE

Depuis le début des années 1980, la démarche documentaire a fait l'objet d'études adoptant une approche cognitiviste (voir par exemple Irving, 1985 ; Kuhlthau, 1993). Kuhlthau a mis en évidence le sentiment d'incertitude qui caractérise le chercheur d'information, particulièrement au début de sa recherche. Elle distingue six étapes dans un processus de recherche d'information (Kuhlthau, 1999) :

1. Lors de la phase d'initialisation, l'individu prend conscience de son manque de connaissance pour résoudre un problème ou accomplir une tâche. Depuis le modèle ASK - Anomalous State of Knowledge - (Belkin, 1982), il est admis que le besoin d'information naît d'une prise de conscience d'un "trou", d'une lacune dans ses connaissances que le sujet doit s'efforcer de combler.
2. Durant la phase de sélection, l'individu cerne peu à peu son sujet, à travers un questionnement qui l'aide à définir sa problématique.
3. Suit une phase d'exploration du sujet durant laquelle l'individu tente de découvrir des informations sur son problème en général.
4. La phase de formulation oblige l'individu à se focaliser sur certaines informations rencontrées lors de l'exploration. La formulation permet de clarifier ses pensées et de se concentrer sur l'objet de sa recherche.
5. La phase de collecte des informations pertinentes est une phase itérative où un dialogue s'instaure entre l'individu et le système qu'il interroge.
6. La phase de présentation termine la recherche : il s'agit de mettre en forme les informations recueillies.

Ce modèle de l'activité du chercheur d'information permet à l'auteur d'élaborer une série de conseils et de stratégies d'assistance qu'un documentaliste peut apporter à une personne en situation de recherche documentaire. Or, dans le contexte de la recherche d'information sur le Web, l'utilisateur est directement confronté à l'outil qu'il utilise. De plus, les SRI ne "voient" que les phases 3 à 5 de ce processus. Ces phases peuvent encore être décomposées en un certain nombre d'étapes au cours desquelles l'utilisateur interagit avec le système (figure 1).

Recherches récentes en Sciences de l'information - convergences et dynamiques, actes du colloque international MICS-LERASS, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410.

