La recherche d'information

Tout ce que je sais, c'est que je ne sais rien..
Socrate

Introduction

Le terme de recherche d'information (*Information Retrieval*) est apparu pour la première fois en 1948 dans le mémoire de MASTER du MIT [1] de Mooers (Mooers, 1948). Dinet et Rouet définissent la recherche d'information (RI) comme « l'activité d'un individu qui vise à localiser et traiter une ou plusieurs informations au sein d'un environnement documentaire complexe, dans le but de répondre à une question ou de résoudre un problème (Dinet et Rouet, 2002) ».

Dans le premier chapitre de *An Introduction to Information Retrieval*, Manning *et al.* (2008) nous offrent la définition suivante de la recherche d'information : « *As*

1. *Massachusetts Institute of Technology*

Figure 3.1: Le modèle « simpliste » Morville et Rosenfeld (2006) de système d'information

an academic field of study, information retrieval might be defined thus :Information retrieval (IR) is finding material (usually documents) (...) that satisfies an information need from within large collections (usually stored on computers)[1]. »

Dans le troisième chapitre de leur livre *Information architecture for the World Wide Web*, Morville et Rosenfeld présentent la vision « grand public » d'une recherche d'information (cf. Fig. 3.1). Cette idée reçue est qu'il suffit de poser une question à un moteur de recherche et la « boîte magique » donnera la réponse ultime relativement à notre besoin de connaissances par rapport au sujet. Toujours selon Morville et Rosenfeld, ce type de résultat fait figure de cas isolé, pour ne pas parler d'exception (Morville et Rosenfeld, 2006, chp. 3). Une requête informationnelle est dépendante du fonctionnement du système d'informations et pas seulement de la personne qui l'initie. Posons les définitions suivantes proposées par l'association des professionnels de l'information et de la documentation (ADBS)[2] :

Définition de Recherche d'Information

> Ensemble des méthodes, procédures et techniques permettant, en fonction de critères de recherche propres à l'usager, de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés ADBS (2012).

1. Proposition de traduction. « En tant que discipline académique, la recherche d'information peut être définie ainsi : La recherche d'information (RI) est le repérage au sein de grandes collections (généralement stockées sur les ordinateurs) de ressources (habituellement des documents) qui répondent à un besoin d'informations. »

2. `http://www.adbs.fr/vocabulaire-de-la-documentation-41820.htm`, accédé le 1er août 2012

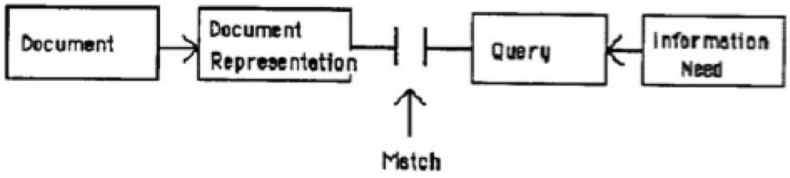

Figure 3.2: Modèle trivial de SRI proposé par Bates

Définition de Recherche documentaire

> Ensemble des méthodes, procédures et techniques ayant pour objet de retrouver des références de documents pertinents (répondant à une demande d'information) et les documents eux-mêmes ADBS (2012).

Définition de Recherche bibliographique

> Ensemble des méthodes, procédures et techniques ayant pour objet de retrouver les références bibliographiques de documents pertinents ADBS (2012).

De ces trois définitions, nous dégagerons une synthèse fonctionnelle de la recherche d'information, qui en tant que processus ne peut être réellement séparée des recherches documentaire et bibliographique. Notre vision sera donc pragmatique, nous voyons la recherche bibliographique comme la finalité de la recherche documentaire, elle-même résultante de la recherche d'information.

Marcia Bates, reprenant les concepts de bases posés par Robertson, proposait un schéma volontairement trivial (cf. Fig. 3.2) modélisant un système de recherche d'information (SRI) dans un contexte idéal (Bates, 1993, Robertson, 1977). Pour elle, il s'agit d'une représentation utopique, bien que répandue, de l'offre documentaire, avec le besoin d'informations formalisé par le biais d'une requête.

Nous ajouterons que le but de cette activité peut être purement cognitif, c'est-à-dire élargir son champ de connaissances dans un domaine (dans notre optique scientifique) pour tenter de le synthétiser pour mieux en saisir la substance. Il ne s'agit pas dans ce cas à proprement parler d'un besoin d'information, mais d'une forme de curiosité, l'envie de mieux appréhender un champ de connaissance.

3.1 Le paradoxe de la RI

Ce qui est paradoxal dans la RI c'est qu'elle peut traduire :

- Un besoin de références bibliographiques pour structurer et étayer sa connaissance et ses idées ;
- Directement un besoin de connaissances, c'est-à-dire combler une lacune.

Cependant dans ce dernier cas, la prise de conscience de ce besoin, ou manque informatif, découle d'une expérience du domaine [1]. Pour comprendre son besoin d'informations, il faut avoir déjà effectué un panorama du champ de connaissances (Boubée *et al.*, 2005). Si nous explicitons différemment les choses, pour comprendre son ignorance, il faut déjà avoir commencé à chercher. Y. F. Le Coadic définissait cet état de connaissance : « nous en savons assez pour savoir que nous avons un besoin d'information, mais nous n'en savons pas assez pour pouvoir poser les bonnes questions (Le Coadic, 2008) ». Cette réflexion peut être illustrée par la citation de l'en-tête de chapitre attribuée à Socrate : *Tout ce que je sais, c'est que je ne sais rien.* L'une des premières difficultés pour les usagers est d'identifier les sources pertinentes et d'avoir une vision claire des contenus.

Nous allons dans cette première partie étudier dans le détail les différents aspects de la recherche d'information dans les systèmes numériques. Nous étudierons les interfaces de recherche d'information et les mécanismes qui y sont associés.

1. Nous voyons le détail du besoin d'information et les aspects psyscho-cognitifs qui y sont liés dans le chapitre 5 « Les écoles de pensées en RI : Processus et Cognition »

3.2 Concepts, modèles et méthodes en RI

Internet offre une vaste collection de documents, et son utilisation comme source d'informations est évidente et est devenue très populaire. Comme l'ont souligné et analysé Dennis *et al.* (2002), il y a pléthore de technologies de recherche d'information en ligne, qui peuvent principalement être classées en quatre catégories :

1. Recherche par mots clés, sans aide. Un ou plusieurs termes de recherche sont entrés et le moteur de recherche renvoie une liste classée des résumés de documents hyperliés.

2. Recherche assistée. Le moteur de recherche produit des suggestions ou recommandations basées sur la requête initiale de l'utilisateur.

3. Recherche par classification. L'espace d'information est divisé en une hiérarchie de catégories, où l'utilisateur navigue du générique vers le spécifique.

4. Requête par l'exemple. L'utilisateur sélectionne un élément intéressant d'un hypertexte, qui est ensuite utilisé comme base d'une nouvelle requête.

Nous ajouterons à cela une cinquième catégorie, la recommandation qui propose de l'information potentiellement intéressante en fonction de l'usager et/ou du contexte de recherche.

Les outils de recherche d'information se divisent en deux catégories radicalement distinctes. Les portails de recherche, ou annuaires web sont des sites Internet qui proposent des liens vers un florilège de sites repérés par des experts d'un domaine pour leur qualité. Les moteurs de recherches sont des systèmes complexes permettant de trouver des ressources dans un corpus numérique. Ce corpus peut être l'Internet dans sa globalité, une base de connaissances ou un seul site web.

3.2.1 Portails de connaissance

Un portail de connaissance est un site de référence dans un domaine précis ou une page hypertexte dédiée à une communauté particulière. Ce site se présente sous la forme d'un ensemble de pages web hyperliées. Un portail peut être perçu comme un point d'entrée sur un panel de ressources autour d'un thème commun. Souvent, ces portails offrent une vingtaine de catégories pour le premier niveau de la classification. Le type des documents référencés et agrégés importe moins que leur spécificité commune : la thématique. Le plus souvent, ces indexations procèdent d'une intervention humaine. C'est le cas des portails spécialisés de Wikipédia . En effet, l'encyclopédie participative en ligne possède un portail d'accès pour chaque grande thématique. Ces thématiques sont animées par des groupes d'intérêt, qui compilent des hyperliens vers les articles au sein des portails. Un autre exemple de portait particulièrement intéressant, parce que géré manuellement, était l'annuaire Google dédié à l'informatique. Une hiérarchie de sujets prédéfinis comme l'informatique, le sport, l'art ou la musique est maintenue et enrichie de manière manuelle. D'après Eissen et Stein (2002), ces hiérarchies statiques ne sont pas satisfaisantes à deux égards :

1. elles nécessitent un effort de maintenance humaine considérable ;
2. pour des sujets particulièrement spécifiques, les catégories de navigation génériques, ou points d'entrée, sont inutiles et allongent considérablement le processus de recherche.

Ces deux points sont indéniables. À moins de justifier le recours à un comité d'experts par sa préexistence (certains sites internet sont créés par un comité d'experts, c'est le cas d'IEEE ou d'ACM), cette méthode est onéreuse et chronophage. De plus, l'accès à l'information pointue est ralenti, voire rendu malaisé si la classification n'est pas triviale. Ce dernier point peut contrevenir avec l'immuable règle des trois clics (Scapin et Bastien, 1997) qui situe le seuil de tolérance d'un usager de système d'informations à un maximum de trois clics pour trouver l'information désirée dans un hypertexte (Sottet *et al.*, 2005). Ainsi, Google a fermé ses services d'annuaire durant l'été 2011. Les services ont été repris depuis par leur initiateur historique Dmoz [1].

1. L'*Open Directory Project*, ou *Directory Mozilla* qui donne son nom au site, Dmoz.org, est un répertoire de sites web créé en 1998, sous licence *Open Directory*. Il est géré par une vaste communauté d'éditeurs bénévoles provenant du monde entier, chacun étant responsable de vérifier l'exactitude et la

3.2.2 Moteurs de recherche

Un moteur de recherche est un outil dont l'interface permet de localiser une information dans une base de données à partir d'une requête. Originellement, les moteurs de recherche étaient des programmes installés localement sur les ordinateurs et ils consultaient, soit des bases locales, soit des bases distantes à travers des protocoles tels le FTP ou Gopher. Avec l'émergence du protocole HTTP, de nouveaux types de moteurs de recherches apparurent. Il s'agit d'une base de données indexant le contenu des pages référençables sur le web visible. Cette base est alimentée par un robot qui parcourt en permanence l'Internet.

Modèle fonctionnel

Curt Franklin expliquait qu'un moteur de recherche est composé de quatre parties principales (cf. figure 3.3 [1]) qui sont détaillées ci-dessous (Franklin, 2000) :

1. Un robot collecteur (web crawler) qui parcoure l'Internet de page en page. Un robot d'indexation est également appelé *crawler* ou *spider*. Il s'agit d'une analogie à une araignée qui parcourt inlassablement la toile, se déplace dans le web de site en site pour en collecter le contenu. Le contenu et les URL sont intégrés à une base de connaissances pour être traités.
2. Un indexeur chargé d'archiver les pages et d'en extraire les termes clés. Les URL des pages et les mots clés seront ensuite intégrés dans une base de données associative. L'indexation des pages repose sur les métadonnées collectées et celles calculées (mots clés calculés par traitement automatique du langage). Le système confronte les mots clés proposés par le webmestre éditorial et les termes émergeant de calcul statistiques. Un système de pondération permettra de faire l'équilibre entre ces données et de proposer des termes représentatifs du contenu de la page.
3. Un index inversé qui est une deuxième entrée sur la base de connaissance qui associe à chaque concept l'ensemble des URL des documents qui sont pertinents.

catégorisation des sites dans une ou plusieurs catégories. Ce service est toujours accessible en français à l'url `http://www.dmoz.org/World/Fran%C3%A7ais/`

1. Crédit image : `http://computer.howstuffworks.com/internet/basics/search-engine1.htm`, accédé le 1 août 2012

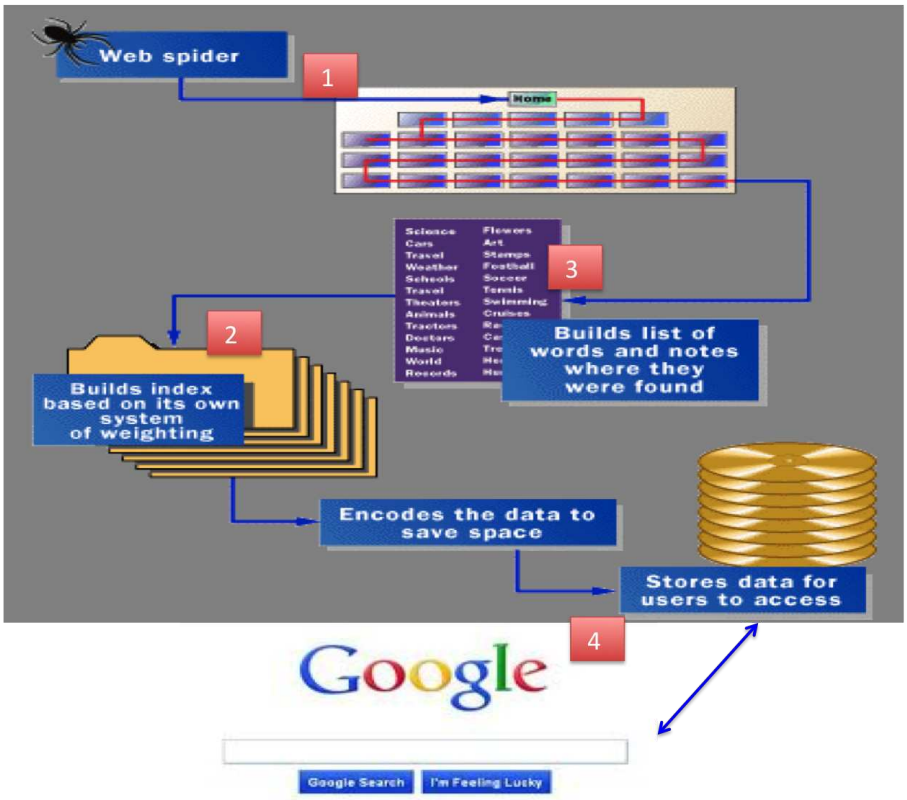

Figure 3.3: Schéma fonctionnel d'indexation d'un moteur de recherche.

