INTERFACE ADAPTATIVE POUR L'AIDE A LA RECHERCHE D'INFORMATION SUR LE WEB

Max Chevalier

N° D'ORDRE : LABORATOIRE IRIT – EQUIPE SIG

THESE

PRESENTEE A

L'UNIVERSITE PAUL SABATIER DE TOULOUSE (SCIENCES)

EN VUE DE L'OBTENTION DU

DOCTORAT DE L'UNIVERSITE PAUL SABATIER

SPECIALITE : INFORMATIQUE

PAR

MAX CHEVALIER

INTERFACE ADAPTATIVE POUR L'AIDE A LA RECHERCHE D'INFORMATION SUR LE WEB

Soutenue le 16/12/2002 devant le jury composé de :

M. C. CHRISMENT,	Professeur à l'Université Paul Sabatier, Toulouse III	(Directeur de thèse)
M. B. ESPINASSE	Professeur à l'Université d'Aix-Marseille	(Rapporteur)
Mme C. JULIEN,	Maître de conférences à l'IUT Informatique, Toulouse III	(Encadrante)
Mme J. MOTHE,	Professeur à l'IUFM Midi-Pyrénées,	(Invitée)
M. J.M. PINON,	Professeur à L'Institut National Sciences Appliquées, Lyon	(Rapporteur)
Mme C. SOULE-DUPUY,	Professeur à l'Université des Sciences Sociales, Toulouse I	(Examinatrice)
M. G. ZURFLUH,	Professeur à l'Université des Sciences Sociales, Toulouse I	(Examinateur)

Institut de Recherche en Informatique de Toulouse

Centre National de la Recherche Scientifique – Institut Polytechnique – Université Paul Sabatier

Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 04

Tél : +33 (0)5.61.55.66.11

A mes parents et amis,

Remerciements

L'achèvement de ce travail mené sur plusieurs années procure une grande satisfaction. Il est l'occasion de se remémorer les différentes embûches qu'il a fallu surmonter mais surtout les personnes qui m'ont permis d'en arriver là.

Je tiens donc à exprimer toute ma reconnaissance à Messieurs Claude CHRISMENT, Jacques LUGUET et Gilles ZURFLUH pour m'avoir accueilli au sein de leur équipe. Qu'ils soient assurés de ma profonde gratitude.

Je tiens également à remercier Monsieur Bernard ESPINASSE, Professeur à l'Université de Marseille et responsable de l'équipe *Information et Connaissances Distribuées* (INCOD) du Laboratoire des Sciences de l'Information et des Systèmes, pour avoir accepté d'être rapporteur de mes travaux et pour ses observations qui m'ont permis d'améliorer la qualité de ce mémoire. Je tiens à lui exprimer mes remerciements pour l'honneur qu'il me fait en participant à ce jury.

Je tiens à remercier Monsieur Jean-Marie PINON, Professeur à l'Institut National des Sciences Appliquées de Lyon et responsable du thème *Ingénierie des Informations Documentaires et des Objets Complexes* à l'INSA Lyon, pour l'honneur qu'il m'a fait en acceptant d'être rapporteur de mon travail et pour ses remarques qui ont permis d'améliorer la qualité de ce mémoire. Je tiens à lui exprimer tous mes remerciements pour l'honneur qu'il me fait en participant à ce jury.

Je remercie Monsieur Claude CHRISMENT, Professeur à l'Université Paul Sabatier pour avoir dirigé mes recherches. Je le remercie également pour la confiance qu'il m'a témoignée tout au long de ces années et pour tous ses conseils et remarques constructives. Son contact a d'ailleurs été très enrichissant tant au niveau humain qu'au niveau de mon travail. Il peut être assuré de mon sincère respect et de ma profonde gratitude.

Je remercie Madame Christine JULIEN, Maître de Conférences à l'IUT Informatique, pour m'avoir encadré. Je la remercie pour toute la patience et la disponibilité dont elle a fait preuve à mon égard. Ses conseils et remarques constructives m'ont permis d'améliorer grandement la qualité de mes travaux et de ce mémoire. Je tiens à lui exprimer mes remerciements pour l'honneur qu'elle me fait en participant à ce jury.

Je remercie Monsieur Gilles ZURFLUH, Professeur à l'Université Toulouse I, pour l'intérêt qu'il a porté à mes travaux en examinant ce mémoire, pour ses conseils avisés et pour l'honneur qu'il me fait en participant à ce jury.

Je remercie également Madame Chantal SOULE-DUPUY, Professeur à l'Université Toulouse I, pour toute l'attention dont elle à fait preuve en examinant ce mémoire et pour l'honneur qu'elle me fait en participant à ce jury. Je tiens également à la remercier pour toutes les remarques et discussions constructives que l'on a pu avoir et pour m'avoir aidé à pousser la voiture ☺ !!

Je remercie Madame Josiane MOTHE, Professeur à l'Institut Universitaire de Formation des Maîtres de Midi-Pyrénées pour ses remarques avisées et sa franchise qui m'ont permis d'améliorer la qualité de ce mémoire. Je tiens à lui exprimer mes remerciements pour l'honneur qu'elle me fait en participant à ce jury.

Je tiens également à remercier tous mes amis et collègues de l'équipe SIG du laboratoire IRIT pour leur aide, leur soutien et leur gentillesse. Je souhaite également les remercier pour la confiance dont ils ont fait preuve à mon égard en me laissant m'impliquer au sein de la vie collective de l'équipe. Les rapports humains dont j'ai profité à leur côté ont fait naître de réels liens d'amitiés qui a mes yeux n'ont pas de prix. Qu'ils soient tous assurés de ma plus profonde gratitude et estime. J'exprime en particulier ma gratitude à Florence, Franck, Gilles, Mohand, Olivier (je te battrais au squash c'est mon objectif ☺) pour toute l'aide qu'ils ont su m'apporter. Je souhaite également remercier tous mes frères d'armes Anis, Ikram & Nawel.

Je ne veux pas oublier les personnes de l'IUT telles que Marie-Noëlle GRANGEON. (« yes I am») qui m'a bien aidé au cours de ces années ainsi que ceux qui ont permis que je garde le contact avec Christine grâce au tableau blanc (Pascal FERNANDEZ...) ☺.

Je ne souhaite pas oublier le personnel du laboratoire (Agathe, Jean-Pierre, Jean-Claude, Michèle...) pour leur attention et leur efficacité qui ont permis que je passe toutes ces années dans de bonnes conditions grâce à leur professionnalisme.

Je voudrais également remercier toutes les personnes extérieures du domaine universitaire qui m'ont, à leur façon, apporté leur aide.

En premier lieu, je remercie mes parents qui ont su croire en moi et qui m'ont apporté toute leur aide quand j'en ai eu besoin. Ce mémoire leur est dédié à 200%. Je remercie également toute ma famille (qui ne cesse de s'agrandir) qui a contribué de près ou de loin à ce que je suis devenu. Je souhaite également exprimer ma sincère gratitude à mes deux frangins Franck et Bruno qui ont eu le droit et le privilège (?) de me supporter pendant plus de vingt ans.

Je souhaite également remercier tous mes amis de grande date qui ont su m'apporter confiance et écoute à tous les moments. Je les remercie pour avoir supporté mes blagues vaseuses et mes jeux stupides, mais bon on ne change pas du jour au lendemain ☺. Ainsi, je remercie particulièrement Dawid & Sasa, Tony & Vanessa, Pierrot, Kane, Maël, Ludo, Ninne et tous les autres pour avoir partagé un grand nombre d'années à mes côtés (il y en aura d'autres c'est sûr).

Je ne veux pas oublier les volleyeurs qui font partie intégrante de ma vie (et même plus), Boris & Valérie ainsi que les 2 petites coco, Fredo & Béa ainsi que la petite Chloé, Raph, Steph & Flo ainsi que la petite Émilie, Fabien & Carole, Nico & David, Matt & Claire, Fifou & Céline, Bernie, Jeannot. Qu'ils soient tous assurés de ma sincère et profonde amitié et gratitude. Je sais, cela ne doit pas toujours être facile de me côtoyer à la fois sur un terrain et dans la vie de tous les jours...

Bien entendu, cette liste n'est pas exhaustive et je remercie tous ceux et celles qui me connaissent et qui me permettent de me sentir exister...

Merci à toutes et à tous.

Résumé

De l'avènement des nouvelles technologies, du « tout numérique », de l'essor d'Internet et plus particulièrement du World Wide Web (ou web) résulte une profusion d'informations à la portée de tous. Néanmoins, la localisation des informations pertinentes au sein de cette masse informationnelle reste posé.

Le contexte de mes travaux est la recherche d'information textuelle sur le Web et s'inscrit dans le cadre du GDR I3 du CNRS. Ma thèse s'intitule « *Interface adaptative pour l'aide à la recherche d'information sur le web* ». Elle concerne la conception et la réalisation d'une interface permettant d'aider l'utilisateur dans sa démarche de recherche d'information afin qu'il puisse trouver plus efficacement des documents pertinents. Le caractère « adaptatif » de cette interface réside dans le fait qu'elle s'adapte aux besoins de l'utilisateur en lui proposant une aide personnalisée. Ce point est d'autant plus important que la vulgarisation de la recherche d'information implique la prise en compte de la différence entre les usagers et de leur spécificités. Il s'agit donc d'étudier et de proposer des outils permettant d'aider l'utilisateur dans sa tâche de recherche d'information en caractérisant notamment sa place au sein d'un tel processus. Nos travaux ont permis la conception et l'implantation d'un système nommé *Easy-DOR* « Easy DOcument Retrieval ». L'aide que nous proposons à l'utilisateur au travers de ce système intervient à tous les niveaux de sa recherche d'information :

- *en amont du processus de recherche*. Le système aide l'utilisateur à faire évoluer son expertise des domaines relatifs à ses centres d'intérêt afin qu'il puisse effectuer de meilleures recherches ultérieures,
- *au cours du processus de recherche*. Le système exploite les informations provenant de l'utilisateur pour tenter d'identifier ses besoins et ainsi lui apporter rapidement des documents pertinents. Par ailleurs, nous proposons une interface de visualisation lui permettant de mieux apprécier de façon globale les résultats de recherche d'information provenant d'un outil de recherche intégré,
- *en aval du processus de recherche*. Le système propose à l'utilisateur une mise à jour ainsi qu'une aide à l'organisation des documents pertinents qu'il souhaite mémoriser au travers de ses signets (ou favoris).

Par ailleurs, l'aide à la recherche d'information sur laquelle repose notre démarche est basée sur un aspect coopératif. Nous privilégions, en effet, le partage des informations pour les diffuser aux utilisateurs possédant les mêmes centres d'intérêt.

Mots-Clés : Aide à la Recherche d'Information, Interface de visualisation pour les résultats de recherche, VIRI (Visual Information Retrieval Interface), Partage d'informations.

Sommaire

./..

./..

Liste des figures

Note : Le symbole * qui succède le numéro d'une figure signifie que celle-ci est présente sur les planches en couleurs situées entre les pages 62 et 63.

INTRODUCTION GENERALE

« Pour chaque être, il existe une sorte d'activité où il serait utile à la société, en même temps qu'il y trouverait son bonheur »
Maurice BARRES, écrivain français (1862-1923)

L'Information est devenue un pilier de notre civilisation. Nul ne peut y échapper et nous la retrouvons partout, dans chaque média... Le pouvoir de l'Information est tellement important que les supports de l'information deviennent aujourd'hui des outils de diffusion de masse offrant un accès à l'information au plus grand nombre. Néanmoins, retrouver une information n'est pas une chose évidente. En effet, compte tenu du nombre important de sources (journaux, radios, chaînes de télévision...), il est quasiment impossible de savoir où retrouver l'information recherchée, à moins de scruter manuellement chacune de ces sources. Ce constat est général et tous les médias souffrent de ce problème. Internet et son fer de lance, qui est le World Wide Web (« la toile d'araignée mondiale »), n'échappent pas à la règle. En réponse à cela, des outils de recherche ont été développés autour des technologies normalisées (HTTP, URL...) pour permettre un meilleur accès aux informations numériques disponibles sur Internet. Cependant, chaque utilisateur est unique alors que les outils proposés sont généralement destinés à un utilisateur « générique ». Cette inadéquation s'ajoute à des problèmes plus généraux (liés à la recherche, à Internet...). De ce fait, les internautes se heurtent à des écueils dans leurs recherches d'informations, cette recherche pouvant devenir fastidieuse lorsque l'usager est peu expérimenté.

Une aide personnalisée a donc été apportée à l'internaute au travers de différentes approches lui permettant de limiter les problèmes rencontrés (problème de formulation de requête, surcharge cognitive lors d'une navigation, ...). Cette aide lui permet également de trouver plus rapidement les informations qu'il recherche.

Par ailleurs, à l'ère de la communication nous pouvons constater que l'internaute, en quête d'informations, n'a étonnamment que peu d'aide de la part de ses amis et collègues... Son seul réel « interlocuteur » étant le moteur de recherche. Les internautes sont ainsi privés de l'expérience et des connaissances de leurs confrères qui ont éventuellement eu le même besoin. De ce fait, ils occultent une information jugée et validée par un humain qui peut être de meilleure qualité que les seules informations issues des moteurs de recherche.

Le but de nos travaux entre dans le cadre de l'aide à la recherche d'informations sur le World Wide Web (web). Ces travaux font l'objet d'un projet nommé *Easy-DOR* proposant à l'utilisateur une aide personnalisée tout au long de sa recherche et dont un prototype a été implanté.

L'aide que nous proposons à l'utilisateur au travers de ce système intervient à différents niveaux de la recherche d'information sur le web :

- *en amont du processus de recherche*. Le système aide l'utilisateur à faire évoluer son expertise relative à ses centres d'intérêt grâce à des recommandations de lecture afin qu'il puisse effectuer de meilleures recherches ultérieures,
- *au cours du processus de recherche*. Le système lui apporte des documents pertinents au cours de sa navigation. Ces documents pertinents sont issus d'une analyse de l'organisation des hiérarchies de signets de l'ensemble des utilisateurs du système. Par ailleurs, nous proposons une interface de visualisation (en trois dimensions) lui permettant de mieux apprécier, de façon globale, les résultats de recherche d'information provenant d'un outil de recherche intégré,

- *en aval du processus de recherche*. Le système propose à l'utilisateur une mise à jour ainsi qu'une aide à l'organisation des documents pertinents qu'il souhaite mémoriser au travers de ses signets. Cette aide repose sur un suivi de l'évolution des documents intéressant l'utilisateur pour l'en informer ainsi que sur une classification hiérarchique des signets qu'il possède.

L'ensemble de ces propositions permet d'assister l'utilisateur dans ses recherches en lui proposant de nouveaux documents liés à ses centres d'intérêts mais également à ses besoins en information ponctuels.