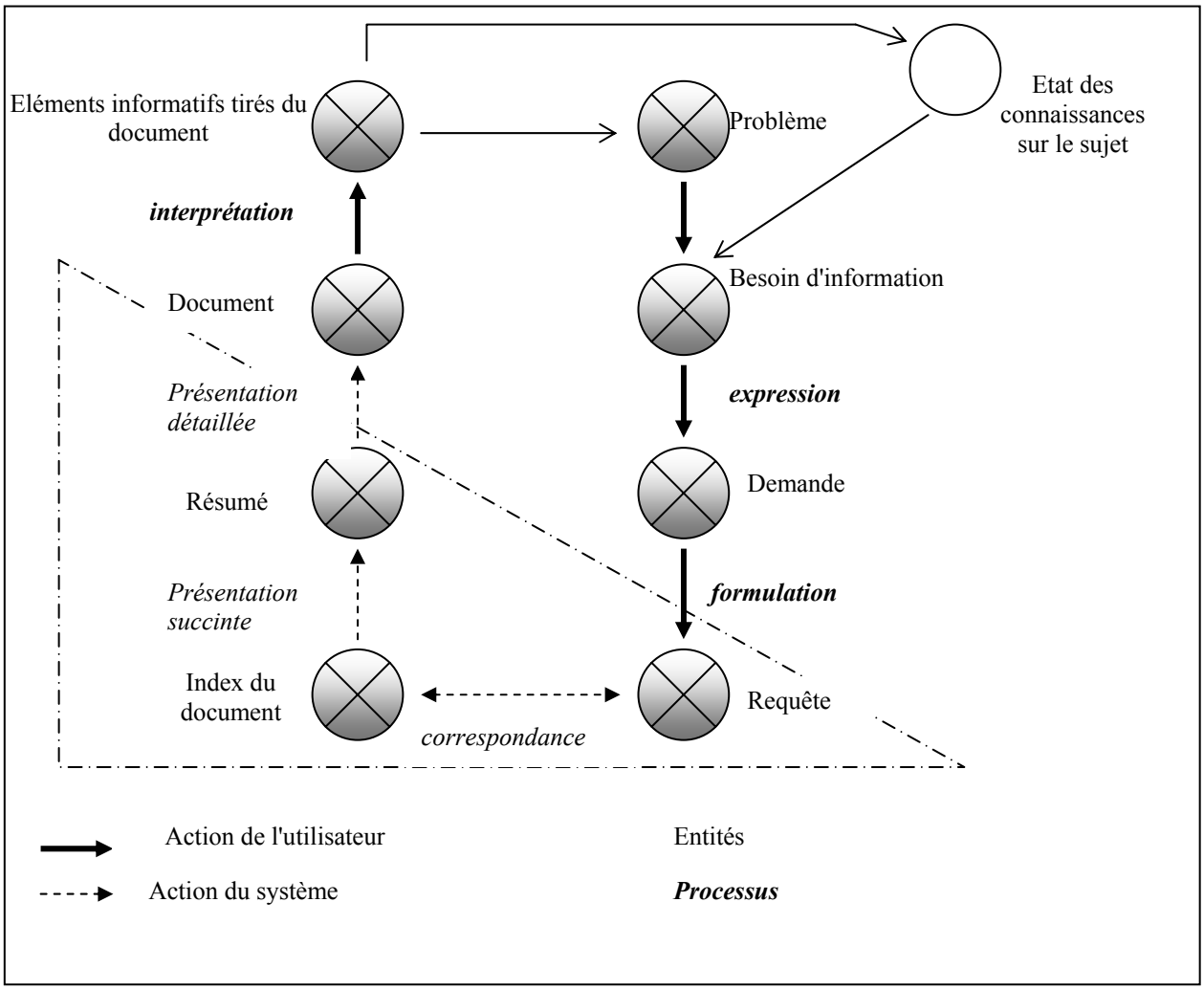

Figure 1 - Les principales étapes d'un processus de recherche d'information

Dans ce schéma, deux groupes d'entités peuvent être identifiés. D'une part, des entités qui sont liées à l'utilisateur du SRI :

- le *problème* qu'il doit résoudre,
- le *besoin d'information* que ce problème fait émerger,
- la *demande*, c'est-à-dire l'expression du besoin en langage naturel,
- et la *requête*, c'est-à-dire la représentation de la demande dans un langage documentaire formalisé pour un outil donné.

D'autre part, sont impliquées des entités liées à la constitution et à l'organisation documentaire de la collection consultée :

- l'*index,* constitué des représentations de documents obtenues par l'indexation, c'est-à-dire l'ensemble des descriptions sur lesquelles s'appliqueront les mesures de comparaison du moteur,
- les représentations ou *résumés* de ces documents qui seront présentés à l'utilisateur suite à sa requête,
- les *documents* eux-mêmes,
- les *éléments informatifs* ou de connaissance dont les documents sont censés être porteurs et que l'utilisateur devra s'approprier pour tirer parti de sa recherche.

Recherches récentes en Sciences de l'information - convergences et dynamiques, actes du colloque international MICS-LERASS, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410.

Le processus de recherche d'information peut donc être décrit comme une succession de relations entre ces différentes entités.

2. LES BESOINS D'INFORMATION

Un besoin d'information est un sentiment subjectif ressenti par un sujet pour améliorer ses connaissances. Les motivations qui amènent un individu à rechercher une information sont de nature fort diverses, surtout dans le cas du Web. Cependant, différents variables peuvent caractériser un besoin d'information :

- le degré de familiarité avec le domaine du sujet,
- le niveau d'études du chercheur d'information, car des transferts méthodologiques sont souvent observés notamment dans la méthodologie de recherche,
- la profondeur du besoin, du basique au besoin approfondi : vérifier une information, trouver une information nouvelle, comprendre une information, mettre à jour ses connaissances, obtenir une sélection d'informations représentatives sur une question, faire une recherche exhaustive sur le sujet.

Croiser ces différentes variables pourrait aider à comprendre la nature du besoin du chercheur d'information. Différentes typologies de besoin ont été proposées (Cluzeau-Ciry, 1988 ; Kuhlthau, 1999), et exploitées pour mettre en évidence des paramètres susceptibles d'adapter automatiquement le fonctionnement des SRI au besoin de l'utilisateur (Simonnot et Smaïl, 1996). Néanmoins, une difficulté subsiste avec ces approches : une recherche documentaire est rarement une démarche isolée dans le temps et les résultats d'une première requête peuvent modifier la nature du besoin. Peu de systèmes sont conçus pour adapter leur stratégie à l'évolution du besoin.

3. LES DIFFÉRENTES CONCEPTIONS DE LA PERTINENCE

Depuis 1958, le concept de *pertinence* a explicitement préoccupé nombre de chercheurs, en science de l'information (Mizzaro, 1997) et dans d'autres domaines comme en atteste la théorie de la pertinence développée par Sperber et Wilson.

3.1. La théorie de la pertinence

Sperber et Wilson (1989) ont élaboré une théorie de la pertinence dans le domaine de la pragmatique linguistique. Pour eux, dans une communication orale, chaque élément du dialogue constitue une hypothèse pour l'interlocuteur. S'opposant aux théories qui avaient cours jusqu'alors, ils montrent que le processus de compréhension prend place dans un contexte qui n'est pas déterminé a priori, mais qui se forme peu à peu, l'individu cherchant à maximiser la pertinence des informations reçues. En effet, le contexte d'interprétation d'une nouvelle hypothèse n'est ni déterminé *a priori* ni immuable. Bien au contraire, les individus espèrent que la nouvelle hypothèse en cours de traitement est pertinente puisqu'ils font l'effort de la traiter : ils s'efforcent de choisir un contexte qui justifiera cet espoir, c'est-à-dire qui maximisera sa pertinence. Selon les auteurs, la donnée du problème est la pertinence et le contexte doit être traité comme la variable.