4. Une interface de consultation qui a pour rôle d'aider les utilisateurs à interroger facilement la base de données sans connaître le langage d'interrogation de la base.

Les principaux robots sont Googlebot, Slurlp de Yahoo et ExaBot du français Exalead. Grâce aux dernières moutures de son algorithme d'indexation (*Panda* et *Penguin*) et aux robots qui parcourent inlassablement internet, le leader Google ambitionne de ne proposer que de la qualité à ses utilisateurs dans les premiers des résultats de recherche. Le moteur va éliminer tous les sites ayant un contenu de piètre qualité, ou résultant d'un copier-coller depuis un autre site. Cette expérimentation est en cours en France depuis la mi-août de l'année 2011 pour *Panda* et avril 2012 pour *Penguin* et nous n'avons pas encore le recul pour juger des résultats, si ce n'est par les baisses de classement des sites trop sémantiquement pauvres. Cependant, comme Google avait appelé dès le début de l'année 2011 à adopter de bonnes pratiques, entre autres l'usage de métadonnées descriptives, il est à espérer une amélioration des résultats.

Traitement du langage naturel

Quand un utilisateur entre une requête dans le formulaire d'interrogation d'un moteur de recherche, il ne se doute pas de la transformation que va subir sa requête. L'utilisateur va peut-être utiliser le moyen d'expression qui lui est familier, à savoir une phrase construite complexe. Cette phrase aura une syntaxe propre et peut être même une orthographe qui lui sera particulière. Nous appellerons ce mode d'expression le langage naturel.

Un des problèmes majeurs auxquels font face les concepteurs de systèmes de recherche d'information (SRI) est la correspondance entre l'information désirée par l'utilisateur avec celle contenue dans les documents indexés. Cette problématique est d'autant plus complexe que la requête exprimée est une étape intermédiaire qui s'intercale entre l'information désirée par projection mentale et les informations disponibles. En 1950, Alan Turing prédisait qu'il ne serait pas possible de communiquer de manière naturelle avec une machine avant la fin du 20^{e} siècle (Turing, 1950) [1]. Malheureusement, à l'heure actuelle aucune solution matérielle ou logicielle n'est à même de réaliser cet exploit (Saygin *et al.*, 2000).

Cependant, grâce à la lemmatisation, à la détection des expressions composées, aux thésaurus et aux réseaux sémantiques, il est possible pour un automate de traiter convenablement une requête en langage naturel pour en extraire les concepts dominants.

De manière générale, avant de commencer à comparer les éléments de la requête avec la base de connaissances, le travail suivant est effectué :

1. Le correcteur orthographique va aligner les termes saisis avec une orthographe cohérente et éliminer les risques d'erreurs de saisie.
2. Les termes composés de plusieurs mots vont être détectés pour être traités comme une seule entité.
3. Le texte va être segmenté en termes (Kan *et al.*, 1998).
4. Une désambiguïsation va tenter de régler les problèmes de polysémie.
5. Un lemmatiseur ou un stemmer va réduire les termes à leur racine.
6. Les mots non discriminants vont être exclus de la requête.

Revoyons en détail ces différents points de traitement de la requête initiale.

1. `http://www.loebner.net/Prizef/TuringArticle.html`, consulté le 22 juin 2012

La correction orthographique

Pour contextualiser la correction orthographique ou grammaticale de termes dans le cadre d'une recherche d'information, Sitbon *et al.* (2007) rappellent que le traitement est une réécriture en vue d'un traitement automatique et non pas une correction complète dans le but de fournir une correction complète tant du point de vue de la grammaire que de l'orthographe. Nous distinguerons la correction grammaticale et celle orthographique. Dans le cadre d'une correction purement orthographique, l'outil compare les mots d'une requête avec ceux d'un dictionnaire. Les problèmes liés à la mauvaise écriture d'un mot dans une requête peuvent avoir des causes multiples :

- Un mauvais usage du périphérique d'entrée, comme une faute de frappe.
- La dysorthographie qui est un trouble de la production écrite généralement associé à la dyslexie ou à l'inattention.

Selon Sitbon *et al.* (2008), dans le cas d'une dysorthographie, les troubles les plus courants sont :

- Segmentation en mots erronée (Gillon, 2004b), exemple :
  re-cherche d'un fort ma Sion au lieu de recherche d'information.
- Erreurs de conversion entre graphème et phonème, exemple :
  Unphormassion au lieu d'Information.
- Confusions de phonèmes, exemple :
  Monné au lieu de Monnaie.

Le traitement de ces troubles est organisé dans le cadre d'un correcteur orthographique par des algorithmes qui calculent la présence d'un mot dans un dictionnaire et à défaut propose une solution de remplacement. Ce peut être par exemple l'utilisation distance de Levenshtein qui calcule la distance minimale entre le mot dysorthographié et un autre mot du dictionnaire (Dice, 1945, Levenshtein, 1966).

Le rapprochement et la segmentation

Le rapprochement est la première étape de la segmentation, ou *tokenisation*. La *tokenisation* consiste à séparer les lemmes entre eux. La détection d'expressions composées de plusieurs mots dans une requête peut être traitée de plusieurs manières. Pour segmenter un texte, Sitbon et Bellot (2005) proposent de le séparer en chaines lexicales cohérentes au niveau du sens, ce qui n'est pas applicable dans ce contexte. En effet, dans le cadre d'une requête, le texte est trop court pour être segmenté en chaines

préfixe 1	lexème 2	suffixe
over	clock	ing
sur	cadence	ment

Tableau 3.1: Exemple bilingue de racinisation

lexicales. Cependant, si cette méthode sert principalement à résumer automatiquement un texte, une recherche d'entités nommées peut être utilisée pour analyser le contenu d'une requête. On appelle entité nommée dans un texte tout ce qui fait référence à un concept unique. En se basant sur les travaux de Chinchor et Robinson (1997), Sitbon et Bellot proposent d'utiliser trois types d'entités nommées à partir d'un lexique fermé : listes de noms de personnes, noms de lieux et noms d'organisations (Sitbon et Bellot, 2005). Originellement, Chinchor proposait un traitement textuel qui offrait en sortie un texte reformaté au format XML et traitait également les unités temporelles (dates, horaires), et les quantitatives (valeurs monétaires et pourcentages). Si un tel travail est fort utile dans une quête de sens, les moteurs de recherche font rarement de la classification, juste de la segmentation.

La lemmatisation, racinisation, troncature et désambiguïsation

En poursuivant notre étude sur les méthodes de traitement d'une requête dans un moteur de recherche, nous allons distinguer trois types de manières de « traiter » les termes qui composent la requête. Après le rapprochement des mots dans les termes composés, le passage de la liste de mots vides, les termes restants vont être traités pour éliminer les variations morphologiques. Chaque terme est réduit à une forme terminologique minimale par la racinisation, lemmatisation ou troncature.

Le processus de racinisation, ou stemming

Dans un cadre de racinisation les deux termes du tableau 3.1 seront réduits au lexème (racine) alors qu'en lemmatisation le mot entier forme une seule entité nommée référencée dans un dictionnaire. La racinisation repose sur une liste d'affixes de la langue et sur un ensemble de règles de dé-suffixation construites *a priori* (Moreau et Claveau, 2006). La base de données Postgres propose Snowball, une solution logicielle basée sur le projet de Martin Porter, inventeur du populaire algorithme de *stemming* en

anglais. Ce système est composé d'un dictionnaire de données et d'un ensemble de règles. Snowball propose maintenant des algorithmes stemming pour un grand nombre de langues, dont le français. L'algorithme sait comment réduire et normaliser les variantes d'un mot (flexions) vers sa base, ou *stem*.

Le processus de lemmatisation

Le processus de lemmatisation est une tâche complexe de réduction des termes à leur forme minimale, ou forme canonique. Un lemme peut être :

- simple : un seul mot ; par exemple : « fûtes » aura pour forme canonique, ou lemme, le verbe « être » quel que soit le contexte.
- composé : un mot composé (mot formé de plusieurs mots) ; par exemple : *peer-to-peer*
- complexe : un syntagme ou expression (groupe de mots placés dans un sens précis et s'organisant autour d'un terme central) ; par exemple : *peer reviewed paper*.

La Bibliothèque du CNAM propose un dictionnaire français de lemmatisation [1].

La lemmatisation peut intervenir, dans sa forme avancée, en contexte. Cette technique identifie la fonction grammaticale d'un mot pour en déduire son lemme. A partir du moment où la fonction grammaticale a pu être détectée, le lemmatiseur recherche dans sa base de connaissance le mot puis retourne le lemme associé à la fonction grammaticale. La solution la plus connue de lemmatisation est *TreeTagger* développée par l'Université de Stuttgart et dont les ressources linguistiques sont disponibles pour de multiples langues dont le français [2].

Le processus de troncature

Une autre méthode de flexion est la troncature, c'est à dire de simplement couper un mot pour n'en garder que le début. La méthode proposée par Enguehard consiste à ne garder de chaque terme que la sous-chaîne de caractères commençant au début du mot jusqu'à atteindre deux voyelles non consécutives (Enguehard, 1992). Cette heuristique qui permet de dé-suffixer les termes par approximation est très rapide et peu coûteuse. La troncature est complètement hors de propos dans un cadre d'indexation, mais peut trouver sa place dans une algorithmique de SRI quand la vélocité est à privilégier.

1. `http://abu.cnam.fr/DICO/mots-communs.html`, accédé le 1er août 2012
2. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`, accédé le 1er août 2012

Les analyseurs morphologiques (raciniseur et troncature) sont généralement plus faciles à mettre en œuvre que les systèmes complexes d'analyse grammaticale. De plus, il fonctionnent plus rapidement, du fait de leur simplicité. Par rapport à une analyse morpho-syntaxique (lemmatisation) des termes dans le cadre d'un processus d'analyse de requête, la précision est donc forcément réduite par analyse uniquement morphologique (racinisation ou troncature). Cependant, pour une intégration dans un moteur de recherche, quand la rapidité doit être privilégiée, la racinisation ou la troncature peuvent être préférées.

La soustraction des mots vides

En informatique appliquée à la recherche d'information documentaire, il existe des mots dits « vides » qui ne doivent ni être indexés dans le cas d'une indexation ni recherché dans le cadre d'une requête, car non discriminants. Ces mots « vides », perturbent le score de recherche en introduisant du bruit [1]. Les mots vides (*stop words* en anglais) sont alors souvent regroupés dans un « anti-dictionnaire » (*stop-list* en anglais). Ces mots sont les déterminants, prépositions, conjonctions et adverbes (Ibekwe-SanJuan, 2007).

Mots clés et critères booléens

Identifier de l'information pertinente pour l'individu dans le déluge informationnel de réponses à une requête est une tâche qui, une fois de plus, requiert une certaine pratique de la recherche d'information. La réduction des réponses ne faisant pas sens dans le contexte, ce que l'on appelle le bruit passe par l'usage de filtres par opérateurs booléens. Or, selon Spink *et al.* (2001) seules 5 % des requêtes comportaient en 2001 au moins un opérateur booléen. Cette pratique est indispensable pour encadrer sa demande informationnelle, mais elle n'est pas forcément maîtrisée par les usagers. Le terme d'algèbre booléenne, ou logique, vient du nom du mathématicien, logicien et philosophe anglais George Boole. Ce dernier publia ses travaux relatifs opérateurs binaires (Boole, 1854) au 19^{e} siècle. L'algèbre booléenne correspond à une grammaire basée sur trois opérateurs : ET, OU, SAUF (en anglais AND, OR, NOT). Ils permettent d'interroger efficacement un outil de recherche d'information.

1. Voir section 4.1.1.

Ces termes sont communs à tous les moteurs de recherche présents sur Internet, d'où la nécessité de bien les maîtriser. Les termes ET et SAUF sont parfois représentés sous la forme de + et -. Il est à noter que dans certains moteurs de recherche, l'opérateur booléen + est implicite. C'est le cas de Google : si l'on ne mentionne rien, un opérateur AND est intercalé à chaque blanc (espace vide).

L'opérateur AND (ET en français) ou +. L'opérateur ET implique que les termes de votre recherche soient contenus dans les pages de résultat. Il faut également garder à l'esprit que les termes ne sont pas forcément contigus, ni même dans le même ordre. Prenons l'exemple d'une requête visant à récolter de la documentation sur le thème de la recherche d'information en bibliothèque :

Information **AND** *Retrieval* **AND** *Library*

Les réponses à cette requête contiendront obligatoirement les mots information, *retrieval* et *library*. Cette méthode permet, en ajoutant progressivement des termes clés, d'affiner la requête et d'obtenir des résultats moins nombreux et plus adaptés (voir section 4.1.1 le concept de bruit).