Notre approche repose également sur un important aspect coopératif (sous la forme d'un partage d'informations entre les différents usagers) pour permettre à un utilisateur de bénéficier des informations collectées par les autres usagers. Ce partage d'information vise à réutiliser les informations d'un utilisateur pour aider un autre usager dans ses recherches. Ainsi, par extension, l'étude que nous avons réalisée s'inscrit dans le cadre général du partage d'informations au sein d'un groupe d'utilisateurs. L'approche proposée est plutôt destinée à un groupe d'utilisateurs au sein d'une organisation (entreprise, laboratoire de recherche) pour laquelle l'information revêt une grande importance.

Le premier chapitre présente dans un premier temps le monde d'Internet et plus particulièrement les fondements ainsi que les caractéristiques du World Wide Web afin de présenter le contexte de nos travaux. Dans un second temps, nous présentons les concepts et fondements de la Recherche d'Information utilisés dans notre approche. Nous soulignons ensuite la Recherche d'Information dans le contexte du web au travers des caractéristiques qui diffèrent d'un environnement « classique » mais aussi des problèmes liés à une telle recherche. Enfin, les solutions issues de la littérature permettant de limiter voire de supprimer ces problèmes sont présentés. Une attention particulière a été portée sur l'aspect coopératif de ces outils d'aide à la recherche.

Le deuxième chapitre présente notre contribution à l'aide à la recherche d'information au travers d'une approche coopérative. Elle a donné lieu à la conception d'un système nommé *Easy-DOR*. Ce système propose différents modules permettant à l'utilisateur d'obtenir une aide précieuse à tous les stades de sa recherche d'information. Dans un premier temps, l'approche générale de l'aide à la recherche d'information que nous proposons est présentée. Nous détaillons, dans un second temps, le projet *Easy-DOR* ainsi que le prototype qui en a découlé.

I

L'AIDE A LA RECHERCHE D'INFORMATION SUR LE WORLD WIDE WEB

L'AIDE A LA RECHERCHE D'INFORMATION SUR LE WEB

./..

I.1 Introduction

Ces dernières années ont vu l'avènement des nouvelles technologies de l'information et de la communication comme Internet. Parmi les services disponibles sur ce média, le web a connu le plus gros essor. Ceci peut s'expliquer par le fait que ce service met à disposition de n'importe quel internaute des informations aisément accessibles et exploitables. Il y a au travers de ce service une profusion d'informations liées les unes aux autres et qui donnent l'image d'une toile d'araignée mondiale.

Pour toutes ces raisons, le web devient peu à peu une source d'informations privilégiée pour quiconque recherche des informations en relation avec ses besoins. En effet, le web regorge de tellement d'informations qu'il est fort à parier que celle que nous recherchons existe réellement. Cependant, du fait de leur grand nombre, la localisation des informations pertinentes est un problème.

Le but de cette section est de faire une rapide présentation d'Internet et plus particulièrement du web. Nous présentons ensuite les concepts de la Recherche d'Information pour enfin introduire les particularités et problèmes de la Recherche d'Information sur le web.

Dans un dernier point, nous détaillons les différentes solutions qui ont été proposées afin de limiter ces problèmes.

I.2 Internet et le web

Cette section vise à présenter Internet et le service web correspondant à notre contexte d'étude. Elle permet également de mettre en évidence les limites et les problèmes liés à ce média concernant la Recherche d'Information.

I.2.1 Internet

L'histoire d'Internet débuta en pleine guerre froide. 1957, les soviétiques lancèrent leur satellite *Spoutnik* et les américains redoutèrent une guerre nucléaire. Le ministère de la défense américain créa alors une agence pour la recherche nommée *ARPA*. Son but était de développer un réseau de communication militaire pouvant fonctionner même avec une partie de ce réseau hors-service. Ainsi, même si des lignes de communications étaient détruites, les militaires devaient garder la faculté de communiquer entre eux. Ce réseau, baptisé *ARPANET*, connut sa première application dans le courrier électronique permettant aux militaires de communiquer et sa première expérimentation eut lieu en 1969 aux Etats-Unis.

Dès le début des années 1980, le réseau échappa de plus en plus aux militaires au profit des universitaires qui virent dans ce réseau un puissant outil de communication et d'échange d'information qui fut rebaptisé Internet (« ***Inter Net***working »).

Dès lors, Internet a connu une perpétuelle évolution en particulier au travers de l'augmentation du nombre de machines connectées (machines hôtes).

Par contre, Internet n'était, à ses débuts, destiné qu'à une poignée d'universitaires qui connaissaient son langage. C'est en Europe (fin 1980) qu'a été simplifié le langage d'Internet, avec la notion d'hypertexte, et développé le premier navigateur permettant de visualiser les différents documents disponibles sur Internet. Basé sur cette approche, le premier navigateur « grand public » nommé *Mosaic* fut développé (1993). Grâce à ce navigateur, chaque utilisateur connecté pouvait accéder et parcourir simplement les documents disponibles sur Internet. Cette fonctionnalité perdure aujourd'hui au travers des navigateurs actuels (Microsoft Internet Explorer, Netscape...).

La figure 1 présentent l'évolution d'Internet au travers du nombre de machines hôtes depuis 1993.

Cet attrait d'Internet peut être expliqué par le fait qu'il permet de partager instantanément des informations entre toutes les machines connectées.

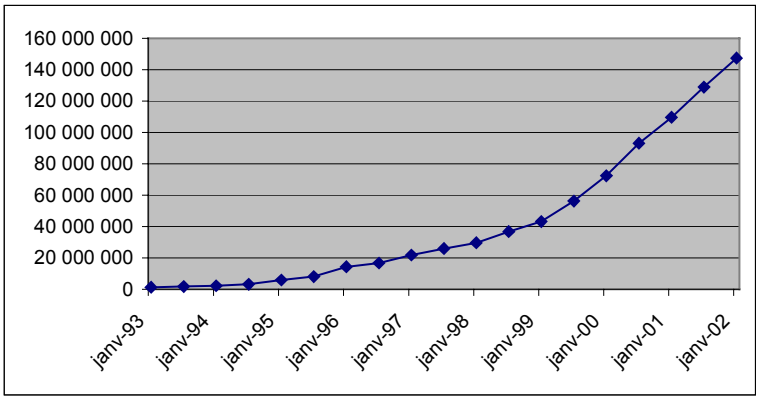


Figure 1 - Nombre de machines hôtes (Source: Internet Software Consortium http://www.isc.org/)

I.2.2 Le World Wide Web

Sur Internet cohabitent un nombre important de services qui n'ont cessé d'évoluer. A ses débuts, Internet proposait des services comme *Gopher* par exemple qui restaient destinés aux universitaires et qui tendent aujourd'hui à disparaître.

Gopher est l'ancêtre du web. Il permettait de visualiser les fichiers contenus sur le serveur au travers d'une arborescence de fichiers. La plupart des serveurs *Gopher* ont aujourd'hui été convertis en serveur web.

Au début des années 1990, Internet a connu un véritable essor du fait de l'avènement du service World Wide Web (ou simplement web). Ce service a permis de simplifier et de démocratiser la mise en œuvre de services multimédia incitant les particuliers ainsi que les entreprises à diffuser leurs informations. Cet essor peut s'expliquer par le fait que le web repose sur des notions peu complexes et parce qu'il permet d'interagir avec les autres protocoles disponibles sur Internet. Cette interaction peut être représentée au travers de la « matrice » (Figure 2). A partir de cette figure, il est aisé de comprendre l'essor d'un tel service. En effet, le web peut être vu comme une « interface » entre les internautes et les différents services d'Internet tout en proposant un outil facile et puissant d'utilisation pour le parcours des différents documents.

« Un document est l'ensemble d'un support d'information et des données enregistrées sur celui-ci sous forme en générale permanente et lisible par l'homme et la machine ». (Définition ISO). Dans le cas du web, la notion de documents revêt un caractère tout particulier et on parle plus couramment de *pages* que de documents.

La technologie sur laquelle repose le web a été développée au CERN (Centre Européen pour la Recherche Nucléaire) en 1989 par Tim Berners-Lee [Berners-Lee, 1994]. L'objectif était la diffusion d'informations scientifiques entre les chercheurs. L'idée sur laquelle repose le web était d'organiser les informations sous forme de documents avec possibilités d'insérer des liens vers d'autre documents autorisant le passage d'un document à un autre sans peine. Cette notion a permis l'émergence, petit à petit, de la toile d'araignée mondiale (« World Wide Web ») faisant référence au nombre très important de liens entre les documents disponibles.

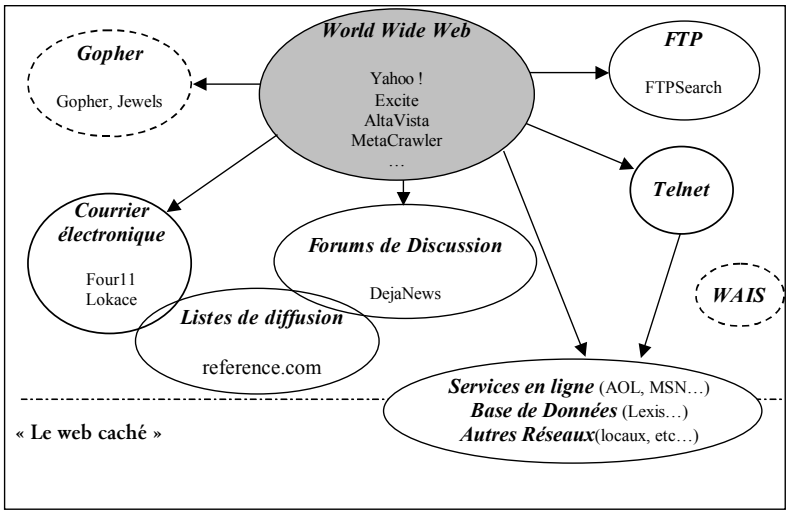


Figure 2 - La « matrice » du web (http://www.cybion.fr)

Le web est donc une illustration parfaite des systèmes hypertextes définis dans [Conklin, 1986]. Il repose sur le protocole *HTTP* (« Hypertext Transfert Protocol ») s'appuyant sur deux normes : la norme *MIME* (« Multipurpose Internet Mail Extensions ») et la norme *HTML* (« Hypertext Markup Langage »).

La norme *MIME* (normes RFC[1] 2045 et suivantes) permet d'encoder des données de types hétérogènes pour être transmis au travers d'Internet. Une liste arborescente de types de documents non exhaustive a été établie. Par exemple, le type *MIME* « text/html » est bien connu sur le web car il représente les documents écrits avec le langage de balisage *HTML*. La norme *MIME* fut, à l'origine, une méthode pour faciliter l'envoi des messages électroniques (mél) multimédia en encodant les données binaires (images, sons) de telle façon qu'un programme de transfert de courrier le plus primitif possible ne les rejette pas. Aujourd'hui, le domaine d'application de cette norme a dépassé le cadre des seuls messages électroniques.

[1] RFC « Request For Comments ». Ces documents fournissent des renseignements précieux sur les standards Internet. La liste générale peut être consultée sur http://www.rfc-editor.org/rfcxx00.html.

La norme *HTML*, développée par le World Wide Web Consortium ou W3C[2], permet à l'utilisateur de décrire les documents qu'il souhaite mettre en ligne sous forme textuelle. Ce langage hypertexte permet également à l'utilisateur d'insérer des liens (ancres) vers tout autre document associé à une *URL*. Un document sous cette forme en *HTML* est communément appelé une *page web*. Par extension, un *site web* correspond à une arborescence de pages web ayant pour racine une page dite d'accueil et se trouvant sur un même serveur.

Ce standard est aujourd'hui remis en cause par des langages où la distinction entre contenu et présentation est beaucoup plus nette comme *XML* (eXtensible Markup Langage) [Bray, 1996] par exemple. L'intérêt de ce langage de balisage est qu'il permet entre autre de créer des documents en distinguant structure logique (contenu sémantique) et structure physique (présentation).

I.2.2.1 Web visible, web caché

Les informations disponibles sur le web peuvent être scindées en deux catégories par rapport aux modes d'accès auxquels elles sont soumises : le ***web caché*** et le ***web visible***.

Le ***web caché*** [Bergman, 2000] correspond à l'ensemble des documents accessibles par l'intermédiaire d'un serveur « dédié » comme un serveur de base de données par exemple. Le seul moyen d'y accéder est d'interroger le serveur grâce à une requête adéquate ou à un formulaire.

Le ***web visible*** correspond à l'ensemble des documents directement accessibles sans avoir besoin de formuler une quelconque requête ou de remplir un quelconque formulaire.

La grande différence entre ces deux modes d'accès réside dans le fait que les informations accessibles par le web caché sont plus nombreuses. D'après [Murray, 2000], le web caché pourrait couvrir la globalité des besoins en information des internautes. Par ailleurs, les informations que contient le web caché sont plus « contrôlées » que celles du web visible.

I.2.2.2 Caractéristiques du web

Le web visible est constitué de plus de 2 milliards de documents web (juillet 2000) et est en plein essor avec une évolution approximative de plus de 7 millions de documents par jour [Murray, 2000]. Du fait de la distinction entre le web caché et le web visible, le web peut donc être vu comme un iceberg, la partie cachée (web caché) étant beaucoup plus importante que la partie visible. En effet, le web caché est évalué comme étant 400 à 550 fois plus important que le web visible [Bergman, 2000].

Il est également à noter que les différentes informations sur le web sont principalement en langue anglaise alors que les informations en français ne sont que peu représentées. En 1998 les documents en anglais représentaient 71% du nombre de documents total alors que ceux en français ne représentaient que 3% [O'Neill, 1998].

[2] http://www.w3c.org

I.2.2.3 Limites des informations sur le web

L'utilisateur a donc accès, au travers du web, à un nombre important de documents contenant des informations aussi diverses que abondantes. Cependant, le web a des limites qui lui sont inhérentes. Ces limites sont [Baeza-Yates, 1999] :

- la *non persistance* de l'information. Le web possède une dynamique très importante et l'information naît, évolue et disparaît rapidement. Un document visité à un moment *t* ne sera pas forcément le même que celui consulté au moment *(t+1)*. Il a d'ailleurs été estimé que 40% des informations disponibles sur le web changent tous les mois [Kahle, 1996],
- l'*instabilité* de l'information. Le web repose sur une architecture informatique qui peut connaître diverses pannes ou dysfonctionnements. De ce fait, l'information n'est pas accessible de façon permanente et il se peut qu'à tout moment celle-ci ne soit plus accessible,
- le manque de *qualité de l'information*. Le web est un média ouvert, dans le sens où il n'y a pas d'organisme régulateur des contenus disponibles. De ce fait, les informations disponibles sont souvent sujettes à des problèmes de véracité, de fautes de langages ou erreurs typographiques,
- la *redondance d'information*. Une expérimentation [Shivakumar, 1998] réalisée à partir d'une collection de 24 millions de pages web montre que plus de 30% de l'information est redondante (page web miroir par exemple). Cette proportion peut être encore plus importante si l'on considère une redondance sémantique ou partielle des informations,
- l'*hétérogénéité de l'information*. sur le web cohabitent des informations dans des médias différents (image, son...), des formats différents (jpeg, mp3...) et même des langues différentes (français, chinois...),
- le *volume d'information* disponible. L'utilisateur a du mal à se repérer, à identifier les documents intéressants au sein de cette masse informationnelle qui évolue sans cesse. De plus, ce volume d'information très important implique que la couverture du web par les outils de recherche est assez faible.