La théorie de la pertinence postule l'existence d'une hiérarchie de contextes d'interprétations, chaque contexte contenant un ensemble d'hypothèses. L'ensemble des contextes est partiellement ordonné par une relation d'inclusion, qui varie dans le même sens que leur accessibilité, en partant du contexte courant - le plus accessible - au contexte le plus général, celui qui contient toutes les hypothèses déjà intégrées par l'individu. Ainsi, la pertinence est définie comme une *relation entre une hypothèse et un contexte*, cette relation étant ordonnée selon deux dimensions : l'effet de l'hypothèse sur le contexte c'est-à-dire sur les connaissances ou les croyances du sujet, mais aussi l'effort que doit fournir le sujet pour l'intégrer dans ses connaissances. Plus le contexte qui permet d'intégrer la nouvelle hypothèse est éloigné du contexte

Recherches récentes en Sciences de l'information - convergences et dynamiques, actes du colloque international MICS-LERASS, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410.

courant, plus l'effort requis sera important. En particulier, un surcroît d'information entraîne un surcroît d'effort pour la traiter, et le traitement d'une même information dans un contexte élargi demande à produire davantage d'effort.

3.2. La nature subjective de la pertinence

Certaines études ont essayé de caractériser les dimensions et les caractéristiques de la pertinence pour les chercheurs d'information à travers une analyse qualitative de l'expression de leurs jugements. Par exemple, Park (Park H., 1997) met en évidence des familles de termes par lesquels les utilisateurs qualifient la pertinence des documents qu'un SRI leur propose. Ces dimensions sont exprimées à l'aide de termes très subjectifs par les utilisateurs, sous une forme positive ou négative[1]. L'étude de l'auteur dégage trois orientations principales pour caractériser la pertinence, de la plus fréquemment invoquée à la moins fréquemment citée :

1. une dimension liée uniquement à la valeur intrinsèque du document,
2. une dimension liée à une variable spécifique du problème,
3. une dimension liée à une variable spécifique de l'usage souhaité de l'information.

Les sujets de l'étude - vingt-quatre étudiants dont la plupart préparaient une thèse - étaient en situation explicite d'apprentissage, ce qui n'est pas le cas de tous les utilisateurs de moteurs de recherche. Il faut donc relativiser la portée du classement obtenu. Néanmoins, ces trois dimensions donnent une première idée de la complexité de la notion de pertinence pour les utilisateurs.

3.3. Le concept de pertinence en recherche d'information

Jusqu'à il y a peu, les chercheurs et les concepteurs de SRI distinguaient deux types de pertinence : la *pertinence système*, c'est-à-dire l'évaluation par un système de l'adéquation entre des documents et une requête, et la *pertinence utilisateur* qui se traduit par des jugements de pertinence sur les documents fournis en réponse à une requête.

Dans une étude approfondie des publications scientifiques portant sur le concept de pertinence en recherche d'information, Mizzaro (Mizzaro, 1997) met en évidence la complexité et la diversité des types de pertinence. Il définit la pertinence de manière générale comme une *relation* entre deux entités, l'une étant liée à la collection de documents et l'autre à l'utilisateur. Il distingue trois dimensions :

- le domaine du sujet, le champ disciplinaire auquel il se rapporte,
- la tâche, c'est-à-dire l'activité que l'utilisateur va réaliser avec les documents retrouvés,
- le contexte, défini par défaut comme tout ce qui n'appartient ni au domaine du sujet, ni à la tâche, mais qui a une influence sur la façon dont se déroule la recherche et l'évaluation des résultats. Le contexte est ici défini comme un ensemble "fourre-tout", qui comprend par exemple les documents déjà connus de l'utilisateur (et qui ne seront donc pas pertinents pour lui) ou le temps voire l'argent disponible pour la recherche. Cette notion de contexte de la recherche demanderait à être définie de manière plus fine.

A la lumière de la théorie de la pertinence, nous pouvons déterminer qu'en recherche documentaire, une information déjà connue ne sera pas pertinente pour l'utilisateur, sauf s'il s'agit du rappel d'un élément enfoui dans sa mémoire. Cela ne signifie pas qu'un document déjà lu ne peut être pertinent : une relecture, avec un point de vue différent, peut en effet donner lieu à une nouvelle interprétation et modifier l'état des connaissances du sujet. D'autre part, le grand nombre de documents qui sont en général proposés suite à

[1] Par exemple : similaire, pertinent, utile, lié, intéressant, important, nouveau, bon, étudié, applicable, conforme au souhait, lu, disponible, approprié, spécifique, rejeté, informatif, basique, spécialisé, inclus, attirant, existant, effectif, su, à propos, de qualité, prometteur, contribuant, manquant, décevant.

l'interrogation d'un moteur nécessite un effort de traitement important, si bien que la majorité des utilisateurs ne vont pas au-delà de la première page de résultats (Jansen, Spink et Saracevic, 2000).

4. PERTINENCE AUX DIFFÉRENTES ÉTAPES DE LA RECHERCHE

En analysant plus finement les relations entre les entités mises en cause aux différentes étapes de la recherche documentaire, nous pouvons dégager différentes catégories de pertinence.