L'opérateur OR (OU en français), parfois noté « | [1] », offre la possibilité de sélectionner l'un ou l'autre des termes d'une recherche dans les résultats. Pour chercher un terme composé de plusieurs mots, il faut intercaler la séquence recherchée dans des guillemets de type anglo-saxon ou *double quote*. Reprenons notre exemple de requête sur les recherches en bibliothèque.

"information retrieval" **OR** *"information seeking"* **AND** library

Cette deuxième requête aboutit sur des réponses qui contiendront soit « *Information Retrieval* », soit « *Information Seeking* », soit les deux. Nous avons ajouté le terme

1. | se prononce *pipe*, tuyau ou tube en anglais.

« *Library* » pour limiter la recherche aux sciences de la documentation. L'usage de OU inclusif permet de faire une recherche impliquant des synonymes. Il est ainsi possible de couvrir un maximum de documents portant sur des concepts identiques (voir plus loin la notion de silence).

L'opérateur NOT (SAUF en français).

L'opérateur **SAUF** propose d'exclure un terme de la liste des réponses proposées par le système de recherche d'information. SAUF peut être noté également « - », selon les outils de recherche d'information. Si nous poursuivons notre exemple, nous pourrons l'adapter ainsi :

"Information Retrieval" **AND** Library **NOT** *"information seeking"*

Dans l'exemple, le résultat portera sur tout les éléments indexés par le système comprenant « *Information Retrieval* » et *Library*, mais pas « *information seeking* ». L'opérateur booléen NOT est ici représenté par le signe mathématique « - ». Une requête comprenant l'opérateur NOT permet de préciser une recherche (voir partie 4.1.1 page 104, le concept de pertinence) en réduisant le nombre de résultats. Il est très utile dans le cadre de la polysémie, de soustraire des termes issus du champ lexical qui ne nous intéresse pas (voir section 4.1.1 la notion de bruit). Pour illustrer notre propos, prenons l'exemple de l'*Association for Computing Machinery* (ACM). Cette association présente une conférence annuelle sur la sécurité des systèmes d'informations. Ce rassemblement, une référence dans le domaine, est sobrement baptisé *ACM conference on Computer and Communications Security*. L'acronyme de cet événement est ACM CCS. Dans notre champ d'intérêt, l'ACM propose également une taxonomie de l'informatique. Ce document est unanimement reconnu comme la référence dans le monde de l'informatique. Les domaines de la recherche et classification d'informations relatives à l'informatique scientifique l'utilisent comme base de classification. Ce système de classification est nommé ACM *Computing Classification System*, son acronyme est également ACM CCS. Si nous mettons en œuvre une recherche simple sur le terme « ACM CCS », au 12 septembre 2011, le moteur de recherche Google offre 110 000 réponses. En utilisant

l'opérateur booléen NOT le résultat est d'environ 26 300 résultats grâce au formulaire de recherche avancée de Google.

- dans un langage spécifique ;
- grâce à l'utilisation de mots-clés.

La requête peut être exprimée dans un langage de requête booléen en langue naturelle au travers de l'interface de recherche.

Les critères avancés de recherche

Des opérateurs spécifiques permettent de construire une requête en délimitant certains aspects. Ces opérateurs ont en général pour objectif de spécifier ou d'élargir une recherche pour encadrer au maximum les phénomènes de bruit et de silence [1].

La distance

Les opérateurs de distance permettent de rechercher des documents au sein desquels les termes t_1 et t_2 seront distant d'un maximum de n mots. Ainsi, plusieurs mots liés par NEAR (moteur Bing) AROUND (Google) doivent apparaître ensemble à une distance limitée (généralement, un maximum de 10 mots). Avec le moteur de recherche Google, l'operateur AROUND peut même devenir une fonction qui prend en argument une valeur numérique de n.

"information retrieval" **AROUND(10)** *"ontology"*

Google offre la possibilité de restreindre sa recherche à un nom de domaine, à un sous élément d'un domaine, voir même à un seul site web. La commande *site :* est l'opérateur de ce type de requête. Il s'agit d'une option accessible à travers le formulaire de recherche avancé de Google ou directement en mode requête.

La commande *define :* offre la possibilité de demander à Google de rechercher une définition d'un terme sur l'Internet. Les résultats retournés sont souvent des définitions

1. Voir section 4.1.1

Fonctions/Moteurs	*Google*	*Bing*
ET	*espace vide,+,AND*	*AND,&,&&*
OU	*| , OR*	*| , OR*
SAUF	*- , NOT*	*- , NOT*
PROCHE	*AROUND(n)*	*NEAR :*
SYNONYME	*∼*	
LIMITER à un espace	*site :*	*site :„domain :,url :,ip :*
DÉFINIR	*define :,définir :*	*define*
Dans le TITRE	*intitle :*	*intitle :*

Tableau 3.2: Résumé des fonctionnalités des deux principaux moteurs de recherche commerciaux en 2011

issues de Wikipédia ou de dictionnaires en ligne. Il arrive également de dénicher des définitions de spécialistes sur des blogs ou des sites de recherches.

L'option Google ∼ (*titlde espagnol*) permet d'ajouter un mot et ses synonymes à une requête. Cet outil est à double tranchant, car si il permet d'éviter des résultats nuls et le silence, il génère également très rapidement du bruit [1]. Il est donc préférable d'utiliser le *tilde* avec parcimonie quand un complément d'information est nécessaire suite à un silence trop important.

1. Voir section 4.1.1

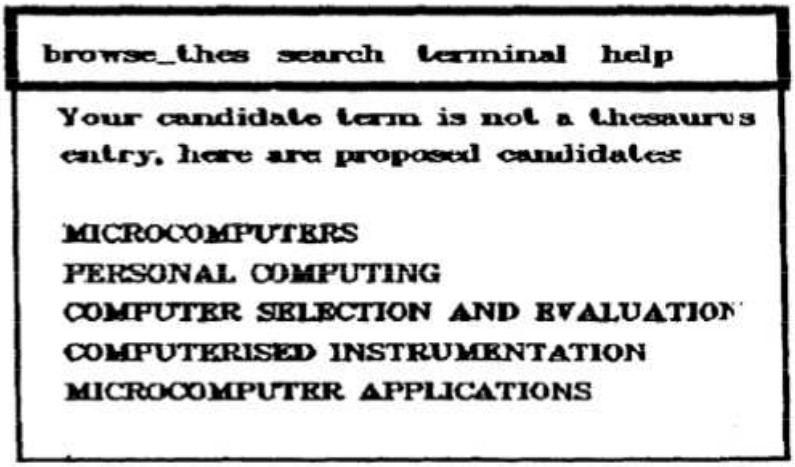

Figure 3.4: Les origines de la recherche à facettes informatisée

Recherche par facettes

Selon Boutin (2008), l'époque où l'on exprimait son besoin d'information uniquement par une requête générique décrivant la thématique générale des documents est révolue. Il s'agit aussi de caractériser son besoin d'information par des dimensions complémentaires (appelées facettes) qui ne renvoient pas seulement au contenu thématique des documents. Nous en avons identifié plusieurs et retenu cinq dans l'implémentation que nous avons proposée : le niveau de polarité d'une page web, le niveau de subjectivité d'une page web, le niveau d'accessibilité d'une page web, le niveau de lisibilité d'une page web et la centralité d'une page web dans son contexte. Chacune de ces dimensions a fait l'objet de développements théoriques et pratiques dans des domaines scientifiques d'appartenance, par exemple la linguistique computationnelle ou la psychologie cognitive. Notre objectif a consisté à aller chercher ces concepts et à étudier dans quelle mesure ils étaient transposables à l'analyse de corpus web. Le concept de recherche à facettes, ou par facettes est nommé ainsi par analogie avec un objet qui dans un monde en 3 dimensions possède de multiples facettes ou angles de visualisation. Comme nous l'explique Boutin (2008), le concept initial de facettes pour une classification bibliographique est imputé à S.R. Ranganathan Ranganathan (1963). L'idée de relier les éléments de métadonnées à l'affichage de résultats dans une RI est attribué à Belkin et Marchetti (1989). À l'époque, l'outil proposé liait un thésaurus à une interface de recherche par terme clé.

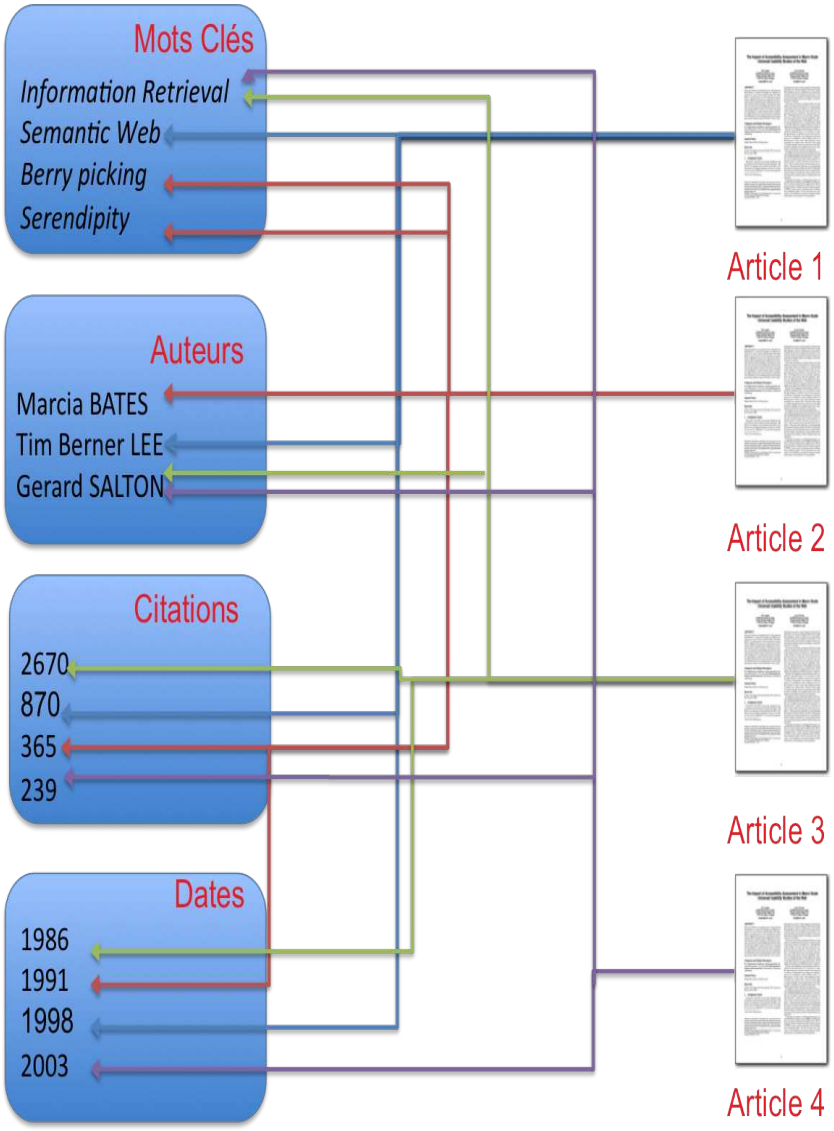

Figure 3.5: Schéma d'une recherche à facettes

Dans l'exemple proposé dans l'article, Belkin et Marchetti (1989) montraient comment rechercher les termes liés à *computer* dans le thésaurus (cf. Figure 3.4). Ce premier exemple illustre les différents aspects d'un même terme.

Ainsi la présentation à facettes permet une classification multiple pour un objet. Chaque facette correspond typiquement à la valeur possible d'une propriété commune à un ensemble d'objets. Dans la figure 3.5, nous proposons d'illustrer par l'exemple la recherche par facettes. Cette recherche est non contractuelle et dans un souci de lisibilité, nous avons limité le nombre des articles et celui des facettes. Si une recherche aboutit à l'affichage de 4 articles, il est possible dans le cas présent de les classer selon 4 critères :

- l'auteur ;
- la date ;
- les mots clés associés ;
- un élément de bibliométrie : ici le nombre de citations.

Pour chaque article, une couleur différente est utilisée pour faciliter la visualisation des facettes en deux dimensions. Il est possible de sélectionner les articles en fonction d'un aspect bien particulier issu des métadonnées. Une sélection par auteur, sans précision retournera ainsi 4 entrées classées par ordre alphabétique alors que si l'on sélectionne l'auteur Salton, seuls deux documents répondront au critère de cette facette. Le critère de sélection peut aussi être appelé contrainte. Une des plus belles réussites dans la présentation et la navigation à facettes dans un corpus multimédia est le projet Flamenco. Les facettes sont préexistantes dans la base de connaissances, il peut s'agir de présenter les résultats d'une requête par auteur, mots clés, la langue, disponibilité ou format de fichier. Un fil d'Arianne (*breadcrumb*) rappelle à l'usager les contraintes de recherches imposées aux facettes.