La plupart de ces problèmes sont difficilement améliorables de façon automatique (stabilité, hétérogénéité de l'information). Certains d'entre eux sont, en effet, intrinsèques à la nature humaine (contenu inexact ou mal formé des documents par exemple).

Du point de vue de l'internaute, le problème principal du web vient de son architecture. En effet, il n'existe aucune organisation spécifique des informations, aucun index général référençant les informations existantes. Les informations peuvent être situées n'importe où sur la toile voire dupliquées, d'où le problème de la localisation de l'information. Ce problème est d'autant plus important que le nombre de documents disponibles est grand. Cependant, ce problème n'est pas récent, il était déjà d'actualité dès les débuts d'Internet avec les premiers outils de recherche tels que *Gopher* mais il ne fait que s'accentuer avec le temps.

Afin de réduire ce problème de localisation de l'information, la Recherche d'Information (RI) s'est développée sur ce média et a permis la mise en œuvre d'outils facilitant la

recherche de l'utilisateur. La RI a donc trouvé, dans le web, un domaine d'application privilégié.

I.3 La Recherche d'Information (RI) textuelle sur le web

I.3.1 Fondements de la Recherche d'Information

La Recherche d'Information (RI) traite de la représentation, du stockage, de l'organisation ainsi que de l'accès à l'information. De la représentation et de l'organisation de l'information dépendent la facilité d'accès à celle qui est pertinente pour l'utilisateur. Un Système de Recherche d'Information (*SRI*) est un ensemble de modèles et de processus permettant la sélection d'informations pertinentes dans une ou plusieurs collections en réponse aux besoins d'un utilisateur.

Le schéma général de la Recherche d'Information est présenté dans la figure (Figure 3).

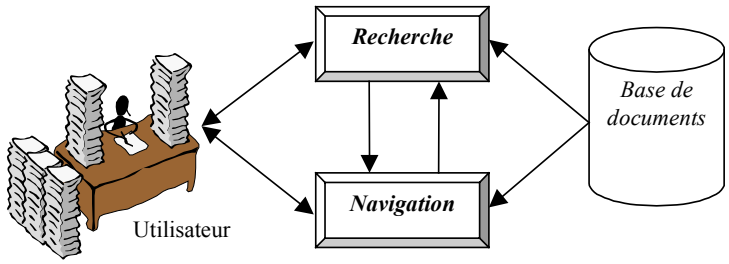


Figure 3 - La Recherche d'Information

La RI intègre deux tâches bien spécifiques : la ***navigation*** et la ***recherche***[3].

La ***navigation*** correspond à l'action de trouver des informations pertinentes au travers d'une base de documents sans connaître, à priori, le contenu et le format des documents contenus dans la base.

La ***recherche*** correspond à l'action de rechercher des informations au travers d'une base de documents à partir des besoins exprimés sous la forme d'une requête.

Dans cette section, nous avons fait une distinction nette entre ces deux tâches car elles mettent en jeu des processus différents auxquels nous allons nous référer tout au long de ce mémoire.

I.3.1.1 La tâche de navigation

La tâche de navigation permet à l'utilisateur de parcourir l'espace des documents de la collection sans devoir formuler ses besoins. Le principal intérêt de cette tâche est qu'elle permet à l'utilisateur d'acquérir des informations sans nécessairement avoir à connaître, a priori, le contenu et/ou la structure (organisation) des informations qu'il va rencontrer.

Trois modèles ont été définis pour caractériser une navigation :

[3] Nous utilisons le terme « recherche » pour représenter le fait de proposer à l'utilisateur des informations en réponse à ses besoins. Ce terme correspond au terme anglais « retrieval ».

- *modèle plat*. Les documents sont présentés dans un plan ou une liste simple,
- *modèle structuré.* Par analogie à un système de fichiers, les documents sont organisés sous la forme d'une arborescence. Ce modèle est intéressant lorsque l'on souhaite proposer à l'utilisateur les documents par rapport aux thèmes qu'ils abordent,
- *modèle hypertexte*. Ce modèle est basé sur la notion d'hypertexte. Cette notion vise à étendre la notion de fichier texte linéaire (ou séquentiel) en permettant une structuration en graphe [Julien, 1988]. Ce concept a été développé pour permettre une consultation non linéaire des documents. Les nœuds sont des contenants (granule ou ensemble de granules) qui ne se limitent pas uniquement à du texte mais peuvent également contenir des images (fixes ou animées) et du son. Un lien hypertexte est un lien *référentiel* établissant des relations non hiérarchiques de sémantique très diverses entre les nœuds. Ils sont généralement orientés et caractérisés par un nœud d'origine et un nœud de destination. L'utilisateur peut ainsi, au travers des liens hypertextes, atteindre de façon transparente une portion du document voire des portions d'autres documents. Les liens sont soit insérés par l'auteur des documents afin de rapprocher des documents (traitant du même thème par exemple) ou automatiquement par un processus informatique comme dans [Agosti, 2000].

Dans le contexte du web, les deux premiers modèles sont fréquemment utilisés tout en servant de base à une navigation hypertexte. La navigation la plus courante sur le web est la navigation hypertexte.

Nous définissons, dans le contexte de la navigation sur le web, un *hypertexte local* comme étant un sous-ensemble de l'hypertexte centré sur le document visité (Figure 4). L'hypertexte est composé de l'ensemble des granules documentaires (documents web) et des liens entre eux. Dans cette figure, la couleur des documents (foncé à clair) représente la distance par rapport au document courant. La notion de distance correspond au nombre de liens qu'il existe, par transitivité, entre deux nœuds. La taille de l'hypertexte local dépendra de la distance maximale prise en compte par rapport au document considéré.

L'hypertexte local évolue donc au fur et à mesure de la navigation de l'utilisateur.

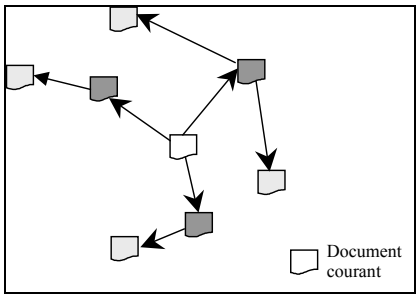


Figure 4 - Exemple d'un hypertexte local

Malgré la facilité d'utilisation, deux limites de la navigation hypertexte doivent être soulignées [Agosti, 1996], [Baeza-Yates, 1999] :

- la *désorientation*. L'utilisateur ne sait plus trop quel chemin suivre. On dit qu'il est « perdu dans l'hypertexte ». Pour tenter de pallier à cet inconvénient les différents outils proposent un mécanisme de retour en arrière,
- la *surcharge cognitive*. L'utilisateur réalise un important effort cognitif pour construire une carte mentale de l'hypertexte reflétant l'organisation de l'hypertexte local. Il se produit une surcharge cognitive lorsque l'utilisateur n'arrive plus à mémoriser la structure de l'hypertexte dans lequel il se trouve. Ainsi, de la conception de l'hypertexte (simplicité, organisation...) dépend le bon déroulement de la navigation.

I.3.1.2 La tâche de recherche

La recherche vise à proposer à l'utilisateur des documents en adéquation avec ses besoins appelés ***requêtes***. Or, pour mesurer cet appariement, le SRI s'appuie sur une représentation commune des besoins de l'utilisateur et du contenu des documents textuels. Ces représentations reposent sur la caractérisation du contenu sémantique des documents et des besoins de l'utilisateur. Ces représentations sont ensuite utilisées au travers d'un modèle de RI permettant de mesurer leur appariement. L'architecture d'un SRI est donné au travers du processus en U présenté dans la Figure 5.

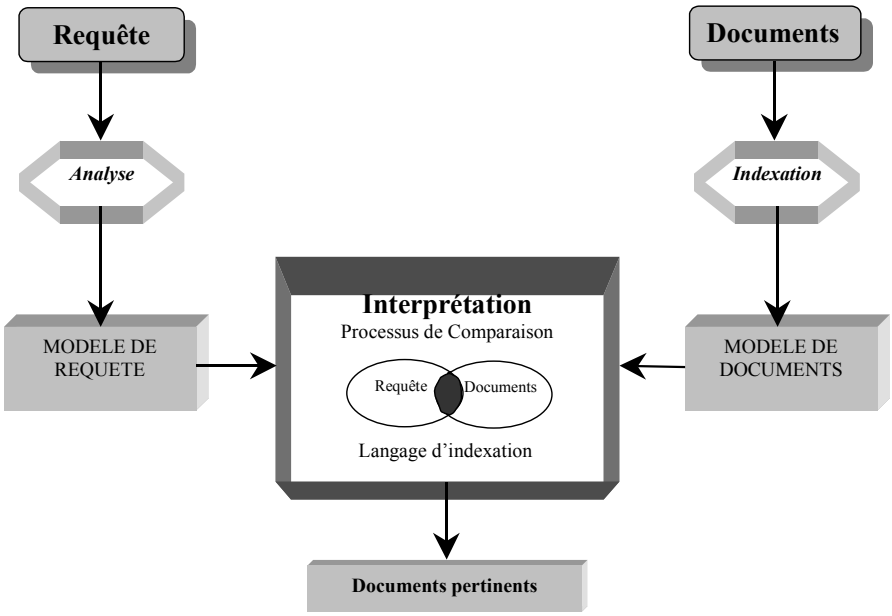


Figure 5 - Architecture d'un SRI

I.3.1.2.1 Représentation des informations textuelles : indexation

Afin de construire des représentations comparables du contenu sémantique des documents et des requêtes, le SRI applique une phase d'***indexation***. Cette phase vise à extraire des caractéristiques sur le contenu sémantique des informations textuelles brutes. Pour cela, le SRI traite les informations textuelles sous la forme d'*unités documentaires*.

Nous définissons une unité documentaire comme étant le plus petit granule d'information correspondant à un niveau de finesse d'étude du contenu informationnel. Ce granule peut correspondre au contenu intégral d'un document ou encore à un paragraphe voire à une phrase du document.

A partir de ces unités documentaires, une phase indexation est réalisée afin d'extraire les caractéristiques de leur contenu informationnel. L'indexation est une phase très importante dans le processus de recherche car de sa qualité dépendra la qualité des réponses lors de l'utilisation du SRI et donc les performances du système. En effet, en terme de performances, l'utilisateur souhaite obtenir toutes les informations répondant à ses besoins (unités documentaires *pertinentes*) et pas d'informations ne répondant pas à ceux-ci (unités documentaires *non pertinentes*).

La phase d'indexation analyse le contenu textuel des unités documentaires en vue de construire un ensemble de termes d'indexation (termes significatifs). Ces termes d'indexation représentent le contenu sémantique de l'unité documentaire et l'ensemble de ces termes est appelé le ***langage d'indexation***.

Ce langage peut être :

- *libre*. Les termes d'indexation sont extraits des textes et peuvent donc inclure toute la variété du langage naturel. Ces langages d'indexation sont construits lors de la phase d'indexation.
- *contrôlé*. Les termes d'indexation utilisés sont prédéfinis et limités. Un thésaurus est consulté lors de l'analyse du texte pour ne conserver que des termes de ce dictionnaire. Ce type de langage d'indexation est construit a priori avant de commencer la phase d'indexation.
- *une combinaison des deux précédents.*

Les termes du langage d'indexation peuvent être sélectionnés par une indexation manuelle, semi-automatique ou automatique.

I.3.1.2.1.1 Indexation manuelle

L'indexation manuelle est réalisée par des documentalistes. Ces experts ont pour tâche de caractériser au mieux les idées contenues dans les unités documentaires. Cette indexation requiert un important effort intellectuel et cognitif pour identifier et décrire l'essence des unités documentaires.

Ce type d'indexation permet d'obtenir une caractérisation performante mais subjective du contenu des unités documentaires car cette approche dépend fortement des connaissances du domaine des documentalistes. L'indexation manuelle trouve ses limites pour de grandes bases de documents qui nécessitent énormément de temps pour être traitées.

I.3.1.2.1.2 Indexation automatique

Pour de grandes bases, la tendance générale s'oriente vers un processus d'indexation automatique permettant d'extraire rapidement les termes représentatifs des unités documentaires. L'intérêt d'une telle indexation réside principalement dans sa rapidité d'exécution qui est tout à fait adaptée à des volumes très importants mais également dans le

fait qu'elle permet de limiter la représentation des documents aux « entrées » utiles permettant de retrouver les informations accessibles. Cette approche repose sur différentes phases correspondant à :

- l'extraction des termes d'indexation,
- la réduction du langage d'indexation,
- la pondération des termes d'indexation.

I.3.1.2.1.2.1 Extraction des termes d'indexation

Afin de construire le langage d'indexation, le système parcourt le contenu du document pour en extraire les termes d'indexation.

Diverses approches pour cette extraction sont envisageables :

- une *approche linguistique*. Cette méthode fait appel à des techniques de traitement du langage naturel pour analyser et comprendre le contenu de l'unité documentaire ou d'une requête. Elle repose sur une recherche du sens même du contenu des unités documentaires,
- une *approche morpho-syntaxique*. Elle repose sur une étude morpho-syntaxique du contenu des diverses unités documentaires,
- une *approche mixte*. Cette approche est un mélange des deux approches précédentes.

Dans une unité documentaire, l'ensemble des termes d'indexation extraits peut être important. Plus il y a de termes, plus on utilise de mémoire de stockage et plus les temps de calculs (notamment lors de l'appariement entre la requête et le document) sont importants. Il est donc nécessaire de limiter le langage d'indexation aux termes les plus représentatifs du contenu d'une unité documentaire.

I.3.1.2.1.2.2 Réduction du langage d'indexation

Pour limiter le nombre de termes d'indexation, il est nécessaire de ne conserver que les termes qui contribuent au mieux à la caractérisation de l'unité documentaire. Ainsi, dans un premier temps, les mots qualifiés de « vides » sont supprimés. Ce processus correspond à la suppression des termes d'indexation qui n'apportent pas de sens réel à l'unité documentaire. Ces mots vides se trouvent généralement dans un antidictionnaire et sont des mots à contribution purement syntaxique comme les pronoms « *à* » ou « *de* » par exemple pour le français.

De plus, les lois de Zipf [Zipf, 1949] et [Luhn, 1958] soulignent que :

- un terme d'indexation apparaissant trop fréquemment dans un texte ne joue qu'un rôle syntaxique (mot vide) et ne doit pas être utilisé dans le langage d'indexation,
- un terme d'indexation présent dans l'ensemble des documents de la base n'apporte aucun pouvoir discriminant à la recherche (le terme est considéré comme un mot vide),
- un terme d'indexation de fréquence intermédiaire est considéré comme significatif, il représente le contenu sémantique de l'unité documentaire et appartient au langage d'indexation.