4.1. Pertinence des mesures de classement

Cette pertinence "système" a longtemps été la seule à être évaluée par les concepteurs de SRI. Elle mesure le degré auquel le système est capable de retrouver et classer des documents appariés à une requête. Les fonctions de similarité entre descripteurs de documents et requêtes diffèrent selon le modèle général qui a guidé l'implantation du système. Les mesures de rappel (proportion de documents pertinents retrouvés dans les documents potentiellement pertinents de la collection) et de précision (proportion de documents pertinents dans les documents retrouvés) sont encore largement utilisées pour évaluer la performance des outils. Outre la difficulté à évaluer le rappel dans le contexte particulier du Web, vu la grande masse de documents potentiellement disponibles[2], il est particulièrement important d'obtenir un classement affiné. Les seules mesures classiques de similarité ne sont pas suffisamment discriminantes pour ce faire.

4.2. Pertinence de l'indexation

Avec l'indexation automatique des textes, la difficulté consiste à trouver des termes descripteurs suffisamment discriminants pour caractériser le document. Différents indices sont sollicités à ce stade.

La *densité des mots* dans le texte est une mesure statistique qui est utilisée par pratiquement tous les moteurs de recherche. Elle représente la fréquence d'apparition d'un terme dans le document. L'hypothèse sous-jacente est qu'un terme qui revient fréquemment dans un texte est propre à bien en caractériser le contenu. Cette mesure peut être normalisée par la fréquence inverse de l'apparition du terme dans l'ensemble de la collection.

La *prise en compte de la structure du document* suppose que la place des mots dans le document est représentative de leur importance. Ainsi, les mots du titre ou du premier paragraphe sont censés être plus significatifs du contenu du document que ceux que l'on trouve à la fin du texte. Un poids d'importance plus fort est donc attribué aux termes descripteurs en fonction de leur position dans le texte.

Lorsque l'auteur d'un document veut attirer l'attention du lecteur sur un terme ou une expression, il les met en relief. La *prise en compte de la mise en forme du texte* est une autre méthode exploitée par les moteurs de recherche pour accorder davantage d'importance à certains mots. La mise en gras, en italique ou le soulignement peuvent être interprétés par analogie avec les intonations de la voix dans la communication orale. La particularité du langage HTML, qui mêle contenu et description de la mise en forme du document, permet à certains moteurs d'exploiter ce critère pour accroître le poids d'importance des termes mis en relief par la typographie.

Deux défauts essentiels caractérisent les documents publiés sur le Web. D'une part leur signalétique échappe à l'indexation automatique plein-texte. Par exemple, la thématique générale abordée, l'auteur, la date de publication sont souvent absents du texte du document. Or, une description de qualité est censée caractériser non seulement le contenu du document, mais donner également un certain nombre

---

[2] En novembre 2001, Google annonçait indexer 3 milliards de documents.

d'informations complémentaires sur le contexte éditorial. D'autre part, l'entité publiée sur le Web (c'est-à-dire la page Web, caractérisée par un URL distinct) ne correspond pas aux entités classiques des collections de documents. La nature hypertextuelle du Web permet de réunir des ressources qui correspondent à un ouvrage entier, un chapitre d'ouvrage, un sommaire, un article de périodique ou un simple paragraphe. Ajouter au document des éléments d'information sur sa nature structurelle peut aider le chercheur d'information à exploiter les résultats de sa recherche.

L'initiative Dublin Core[3] a élaboré une définition de balises spéciales du langage HTML, les balises META, grâce auxquelles il est possible d'introduire dans l'en-tête du document des méta-descripteurs. Les méta-données prennent la forme de couples (attribut, valeur), les attributs pouvant par exemple décrire le ou les auteurs, l'éditeur, le format du document, sa date de publication, sa langue, le champ disciplinaire concerné, etc. Le format RDF permet de superposer différents niveaux de méta-données et de raffiner la description à l'envi.

Cependant, l'introduction des balises META dans les en-têtes de documents a subi un avatar propre au contexte éditorial du Web. En effet, Internet est le lieu par excellence où les utilisateurs s'approprient voire détournent les outils de l'usage pour lequel ils ont été conçus initialement. Ainsi, certains auteurs de pages Web ont trouvé là un moyen de promouvoir leur site par l'intermédiaire des moteurs de recherche. Le *spamdexing* consiste à détourner les méta-descriptions dans le but d'augmenter l'audience d'un site, par exemple, en multipliant les méta-descripteurs pour leur donner plus de poids, ou y incluant les termes qui apparaissent le plus souvent dans les requêtes des internautes. Ces pratiques, même si elles sont réprimées par les moteurs qui les détectent, provoquent énormément de bruit dans les recherches, à tel point que certains moteurs, comme Google[4] ou NorthernLight[5], ont abandonné l'exploitation des méta-descripteurs.
La pertinence de l'indexation est à classer dans la pertinence "système". Elle doit caractériser le document en prévoyant quels termes seront utilisés pour le retrouver.