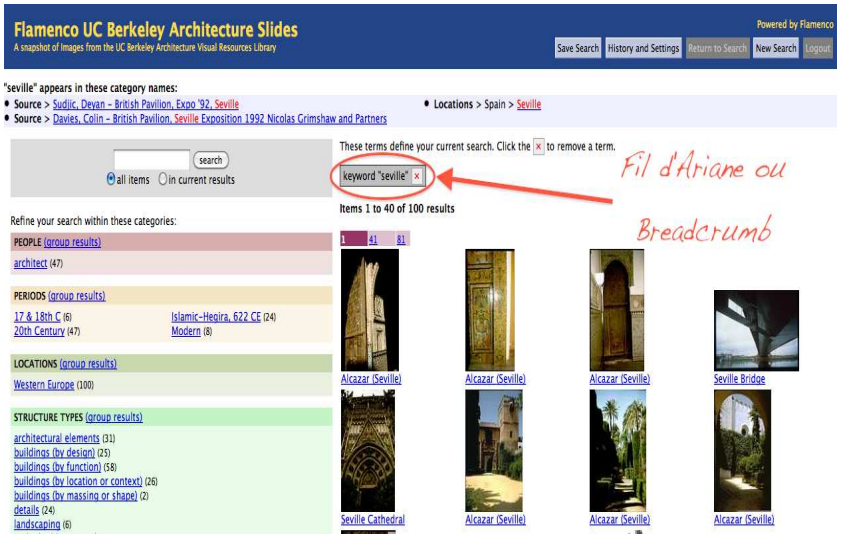

Figure 3.6: Projet Flamenco

Moteur à curseur

Pour aller plus loin dans le concept de facettes, Boutin déclare que « l'expression du besoin pourrait être affinée par l'internaute à travers l'expression de dimensions complémentaires au sujet de la recherche (Boutin, 2008) ». Son propos sort la recherche à facettes de la simple réorganisation de la présentation des résultats sur des critères purement objectifs comme ceux fournis par les métadonnées. Boutin déclare qu'un contenu de l'hyperespace est également possible à mettre en exergue sous des aspects subjectifs tels la tonalité du discours, le degré de subjectivité, son niveau de langage. Cette catégorisation supplémentaire offre à l'usager un moyen original de spécifier sa recherche pour limiter le bruit [1]. Cette méthode s'illustre le plus souvent par des curseurs qui permettent de régler avec précision la focale d'une recherche sur un ou plusieurs aspects. Boutin citait en exemple les moteurs clush.com et le mindset de Yahoo. L'initiative de Yahoo offrait un curseur horizontal axé de la publicité (*shopping*) vers l'information (*researching*). Ce curseur est à la disposition de l'usager qui peut choisir entre des documents à caractère plus ou moins commerciaux. Clush permettait de privilégier des contenus contenant du texte ou des hyperliens grâce à un curseur. Ces deux initiatives ne sont plus en ligne, nous avons donc sélectionné des sites actuels avec

1. Voir section 4.1.1

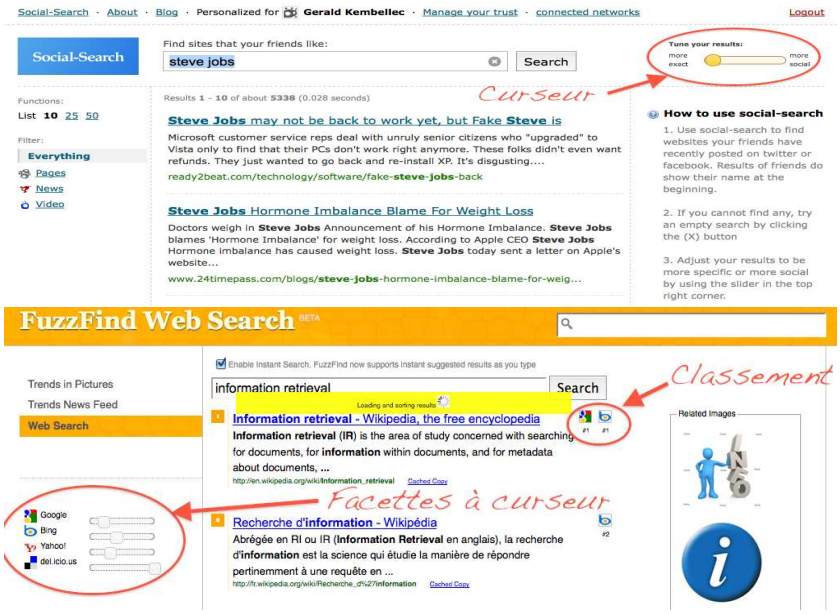

Figure 3.7: 2 exemples de moteur à curseur : Social Search et FuzzFind

des fonctionnalités toujours d'actualité. Les deux illustrations de moteurs à curseurs montrent comment atteindre un équilibre dans une recherche sous des aspects divers.

La première partie de l'illustration propose une recherche d'information avec une possibilité de choisir d'orienter sa recherche entre information traditionnelle et sociale sur les réseaux sociaux [1].

La deuxième moitié de l'illustration propose un modèle d'outil de recherche qui offre un accès aux informations de plusieurs moteurs (méta-moteur) [2]. L'originalité de cet outil est de pourvoir doser le crédit accordé à chaque moteur pour proposer un affichage personnalisé. Comme le montre la capture d'écran, chaque résultat est noté par rapport au classement qui lui est attribué sur les moteurs de recherches commerciaux. Ce classement est pondéré par rapport à la valeur attribuée à chaque moteur par le curseur. Cet outil, associé à une bonne connaissance des politiques de classement et des objectifs commerciaux des outils de recherches, offre un méta-moteur particulièrement efficace pour croiser le meilleur des algorithmes de recherche.

1. Social Search : `http://www.social-search.com`
2. Fuzz Find : `http://www.fuzzfind.com/v2/`

Moteurs	*Dir*	*Exalead*	*Google*	*MSN*	*Voilà*	*Yahoo*
Toutes positions	46,5%	34,5%	24,8%	31,2%	49,1%	21,7%
Première position	43,3%	29,7%	16,2%	29,0%	72,3%	17,9%

Tableau 3.3: Bruit généré par les moteurs de recherche, Véronis

Pour conclure sur les moteurs de recherche

Si l'on reprend l'étude française menée par Véronis (2006) pour comparer l'usage et les performances des moteurs de recherche en France, les résultats sont sans appel. Il est étonnant de croiser quelques chiffres issus de cette étude. [1] Premièrement, le fort taux de documents non pertinents retournés par le système (bruit [2]), quel que soit le moteur de recherche. La proportion de bruit généré est élevée puisqu'elle atteint pratiquement la moitié pour certains moteurs, et le cinquième pour Yahoo qui réalise la meilleure performance sur ce critère.

Deuxièmement, il est à noter le degré de satisfaction très médiocre des utilisateurs. Pour quantifier la pertinence perçue, Véronis a demandé de noter la pertinence des résultats retournés de 0 à 5, en fonction du rang occupé par la réponse lors de l'affichage (premier rang ou tous rangs confondus). Pour les meilleurs moteurs (Yahoo, Google), la note moyenne pour les dix premiers résultats affichés se hisse péniblement à 2,3 sur l'échelle suivante :

0. Entièrement insatisfait du résultat ;
1. pas satisfait du résultat ;
2. plutôt pas satisfait du résultat ;
3. généralement satisfait du résultat ;
4. satisfait du résultat ;
5. entièrement satisfait du résultat.

À l'heure actuelle, fin 2011, les principaux moteurs de recherche commerciaux sont l'incontournable Google, Yahoo et Bing (cf. Fig 3.8. [3]). MSN search est devenu Bing,

1. Pour information *Dir* est un moteur expérimental proposé par le groupe Iliad : `http://fr.dir.com/`
2. Voir section 4.1.1
3. Illustration issue du site web du spécialiste international en référencement Greenlight à l'URL : `http://www.greenlightsearch.com`

Moteurs	*Dir*	*Exalead*	*Google*	*MSN*	*Voilà*	*Yahoo*
Toutes positions	1,4	1,8	2,3	2	1,2	2,3
Première position	1,5	2,2	2,9	2,3	0,5	2,8

Tableau 3.4: Indice de pertinence perçue par moteur de recherche dans l'étude de Véronis (2006)

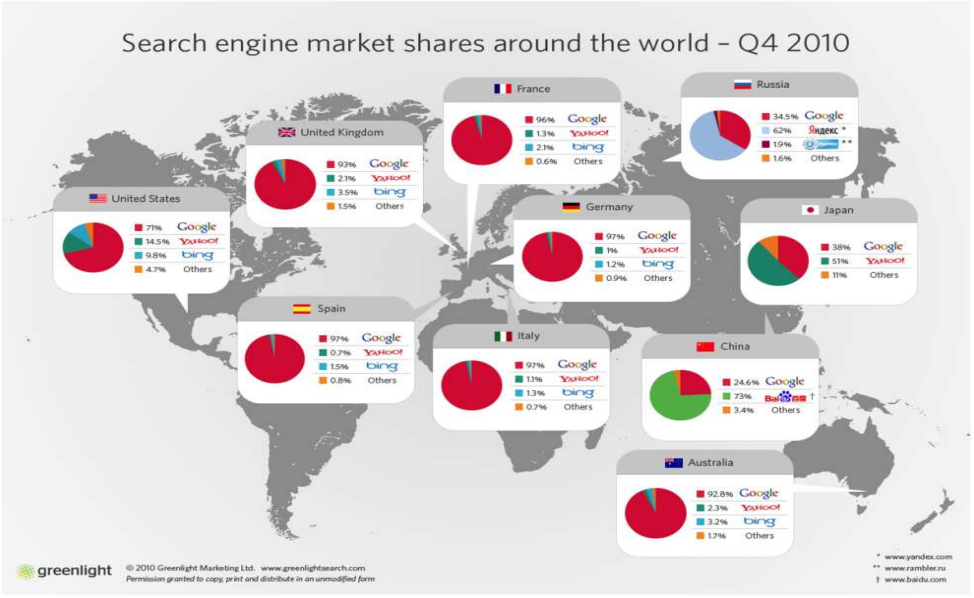

Figure 3.8: Parts de marché des moteurs de recherche commerciaux à travers le monde

mais depuis 2006, l'utilisation des moteurs de recherche n'a pas fondamentalement changé, et les statistiques d'utilisation restent comparables en terme de part de marché. Avec l'émergence de la Chine comme puissance économique, la donne va peut-être évoluer. Si l'on considère la forte implantation de moteurs locaux, il est possible d'imaginer une ouverture de ces moteurs vers l'étranger. Plus probablement, la Chine et ses centaines de millions d'usagers de l'Internet vont rejoindre massivement un moteur à forte connotation capitaliste, comme pour les biens de consommation classiques. Il faut cependant, pour cela, que l'accès à l'information s'assouplisse.

3.2.3 Les méta-moteurs

Un méta-moteur ou un méta-chercheur est une interface de recherche dont l'aspect est souvent identique à celui d'un moteur classique. La différence majeure entre un méta moteur et un moteur de recherche réside dans le contenu indexé. En effet, là où un moteur de recherche indexe des millions, voire des milliards de pages, un méta-moteur se contente d'interroger les moteurs. Le principe du méta-moteur est d'utiliser les résultats fournis par les moteurs classiques, d'en retraiter le contenu et d'en faire une présentation personnalisée, éventuellement épurée de tout aspect commercial. Pour instaurer un canal de communication avec un moteur classique, un méta-moteur utilise soit une API, soit un *wrapper* développé à cet effet.
Un *wrapper* est une « rustine » logicielle développée pour exploiter une application tierce lorsque l'on a une visibilité réduite sur son fonctionnement. Le plus souvent, une URL est « forgée » et encodée en HTML avec l'adresse d'un moteur de recherche, mais aussi les variables recherche et le contenu de la requête pour simuler une requête » manuelle »
Il est donc plus avantageux d'utiliser une , comme celles fournies par Google et Yahoo qui permettent de parser des arguments personnalisés à la requête, mais également d'avoir un flux de retour normalisé, le plus souvent en XML. De plus, cette option légalise le processus d'utilisation, ce qui s'accompagne malheureusement d'une publicité ciblée sur la requête. Lors d'une requête effectuée par un utilisateur, le processus de recherche se décompose ainsi :

1. analyse syntaxique de la requête (séparation des expressions booléennes et des termes de la recherche.
2. Liste de mots vides ou *Stop list* (l'étape de lemmatisation / racinisation est effectuée par les moteurs)

De manière concrète, le méta-moteur envoie ses requêtes à plusieurs moteurs de recherche, et retourne les résultats de chacun d'eux, en les classant, tout en éliminant les doublons. Il est souvent possible de paramétrer le méta-moteur de façon à sélectionner en amont ses moteurs favoris. De même, il est possible de choisir la manière dont les flux d'information seront traités et affichés.