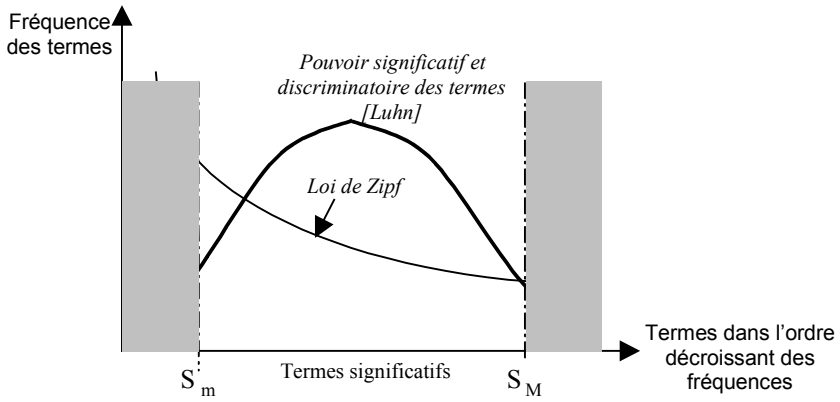


Figure 6 - Pouvoir discriminatoire des termes d'indexation

I.3.1.2.1.2.3 Lemmatisation

Un mot peut revêtir différentes formes au sein des différentes unités documentaires (adjectif, verbe...). La lemmatisation tend à regrouper toutes les formes d'un mot sous un même radical (forme neutre d'un mot) permettant ainsi de remplacer un ensemble de termes d'indexation par leur radical. Ce regroupement permet de retrouver des unités documentaires même si les termes de la requête n'ont pas forcément la même forme que dans les unités documentaires. Elle permet également de réduire le nombre de termes d'indexation. Les méthodes de lemmatisation les plus courantes sont :

- *la troncature*. Seulement un nombre fixe de caractères des termes d'indexation est conservé. Par exemple, pour la langue française, une troncature à 7 caractères est communément effectuée (référence, référencement → référen) [Soulé-Dupuy, 1990],
- *l'extraction des radicaux par utilisation de règles*. L'algorithme de Porter [Porter, 1980] est par exemple utilisé pour l'anglais.

Afin de mesurer un appariement graduel entre une unité documentaire et une requête, les termes d'indexation sont pondérés afin de refléter leur degré d'importance ou de représentativité dans l'unité documentaire.

I.3.1.2.1.3 Pondération des termes d'indexation

Le calcul de la représentativité d'un terme d'indexation repose sur sa fréquence d'apparition dans le texte en langage naturel [Salton, 1983], [Zipf, 1949]. Afin de caractériser le pouvoir de représentativité (poids) des termes d'indexation, différentes mesures ont été proposées.

Elles reposent sur des mesures statistiques dont les plus utilisées sont :

- la *fréquence relative* d'un terme d'indexation (*tf*). Il s'agit de la fréquence d'apparition du terme d'indexation dans l'unité documentaire,
- la *fréquence absolue* d'un terme d'indexation dans la collection globale d'unités documentaires (*idf*). Il s'agit de la fréquence inverse d'apparition du terme d'indexation dans l'ensemble des unités documentaires de la collection.

Le poids d'un terme d'indexation *i* dans une unité documentaire peut être défini par l'équation générale (1) [Sparck Jones, 1972] :

$$Poids_i = tf_i.idf_i \qquad \textbf{1}$$

Où $idf_i = \log(\frac{N}{N_i}) + 1$ avec N représentant le nombre d'unités documentaires dans la collection et N_i le nombre d'unités documentaires possédant le terme d'indexation *i*.

Plusieurs travaux se sont intéressés à l'amélioration de ce schéma de pondération de base (tf.idf). Ainsi, les exemples suivants proposent des variantes à ce schéma :

[Salton, 1990]
$$w_{ij} = \frac{\log(tf_{ij}) + 1}{\sqrt{\sum_{i=1}^{N} (\log(tf_{ij}) + 1)}} \qquad \textbf{2}$$

Où tf_{ij} : fréquence du terme d'indexation *j* dans l'unité documentaire *i*,

N : nombre d'unités documentaires dans la collection.

$(\log(tf_{ij})+1)$ est utilisée afin de réduire l'intervalle des fréquences d'apparition des termes d'indexation d'une unité documentaire et ainsi d'éviter l'impact des termes d'indexation trop fréquents lors du processus de comparaison entre les requêtes et les unités documentaires.

Une restriction de ces mesures doit être formulée dans le sens où ces approches ne prennent pas en compte la longueur des unités documentaires indexées. Pour des unités documentaires de longueur variable, le pouvoir d'un terme sera faussé car la fréquence relative sera intuitivement plus importante pour de longues unités documentaires. Ainsi, divers autres travaux se sont intéressés à la prise en compte de la longueur des unités documentaires pour la pondération des termes d'indexation.

Nous pouvons citer, par exemple, les travaux de :

[Singhal, 1997]
$$w_{ij} = \frac{1 + \log(tf_{ij})}{0.7 + 0.3 * \frac{l_i}{L_M}} \qquad \textbf{3}$$

Où l_i : longueur de l'unité documentaire *i*,

L_M, longueur moyenne des unités documentaires de la collection.

[Boughanem, 2000]

$$w_{ij} = \frac{1 + \log(tf_{ij}) * \left(h_1 + h_2 * \log\left(\frac{M}{n_j}\right)\right)}{h_3 + h_4 * \frac{dl_i}{\Delta}} \qquad \textbf{4}$$

Où M : nombre d'unités documentaires dans la collection,

n_j : nombre d'unités documentaires contenant le terme d'indexation j,

dl_i : nombre de termes d'indexation de l'unité documentaire d_i,

Δ : nombre moyen de termes d'indexation dans les unités documentaires de la collection,

h1, h2, h3, h4 : Ces paramètres doivent être ajustés en fonction des collections de documents utilisées. Par exemple, sur la collection TREC-5, les valeurs optimales sont h1=0.8, h2=0.2, h3=8.8, h4 = 0.2.

I.3.1.2.1.4 Indexation semi-automatique

Ce type d'indexation combine les méthodes d'indexation manuelle et automatique en privilégiant toutefois l'intervention humaine. Ainsi, les experts caractérisent les idées contenues dans une unité documentaire sous la forme de méta-informations. Une indexation automatique est ensuite réalisée pour l'unité documentaire en tenant compte de ces méta-informations.

L'indexation permet donc au SRI d'obtenir l'essence même des unités documentaires par le biais d'un langage d'indexation. Cependant, il est nécessaire d'utiliser un modèle unique de représentation (modèles de recherche) pour la requête et pour les unités documentaires afin d'en apprécier l'appariement.

I.3.1.2.2 Modèles de recherche d'information

Ces modèles peuvent être divisés en deux catégories : les modèles dits « exacts » qui ne retournent que des documents répondant exactement à la requête (modèle booléen) ou les modèles dits « partiels » (probabiliste, vectoriel...) qui retournent des documents répondant à tout ou partie de la requête. Ces derniers utilisent une *valeur réelle (degré de pertinence système)* pour rendre compte du degré de l'appariement entre la requête et une unité documentaire. Il est à noter que cette pertinence système est une valeur calculée et qu'elle peut être différente de la pertinence réelle qui découle du jugement de pertinence réalisé par l'utilisateur.

Dans cette section, nous ne présentons que le modèle vectoriel car la plupart des approches reposent sur ce modèle. Pour plus d'informations concernant les autres modèles se référer à [Baeza-Yates, 1999], [Salton, 1983]. Pour une étude comparative de ces modèles ainsi que des extensions possibles peuvent être trouvées dans [Soulé-Dupuy, 2001].

Le modèle vectoriel représente la requête ainsi que le contenu des unités documentaires dans un espace vectoriel engendré par les termes d'indexation [Salton, 1983].

Soient $\vec{Q}_i$ le vecteur représentant la requête *i* et $\vec{D}j$ le vecteur représentant l'unité documentaire *j* dans l'espace vectoriel :

$$\vec{Q}_i = \begin{pmatrix} q_{i1} \\ q_{i2} \\ q_{i3} \\ \cdots\cdots \\ q_{in} \end{pmatrix} \qquad \vec{D}j = \begin{pmatrix} d_{j1} \\ d_{j2} \\ d_{j3} \\ \cdots\cdots \\ d_{jn} \end{pmatrix}$$

Où n représente le nombre de termes d'indexation de la base,
q_{ik} représente le poids du terme *k* dans la requête *Qi*,
d_{jk} représente le poids du terme *k* dans l'unité documentaire *Dj*.

L'appariement entre une requête et une unité documentaire est aisément calculé à partir d'un calcul de similarité entre les vecteurs $\vec{Q}_i$ et $\vec{D}j$. Pour cela, différentes mesures ont été proposées comme :

- le produit scalaire,
- la distance métrique,
- la mesure cosinus.

Le ***produit scalaire*** est une fonction mathématique utilisant les coordonnées des vecteurs (poids des termes). La mesure de similarité par le produit scalaire entre un vecteur document $\vec{D}_j = (d_{j1}, d_{j2}, ..., d_{jn})$ et un vecteur requête $\vec{Q}_i = (q_{i1}, q_{i2}, ..., q_{in})$ est donnée dans l'équation ***5***.

$$Sim(\vec{Q}_i, \vec{D}_j) = \sum_{k=1}^{n} q_{ik} . d_{jk} \qquad \textbf{5}$$

Où d_{jk} est le poids du terme t_k dans l'unité documentaire *j*,
q_{ik} est le poids du terme t_k dans la requête *i*.

Un degré de similarité maximal de 1 indique que l'unité documentaire correspond exactement à la requête, un degré égal à 0 indique qu'elle ne correspond pas du tout à la requête.

La ***distance métrique*** entre un vecteur document et un vecteur requête est donnée par l'équation ***6***.

$$Dist(\vec{Q}_i, \vec{D}_j) = \left\| \vec{Q}_i, \vec{D}_j \right\| = \left(\sum_{k=1}^{n} \left| d_{jk} - q_{ik} \right| \right)^{1/p} \quad \text{avec } p \geq 1 \qquad \textbf{6}$$

Ce qui implique que $Dist(\vec{Q}_i, \vec{D}_j) = 0 \Leftrightarrow \vec{Q}_i = \vec{D}_j$ et que $Dist(\vec{Q}_i, \vec{D}_j) = Dist(\vec{D}_j, \vec{Q}_i)$.

La ***mesure cosinus*** est la plus répandue des mesures de similarité utilisées pour ce modèle de recherche. Elle mesure l'angle entre les vecteurs $\vec{Q}_i$ et $\vec{D}_j$. Elle est équivalente au produit scalaire normalisé des vecteurs. La mesure de cosinus est donnée par l'équation ***7***.

$$Sim(\vec{Q}_i, \vec{D}_j) = \frac{\sum_{k=1}^{n} q_{ik} . d_{jk}}{\left(\sum_{k=1}^{n} q_{ik}^2\right)^{1/2} \left(\sum_{k=1}^{n} d_{jk}^2\right)^{1/2}} \quad 7$$

Le modèle vectoriel permet à l'utilisateur de formuler sa requête en langage (pseudo) naturel. Il permet de retrouver des unités documentaires plus ou moins pertinentes par rapport à la requête et ainsi de restituer une liste ordonnée par pertinence système ou bien une liste limitée aux *k* documents les plus pertinents. Par contre, l'inconvénient majeur de ce modèle vectoriel est qu'il ne permet pas de modéliser les associations entre les termes d'indexation : chacun des termes d'indexation est considéré comme indépendant des autres (variables indépendantes dans l'espace du langage d'indexation).

I.3.1.2.3 Environnement de recherche

La tâche de recherche peut être réalisée dans un environnement adhoc ou dans un environnement dynamique qui conditionne le processus de recherche.

Dans un environnement adhoc, l'outil de recherche repose sur une ou plusieurs collections de documents stables et des besoins en informations momentanés et dynamiques.

Dans un environnement dynamique, l'outil de recherche repose non plus sur une collection stable de documents mais sur un flux dynamique de documents. Les besoins de l'utilisateur, contrairement à l'environnement adhoc, sont relativement stables (profil de filtrage).

Dans la suite du document, afin de faire la distinction entre un outil de recherche dans un environnement adhoc ou un environnement dynamique, nous avons utilisé respectivement les noms « d'outils de recherche adhoc » et « d'outils de filtrage ».

Une présentation des spécificités de chacune de ces deux catégories d'outils est réalisée pour en souligner la dualité.

I.3.1.2.3.1 Outils de recherche adhoc

La Figure 7 présente l'approche générale d'un outil de recherche dans un contexte adhoc. Cette approche peut être résumée en deux étapes :

- d'une part, les unités documentaires devant servir de support à la RI sont traitées pour être insérées dans la base d'indexation. L'ensemble des unités documentaires traitées par l'outil de recherche constituent la base d'indexation,
- d'autre part, l'utilisateur formule ses besoins sous la forme d'une ***requête*** (généralement sous la forme d'une liste de mots-clés).

L'outil de recherche identifie dans sa base d'indexation les unités documentaires pertinentes pour les besoins de l'utilisateur. Les unités sont proposées à l'utilisateur pour qu'il puisse à son tour les exploiter.

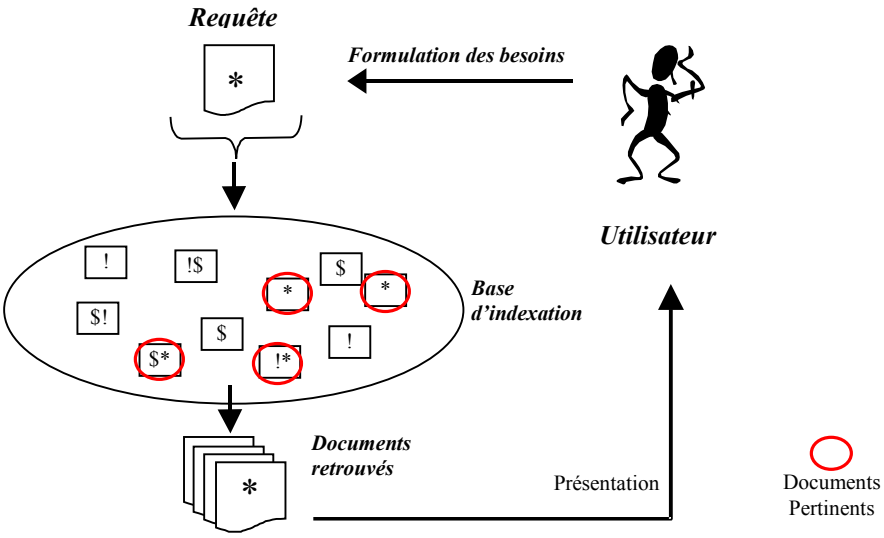


Figure 7 - Outil de recherche adhoc

Évaluation des outils de recherche adhoc

Pour une requête $\vec{Q}_i$ donnée, les unités documentaires de la base d'indexation peuvent être classées en quatre catégories (Figure 8) :

- les unités documentaires pertinentes pour la requête retournées par l'outil (DPR),
- les unités documentaires non pertinentes pour la requête retournées par l'outil (DNPR),
- les unités documentaires pertinentes pour la requête non retournées par l'outil (DPNR),
- les unités documentaires non pertinentes pour la requête non retournées par l'outil (DNPR),

De façon intuitive, il est aisé de comprendre que la meilleure recherche possible correspond à un résultat qui maximise le nombre de documents pertinents retrouvés (DPR) et minimise le nombre de documents non pertinents retrouvés (DNPR).