4.3. Pertinence de la formulation de la requête

La qualité des résultats de la recherche dépend fortement de la qualité de la formulation de la requête. Or, pour un utilisateur non averti, formuler une requête qui représente correctement son besoin est un challenge. La plupart des moteurs proposent deux types d'interfaces d'interrogation : recherche simple et recherche avancée.

Dans les interfaces de recherche simple, la requête est une juxtaposition de termes. Cette juxtaposition est interprétée, par la majorité des outils à l'heure actuelle, comme une disjonction booléenne. Par exemple, une requête formulée par "`conférence recherches récentes science information`" rapportera tous les documents qui contiennent au moins l'un de ces termes. Cette interprétation disjonctive des requêtes explique le grand nombre de pages généralement retrouvées. D'autres outils (comme Google ou NorthernLight) font une interprétation plus restrictive de la requête en formant une conjonction booléenne à partir des mots : seuls les documents où tous les mots apparaissent seront retrouvés.

L'usage des opérateurs '+' et '-' se standardise pour pondérer en quelque sorte les termes de la requête. Ainsi, "+terme" indique que ce terme doit obligatoirement apparaître dans le document, "-terme" exprime qu'il faut exclure les documents comportant ce terme. Une difficulté observée chez les utilisateurs des outils est qu'ils ont tendance à considérer ces opérateurs comme des opérateurs binaires alors qu'il s'agit d'opérateurs unaires, ce qui conduit à de légers contresens dans l'expression des requêtes : le premier mot n'est souvent pas pondéré et les requêtes sont souvent limitées à deux mots. En tenant compte de l'ordre des mots dans la requête, certains moteurs contrebalancent ce défaut : un '+' est ajouté devant le terme qui arrive en tête dans la requête. Cependant, cette heuristique ne permet pas de traiter correctement les

---

[3] Dublin Core : http://dublincore.org/ (consulté le 14/12/2001)

[4] www.google.fr ou www.google.com (consulté le 29/01/2002)

[5] www.northernlight.com (disparu en janvier 2001)

requêtes formulées en langage naturel. Par exemple, la requête " `je cherche le site de l'INIST` " sous Google ramène 47 résultats, aucun n'est la page d'accueil de l'Inist, ni même une page interne de ce site, alors qu'avec la requête " `inist` " : le site officiel de l'Institut apparaît en tête.

Des études récentes montrent que, dans la grande majorité des cas, les requêtes formulées comportent deux termes ou moins (Jansen, Spink, Saracevic, 2000), ce qui est bien peu pour représenter un besoin d'information. L'ergonomie des interfaces y est peut être pour quelque chose : la taille réduite des fenêtres prévues pour formuler la requête pourrait inciter les utilisateurs à saisir peu de mots (Brangier et Zimmer, 2001).

Les interfaces de recherche avancée permettent de raffiner l'expression de la requête soit en utilisant des formules booléennes avec les opérateurs AND, OR, NOT (exemple : Alta Vista), soit en remplissant un formulaire. On sait que la formulation d'une requête à l'aide d'opérateurs booléens est contre-intuitive pour les utilisateurs non spécifiquement formés à cette démarche (Avrahami et Kareev, 1993). Cependant, la mise à disposition d'un formulaire "ad hoc", comme celui proposé pour une recherche avancée sous Google, ne permet pas de formuler une vraie requête booléenne, puisqu'il n'autorise qu'une seule disjonction et une seule conjonction, et non la combinaison des opérateurs entre eux.

L'exclusion de termes dans les requêtes grâce à l'opérateur '-' ou AND NOT est très peu utilisée[6] (voir sur ce point aussi Jansen, Spink, Saracevic, 2000), alors qu'elle permet, en particulier, de désambiguiser certains cas d'homonymie et de synonymie. Cette démarche nécessite en effet une pensée "documentaire" c'est-à-dire une vision des termes descripteurs sous forme de listes inverses de documents pour comprendre qu'elle peut éliminer un grand nombre de résultats. Peut être certains facteurs psychologiques interviennent-ils également : on sait par exemple, dans le domaine des sondages d'opinion, que les personnes préfèrent répondre par des formes positives que par des formes négatives d'expression aux questions posées.

Enfin, les interfaces de recherche avancée demandent un effort supplémentaire à l'utilisateur qui doit s'adapter à la diversité de ces présentations, effort dont l'internaute n'est pas sûr d'être récompensé. Un pas important en avant reste à faire pour l'amélioration de l'ergonomie de ces interfaces.

4.4. Pertinence liée à la valeur des documents

Un des problèmes essentiels du à l'absence de chaîne éditoriale sur le Web est la crédibilité et la valeur des informations qui y sont publiées. Mettre un document en ligne est en effet à la portée de tout un chacun. Pour estimer la crédibilité des sources, les moteurs utilisent largement un indice de popularité des pages (IPP), inspiré du facteur d'impact conçu en bibliométrie pour l'information scientifique et technique. Le facteur d'impact d'un article, calculé comme le nombre moyen de citations dont il a fait l'objet durant une période de référence, est considéré comme un indicateur de qualité de la recherche scientifique. Exploitant la nature hypertextuelle du Web, certains ont imaginé l'appliquer à ces ressources : l'IPP d'une page mesure la proportion de pages Web présentes dans l'index du moteur qui possède un lien vers cette page.