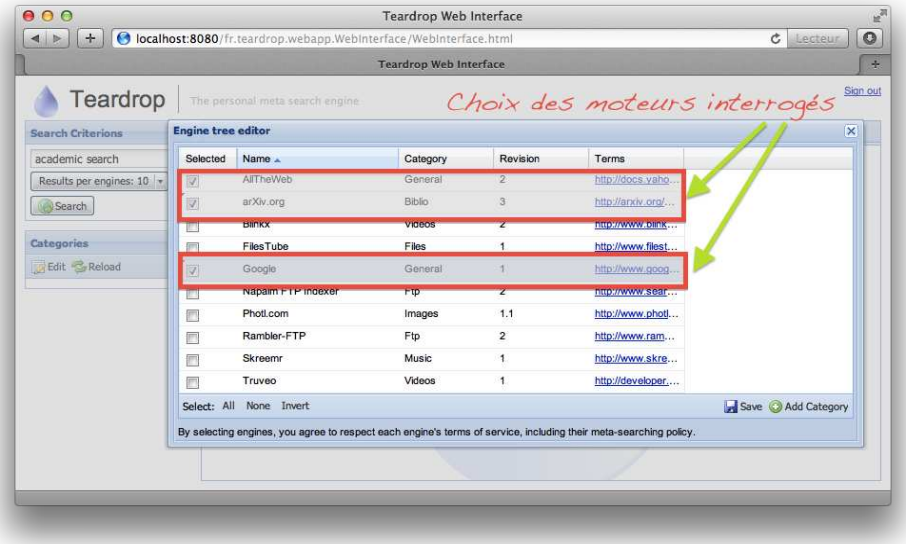

Figure 3.9: Paramétrage d'un méta-moteur, ici Teardrop

Voici quelques exemples de méta-moteurs :

1. Copernic agent, un logiciel pour Windows, technologie rachetée et utilisée par le site mamma.com [1]
2. Teardrop, un logiciel java pour toutes les plateformes (cf. figure 3.9) [2].
3. HooSeek [3], un méta-moteur solidaire (finance des associations).
4. Ixquick [4], un méta-moteur qui ne conserve pas les adresses IP des utilisateurs.
5. Seek [5], un méta-moteur francophone.
6. Seeks [6], un méta-moteur libre, sous licence GPL [7].

Le principe de méta-moteur offre un certain nombre d'avantages qui sont malheureusement contrebalancés par quelques inconvénients de tailles. Quand une requête est soumise au métamoteur, ce dernier interroge simultanément d'autres moteurs et

1. `http://www.copernic.com/fr/` et `http://www.mamma.com/`, accédés le 1er août 2012
2. `http://www.teardrop.fr/`, accédé le 1er août 2012
3. `http://www.hooseek.com/`, accédé le 1er août 2012
4. `https://www.ixquick.com/fra/`, accédé le 1er août 2012
5. `http://www.seek.fr/`, accédé le 1er août 2012
6. `http://www.seeks.fr/`, accédé le 1er août 2012
7. Gnu Public Licence : `http://www.gnu.org/licenses/agpl.html`, accédé le 1er août 2012

reformate les résultats par ordre de pertinence. Par exemple, comme les méta-moteurs ne possèdent pas leur base de connaissances propre, il n'est pas possible d'utiliser des technologies de suggestion (sauf à procéder à un enregistrement systématique des requêtes utilisateurs). Les requêtes à facettes ne sont également pas utilisables, faute de métadonnées indexées. Plus de réponses ne signifient pas plus d'informations exploitables, mais plus de données retournées. Le rappel est fortement impacté par le bruit généré en raison de l'abondance de résultats [1]. L'ordre d'affichage est le résultat d'une moyenne des classements des moteurs de recherche pour les résultats proposés. L'intérêt principal du méta-moteur est de réordonner les résultats fournis par les moteurs de recherche classiques en supprimant les doublons.

3.3 Concepts avancés de recherche d'information

3.3.1 Interfaces de références virtuelles

L'offre de références bibliographiques par le biais de l'outil informatique est un phénomène antérieur à l'avènement de l'Internet. Ce type d'offre est tout simplement l'extension électronique des offres postales ou téléphoniques. Ce service a suivi une évolution parallèle à la révolution de l'Internet des années 90. Nicolas Morin (2003) explique que le premier modèle recensé de ce type de service date du milieu des années 1980, avant même l'émergence du réseau Internet (Howard et Jankowski, 1986). Des échanges entre usagers et documentalistes avaient lieu par courriel. En 2002, des bibliothèques de toute la Floride ont mis au point un projet collaboratif de Service de Référence Virtuel, sobrement intitulé « *Ask a Librian* ». Ce projet permet de poser des questions directement à un bibliothécaire au travers d'outils tels que les formulaires ou les chats. La charte de ce projet se proposait de répondre sous 72 heures. Depuis la réussite de ce projet, de nombreuses institutions ont repris le concept, parmi lesquelles la bibliothèque du congrès ou l'Université de Cornel. Askal est devenu un logiciel à part entière, maintenu par l'université du Nebraska à Omaha [2]. Ce logiciel est utilisé, entres autres par le SCD de l'université d'Angers (voir figure 3.10 [3]). Même si le

1. Voir section 4.1.1
2. `http://library.unomaha.edu/askal` consulté le 22/12/2011
3. `http://bu.univ-angers.fr` consulté le 22/12/2011

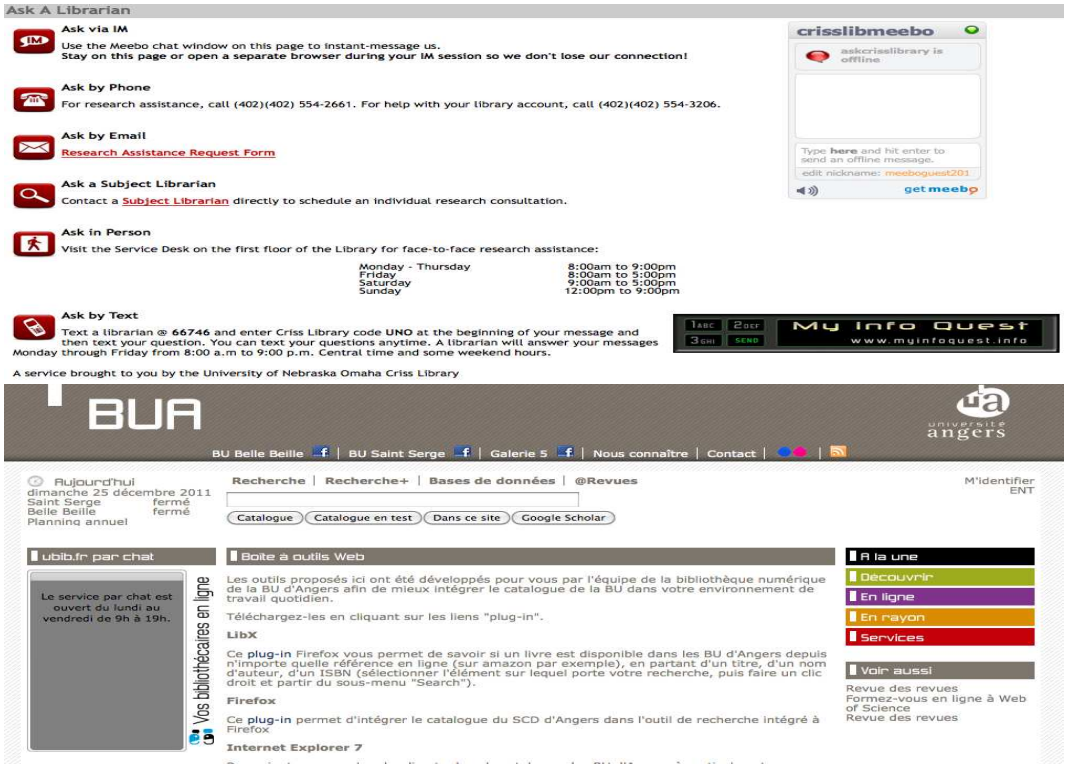

Figure 3.10: Modèle original de ASKAL et son adaptation à la BU d'Angers

projet « *Ask a Librian* » est initialement prévu pour des questions ouvertes, l'objectif principal est la constitution de bibliographies thématiques. Ainsi, des bibliographies « pointues » sont générées sur des sujets de connaissance bien définis et cadrés. Il s'agit de l'ancêtre de l'actuelle « Rue des facs [1] » ouvert début 2009 avec la charte du service d'information à la demande SI@DE [2]. Les documentalistes effectuent une curation sur les demandes et les transforment en requêtes compatibles avec les moteurs d'interrogation des bases de connaissances spécialisées. La recherche étant mutualisée sur des dizaines de SCD de grands établissements et d'universités d'Île-de-France, « Rue des Facs » redistribue les questions en fonction des spécialités de chaque établissement(Huyghe, 2010a). Des bibliographies sont générées par des spécialistes pour répondre aux besoins informationnels. Dans le cadre du projet « Le guichet du Savoir », proposé par la Bibliothèque municipale de Lyon, Calenge et di Pietro (2005) Ce service offre également des méthodes conviviales d'accès à l'information comme un nuage de mots clés (tag

1. `http://www.ruedesfacs.fr/` consulté le 22/12/11
2. SI@DE (Services d'information @ la demande) : `http://www.bnf.fr/fr/collections_et_services/poser_une_question_a_bibliothecaire/s.charte_siade.html` consulté le 22/12/11

cloud), « mur » Facebook ou flux RSS. En 2005, après un an d'existence, un premier bilan a été dressé. 4800 questions avaient été traitées et la page questions/réponses a été consultée 452 000 fois. Au 26 décembre 2011, 39 976 questions ont été traitées, ce qui montre une augmentation du traitement annuel des questions, si ce n'est de la demande. L'ENSSIB propose depuis 2009 le projet « Q ? R ! », pour « Question ? Réponse ! ». Cette interface est inspirée du projet « question point » de l'OCLC. Elle permet aux usagers de poser des questions à une équipe dédiée de spécialistes. Une modératrice réceptionne les questions, effectue des corrections orthographiques ou grammaticales et un classement des questions. La répartition effectuée, les analystes de l'équipe sont interrogés en fonction de leur spécialité. Catherine Jackson qui gère « Q ? R ! » note comme avantage certain de ce système le fait de faire émerger les besoins de formation des analystes. Les utilisateurs ont également l'occasion – unique – d'accéder à des réponses issues de la collection « papier » des documents primaires de la bibliothèque, ces collections étant généralement délaissées au profit des recherches sur le web (Jackson, 2009). La dernière offre française que nous décrirons est issue d'une collaboration entre l'université numérique Paris Ile-de-France (UNPIdF), la mairie de Paris et les SCD des universités de la région parisienne. Il s'agit de « Rue des facs », service ouvert début 2009 avec la charte SI@DE (déjà mise en œuvre par SINDBAD). Les documentalistes effectuent une curation sur les demandes et les transforment en requêtes compatibles avec les moteurs d'interrogation des bases de connaissances spécialisées. La recherche étant fédérée sur des dizaines de SCD de grands établissements et d'universités d'Île-de-France, Rue des Facs redistribue les questions en fonction des spécialités de chaque établissement (Huyghe, 2010b). Des conseils de recherche et des références bibliographiques sont proposés par des spécialistes pour répondre aux besoins informationnels spécifiques (cf. Figure [16]. Capture d'une réponse sur « Rue des Facs ») Discussion autour des services de références virtuels (SRV) Face à la pléthore de sources et de méthodes d'accès direct aux sources numériques de documentation, est-il encore utile d'introduire une étape de médiation documentaire ? Cette question est légitime, surtout si l'ensemble des usagers est autonome relativement à des outils parfois complexes, aux méthodes d'interrogation pointues. Comme le rappelait C. Nguyen dans son article sur les services de renseignements virtuels, les « sites web des bibliothèques (. . .) donnent aux étudiants un accès immédiat aux collections » (Nguyen, 2006).

SRV comme une béquille documentaire ?

Cependant, tout le monde n'a pas bénéficié d'initiation à la recherche documentaire. Jean Bouyssou, de Rue des Facs, s'est précisément posé la question de la maîtrise du panel d'outils offerts par les SCD par les étudiants. Sont-ils à l'aise avec l'interface d'interrogation et l'analyse des réponses ? Il semble que dans les usages classiques des OPAC 65 % des étudiants ne consultent que les résultats de la première page de résultats. Les 33% restant poussent leur investigation documentaire jusqu'à la deuxième page de résultats (de Saxcé, 2010). Il s'agit donc d'une recherche que l'on pourrait qualifier de « surface ». Dans ce cas, une médiation de la part d'un professionnel de la recherche documentaire est un atout majeur. Le documentaliste va interroger les bonnes sources et apprécier les meilleures références, et pas simplement les premières. Cela est d'autant plus vrai, que dans le cadre de « Rue des Facs », les documentalistes sont souvent des spécialistes du sujet traité. La documentation électronique offerte sur les sites des SCD, notamment les abonnements aux revues scientifiques, souvent anglophone, a pour public cible les doctorants et les chercheurs (de Saxcé, 2006). Les étudiants de licence et MASTER, pourtant majoritaires en université sont peu intéressés par cette documentation [ibid]. De plus, ils ne possèdent pas encore tous la connaissance des bonnes pratiques documentaires. Les services de type « questions et réponses » pourraient être particulièrement adaptés à cette population.