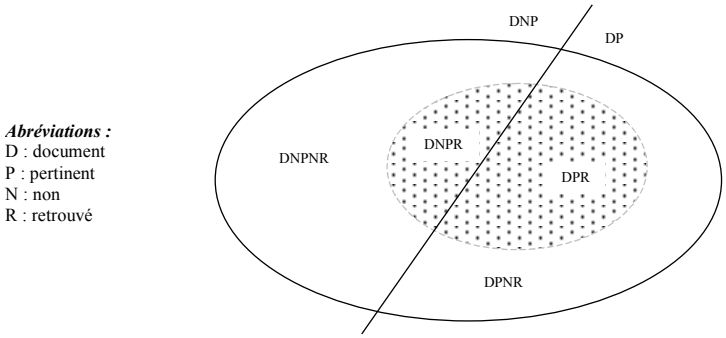


Figure 8 - Collection de documents selon leur pertinence et leur restitution par le SRI

Pour mesurer la qualité de la recherche, deux indicateurs sont généralement utilisés : le taux de rappel et le taux de précision [Salton, 1983].

$$\text{Taux de Rappel} = R = \frac{\text{Nombre de documents pertinents restitués}}{\text{Nombre de documents pertinents de la collection}} = \frac{DPR}{DPR + DPNR}$$

$$\text{Taux de Précision} = P = \frac{\text{Nombre de documents pertinents restitués}}{\text{Nombre de documents restitués}} = \frac{DPR}{DPR + DNPR}$$

Le taux de rappel évalue le pourcentage de documents pertinents qui ont effectivement été restitués par le système. Le taux de précision évalue le pourcentage de documents restitués qui sont pertinents. Une recherche optimale correspond à une valeur de 1 pour chacun de ces taux. Cependant, ces deux taux évoluent en sens inverse. En effet, nous pouvons, de façon intuitive, souligner que le taux de rappel peut augmenter avec un nombre de résultats plus important mais au prix d'un éventuel taux de précision plus faible.

Ces deux indicateurs sont liées à d'autres mesures complémentaires comme le bruit (complémentaire du taux de rappel) et le silence (complémentaire du taux de précision).

$$\text{Bruit} = \frac{\text{Nombre de documents non pertinents restitués}}{\text{Nombre de documents restitués}} = \frac{DNPNR}{DPR + DNPR}$$

$$\text{Silence} = \frac{\text{Nombre de documents pertinents non restitués}}{\text{Nombre de documents pertinents}} = \frac{DPNR}{DPR + DPNR}$$

Ces deux mesures peuvent être combinées en une seule valeur au travers de la mesure F_β.

$$F_\beta = \frac{(\beta^2 + 1)P.R}{\beta^2 P + R}$$

Où β est une valeur entre 0 et ∞. F_0 correspond à la précision et F_∞ correspond au rappel. Les valeurs les plus courantes de β sont 0.5 (la précision est privilégiée par rapport au rappel), 1 (le rappel et la précision sont d'égale importance), 2 (le rappel est privilégié par rapport à la précision).

Dans notre mémoire, nous avons fréquemment fait appel à la fonction F_1 qui correspond à :

$$F_1 = \frac{2P.R}{P + R}$$

La Figure 9 montre la forme générale des courbes de précision, rappel et F_1.

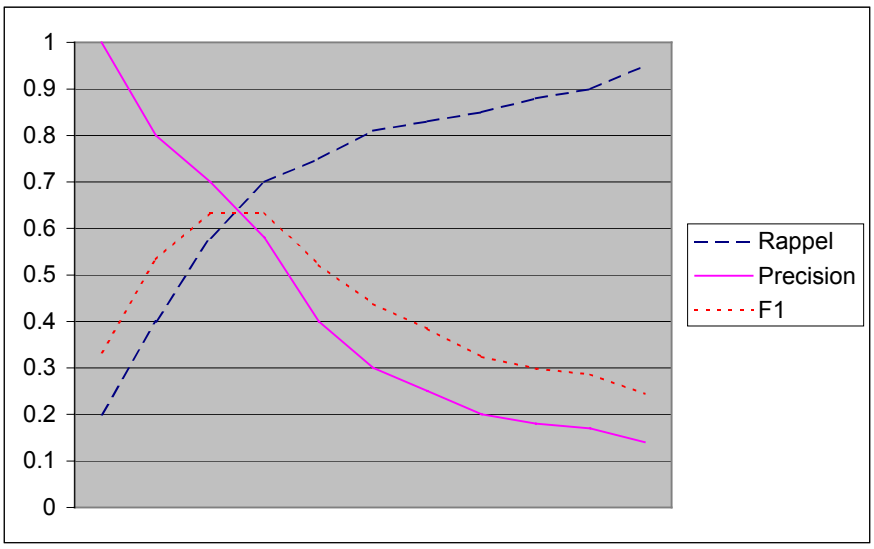


Figure 9 - Forme des courbes F1, Précision et Rappel

I.3.1.2.3.2 Outils de filtrage

Le but d'un outil de filtrage est, à partir d'un flux d'unités documentaires, d'identifier celles qui sont susceptibles d'être pertinentes pour un utilisateur par rapport à son profil et une fonction de décision. La démarche générale d'un outil de filtrage est présentée dans la Figure 10.

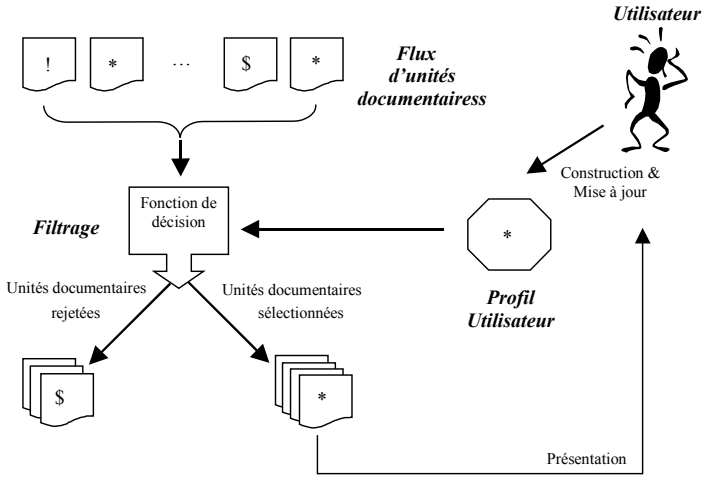


Figure 10 - Processus général d'un outil de filtrage

Au début du processus, l'utilisateur définit ses besoins en information qui sont traduits par le système en un *profil utilisateur*. Pour [Korfhage, 1997], un profil utilisateur est un indicateur des besoins en information, durables ou récurrents, qui sont représentés communément sous la forme d'une liste de mots-clés pondérés.

A partir du flux de documents entrants, le système apprécie l'appariement entre le profil utilisateur et les documents. Grâce à une fonction de décision, le système est en mesure de décider, par rapport à l'appariement mesuré, si un document est pertinent ou non pertinent

pour l'utilisateur. Les documents jugés comme pertinents seront ensuite proposés à l'utilisateur.

Malone dans [Malone, 1987] a mis en évidence trois modes de filtrage :

- le filtrage coopératif. Le filtrage coopératif se base sur des règles de sélection permettant d'évaluer un document selon les jugements de pertinence émis par d'autres utilisateurs,
- le filtrage économique. Le filtrage économique se base sur les notions de coût et d'intérêt relatifs à la production et la lecture d'un document. Le but est de filtrer l'information pour limiter le coût et maximiser l'intérêt. Le coût et l'intérêt sont représentés par des constantes définies par l'utilisateur,
- le filtrage cognitif. Le filtrage cognitif (ou basé sur le contenu) suppose que chaque utilisateur opère indépendamment les uns des autres. Par ailleurs, le système n'utilise que le contenu des documents pour les représenter.

Le profil utilisateur ainsi que la fonction de décision sont des points essentiels dans le filtrage. Or, à l'initialisation du programme, le système ne possède aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision. De même, le système ne dispose pas, au départ, d'une représentation exacte des besoins de l'utilisateur.

Ainsi, pour obtenir ces éléments, deux types de filtrages ont été proposés :

- type synchrone ou adaptatif. Les éléments sont déduits des unités documentaires filtrées et cumulées dans le temps,
- type asynchrone ou différé. Les éléments sont déduits d'une collection de documents existante. Cette collection fournit des exemples de documents pertinents pour les besoins de l'utilisateur.

I.3.1.2.3.3 Dualité entre les outils de recherche adhoc et les outils de filtrage

Cette dualité entre les outils de recherche adhoc et les outils de filtrage repose sur le fait que :

- un outil de recherche adhoc suppose l'existence d'une collection de documents alors qu'un outil de filtrage repose sur un ensemble de profils utilisateurs,
- un outil de recherche adhoc est utilisé de façon ponctuelle, le besoin en information est de même unique et temporaire tandis qu'un outil de filtrage utilise des besoins à long-terme,
- un outil de recherche adhoc utilise et organise des informations tandis qu'un outil de filtrage vise à les diffuser,
- un outil de recherche adhoc repose sur une base d'indexation statique alors qu'un outil de filtrage utilise des informations provenant d'un flux dynamique,
- un outil de recherche adhoc permet de décider si un document est intéressant ou non plutôt que d'aller chercher les documents intéressants,
- un outil de recherche adhoc repose, comparativement, sur une interaction importante avec l'utilisateur qui consulte les résultats de recherche, juge ces résultats... Au contraire, un outil de filtrage est peu interactif puisque l'utilisateur consultent les documents proposés par le système de façon périodique,

- un outil de recherche adhoc peut proposer à l'utilisateur la liste de tous les documents ordonnés par pertinence système alors que l'outil de filtrage doit décider si un document est pertinent ou non.

Le Tableau 1 résume cette dualité entre les outils de recherche adhoc et les outils de filtrage [Belkin, 1992].

	Adhoc	**Filtrage**
Besoin en information	*Momentané*	*Permanent*
Collection de documents	*Statique*	*Dynamique*
Interaction	*Très intéractif*	*Peu interactif*

Tableau 1 - Dualité entre les outils de recherche adhoc et les outils de filtrage

I.3.1.2.4 Plates-formes d'évaluation

Afin d'évaluer les systèmes les uns par rapport aux autres, diverses plates-formes proposent un cadre d'évaluation entre les différents systèmes. Parmi les plus importantes, nous pouvons citer la plate-forme *TREC* (Text Retrieval Conference) [Voorhees, 2001] qui propose un cadre expérimental afin d'évaluer différentes applications de la Recherche d'Information (filtrage, recherche adhoc...) ou la plate-forme *CLEF* (Cross-Language Information Retrieval and Evaluation) [Peters, 2000] qui propose un cadre d'évaluation spécialisé dans la RI multilingue.

I.3.2 La Recherche d'Information sur le web

Après avoir décrit les concepts généraux de la RI, nous présentons dans cette section les particularités de celle-ci appliquée au web.

Sur le web, les outils de recherche adhoc correspondent aux moteurs de recherche. Ils reposent sur une méthode d'accès à l'information de type PULL : l'utilisateur suit une démarche active pour retrouver des documents répondant à ses besoins. Il existe toute une panoplie de moteurs de recherche sur le web qui se différencient notamment par la taille de leur base d'indexation, leur langage d'interrogation, le type d'indexation utilisée...

Les outils de filtrage, quant à eux, sont généralement nommé des outils PUSH : à l'instar des moteurs de recherche, les outils PUSH proposent automatiquement des documents pertinents à un utilisateur passif ayant initialement formulé ses besoins.

I.3.2.1 Caractéristiques de la recherche d'information sur le web

L'utilisateur est au centre du processus de recherche d'information. Il intervient à différents niveaux (formulation de la requête, étude des résultats...) et de lui dépend en partie le résultat de la recherche. Or, chaque utilisateur est différent et certaines aptitudes sont nécessaires pour le bon achèvement de sa tâche de recherche.

D'un point de vue général, souligne l'impact de la diversité humaine [Shneiderman, 1998] sur l'utilisation d'une application informatique aux travers des aspects :

- physiques et lieu de travail (relatifs à l'utilisation de l'interface),
- cognitifs et sensoriels (mémorisation, apprentissage...).

L'intelligence humaine n'est pas citée ici car cette notion est très controversée et difficile à évaluer. Par ailleurs, les capacités de l'utilisateur à effectuer une tâche donnée peuvent être altérés par des facteurs comme la fatigue, la monotonie.... D'autres aspects comme la personnalité, la culture peuvent également influer sur les capacités d'un utilisateur à effectuer convenablement une tâche donnée.

Dans le contexte de la RI, ces aspects sont difficilement pris en compte du fait de l'inexistence de capteurs et d'indicateurs permettant de les mesurer. Cependant, nous pouvons souligner deux éléments conditionnant une RI sur le web c'est-à-dire la *connaissance pratique* et la *connaissance du domaine.* Ces deux connaissances jouent un rôle important dans l'apprentissage et les performances de l'utilisateur. Höschler [Höschler, 2000], souligne que ces connaissances sont les caractéristiques humaines essentielles de la RI sur le web.

La *connaissance pratique* représente la connaissance du web avec tout ce que cela comporte. Nous assimilons à cette catégorie la maîtrise du navigateur, l'utilisation des liens hypertextes, des fonctionnalités offertes par les outils de recherche... [GVU, 1998] souligne également le fait qu'il y a une grande différence entre les utilisateurs novices et experts notamment due à leur *connaissance pratique*. Cette étude souligne que plus un utilisateur est expert plus il utilise la quantité d'outils mis à sa disposition sur le web tandis qu'un utilisateur novice se limite à un moteur de recherche par exemple. Cette connaissance pratique s'acquiert et progresse essentiellement par la pratique.

La *connaissance du domaine* correspond à la connaissance que possède un utilisateur sur les thèmes relatifs à ses besoins. Elle permet une bonne formulation des requêtes (sélection des termes les plus appropriés pour trouver des documents pertinents) ainsi qu'une meilleure évaluation des documents visités [Pejtersen, 1998]. Par exemple, le passage des besoins « mentaux » aux besoins « explicites » (sous forme de mots-clés) est un réel problème puisque le vocabulaire utilisé influe directement sur les résultats de la recherche. Des termes trop généraux risquent de générer un nombre de résultats trop important (probabilité d'obtenir un fort bruit). A l'inverse, des termes trop spécifiques risquent de générer un nombre de résultats faible voire nul (probabilité d'obtenir un fort silence). Par ailleurs, un vocabulaire en inadéquation avec les besoins réels de l'utilisateur risque donc tout simplement de produire des documents inadaptés aux besoins réels de l'utilisateur. [Pejtersen, 1998] souligne également le fait que les utilisateurs n'ayant pas ou peu de connaissances dans un domaine ont du mal à définir les termes le caractérisant et à concevoir une stratégie de recherche. De plus, cette étude montre que les usagers ont du mal à évaluer si un document correspond ou non au thème recherché.