Ce critère avait déjà été contesté dans le domaine de l'information scientifique et technique, pour deux raisons essentielles. D'une part, il permet d'observer surtout les pratiques qui ont cours dans certains milieux scientifiques : certains auteurs privilégient explicitement les citations d'un petit groupe de chercheurs, chacun se rendant la politesse. Cette tentation existe également dans le contexte du Web, les individus créant un grand nombre de pages "vides" pointant sur leur site pour en augmenter artificiellement la popularité. Les auteurs de ressources peuvent également se lancer dans des campagnes d'échanges de liens pour valoriser leurs productions. C'est pourquoi les moteurs pondèrent généralement l'IPP en tenant compte de la popularité de la source des liens, estimée par l'IPS (indice de popularité d'un site) c'est-à-dire la proportion de liens qui pointent vers une des pages du site.

---

[6] Un seul étudiant sur les 120 observés en formation l'a utilisé spontanément, et pour une seule requête sur les 10 sujets proposés.

Une autre réserve porte sur les grandes différences sémantiques observées dans les citations. Dans un contexte scientifique, une citation peut représenter des relations de nature très diverses. Un document cité peut être l'objet d'une critique positive ou négative, d'un approfondissement ou de détails supplémentaires sur le sujet, de l'application d'un même concept à un autre domaine, etc. Dans le domaine du Web, il reste à étudier la diversité sémantique des liens proposés dans les pages. L'initiative Xlink[7] pourrait amener davantage d'investigations dans ce domaine.

Enfin, l'IPP pénalise les ressources nouvelles, puisqu'il faut un certain temps pour qu'une ressource soit connue et citée par d'autres. Dans le contexte du Web où la rapidité de publication est une caractéristique importante, cet aspect peut être négatif.

L'IPP et l'IPS sont, à l'heure actuelle, les seules mesures automatisables pour approximer la crédibilité d'un site dans un domaine particulier.

L'*indice de clic*, implanté par DirectHit[8] s'inspire du principe de rétroaction de pertinence implanté dans les SRI interactifs : il renforce l'importance des liens le plus souvent cliqués par les internautes dans les pages de résultats qu'ils obtiennent en réponse à leur requête, c'est-à-dire les pages qu'ils ont consultées réellement et sur lesquelles ils sont restés un certain temps. Un des problèmes posés par les jugements de pertinence des individus est leur inconsistance : en effet les jugements de pertinence individuels varient énormément pour une même requête d'un individu à l'autre, et pour un même individu ils varient dans le temps. Cependant, certaines études tendent à montrer que l'évaluation moyenne est relativement stable, lorsque l'on prend en compte les jugements d'un grand nombre d'individus (Vorhees, 2000), avec quelques réserves : pour les requêtes qui rapportent très peu de documents pertinents (moins de 5), la précision moyenne n'est pas stable. De même, les SRI interactifs, qui permettent l'expression de jugements de pertinence lors de nombreuses interactions montrent davantage de variabilité dans les résultats. Ce critère doit donc être utilisé avec prudence.

4.5. Pertinence de la présentation des résultats

Les résultats d'une requête sont présentés généralement sous la forme d'une page où apparaissent, de manière condensée, un certain nombre de liens vers des documents répondant potentiellement à la demande. Cette présentation succincte doit permettre à l'utilisateur de juger de l'opportunité de consulter l'intégralité de la ressource.

Pour la composition de cette présentation succincte, deux grandes approches se distinguent à l'heure actuelle. L'approche documentaire classique (exemple : Alta Vista[9]) propose un résumé du document extrait soit de l'en-tête (balises META description), soit du début du texte. Une autre approche (exemple : Google) sélectionne des extraits du texte où apparaissent en gras les termes de la requête. Si les phrases hachées et les extraits trop brefs ne permettent pas toujours de se faire une bonne idée de la ressource, cette présentation rassure le chercheur d'information sur le bon fonctionnement du moteur : il peut constater que le document proposé contient bien les termes qu'il a utilisés pour sa requête et il est invité implicitement à formuler une meilleure requête s'il n'est pas satisfait des résultats obtenus. Kuhlthau (Kuhlthau, 1999) souligne que l'approche traditionnelle des systèmes est de considérer que l'incertitude de l'utilisateur est un sentiment négatif qu'il faut réduire aussi vite que possible. Lors de sessions de formation à l'utilisation des moteurs de recherche, il est courant de voir des novices déconcertés par les résultats parce que les documents sont loin de leur besoin et ne contiennent pas de manière explicite les termes de leur requête (quelquefois, ceux-ci se trouvent dans l'en-tête du document, cachée à l'utilisateur). A l'incertitude liée à l'objet de la recherche s'ajoute alors celle liée au fonctionnement de l'outil.

---

[7] XLINK : XML Linking Language http://www.w3.org/TR/xlink (consulté le 14/12/2001).