La dimension pédagogique et humaine des SRV

Cependant, une question posée par Agosto est la dimension pédagogique, à savoir si le documentaliste doit offrir un accompagnement dans la méthodologie de recherche ou effectuer lui-même la recherche et en offrir le fruit (Agosto *et al.*, 2011). Elle oppose deux visions du rôle du documentaliste au sein des systèmes de renseignement virtuel liés à l'enseignement. La première, dite « libérale », défend le point de vue de l'absence de responsabilité éducative dans le contexte du SRV. Le documentaliste doit se concentrer sur la recherche d'information pertinente pour répondre à la demande. L'autre vision du système, qualifiée de « conservatrice », soutient la thèse que même dans un système de SRV, la dimension éducative reste prioritaire. Sur cette question, dans un cadre universitaire, nous pensons comme Fritch et Mandernack (2001) et comme Galvin (2005) qu'au cours d'une session de référence virtuelle, le documentaliste peut profiter du

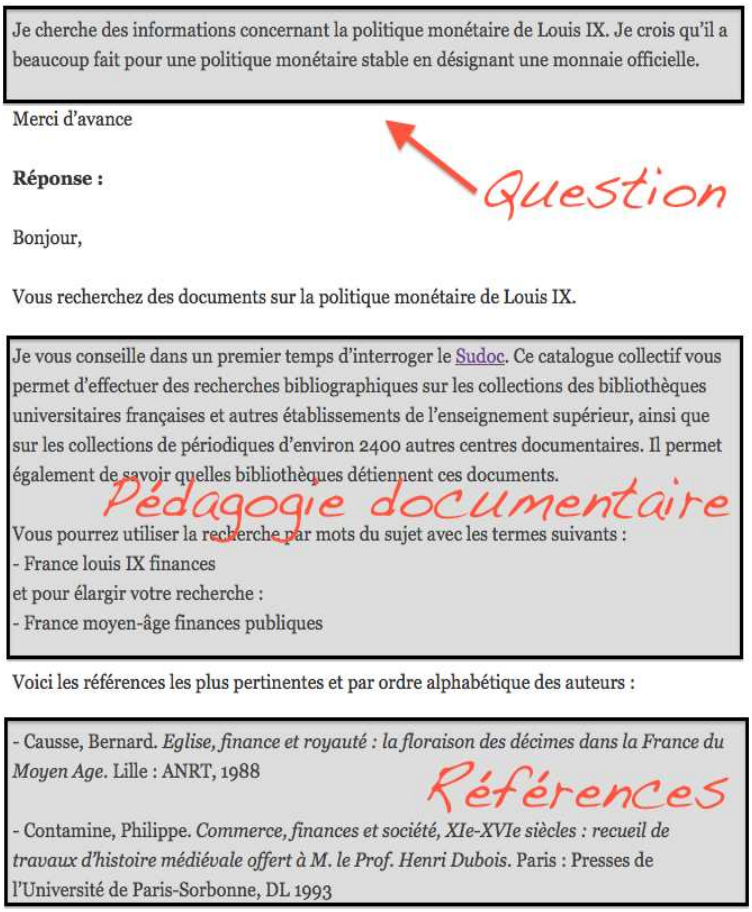

Figure 3.11: Capture d'une réponse sur « Rue des Facs ».

contact pour essayer de transmettre des compétences en recherche documentaire. Ainsi le choix entre offre de service et pédagogie n'a pas forcément lieu d'être. C'est ce que montre l'exemple tiré de rue des Facs (cf. 3.11. Capture d'une réponse sur « Rue des Facs »). Cette opinion est également partagée par Claire Nguyen qui déclare que dans ce cadre les médiations peuvent être « collectionner, sélectionner, (ré)orienter, et proposer les documents ; (...) informer et former » (Nguyen, 2012). Nous pensons également que la relation personnelle établie dans le cas d'une prescription au travers d'un SRV permet d'adapter la réponse de mieux profiler la sélection de références en fonction du niveau du demandant [ibid.]. On ne répondra pas de la même façon à un doctorant qu'à un étudiant en première année de licence.

Légitimité des résultats offerts par les SRV académiques

L'autre question posée par les systèmes de références virtuels est celle du rôle du documentaliste comme prescripteur. Nous pouvons nous interroger sur le bien-fondé du positionnement du documentaliste comme sélectionneur, évaluateur et prescripteur de documentation. Cette question, dans le cadre de bibliothèques spécialisées, avec des documentalistes formés dans le domaine du champ disciplinaire de l'établissement, rejoint la précédente problématique. La légitimité existe, mais la question pédagogique demeure, il faut apprendre à sélectionner les sources, un accompagnement difficile à effectuer à distance, surtout lors d'une session asynchrone (Tyckoson, 2003). La réponse à cette dernière objection peut trouver réponse dans les méthodes de navigation partagée (*co-browsing*), au cours de laquelle le documentaliste pourra effectuer une présélection des documents et expliquer ses choix (Nguyen, 2012) .

Conclusion sur les SRV

Nous avons également discuté sur l'éthique et la pédagogie associées aux systèmes de références virtuels. Si les vecteurs de communication sont multiples, le principe reste le même. Il s'agit de poser une question précise à un service de documentation, qui tentera d'y répondre dans un temps imparti, avec le plus souvent un conseil en méthodologie de recherche associé au contexte de recherche. Même si l'on peut discuter de l'intérêt ou de la validité pédagogique d'un tel service, il n'en est pas moins que les statistiques d'utilisation indiquent l'engouement du public. Par ailleurs, ces services en milieu universitaire semblent être un bon accompagnement méthodologique pour les étudiants de premiers cycles. De plus, un chercheur peut également apprécier l'aide d'une bibliographie « cousue main » par un documentaliste spécialiste d'un domaine de recherche dans lequel il s'aventure au cours d'une recherche pluridisciplinaire.

3.3.2 Introduction aux systèmes de recommandation

Traditionnellement, les individus avaient l'habitude de se recommander des produits ou services par le « bouche à oreille ». Aujourd'hui, l'offre (que ce soit d'informations ou de produits) augmente de jour en jour. Elle est proposée principalement à travers le vecteur d'Internet. Au-delà d'un seuil, plus d'informations conduisent à dégrader la qualité de l'information (Chen *et al.*, 2009). Le développement de systèmes automatisés

de recommandations (*Recommend System* ou RS) était donc un phénomène inéluctable dans l'optique de trouver des informations de qualité provenant de sources hétérogènes. Dès 2000, Burke faisait remarquer que de nombreux sites commerciaux comme Amazon ou encore e-Bay avaient compris l'intérêt (commercial) de contextualiser des offres d'hyperliens périphériques à l'hypertexte consulté par l'usager (Burke, 2000). Les moteurs de recherche commerciaux ont même créé des produits dérivés comme le « *Google AdSense* » afin d'optimiser leurs profits publicitaires en exploitant le RS. Le principe est simplement de proposer à des annonceurs privés de fournir des hyperliens vers leur site en marge des recherches des usagers. Le gros avantage de ce type de publicité est qu'elle est forcément ciblée autour des centres d'intérêt de l'usager. Les utilisateurs du portail messagerie en ligne Gmail auront forcément remarqué que les messages publicitaires en marge de leur outil est toujours en relation direct avec le contenu de leur courriel ouvert. La FAQ[1] officielle de Google :

« Notre objectif est de proposer aux utilisateurs de Gmail des annonces utiles qui correspondent à leurs centres d'intérêt (...) Le système que nous avons développé pour les annonces est similaire : en utilisant certains des critères qui permettent d'identifier les messages potentiellement importants à vos yeux, Gmail est en mesure de déterminer quelles sont les annonces susceptibles de présenter un intérêt pour vous. Par exemple, si vous avez récemment reçu un grand nombre de messages sur la photographie ou les appareils photo, il se peut que les offres d'un magasin d'appareils photo proche de chez vous et qui vous intéressent En revanche, si vous avez signalé ces messages comme étant du spam, vous ne souhaitez probablement pas les voir s'afficher[2]. »

Certains webmestres offrent contre rémunération des encarts publicitaires vides qui sont dynamiquement remplis par Google en fonction du contenu global de la page. Le but d'un système de recommandations est de réduire la surcharge d'informations(Herlocker *et al.*, 1999) en sélectionnant un sous-ensemble des éléments d'un ensemble universel basé sur les préférences des utilisateurs. Dans le domaine scientifique, nous assistons à une croissance rapide et continue des contenus des bibliothèques numériques. Cette constatation amène les utilisateurs à rechercher des outils et des services qui ne sont pas seulement adaptés à leurs besoins spécifiques de recherche, mais également à la

1. Foire aux questions, sorte de mode d'emploi sous forme de questions/réponses

2. Accédé en ligne le 17 Septembre 2011 sur la foire aux questions officielle de Google à l'URL suivante : `http://mail.google.com/support/bin/answer.py?hl=fr&ctx=mail&answer=6603`

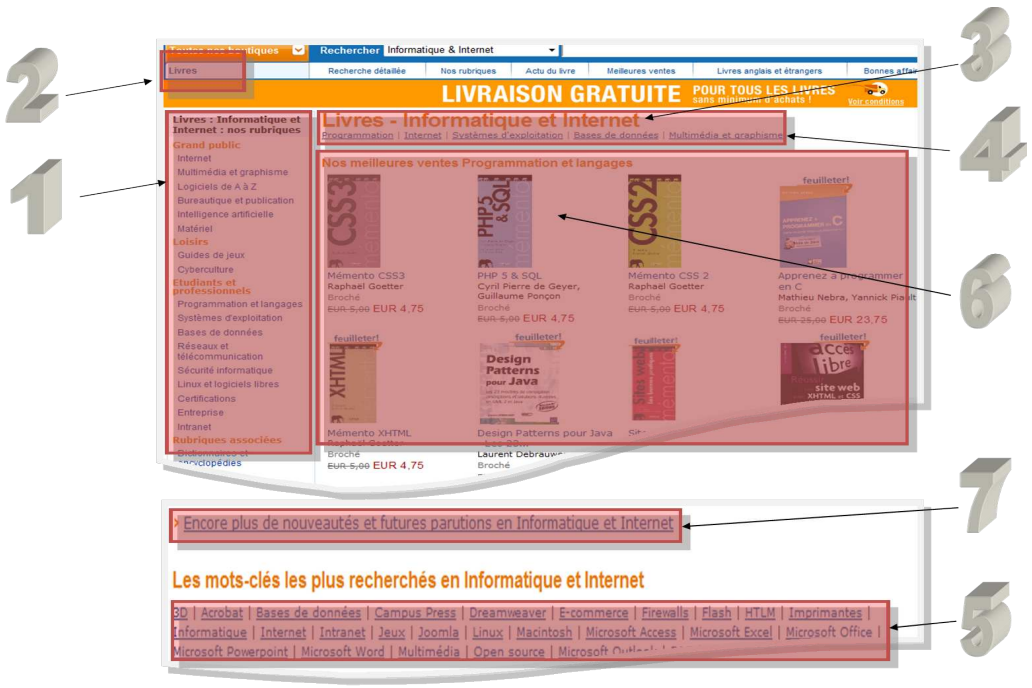

Figure 3.12: Concept de QBE sur le site commercial Amazon.

veille technologique de leur domaine (Kapoor *et al.*, 2007).

3.3.3 Le concept de requête par l'exemple (QBE)

Pour expliciter le modèle conceptuel populaire de *Query By Exemple* (Zloof, 1977), prenons une démonstration générique avec un des portails commerciaux de type « deep web » ou Internet profond [1]. Des portails, comme celui du libraire Amazon [2], peuvent se prévaloir du concept QBE pour un domaine de connaissances. La surcharge d'information déroute l'utilisateur du système d'information (ici le consommateur). Les résultats de l'étude de Chen *et al.* (2009) indiquent que la surabondance d'informations engendre une perception de surcharge d'informations ce qui conduit les consommateurs à ne pas acheter. Accablées par ce manque à gagner, les enseignes de vente en ligne ont dû contre attaquer et mettre au point des stratégies pour accompagner et assister le client dans sa quête de produit. Le site Amazon construit ses recommandations uniquement à l'aide

1. Cette notion est explicitée dans la section 2.1.1
2. urlwww.amazon.com, accédé le 1er août 2012

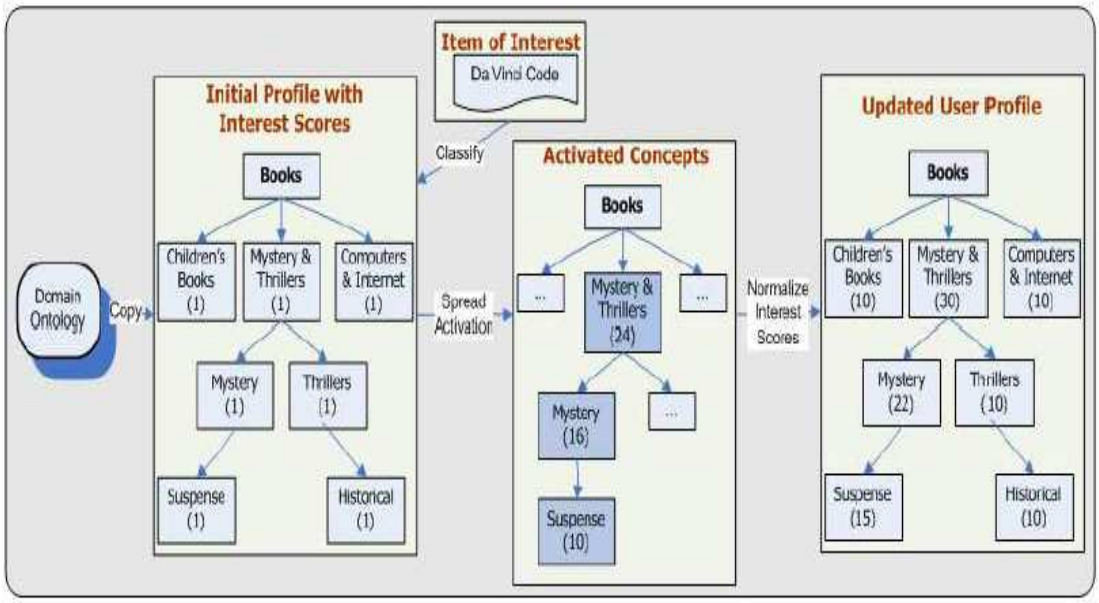

Figure 3.13: Exemple d'ontologie de la RS sur un site marchant Sieg *et al.* (2010)

des informations apprises sur ses propres clients par l'observation de leurs usages. Ainsi, sur le site d'Amazon, en sélectionnant « Livres » puis « Informatique et Internet » des requêtes sont automatiquement générées pour :

1. Parcourir une ontologie du domaine Amazon ;
2. Sélectionner le type « livre » de support de diffusion ;
3. Sélectionner la catégorie de connaissance « informatique et Internet » ;
4. Proposer les sous catégories de connaissance « Programmation, Internet, Systèmes d'exploitation, Bases de données, Multimédia et graphisme » ;
5. Mettre en exergue les termes clés les plus représentatifs de la catégorie ;
6. Calculer les catégories les plus consultées, et proposer les livres les plus populaires de ces catégories.
7. Proposer un hyperlien pour l'affichage des nouveautés de la catégorie

Ces différentes requêtes sont automatiquement générées par une navigation initiée par la proposition d'un contexte sémantique formel strict basé sur une ontologie de domaine

de la vente en ligne (cf. illustration 3.13), incluant des taxonomies des sujets proposés et thésaurisant des vocabulaires contrôlés pour chacun. Ensuite, l'utilisateur choisit un livre parmi ceux proposés. Ce choix va servir d'exemple au système et permettre de recréer une requête plus fine en partant du postulat que l'usager est intéressé par les « objets » dont les caractéristiques, dans la base relationnelle, sont proches de celles de l'objet déjà sélectionné. Cette navigation s'effectue sans que l'utilisateur n'ait eu à user d'opérateurs booléens, de mots clés et encore moins d'un quelconque langage d'interrogation de base de connaissances.