Concernant la recherche en elle-même, une étude menée par *CommerceNet* et *Nielsen Media*, en juillet 1997, démontre que 70% des informations trouvées proviennent d'un moteur de recherche.

Par ailleurs, les outils de recherche sur le web utilisent une collection de documents « locale ». En effet, le nombre de documents est si important sur le web qu'il n'est pas

envisageable de stocker le contenu de tous les documents. C'est pour cela que les outils de recherche adhoc sur le web utilisent généralement une collection virtuelle de documents. C'est-à-dire qu'ils ne conservent qu'un minimum d'informations concernant les documents (URL, termes d'indexation...). De plus, la structure hypertexte sur laquelle repose le web peut être intégrée au niveau du processus d'indexation [Li, 1997], [Gery, 1999]. Cette indexation repose généralement sur une indexation automatique grâce à des robots d'indexation nommés « crawlers » ou « spiders » qui parcourent le web à la recherche de nouveaux documents à indexer. Le modèle de recherche communément utilisé est le modèle vectoriel.

De ces caractéristiques découlent des problèmes de la RI sur le web. Par exemple, le fait d'utiliser une collection virtuelle implique que le moteur de recherche peut proposer des documents qui n'existent plus ou qui ont été modifiés.

La section suivante présente les différents problèmes liés à la RI sur le web.

I.3.2.2 Les problèmes de la RI sur le web

[GVU, 1998] souligne les problèmes auxquels les utilisateurs peuvent être confrontés lorsqu'ils utilisent le web. Selon eux, les problèmes rencontrés par les internautes et présentés dans la Figure 11 sont :

- l'impossibilité de trouver des informations recherchées (**1**),
- l'impossibilité d'organiser efficacement les informations retrouvées (**2**),
- l'impossibilité de trouver une page dont on connaît l'existence (**3**) ,
- l'impossibilité de revenir à un document déjà visité (**4**),
- l'impossibilité de déterminer où l'utilisateur se situe (perdu dans l'hyper-espace) (**5**),
- l'impossibilité de visualiser où l'utilisateur est allé, où il peut aller (visualisation de portions du site web visité par exemple) (**6**),
- la rencontre de liens ne fonctionnant pas (liens morts) (**7**).

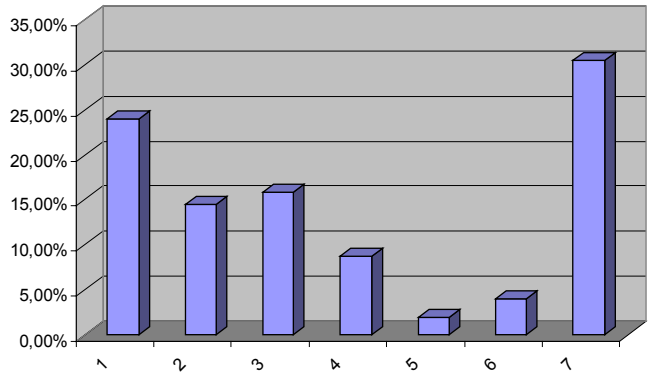


Figure 11 - Problèmes liés à l'utilisation du web [GVU, 1998]

Nous n'avons pas fait apparaître, dans la Figure 11, les problèmes soulevés par l'étude qui sont liés à la technologie (coût de la communication, vitesse de téléchargement...) ou liés à la conception des documents (code HTML non valide, documents non compatibles avec le navigateur...).

Cette étude nous montre qu'environ un quart des utilisateurs ne sont pas satisfaits de leur recherche dans le sens où ils ne retrouvent pas les informations recherchées (1 & 3). De

plus, l'utilisateur a du mal à organiser et à réutiliser les informations retrouvées (2). Quant aux problèmes 5 et 6, ils sont principalement dûs à la surcharge cognitive qu'implique une navigation hypertexte. Le problème 7 provient directement de la structure du web et surtout de l'évolution rapide à laquelle sont soumis les documents sur ce média (les auteurs modifient le contenu, déplacent les documents...).

Cependant, cette étude ne nous permet pas d'identifier les réels problèmes de la recherche d'information sur le web. Elle ne nous permet de répondre qu'en partie à la question « ***Que faut-il prendre en compte pour effectuer une recherche d'information efficace ?*** ». Nous qualifions une recherche d'efficace si elle permet à l'utilisateur de trouver rapidement un maximum d'informations répondant à ses besoins.

Pour cela, nous nous proposons d'identifier les différents problèmes et facteurs influant sur une recherche sur le web pour ensuite souligner les différentes approches visant à limiter voire à supprimer ces problèmes. La Figure 12 reprend le processus de recherche d'information avec les divers éléments pouvant poser problème. Ainsi, la connaissance du domaine de recherche, le processus de recherche mais également la gestion des informations retrouvées met l'utilisateur face à un grand nombre d'écueils qu'il est important d'éviter.

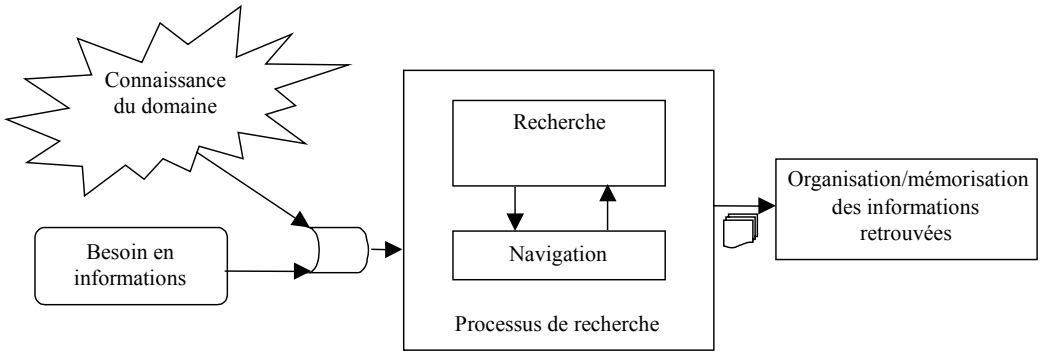


Figure 12 - Processus général de la recherche sur le web

Dans un premier temps, nous rappelons que la principale limite de la RI correspond au facteur humain (connaissance pratique et connaissance du domaine). En effet, ces connaissances ont un fort impact sur les résultats d'une recherche d'information.

Dans la section suivante, nous présentons les problèmes liés aux outils de RI sur le web ainsi que le problème d'ordre plus général qui est la gestion de l'information mémorisées.

I.3.2.2.1 Problèmes liés au processus de recherche sur le web

Nous avons vu précédemment que pour effectuer une recherche, l'utilisateur s'appuie sur une navigation hypertexte et sur une recherche adhoc utilisées de façon alternées. Nous n'avons pas détaillé les problèmes liés à la navigation hypertexte car nous en avons déjà souligné les limites (surcharge cognitive et désorientation). Par contre, les difficultés qui peuvent être rencontrées lors du processus de recherche par interrogation nécessitent un approfondissement.

Les problèmes liés à la tâche de recherche sont les suivants :

- *la couverture du web.* Les outils de recherche n'indexent qu'une partie limitée du web [Lawrence, 1999], [Sullivan, 2001], [Notess, 2002]. Une première raison à cette restriction provient du nombre très important de documents disponibles. Une seconde raison est l'incapacité qu'ont les robots d'indexation à indexer les documents du web caché. En effet, ces documents ne sont accessibles généralement que par le biais de formulaires que le robot ne peut pas automatiquement remplir [Seltzer, 1997]. Les robots ne se concentrent donc que sur le web visible (portion limitée du web global),
- *le chevauchement des bases de documents des différents outils de recherche.* Le chevauchement entre les bases d'indexation des différents outils de recherche est relativement faible [Notess, 2000], ce qui signifie que chacun des outils a préalablement indexé des documents différents et que pour une même requête il va retourner des documents différents des autres outils. Par exemple, l'étude de Notess [Notess, 2000] indique, suite à une expérimentation sur quatorze moteurs de recherche, que plus de 35% des résultats n'a été retrouvé que par un seul moteur,
- *la mise à jour des bases d'indexation des outils de recherche*. Un facteur important dont pâtissent les outils de recherche est l'évolution rapide des documents. Par ce fait, les outils de recherche n'ont pas une base d'indexation à jour et ils proposent à l'utilisateur un grand nombre d'URLs de documents déplacés, supprimés voire obsolètes. Ce point fait référence au problème numéro 7 de la Figure 11,
- *la présentation des résultats à l'utilisateur*. Même s'ils ne couvrent qu'une partie du web, les outils de recherche indexent plusieurs millions de documents. De ce fait, suite à une recherche, l'utilisateur se retrouve souvent avec des milliers de documents pertinents (du point de vue système) pour ses besoins. Or, les outils actuels utilisent communément des listes de résultats pour présenter ces derniers. Ce mécanisme ne s'avère pas adapté à la compréhension du résultat dans sa globalité car une liste de résultats ne présente que quelques dizaines de résultats par page.

Outre ces difficultés de recherche liées à la technologie, nous avons identifié les difficultés d'ordre général liées à la gestion et à l'organisation des résultats retrouvés qui ont une incidence sur la RI.

I.3.2.2.2 Problèmes rencontrés lors de la gestion des informations mémorisées

L'être humain éprouve une nécessité à organiser les informations qui l'intéressent afin de pouvoir les retrouver et les réutiliser plus facilement. Ainsi, l'utilisation de marque-pages, d'annotations dans un article de « papiers à coller » est omniprésente dans la vie courante. Nous pouvons également voir cela dans le besoin qu'ont les personnes de posséder une bibliothèque de documents classée selon une organisation qui leur est propre.

La RI doit répondre à cette attente car l'internaute éprouvent également dans ce contexte le besoin d'organiser les informations retrouvées sur le web afin d'y accéder facilement de nouveau, de les partager etc.

I.4 Approches existantes pour l'aide à la RI sur le web

Au regard des différents problèmes évoqués précédemment, nous avons, dans cette section, dresser un panorama des différentes approches développées visant à limiter voire annihiler les problèmes rencontrés lors d'une RI sur le web.

Dans un premier temps, nous nous sommes intéressés aux les approches qui tiennent compte des facteurs humains c'est-à-dire celles qui visent à aider l'utilisateur à acquérir de meilleures connaissances (pratique et du domaine). Dans un second temps, nous avons présenté les outils permettant d'aider l'utilisateur lors du processus de RI sur le web (navigation, recherche adhoc). Nous avons détaillé, dans une troisième partie, les approches permettant à l'utilisateur de mémoriser des documents web. Enfin, pour conclure, nous avons souligné l'aspect coopératif qu'il est possible de mettre en œuvre lors de la Recherche d'Information sur le web afin que celle-ci devienne moins solitaire.

I.4.1 Les facteurs humains

Nous avons précédemment souligné que l'utilisateur doit posséder une bonne connaissance pratique et une bonne connaissance du domaine afin de réaliser des recherches efficaces.

La connaissance pratique peut être améliorée au cours de formations, grâce à la lecture d'ouvrages spécifiques ou de sites web. Par exemple, le site web *LearnTheWeb* (http://www.learnthenet.com/french/index.html) peut être utilisé pour parfaire cette connaissance.

Contrairement à la connaissance pratique, la connaissance du domaine est sujette à une forte évolution. En effet, les centres d'intérêt de l'utilisateur évoluent au même titre que les informations s'y référant.

La solution idéale pour augmenter et affiner la connaissance d'un centre d'intérêt est de faire régulièrement des recherches sur le web pour être au courant des dernières évolutions dans le domaine. Cependant, cette démarche demande de la part de l'utilisateur un lourd investissement.

Une alternative pour réduire en partie cet investissement, consiste à utiliser des outils basés sur le principe de *Push* ou autres agents intelligents. En effet, ces outils permettent de présenter automatiquement et de manière permanente des documents répondant aux centres d'intérêt de l'utilisateur. Des exemples d'applications *Push* sont disponibles pour les articles de journaux tels que [Kamba, 1995]. Dans le contexte de la recherche d'information, nous pouvons citer l'exemple du méta-moteur de recherche *ProFusion* (http://www.profusion.com/) qui propose à l'utilisateur un système d'alertes qui le prévient dès qu'une nouvelle information relative à ses centres d'intérêt apparaît au sein des bases d'indexation des moteurs utilisés. Ces avertissements sont effectués par courrier électronique.

Du fait que la connaissance pratique est peu sujette à évolution, nous ne nous sommes consacrés, dans ce document, qu'à la connaissance du domaine afin d'aider un utilisateur lors d'une recherche d'information.

I.4.2 Le processus de recherche

Nous avons présenté dans cette section les approches proposées afin d'améliorer le processus de recherche. Dans un premier temps, nous avons détaillé les approches visant à améliorer la navigation au travers de :

- la réduction de l'effort cognitif nécessaire,
- l'aide à l'orientation.

Dans un second temps, nous avons présenté les approches liées à la recherche adhoc au travers :

- des aides à la formulation des besoins,
- des aides à la sélection des outils de recherche,
- des méta-moteurs de recherche,
- des interfaces de visualisation des résultats de recherche.

Enfin, nous avons souligné l'approche des agents de recherche et de recommandation.

I.4.2.1 La tâche de navigation

La navigation souffre essentiellement de problèmes provenant de l'architecture même sur laquelle elle repose c'est-à-dire sur la notion d'hypertexte. Ces problèmes sont la surcharge cognitive et la désorientation. Pour chacun d'eux, nous avons présenté les approches visant à limiter ces problèmes

I.4.2.1.1 La surcharge cognitive

Afin de limiter l'effort cognitif induit par l'hypertexte, divers outils ont été mis en œuvre pour garder une trace et un cheminement des documents visités.

I.4.2.1.1.1 Liste « historique »

Les principaux navigateurs proposent un historique des différents documents visités lors de la navigation de l'utilisateur présenté sous forme d'une liste. Cette liste peut être organisée par site, par jour ou encore par ordre de visite. Elle permet à un utilisateur de limiter l'effort cognitif nécessaire lors de la navigation car elle offre une vision des différents points de passage de l'hypertexte par lequel il est passé. Ainsi, l'utilisateur peut aisément revenir à un document précédemment visité dans l'hypertexte.

Un projet intéressant lié à ce problème a été proposé au travers de *BookMap* [Hascoët, 2000] (Figure 13).