[8] www.directhit.com (consulté le 29/01/2002)

[9] www.altavista.fr (consulté le 29/01/2002)

De nouvelles formes de présentations, plus visuelles, commencent à apparaître, par exemple sous forme de réseaux sémantiques (Kartoo[10]). L'impact de ce type de présentation reste à être évalué de manière approfondie

.

4.6. Pertinence du document par rapport au besoin

Juger de la pertinence du contenu du document par rapport à la demande fait intervenir différents paramètres liés à l'utilisateur : l'état de ses connaissances sur le sujet de sa recherche, ses savoir-faire - lecture rapide, lecture approfondie - et son niveau d'éducation - niveau de langage, richesse du vocabulaire, aptitude à décrypter des tournures syntaxiques, aptitude à transférer une méthodologie d'un domaine à un autre -vont influencer la pertinence de son interprétation. Le temps dont il dispose influera sur la longueur des documents choisis.

Il ne semble pas réaliste de demander à un utilisateur de décrire son niveau de connaissance et ses savoir-faire avant de lui permettre de poser une question. Ce problème reste donc entier pour les moteurs. Par contre, l'avènement de portails regroupant un grand nombre de ressources ciblées pour des publics particuliers peut répondre en partie à ce problème.

4.7. Pertinence du choix de l'outil

Il existe nombre de moteurs ou d'annuaires différents sur le Web. Choisir un outil adéquat pour son type de besoin fait partie d'une bonne démarche de recherche documentaire. Certains moteurs proposent des interfaces spécifiques pour aider à trouver tel ou tel type de document (par exemple des images, des séquences vidéo). Malheureusement, c'est souvent la loi du moindre effort qui est appliquée par les usagers : l'outil est choisi parce que l'utilisateur le connaît, qu'il lui est familier, et non pas parce qu'il est plus adapté qu'un autre au problème à résoudre.

On pourrait penser que les méta-moteurs résolvent ce dilemme, mais la contrepartie est lourde : la plupart d'entre eux ne permettent pas d'exploiter finement les particularités de chacun des moteurs qu'ils sollicitent. Une nouvelle génération d'outils, qui orienteraient l'utilisateur vers un moteur particulier en fonction de certaines caractéristiques de sa recherche, reste à imaginer. La mise en œuvre de techniques sophistiquées d'analyse de données, en aval de l'interrogation des moteurs, est une autre approche prometteuse.

5. PERTINENCE OU UTILITÉ ?

Le concept de pertinence, si flou et multidimensionnel, ne pourrait-il pas être remplacé simplement par celui d'utilité ? Kemp n'avait-il pas déjà défini la pertinence comme " *l'utilité de l'information contenue dans les documents pour un utilisateur particulier ayant un besoin d'information particulier* " (Kemp, 1974) ?

Dans le sens commun, l'utilité est la qualité, le caractère de ce qui est utile, c'est-à-dire propre à satisfaire un besoin, avantageux ou profitable, voire opportun[11]. On retrouve dans cette définition les dimensions subjective et temporelle mises en évidence dans les études de la pertinence, mais aussi tout le flou et l'ambiguïté du concept. Comme la pertinence, l'utilité est une mesure ordinale, qui permet de comparer les éléments entre eux, et non cardinale, c'est-à-dire ayant une valeur dans l'absolu.

---

[10] www.kartoo.com (consulté le 29/01/2002)
[11] Dictionnaire Hachette, 1999.

C'est principalement dans le domaine des sciences économiques que le concept d'utilité a été étudié. Dans ce domaine, l'utilité générale d'un bien est l'aptitude que les hommes lui reconnaissent de correspondre à leurs désirs. Les marginalistes opposaient l'utilité et la quantité. L'utilité subjective est définie comme l'importance que le sujet attribue à un bien disponible en quantité limitée. En situation de pénurie, l'utilité subjective s'accroît alors qu'elle diminue en situation d'abondance. On peut se demander si l'explosion du nombre de ressources disponibles sur le Web ne va pas banaliser la valeur de l'information. A l'heure actuelle, les internautes ne passent-ils pas déjà plus de temps à chercher, voire à butiner, qu'à lire et à exploiter les informations retrouvées ?

L'utilité d'un document renvoie moins aux connaissances du chercheur d'information qu'à l'usage qu'il prévoit pour l'information retrouvée. Or, la vraie valeur de l'information, c'est sa valeur d'usage, constituée par les interprétations sans cesse renouvelées qui en sont faites (Mayère, 1997). La capacité à bien lire les documents, analyser et synthétiser, relier les éléments informatifs à ses connaissances propres est déterminante dans ce domaine. Si le concept d'utilité rejoint celui de pertinence, il souligne moins l'importance de la démarche intellectuelle complexe liée au traitement maîtrisé de l'information.
Si le Web facilite les échanges et l'accès aux ressources documentaires, il reste beaucoup à faire pour améliorer les comportements d'usage de cette masse de ressources. C'est pourquoi, s'il est évidemment important de développer la performance technique des outils de recherche et leurs qualités ergonomiques, la formation des utilisateurs doit également être généralisée. Notamment avec l'avènement des techniques de plus en plus sophistiquées d'analyse de données qui commencent à être mises en œuvre pour permettre le classement des documents et l'exploitation des résultats obtenus suite à une recherche.