Cet exemple trivial [1] nous amène au travail plus scientifique l'équipe de Petroupolos dans le projet Clide (Petropoulos *et al.*, 2007a,b). Petropoulos adopte un contexte d'interaction visuelle dans l'optique de permettre aux utilisateurs d'amener le système à formuler des requêtes comme suite à une « navigation » graphique. Dans ce contexte, l'objet est l'interrogation d'une base de connaissances relative à l'état d'un système d'informations (ordinateurs et matériels actifs). Après le choix par navigation, l'utilisateur se voit également proposer des éléments approximativement similaires suivant leur état ou les caractéristiques techniques. Ainsi, dans l'exemple exposé dans la figure 3.14, après une sélection des tables *Com1* et *Net1* (respectivement *Computers* et *NetInterface*), l'utilisateur sélectionne le type de processeur (CPU) Pentium4 et la vitesse de transmission 54 MB. Puis l'usager coche les champs la mémoire vive (RAM), le prix et le type d'interface. Une requête est immédiatement générée pour afficher le prix et la quantité de mémoire des machines dont le processeur est un Pentium 4 et dont le débit maximal de l'interface réseau est à 54 MB. Comme indiqué en haut à droit par un indicateur vert, la requête aboutit. Selon les auteurs, la principale motivation de l'architecture Clide est de déterminer quelles requêtes donneraient des résultats par rapport à celles qui produisent des résultats vides ou une erreur. L'originalité de ce travail est de montrer les dessous du mécanisme en affichant les requêtes générées et les tables impactées. Malgré l'intérêt réel d'assister l'usager dans sa demande d'informations, il est indéniable que la représentation des tables de la base requiert une connaissance du SQL, à tout le moins de la modélisation Merise.

1. Cet exemple est une adaptation francophone du concept proposé par Sieg *et al.* (2010).

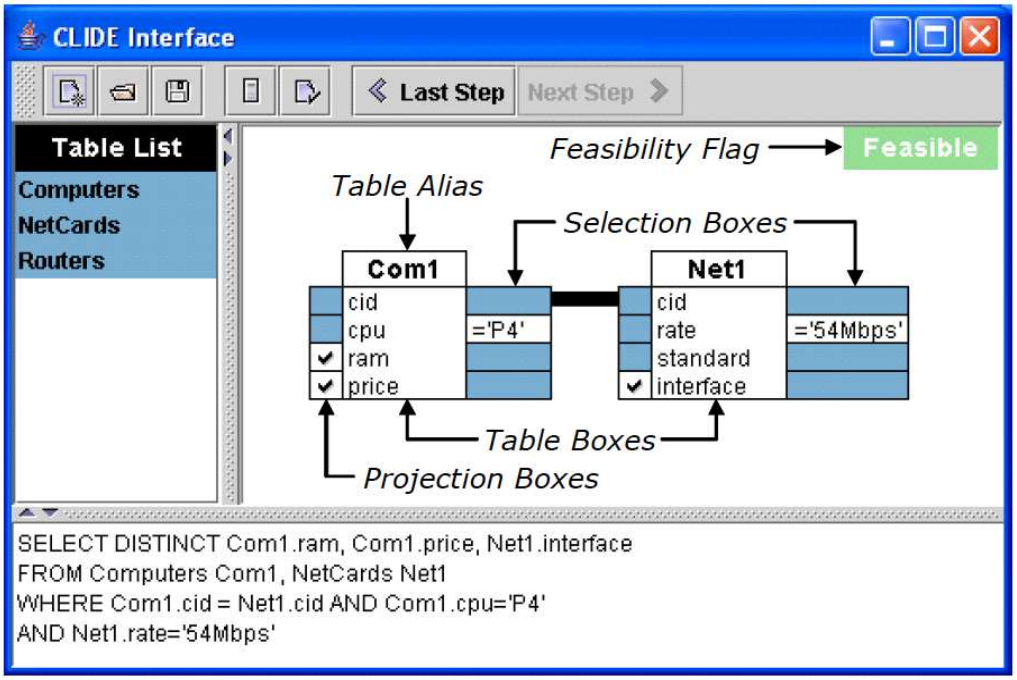

Figure 3.14: Exemple d'utilisation du système Clide.

3.3.4 Services rendus par les systèmes de recommandation

Les catalogues de bibliothèque n'exploitent pas encore systématiquement les fonctionnalités de recommandation. Pourtant, les utilisateurs semblent réclamer ce type de services, y compris pour la recherche de documents à travers un catalogue de bibliothèque. Cette étude se focalise sur 2 types de services de recommandation : « Les lecteurs de cet ouvrage ont aussi emprunté tel autre ouvrage » et « Plus de résultats comme celui-ci ».

3.3.5 Méthodologies des systèmes de recommandation

Dans cette partie nous introduirons les systèmes de recommandation d'un point de vue méthodologique. Les points abordés seront principalement la recommandation sociale , celle basée sur le contenu et les systèmes mixtes.

Votes	Utilisateur A	Utilisateur B	Utilisateur C	Utilisateur D	Utilisateur E
Objet 1	9	10	7	10	9
Objet 2	7	6	2	1	1
Objet 3	5	1	5	5	4
Objet 4	3	5	3	2	2
Objet 5	1	3	5	7	6

Tableau 3.5: Exemple d'échantillon de votes

Filtrage collaboratif

Une fois un document proposé par le système à l'usager, ce dernier se voit proposer la possibilité de lui attribuer une valeur. Cette note peut donner une appréciation intrinsèque pour le document, ou juger de l'adéquation de ce document avec le contexte de recherche. Cette note sera conservée dans le système pour être réutilisée. Selon la méthode de filtrage collaboratif dite « *Memory based* », ou heuristique, les notes peuvent aider à prédire l'appréciation d'un usager α sur un objet en se basant sur celle d'un autre utilisateur β ayant voté régulièrement de manière similaire. Pour déterminer quel sera l'utilisateur β le plus similaire à α, la corrélation de Pearson peut être utilisée (Resnick *et al.*, 1994). Cette méthode est également nommée « *Word of Mouth* » ou bouche à oreille (Shardanand et Maes, 1995) ou « *People-to-people correlation* » (Schafer *et al.*, 1999).

$$r = \frac{\sum (\alpha - |\alpha|)(\beta - |\beta|)}{\sqrt{\sum (\alpha - |\alpha|)^2 \sum (\beta - |\beta|)^2}} \tag{3.1}$$

Exemple de calcul de proximité entre usagers ayant voté pour un ensemble d'objets :

Le tableau ci-dessus donne les votes des utilisateurs pour les objets. Les corrélations calculées deux à deux donnent les résultats suivants :

Cela signifie que dans l'exemple chaque utilisateur pourra bénéficier des appréciations d'au moins un autre usager au profil similaire au sien (corrélation tendant vers 1). Une fois la base de vote utilisateur abondamment pourvue, elle peut être utilisée pour offrir une méthode de prédiction plus fine dite « *Model based* » basée sur des profils d'usagers (Breese *et al.*, 1998). Dans cette seconde méthode, les profils types sont établis à partir

Corrélation	Utilisateur A	Utilisateur B	Utilisateur C	Utilisateur D	Utilisateur E
Utilisateur A	X	0,699	0,243	0,215	0,246
Utilisateur B	0,699	X	0,265	0,341	0,413
Utilisateur C	0,243	0,265	X	0,977	0,669
Utilisateur D	0,215	0,341	0,977	X	0,996
Utilisateur E	0,246	0,413	0,669	0,996	X

Tableau 3.6: Proximité des usagers basée sur la corrélation de Pearson

de regroupement de ceux qui ont effectué des notations similaires. Ce sont les profils types ou modèles qui seront utilisés pour prodiguer des recommandations.

Avantages et inconvénients du filtrage collaboratif

Le tout premier avantage de la recommandation basée sur le filtrage collaboratif est que la connaissance du domaine de connaissance n'est pas un pré requis à la recherche d'information (Burke, 2002). Ce système permet également d'élargir la recommandation à des sujets transverses au domaine initial de connaissance en utilisant les autres centres d'intérêt des profils similaires. Cette sérendipité provoquée est appelée par Burke « *cross-genre niches* » (Burke, 2002). Selon Poirier *et al.* (2010), grâce à son indépendance vis-à-vis de la représentation des données, cette technique peut s'appliquer dans les contextes où l'analyse du contenu est difficile à automatiser. Nous rajoutons que pour des documents de type image, audio et vidéo les métadonnées ne sont pas systématiquement renseignées. Dans ce cadre, en dehors du filtrage collaboratif (ou d'un important travail de *crowdsourcing* descriptif préalable), il n'y aurait pas de méthode alternative de recommandation. Le dernier point positif est que la qualité de la recommandation proposée par filtrage collaboratif croit avec l'utilisation du système.

Claypool et al ont pointé un certain nombre de problèmes issus des méthodes initiales de recommandation (Claypool *et al.*, 1999). Par exemple, à l'état initial, le système de recommandation basé sur le filtrage collaboratif est inutilisable pour cause de « *cold start* ». Ce problème de démarrage à froid s'exprime de la manière suivante : sans note, pas de recommandation possible. Cette difficulté est reproduite lors de l'ajout d'items ou d'usagers. Avec un nombre trop faible d'évaluations pour un corpus vaste,

les données sont trop éparses pour établir des corrélations suffisantes. Ce phénomène est appelé « *sparsity* », ou éparpillement (Claypool *et al.*, 1999). Il arrive, dans le cadre d'une tentative de classification des individus, qu'un élément soit à la frontière de plusieurs groupes. Par exemple, dans un système de mise à disposition et de notation de littérature scientifique, il peut arriver qu'un usager α soit autant intéressé par la science $S1$ que par la science $S2$. Malheureusement, les recommandations de sur les documents relatifs à ces deux sciences sont notés par des individus appartenant à deux groupes distincts. Comme α vote de manière atypique, il ne se verra intégré dans aucun des deux groupes. Ce phénomène connu sous le nom de « *grey sheep* », ou mouton gris (Burke, 2002, Claypool *et al.*, 1999). Si d'autres personnes adoptent son comportement, ils formeront un groupe à part qui produira de la recommandation pour la nouvelle « *cross-genre niche* ». Il est également indéniable que le principe de popularité sera privilégié par le filtrage collaboratif. Plus un objet sera noté positivement, plus il sera recommandé et donc sera de nouveau évalué. Ce principe de notoriété auto engendrée sera peut-être plus dû à l'ancienneté qu'à la qualité réellement perçue par les usagers. Ce problème peut être contrebalancé ou au contraire intensifié par une faille du système de recommandation sociale : la fraude au vote avec des identités multiples. Il peut être tentant de modifier les recommandations dans une optique marchande en votant sous plusieurs identités. Cette technique est appelée « *shilling* » et fait l'objet de nombreuses études (Lam et Riedl, 2004, Williams *et al.*, 2006).