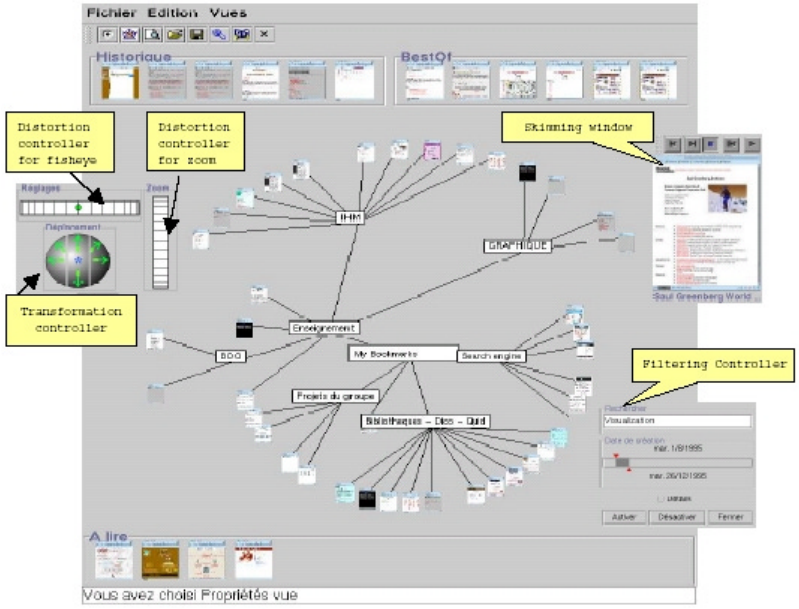


Figure 13 - *Bookmap* [Hascöet, 2000]

Ce projet permet de façon intégrée :

- de visualiser sous forme de graphe les hiérarchies de signets,
- de mémoriser les documents « à lire »,
- de visualiser directement l'historique,
- d'avoir un accès direct aux documents les plus visités via leur URL (« best of »).

Ce projet est intéressant par le fait qu'il utilise des « thumbnails » (images réduites des documents) au lieu des traditionnelles URLs ou titres des documents. Par ce biais, l'utilisateur peut avoir visuellement un aperçu d'un document, ce qui en facilite ainsi la remémoration.

Une limite de la liste historique est qu'elle ne permet pas à l'utilisateur de s'affranchir totalement de l'effort cognitif lié à la navigation. En effet, les différents documents sont présentés de façon indépendante.

I.4.2.1.1.2 Visualisation de la navigation

La visualisation de la navigation, est une évolution des historiques. Elle permet de représenter graphiquement les documents visités ainsi que les liens éventuels entre ces documents. Nous pouvons citer comme exemple *WebMap* [Dömel, 1994]. Grâce à de tels outils, l'utilisateur peut visualiser non seulement les documents qu'il a précédemment visités mais également l'organisation de ces documents.

I.4.2.1.2 Désorientation

Le fait de se perdre dans l'hypertexte provient du fait que l'utilisateur perd le cheminement de sa navigation car il ne comprend plus l'organisation de l'hypertexte local. Pour éviter cela, une cartographie de l'hypertexte local a été proposée. Au lieu de ne présenter que les documents visités, ces cartographies visent à présenter les documents visités au sein de leur hypertexte local. Ainsi, les documents visités apparaissent avec les documents liés par des liens hypertextes afin que l'utilisateur puisse avoir une vision globale de

l'hypertexte « local » dans lequel il se trouve. Comme exemple d'outils de cartographie citons *Hyperspace* [Wood, 1995] ou *Internet Cartographer* (http://www.inventix.com/).

Grâce aux différentes approches proposées dans la littérature, la tâche de navigation peut être réalisée dans de meilleures conditions. Mais la RI sur le web fait également appel à une tâche de recherche. La section suivante présente les approches visant à améliorer cette tâche de recherche.

I.4.2.2 La tâche de recherche

I.4.2.2.1 La formulation des besoins

Le premier point qu'il est important de prendre en compte lors d'une RI est la formulation des besoins sous la forme d'une requête. En effet, quel que soit l'outil de recherche utilisé, quel que soit le domaine, il est nécessaire pour l'utilisateur d'expliciter au mieux ses besoins pour obtenir des résultats pertinents. Une inadéquation entre les besoins réels et la requête peut être la cause d'un grand nombre d'échecs en RI par interrogation. De façon générale, la formulation de la requête repose sur l'utilisation d'un pseudo langage naturel. Ainsi, l'utilisateur utilise des mots simples (mots-clés) visant à représenter ses besoins.

Les limites de cette formulation proviennent essentiellement du fait que :

- l'utilisateur n'a pas une bonne connaissance du domaine de recherche,
- l'utilisateur ne connaît pas le contenu a priori de la base d'indexation des outils de recherche,
- l'utilisateur peut avoir du mal à traduire la représentation mentale de ses besoins en une représentation sous la dorme d'une liste de mots-clés,
- l'utilisateur n'utilise que peu de termes pour représenter ses besoins. Différentes études [Silverstein, 1998], [Jansen, 2000], [Spink, 2002] montrent qu'en moyenne moins de trois mots-clés sont utilisés pour formuler la requête. Ce petit nombre de termes ne suffit généralement pas pour représenter un besoin en information de l'utilisateur.

Pour aider l'utilisateur à franchir ce premier obstacle que représente l'expression de ses besoins, il peut faire appel à divers procédés tels que l'interrogation par médiation, la classification thématique, la requête dynamique.

I.4.2.2.1.1 Interrogation par médiation

Une première approche est l'interrogation des outils de recherche par médiation. L'idée réside dans le fait que plutôt que de demander à l'utilisateur de formuler ses besoins sous la forme de mots-clés, celui-ci choisit dans un ensemble de classes de documents celles qui correspondent à ses besoins. A partir des classes sélectionnées, le système génère automatiquement la requête correspondante. Comme application de ce principe de médiation, le système *WebCluster* [Mechkour, 1998] permet de générer une requête qu'il propose soit directement à un outil de recherche, soit à l'utilisateur pour lui permettre de la modifier. Par ce biais, l'utilisateur peut s'affranchir de la formulation d'une requête.

I.4.2.2.1.2 Classification thématique

Tout outil de recherche propose généralement aujourd'hui à la fois un service de recherche adhoc basé sur une requête et une classification thématique des documents (ex. Yahoo !). L'utilisateur peut ainsi trouver dans cette approche une alternative agréable à la recherche via une requête en parcourant les thèmes l'intéressant.

Le système *Cat-a-cone* [Hearst, 1997] est un exemple d'application qui offre à l'utilisateur la possibilité de soit formuler une requête à l'aide d'un ensemble de mots-clés, soit naviguer visuellement, au travers d'une interface 3D, dans une hiérarchie de thèmes décrivant les documents de la base d'indexation.

I.4.2.2.1.3 Requête dynamique

Une approche attrayante de l'aide à la formulation de requêtes est également proposée au travers des outils de requêtes dynamiques par interactions. Par leur biais, l'utilisateur peut interroger l'outil de recherche par tâtonnement. A partir d'une requête, l'utilisateur peut modifier les composantes de celle-ci et visualiser immédiatement le résultat des modifications apportées. Le projet *Dynamic HomeFinder* [Jain, 1994] qui permet de rechercher des maisons à vendre peut être cité à titre d'exemple.

Malheureusement, ces outils sont plutôt destinés à des informations factuelles car elles nécessitent de lourdes structures de données. Cette approche est difficilement adaptable à la RI textuelle.

I.4.2.2.2 Reformulation de requête

Même si la requête de l'utilisateur n'est pas optimale, diverses approches tentent de l'améliorer. En effet, les performances d'un SRI dépendent de façon non négligeable des requêtes formulées par l'utilisateur.

Or, le plus souvent, l'utilisateur formule ses requêtes en des termes qui lui sont propres, mais qui ne correspondent pas forcément à ceux utilisés pour indexer les documents pertinents de la base d'indexation du SRI. Alors que pour sélectionner le maximum de documents pertinents en limitant le bruit, il faudrait que l'utilisateur formule ses besoins à partir de termes pertinents directement issus du langage d'indexation du SRI. Cette tâche s'avère difficile dans la mesure où, en règle générale, sur de gros corpus, il est impossible de connaître le langage d'indexation utilisé.

On peut en conclure que, compte tenu des volumes croissants des bases d'informations, retrouver les informations pertinentes en utilisant seulement la requête initiale est une opération quasi-impossible : le taux de précision obtenus dans la tâche adhoc de TREC ne dépassent pas 30% [Voorhees, 1999].

En conséquence, de nombreuses recherches passées et actuelles visent à concevoir des SRI capables de s'adapter aux besoins de l'utilisateur (via le concept de profil par exemple). Ces systèmes devraient également être capables de déterminer son but de recherche afin de l'aider à cibler son besoin.

La reformulation de requête est un processus ayant pour objectif de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur. Cette reformulation permet de coordonner le langage de recherche (utilisé par l'utilisateur dans sa requête) et le

langage d'indexation du SRI. Par conséquent, elle limite le bruit et le silence dus à un mauvais choix des termes d'indexation dans l'expression de la requête d'une part, et les lacunes du processus d'indexation d'autre part.

Nous distinguons principalement deux approches de reformulation de requêtes, (1) selon qu'elles utilisent les associations entre les termes, ou (2) la pertinence et non pertinence des documents restitués en réponse à une requête initiale. Les deux principales techniques utilisées sont respectivement l'expansion de requête et la réinjection de la pertinence.

I.4.2.2.2.1.1 Expansion de requête

Ce mécanisme se base sur le principe suivant : la simple comparaison du contenu de la requête et des documents de la base d'indexation ne permet pas d'avoir tous les documents correspondant à la requête. Il reste toujours des documents pertinents non restitués par le SRI.

Des travaux de recherche ont proposé de reformuler la requête initiale par l'ajout des termes sémantiquement proches. Ces derniers sont issus :

- soit d'études sur le langage naturel (variantes morphologiques...). Il est ainsi possible d'ajouter à la requête des variantes morphologiques des différents termes employés par l'utilisateur. le but de ce mécanisme est d'assurer la restitution des documents indexés par des variantes des termes composant la requête. Dans ce cadre, on utilise des algorithme de radicalisation et de troncature,
- soit d'études statistiques et d'analyses sur les contenus des documents de la base. On peut ainsi choisir d'ajouter un certains nombre de termes les plus pertinents des documents sélectionnés, ou de n'en conserver qu'un nombre limité parmi les termes initiaux et rajoutés.

De même, une autre méthode propose d'ajouter des termes voisins ou des termes associés à ceux de la requête. Il s'agit de chercher des associations inter-termes (corrélation entre les termes, classification des termes...).

I.4.2.2.2.1.2 Réinjection de la pertinence et/ou non pertinence

Lorsque les documents sont restitués en réponse à une requête initiale formulée par un utilisateur, ce dernier peut fournir des jugements de pertinence les concernant en indiquant ceux qu'il juge pertinents et ceux qu'il juge non pertinents. Ces jugements sont alors utilisés dans le but de reformuler la requête. Cette méthode est plus connue sous le nom réinjection de pertinence (ou « relevance feedback » en anglais) [Salton, 1983]. C'est un processus évolutif et interactif. Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche, puis modifier celle-ci à partir des jugements de pertinence et/ou non pertinence de l'utilisateur sur les documents restitués, soit pour repondérer les termes de la requête initiale, soit pour y ajouter (respectivement supprimer) d'autres termes contenus dans les documents pertinents (respectivement non pertinents). La nouvelle requête, obtenue à chaque itération du feedback, permet de corriger la direction de la recherche dans le sens des documents pertinents.

Le processus de réinjection de pertinence peut être adapté aux différents modèles de recherche comme le modèle vectoriel avec l'approche de Rocchio [Rocchio, 1971], le modèle

connexionniste [Boughanem, 2000] mais peut également être réalisé par l'approche des algorithmes génétiques [Tamine, 2000].

I.4.2.2.3 Sélection de l'outil de recherche

Après avoir pris conscience de la façon dont ses besoins peuvent être formulés, l'utilisateur a la lourde tâche de sélectionner l'outil de recherche qu'il souhaite interroger. Or, les différents outils de recherche (moteurs de recherche, annuaires) disposent, au sein de leur base d'indexation, des mêmes documents. Ceci se traduit par un faible recouvrement des bases d'indexation. La sélection de l'outil de recherche conditionne les résultats de la recherche : il est intuitivement facile à concevoir qu'utiliser un outil de recherche généraliste pour une requête dans un domaine spécifique donnera vraisemblablement des résultats moins « bons » que la même requête posée sur un outil de recherche spécialisé dans le domaine. L'utilisateur doit donc choisir au mieux l'outil qu'il va interroger. *GloSS* [Gravano, 1999], par exemple, est un système permettant à partir d'une requête de rechercher les sources d'informations textuelles spécialisées dans le domaine de recherche. Ce système se limite à identifier les sources d'informations les plus pertinentes pour la requête. L'utilisateur doit ensuite interroger la source d'informations choisie(s) pour obtenir les documents répondant à ses besoins. La principale limite de cette approche est que le système a besoin d'un accès aux bases documentaires des différentes sources pour pouvoir qualifier leur contenu.

Cependant, quel que soit l'outil de recherche interrogé et du fait de cette faible couverture des bases d'indexation, l'utilisateur ne peut se contenter d'interroger un seul outil de recherche. Ainsi, un problème de la recherche adhoc réside dans le fait de bien savoir choisir les outils de recherche pour obtenir les meilleurs résultats puis de réaliser une synthèse des résultats obtenus. Si elle est réalisée manuellement, cette tâche s'avère longue et fastidieuse surtout si chacun des outils de recherche retourne des milliers de documents, ce qui est assez courant à l'heure actuelle. Cette interrogation multiple est d'autant plus difficile à réaliser que chacun des outils de recherche propose son propre langage de requête et qu'il est nécessaire d'intercaler manuellement les résultats fournis par chaque moteur. Pour l'aider dans cette tâche, l'utilisateur peut faire appel à un *méta-moteur* de recherche d'information.

I.4.2.2.4 Les méta-moteurs de recherche d'information

Un méta-moteur de recherche d'information se présente à l'utilisateur comme un outil de recherche « classique ». Cependant, à partir d'une requête, le système crée une multitude de requêtes qu'il soumet en parallèle à un ensemble d'outils de recherche prédéfinis. Chacune d'elles correspond à la traduction de la requête initiale dans le langage spécifique de l'outil interrogé.

Du point de vue du résultat obtenu au travers de ces outils, les méta-moteurs peuvent être classés en diverses catégories [Andrieu, 1998] :

- les *aides à la saisie*. Ces outils proposent uniquement une traduction de la requête initiale pour un ensemble d'outils de recherche prédéfinis. Cependant les différents outils de recherche restent indépendants. L'utilisateur doit manuellement procéder à

l'interrogation des outils de recherche. *MetaSearch* (http://www.metasearch.com/) illustre bien cette catégorie,

- les *listes de résultats*. Ces outils soumettent la requête originale parallèlement à un ensemble d'outils de recherche. Les résultats sont présentés pour chaque outil de recherche indépendamment. Nous pouvons citer l'outil *Internet Sleuth* (http://www.isleuth.com),
- les *listes synthétisées*. Ces outils sont des listes de résultats qui fusionnent les résultats issus des différents outils interrogés en une seule liste de résultats. Les résultats en double sont supprimés et les résultats réorganisés. Pour illustrer cette catégorie nous pouvons citer l'outil *Copernic* (http://www.copernic.com) ou encore le service en ligne *MetaCrawler* (http://www.metacrawler.com/index.html).