Conclusion

Nous avons essayé, de manière assez systématique, d'analyser les différentes formes de la pertinence pour chacune des relations qui s'établit durant le processus de recherche d'information sur le Web. Cette analyse ne se prétend pas exhaustive, mais elle met en évidence la diversité des critères qui interviennent et la complexité de leur imbrication.

Grâce à l'historique de leurs interrogations, les moteurs peuvent fournir pour les chercheurs une mine d'informations intéressantes, qui commencent tout juste à être exploitées pour l'évaluation de la performance des outils mais aussi pour des études comportementales en sciences sociales. Parallèlement, il reste à mener des analyses qualitatives plus complètes dans ce domaine. Edgard Morin (Morin, 1990) soulignait que l'information est principalement analysée sous son aspect statistique et son aspect communicationnel. "*L'aspect communicationnel ne rend absolument pas compte du caractère polyscopique de l'information[…]. L'aspect statistique ignore, y compris même dans le cadre communicationnel, le sens de l'information, il ne saisit que le caractère probabilitaire-improbabilitaire, non la structure des messages et, bien entendu, ignore tout de l'aspect organisationnel."*

Dans ce contexte actuel de développement du Web, une collaboration interdisciplinaire accrue entre chercheurs est devenue indispensable pour explorer plus à fond toutes les dimensions de l'information.

Bibliographie

Avrahami J., Kareev Y., 1993, What do you expect when you ask for "a cup of coffee and a muffin or a croissant"? On the interpretation of sentences containing multiple connectives, *International Journal on Man-Machine Studies*, vol. 38, p. 429-434.

Belkin N. J., Oddy R. N., and Brooks H. M., 1982, ASK for information retrieval : Part I. Background and theory, *Journal of Documentation*, 38(2), p. 61-71.

Brangier E., Zimmer P., 2001, Quelques principes d'amélioration de l'utilisabilité des systèmes de recherche d'informations : note bibliographique, *Revue d'Interaction Homme-Machine*, vol 2, n° 1.

Recherches récentes en Sciences de l'information - convergences et dynamiques, actes du colloque international MICS-LERASS, 21-22 mars 2002, Toulouse ; ADBS Éditions, collection Sciences de l'information, série Recherches et Documents, Paris, 2002, pp. 393-410.

Burnkrant R.E., 1976, A motivational model of information-processing intensity, *Journal of Consumer Research*, n°3, p. 21-30.

Cluzeau-Ciry M., 1988, Typologie des utilisateurs et des utilisations d'une banque d'image, *Documentaliste-Sciences de l'information*, vol. 25 n°3, p.120-155.

Hawking D., CrasWell N., Bailey P., Griffiths K., 2001, Mesuring Search Engine Quality, *Information Retrieval*, n° 4, p. 339.

Ingwersen P., 1992, *Information retrieval interaction*, Taylor Graham Publishing, Cambridge, UK.

Irving Ann, 1985, *Study and information skills across the curriculum*, London : Heinemann Educational Books.

Jansen B.J., Spink A., Saracevic T., 2000, Real life, real users, and real needs: a study and analysis of user queries on the web, *Information Processing & Management*, vol. 36, n° 2, p. 207-227.

.

Kemp D.A.,1974, Relevance, pertinence and information systems development, *Information Storage and Retrieval*, vol. 10 N° 2, p.37-47.

Kuhlthau, C. C., 1993, *Seeking Meaning : A Process Approach to Library and Information Services*, Norwood, N.J.: Ablex Publishing Corp.

Kuhlthau C. C., 1999, Accomodating the User's information search process : challenges for information retrieval system designers, *Bulletin of the American Society for Information Science*, vol. 25, n° 3.

Mayère A., 1997, Produits et services d'information : proposition de dépassement de la théorie standard de l'information, *in* Anne Mayère dir., *La société informationnelle*, L'Harmattan, communication, ISBN 2-7384-5453-4, p. 125-131.

Mizzaro S.,1997, Relevance : the whole history, *Journal of the American Society for Information Science*, vol. 48, n°9, p.810-832.

Morin E., 1990, *Introduction à la pensée complexe,* Collection Communication et complexité, ESF éditeur.

Park H., 1997, Relevance of science information: origins and dimensions of relevance and their implications to information retrieval, *Information Processing & Management*, vol 33, n° 3, p.339-352.

Simonnot B., Smaïl M., 1995, *Modèle flexible pour la recherche interactive de documents multimédias,* actes du XIIIe Congrès Inforsid, Grenoble (France).

Sperber D., Wilson D., 1989, *La pertinence : communication et cognition*, (traduction française de *Relevance : Communication and cognition*, 1985*),* Editions de Minuit.

Voorhees E. M., 2000, Variations in relevance judgements and the mesurement of retrieval effectiveness, *Information Processing and Management*, vol. 36, p. 697-716.