L'indexation et la catégorisation

L'autre méthode traditionnelle de filtrage est basée sur la description et l'analyse des contenus proposés par le système. Ce procédé est principalement basé sur des techniques d'analyse textuelle, mais peut être étendu à des contenus divers contenant des métadonnées. Les photos illustrent le propos, composées d'un contenu binaire, elles offrent la possibilité de contenir des métadonnées auto générées comme les coordonnées géographiques, l'exposition, l'orientation ou la date grâce au format EXIF (*Exchangeable Image File Format*). La technique de recommandation sur la base du contenu se fonde sur la relation entre le profil de l'usager et les métadonnées associées aux objets stockés dans la base de connaissances (Boutell et Luo, 2004, Lee *et al.*, 2006) . L'utilisateur peut entrer volontairement ses préférences lors de son inscription au service, elles seront

dites « fournies ». L'autre possibilité est de calculer les préférences par l'observation de son comportement Adomavicius et Tuzhilin (2005), dans ce cas elles seront « calculées » et vectorisées. Les préférences de l'utilisateur sont représentées sous forme d'un vecteur contenant les termes les plus représentatifs des goûts de l'usager. Ces termes clés peuvent avoir une valeur déterminée statistiquement en fonction de leur fréquence dans les documents consultés et/ou notés par l'usager au sein du corpus (Balabanović et Shoham, 1997). Par exemple, il est possible d'utiliser l'algorithme $tf.idf$ pour pondérer termes clés issus de textes (Salton et Waldstein, 1978).

$$tf_{(m,d)} = \frac{n}{card(d)} \tag{3.2}$$

Exemple de calcul de *term frequency*

> Considérons un document d contenant 100 mots dans lequel le terme m apparaît n fois avec $n = 3$. La fréquence du terme (tf) pour m au sein du document d est alors le quotient entre le nombre d'occurrences de n du mot m dans le document d et le nombre total de mots dans d. Ce qui appliqué à l'exemple donne $\frac{3}{100}$.

L'inverse de la fréquence de documents (Jones, 1972) est calculée ainsi par le logarithme du quotient entre le cardinal de l'ensemble du corpus C et le cardinal du sous corpus C' des documents de C qui contiennent le terme m. Nous ajoutons 1 au dénominateur pour généraliser la fonction au cas de l'absence du terme dans le corpus.

$$idf_m = log(\frac{card(C)}{1 + C'_{m,C}}) \tag{3.3}$$

Exemple de calcul de *inverse definition frequency*

> Maintenant, supposons que nous avons 10 millions de documents dans le corpus C et que le terme m apparaît dans un millier de ceux-ci. Appliqué à notre exemple le résultat de idf est $log(10000000/1000)$ soit 4.

Finalement, le poids pondéré d'un terme dans un document par rapport à un corpus s'obtient en multipliant les deux mesures tf et idf

$$tf.idf_{m(C',C)} = \frac{n}{card(d)}.log(\frac{card(C)}{1 + C'_{m,C}}) \tag{3.4}$$

Exemple de calcul de *term frequency . inverse definition frequency*

> La valeur $tf.idf$ dans notre exemple précédent est le produit de ces quantités : 0,03 x 4 = 0,12. Ainsi le terme m sera statistiquement pondéré avec un coefficient de 0,12 dans le document d du corpus C.

Cet algorithme basique est rarement utilisé seul, remplacé par des générations plus récentes et sophistiquées de combinaisons, comme Terrier (Ounis *et al.*, 2005), avec notamment okapi BM25, mais reste le fondement de la pondération de termes représentatifs de documents dans des corpus textuels. Les méthodes basées sur la vectorisation de requêtes montrent des résultats prometteurs. Berry et al suggèrent la récupération de la requête sous forme matricielle par l'algorithme populaire LSI ou indexation sémantique latente. Dans essence, l'algorithme crée un espace vectoriel de dimensions réduites qui offre une représentation à n dimensions d'un ensemble de documents (Dumais *et al.*, 1988). Quand une requête est entrée, sa représentation numérique est comparée avec les cosinus d'autres documents de la base, et l'algorithme retourne les documents pour lesquels la distance est la plus faible. Cette méthode peut être adaptée pour recommander des documents en fonction des besoins des usagers. Dans le cas de données non textuelles, il est cependant possible de mettre en exergue et d'évaluer statistiquement les centres d'intérêt de l'usager. Il est ainsi envisageable de faire des évaluations statistiques sur les coordonnées géographiques des photos préférées ainsi que sur leur orientation (portrait ou paysage) ou encore type d'équipement utilisé. Ces données seront à croiser avec les préférences utilisateurs, qu'elles soient renseignées par l'utilisateur ou calculées sur les statistiques d'utilisation de l'usager.

Avantages et inconvénients du filtrage sur le contenu

Les avantages du filtrage sur le contenu sont similaires à ceux observés par le filtrage collaboratif (Burke, 2000). Ainsi, la connaissance du domaine n'est pas obligatoire pour l'usager, car les recommandations seront issues des données du corpus. La finesse des recommandations système évoluera également avec la taille du corpus.

Cependant, le système basé sur les seules données du corpus ne pourra pas proposer de « sérendipité » faute de corrélation avec les usagers. De plus, comme le fait remarquer Poirier, chaque utilisateur est absolument indépendant des autres. Ainsi, un usager qui aura correctement rempli son profil avec ses thématiques de prédilection recevra des propositions même s'il est le seul inscrit (Poirier *et al.*, 2010).

L'inconvénient majeur d'un moteur de recommandation orienté données sera, dans un premier temps, comme pour le type collaboratif le problème posé par le nouvel utilisateur qui n'a pas encore de profil établi et donc pas de données de référence « observées ». Ensuite, il est évidemment plus complexe d'indexer des données non textuelles. De plus, l'usager sera « sclérosé » dans un contexte de recherche, celui qu'il a déjà positionné comme étant son centre d'intérêt. Ce problème est identifié comme étant l' « *overspecialization* », ce qui annihile toute possibilité de sérendipité par proposition de sujets connexes.

Les méthodes hybrides de recommandation

De manière triviale, l'hybridation de systèmes de recommandation résulte de la combinaison de méthode de filtrage collaboratif avec celle basée sur le contenu. Cette vision de l'hybridation a été affinée par Burke puis par Adomavicius et Tuzhilin (Adomavicius et Tuzhilin, 2005, Burke, 2002).

Burke recense les sept techniques d'hybridation suivantes (Burke, 2002) :

1. *Weighted* / Pondération : La valeur de recommandation d'un item est issue de la somme des méthodes disponibles. Par exemple *P-Tango* Claypool *et al.* (1999) donne une valeur égale au filtrage collaboratif et à aux prédictions basées sur le contenu. Cette valeur est ensuite pondérée par une confirmation des usagers.
2. *Switching* / Bascule : Le système choisit d'appliquer soit une méthode orientée données, soit un filtrage social selon le contexte de recherche de l'utilisateur.
3. *Mixed* / Mixte : Cette technologie permet de proposer des recommandations provenant des méthodes traditionnelles dans l'optique de limiter les inconvénients de chaque méthode classique.

4. *Features combination* / Combinaison : Cette méthode offre la possibilité d'enrichir les données intégrées a priori dans le système avec les votes des utilisateurs, qui enrichissent la base a posteriori. Le calcul de recommandation se fait sur l'ensemble des données.
5. *Cascade* / Cascade : Ce procédé consiste à une double analyse des profils utilisateurs. La première passe sert à faire émerger des candidats, Le deuxième à affiner la sélection d'utilisateurs.
6. *Features augmentation* / augmentation : Il s'agit d'une technique similaire à la précédente pour le premier passage. Si le nombre de candidats est trop élevé au premier passage, alors un deuxième fera une discrimination supplémentaire en intégrant les données des objets recommandés.
7. *Meta level* / niveau modèle : Comme pour les deux dernières méthodes, il s'agit de passer filtrer deux fois les usagers pour déterminer des similarités. La différence est que le premier passage permet de générer un modèle ou profil type d'utilisateur.

Adomavicius et Tuzhilin proposent une classification des méthodes hybrides de recommandation reposant sur quatre axes (Adomavicius et Tuzhilin, 2005) :

1. *Combining separate recommenders* / Combinaison des résultats séparés : La méthode collaborative et la méthode basée sur le contenu sont appliquées séparément, puis leurs prédictions sont combinées.
2. *Adding content-based characteristics to collaborative models* / Ajout de résultats issus du contenu aux prédictions collaboratives : Ce système utilise l'approche collaborative traditionnelle entre individus « *People-to-people correlation* », à laquelle il ajoute des recommandations basées sur la classification du contenu et des goûts renseignés par les usagers.
3. *Adding collaborative characteristics to content-based models* / Ajout de prédictions collaboratives aux groupes d'intérêt issus du contenu : Le principe de ce modèle n'est pas d'inverser par rapport au précédent, mais d'incorporer quelques caractéristiques de la méthode collaborative par profil de groupe « *Model based* » dans l'approche à base de contenu.

4. *Single unifying recommendation model* / Modèle de recommandation unifié : Construction d'un modèle général qui incorpore les caractéristiques des deux modèles au sein d'un même algorithme.

Conclusion sur les modèles de recommandation historiques

Les deux premiers types de modèles de recommandation se chevauchent sur un axe historique dans les années 90. Cette première partie a présenté les méthodes et algorithmes associés à la base des systèmes de recommandation, à savoir les systèmes basés sur le filtrage collaboratif et ceux issus d'un traitement par catégorisation du contenu. Nous avons exposé que les systèmes de recommandation par filtrage collaboratif sont issus d'un traitement statistique sur l'opinion exprimée des usagers. Il est apparu que les méthodes basées sur les données sont adaptées des règles de traitement automatique du langage, notamment de l'indexation automatisée et de la pondération de termes représentatifs. Pour pallier aux faiblesses inhérentes à ces modèles initiaux, dès la fin des années 90, des méthodes hybrides sont apparues.

Les folksonomies

Les systèmes d'annotation sociale permettent aux utilisateurs d'annoter des ressources avec des étiquettes personnalisées. Ces étiquettes ou *tags* servent à naviguer des espaces d'informations vastes et complexes, sans la nécessité de s'appuyer sur des hiérarchies prédéfinies comme les taxonomies. Ces systèmes permettent aux utilisateurs d'organiser et de partager leurs ressources propres, ainsi que d'en découvrir de nouvelles annotées par d'autres utilisateurs. Des recommandeurs de *tags* dans de tels systèmes permettent d'aider les utilisateurs à trouver des balises appropriées pour les ressources qu'ils déposent, référencent ou consultent. Cela contribue à la consolidation de l'ensemble du système d'annotation et bénéficie à tous les utilisateurs et à toutes les ressources (Gemmell *et al.*, 2010). Ce système participatif par filtrages collaboratifs permet de mettre en place à moindre coût des systèmes de recommandation efficace sur les espaces de partage de signets génériques comme *del.icio.us* (cf. Figure 3.15) ou scientifiques avec *citeulike*. Dans les deux cas, les utilisateurs proposent des liens sur vers des documents présentant un intérêt à leurs yeux et ajoutent des mots clés. Un même document est souvent proposé et annoté par plusieurs utilisateurs.

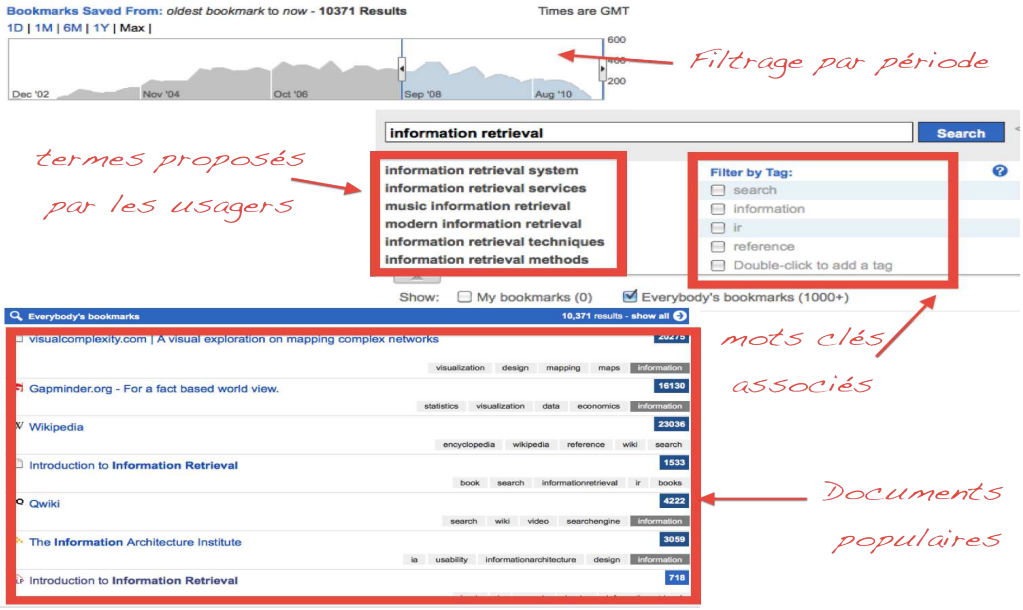

Figure 3.15: Système de recommandation folksonomique : del.icio.us

Les statistiques globales d'utilisation

En appliquant des techniques et méthodes du domaine de l'aide à la décision, une nouvelle approche pour pallier l'absence de l'utilisateur du système d'interaction dans les systèmes de recommandations existantes est proposée.

3.3.6 Conclusion

Les outils de recherche existants sont génériques, pragmatiques et offrent un accès à l'information acceptable pour des sujets basiques. Les moteurs (et méta-moteurs) de recherche nous relient à des milliards de documents qui peut être consultés rapidement grâce des mots clé. La recherche par mot clé plein constitue le point de départ pour la majorité des utilisateurs. Cependant, si cette stratégie fonctionne bien pour une minorité de requêtes, l'utilisateur type est souvent confronté soit avec une liste de résultats vide ou avec une liste contenant des milliers, voire des millions, de réponses possibles (Eissen et Stein, 2002). Il devient donc évident que la connaissance de l'usage des outils et méthodologies basiques de recherche d'information forment un pré-requis nécessaire, mais non suffisant à l'activité de collecte d'informations pertinentes, particulièrement pour le domaine de la science.