Les deux premières catégories de méta-moteurs tendent aujourd'hui à disparaître au profit des listes synthétisées qui fournissent à l'utilisateur un confort d'utilisation optimum dans le sens où elles lui fournissent un résultat synthétique.

La plupart des méta-moteurs possèdent une liste de moteurs à interroger prédéfini que l'utilisateur peut sélectionner ou non pour la recherche. Cependant, la plupart des outils disponibles interrogent toutes les sources sélectionnées sans, à priori, vérifier si ces sources sont pertinentes pour la requête. Cette vérification permet de limiter le bruit éventuel. Une approche combinant l'aspect méta-moteur mais également l'aspect sélection des sources d'informations a été proposée par [Dreilinger, 1997].

Malgré ces outils de recherche « améliorés », certains problèmes liés à la RI persistent. En effet, comment l'internaute réagit-il face à un nombre de résultats dépassant le millier voire le million de documents ? Ce cas est plus que fréquent sur Internet du fait du nombre très important de documents disponibles mais aussi par les modèles de recherche inexacts utilisés (modèle vectoriel). Différentes études montrent que l'utilisateur ne parcourt que les premières pages de résultats (environ 30 documents) [Silverstein, 1998], [Spink, 2002] alors que des documents intéressants peuvent se situer au-delà. Il est donc important de prendre en compte la présentation à l'utilisateur des résultats de recherche (des millions de documents) dans le processus de la RI pour optimiser et faciliter la tâche de l'internaute. En réponse à cela, nous pouvons souligner l'intérêt des approches qui visent à étudier les résultats pour personnaliser la réponse proposée à l'utilisateur. Ainsi, le projet *Profildoc* [Lainé-Cruzel, 1999] permet de filtrer les documents retrouvés par rapport au profil de l'utilisateur. Ce système repose sur un profil utilisateur contenant des informations telles que le niveau éducationnel, le champ disciplinaire (Sciences de l'information, agronomie...), le type de recherche (recherche généraliste ou pointue)... Ce profil est utilisé afin d'identifier au sein des documents retrouvés ceux qui correspondent au profil de l'utilisateur et ainsi réduire le nombre de documents en éliminant ceux qui ne seraient pas pertinents pour l'utilisateur.

Malgré tout, le nombre de résultat reste très important et il peut s'avérer intéressant de présenter les résultats de façon globale au travers d'une interface de visualisation.

I.4.2.2.5 La visualisation des résultats de Recherche d'Information

Les résultats issus d'un moteur ou encore d'un méta-moteur de recherche d'information sont communément présentés sous la forme d'une liste (liste d'URLs).

Cette liste présente les différents résultats au travers (cf Figure 14) :

- d'un numéro de classement ou une appréciation de la pertinence système,
- d'un nom ou d'une URL associé à un lien hypertexte pour permettre à l'utilisateur d'accéder au document,
- d'un court descriptif présentant les premières lignes du document dans lesquelles apparaissent les termes de la requête.

Figure 14 - Liste de résultats issue de Yahoo ! (http://www.yahoo.fr**)**

Cet affichage est facile à mettre en œuvre et à utiliser mais n'est efficace que pour un nombre réduit de résultats (< 20) [Cugini, 2000]. Pour un nombre important, les listes de résultats souffrent essentiellement des limites suivantes [Dubin, 1995], [Zamir, 1998] :

- la position d'un document dans la liste ne permet pas explicitement de déduire sa similarité avec les autres documents de la liste,
- l'utilisateur peut avoir du mal à comprendre pourquoi le document a été inséré à cet endroit dans la liste et quelle est la relation avec la requête sans avoir visualisé son contenu.

A cela s'ajoute le fait que les résultats sont affichés page par page, ce qui ne facilite pas la vision globale du résultat. Pour obtenir une vision globale des résultats, l'utilisateur doit visiter chacun des documents, un à un, pour en apprécier la réelle pertinence et identifier les liens potentiels entre ceux-ci. De ce fait, il est compréhensible que l'utilisateur se contente en moyenne des 30 premiers résultats en occultant le reste des résultats (éventuellement pertinents) si le nombre de résultats est très important. Cependant, il occulte également un ensemble de documents potentiellement pertinents pour ses besoins.

C'est pour remédier à cet état de fait que des travaux ont été réalisés dans le domaine des interfaces de visualisation. Leur but est de proposer à l'utilisateur un moyen efficace pour apprécier et manipuler les résultats de recherche d'information dans leur globalité.

Diverses visualisations, que ce soit en mode texte, sous forme de représentation graphique dans un espace à deux ou trois dimensions, tentent de pallier aux limites des listes de résultats. Diverses classifications des interfaces de visualisation ont été proposées comme dans [Zamir, 1998] qui traite des visualisations pour la RI ou encore [Chi, 2000] et [Hascoët, 2001] qui effectuent une taxonomie des différentes techniques de visualisation générale d'informations.

Nous ne présentons dans cette section que la classification proposée par Zamir car elle traite du domaine de notre étude qui est celui de la Recherche d'Information.

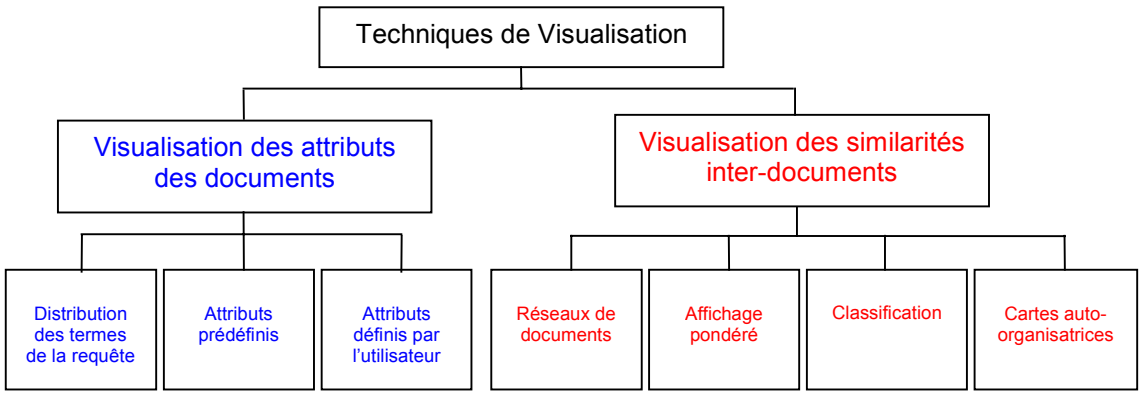


Figure 15 - Classification des techniques de visualisation [Zamir, 1998]

Cette classification met en évidence deux types de techniques selon leur but :

- les visualisations des attributs des documents (informations issues des documents). Il y a trois sous-catégories de techniques.
  - la distribution des termes de la requête. Elle permet de savoir comment chaque mot-clé utilisé dans la requête est réparti dans les documents,
  - les attributs prédéfinis. Elle permet de montrer la relation qu'a le document avec des attributs tels que la taille, l'auteur, etc.,
  - les attributs formulés par l'utilisateur. Elle permet de montrer la relation qu'a le document avec des critères choisis par l'utilisateur (requête par exemple...),
- les visualisations de similarité inter-documents. Il y a quatre sous-catégories.
  - les réseaux de documents. Les documents sont reliés entre eux selon leur similarité,
  - « les affichages pondérés ». Les documents sont répartis visuellement selon des forces qui les repoussent ou les rapprochent des autres par rapport à leur similarité,
  - les « classifications ». Ces visualisations représentent les documents sous forme de groupes de documents (par similarité de contenu, selon les liens hypertextes...),
  - les cartes auto-organisatrices (ou SOM). Ces techniques permettent d'afficher sur une « carte » 2D les documents par rapport à leur similarité de contenu.

Cette classification ne permet toutefois pas une catégorisation globale des interfaces de visualisation. En effet, la catégorie de la visualisation des similarités inter-documents n'est pas relative aux éléments visualisés mais aux techniques employées. Pour pallier à cela, nous avons modifié cette classification. Cette nouvelle classification fait abstraction délibérément des techniques de visualisation car leur nombre est très important et elles évoluent continuellement. Ainsi, la catégorie des visualisations des similarités inter-documents a été reconsidérée en s'appuyant sur la visualisation des documents les uns par rapport aux autres ou sur la visualisation des relations entre classes de documents. La branche concernant la visualisation a été conservée telle quelle (Figure 16).

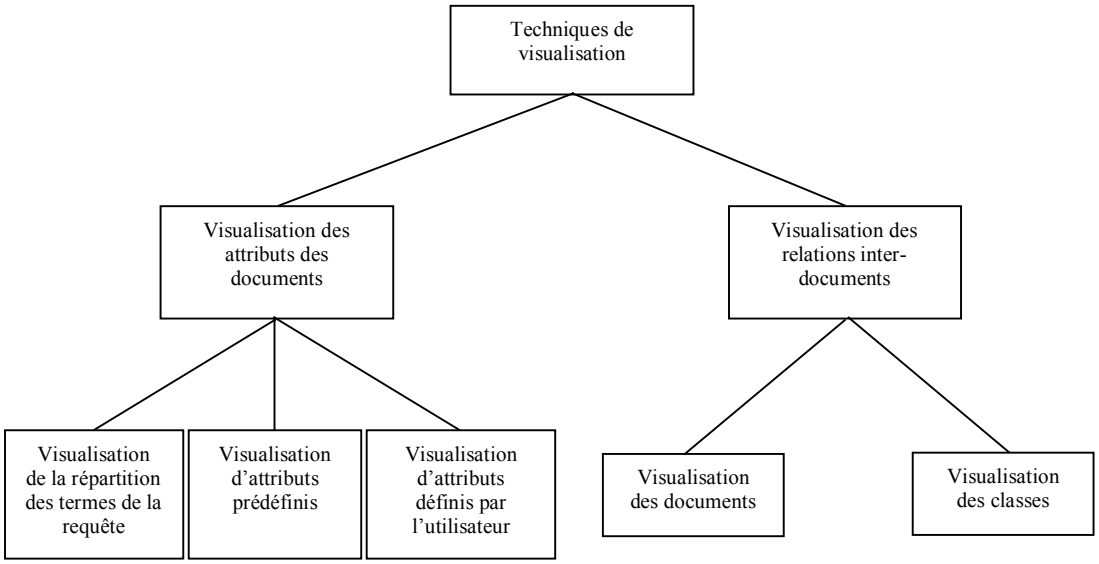


Figure 16 - Une nouvelle classification des techniques de visualisation

Pour chacune des catégories de cette dernière classification, nous avons donné succinctement dans ce qui suit une description ainsi que des exemples d'interfaces de visualisation. Cette présentation ne prétend, en aucun cas, fournir une liste exhaustive des différents projets menés en visualisation.

I.4.2.2.5.1 Visualisation des attributs des documents

I.4.2.2.5.1.1 Répartition des termes de la requête

Cette technique de visualisation vise à présenter la répartition des différents mots-clés de la requête au sein des documents retrouvés. L'exemple le plus représentatif est *TileBars* [Hearst, 1995] qui présente les résultats dans une liste de résultats mais accompagnée, pour chaque document, d'un bloc visuel représentant la répartition des termes de la requête. Le but recherché est de pouvoir suggérer simultanément à l'utilisateur :

- la longueur relative du document (taille du bloc),
- la fréquence des termes de la requête dans le document (par la nuance de gris),
- la distribution visuelle des termes de la requête dans le document, mais aussi dans les autres résultats.

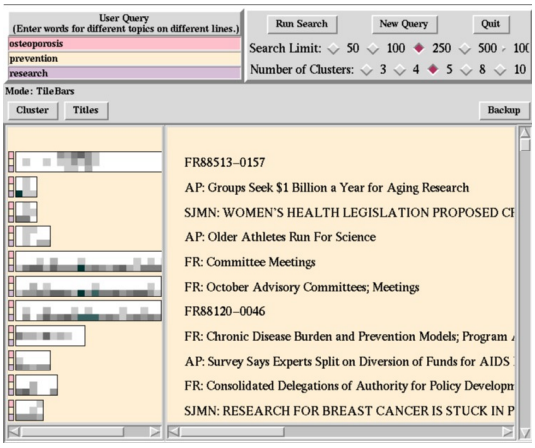


Figure 17* - *TileBars* [Hearst, 1995]

Nous pouvons également citer comme exemple l'interface *Seesoft* [Eick, 1994].

I.4.2.2.5.1.2 Attributs prédéfinis

Dans cette catégorie, les documents sont visualisés par rapport à des attributs prédéfinis comme des auteurs ou encore des thèmes prédéfinis par exemple. Ainsi *Cougar* [Hearst, 1994] permet de visualiser un ensemble de documents par rapport aux thèmes qu'ils abordent. La technique utilisée repose sur les diagrammes de Venn[4] (Figure 18). Chaque thème est représenté par un cercle et chaque document est situé dans les cercles correspondant aux thèmes qu'il aborde. Il n'y a au maximum que 3 thèmes sélectionnés dans cette interface car l'affichage proposé est en 2D. *InfoCrystal* [Spoerri, 1993] fait également partie de cette catégorie d'interfaces.

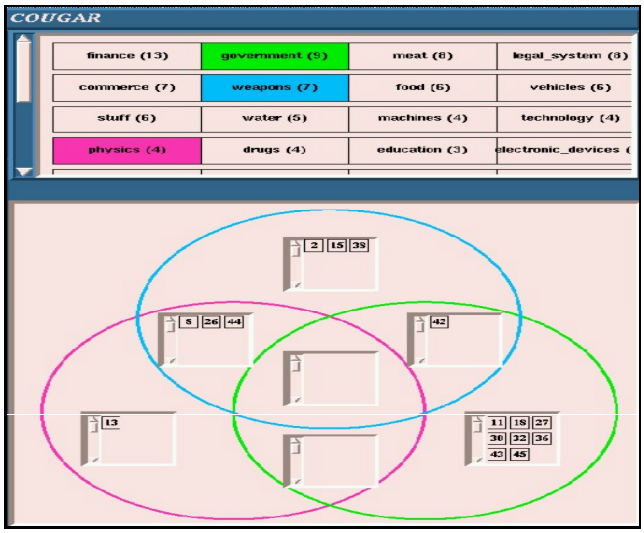


Figure 18* - *Cougar* [Hearst, 1994]

I.4.2.2.5.1.3 Attributs formulés par l'utilisateur

Dans cette catégorie, l'utilisateur peut choisir les attributs des documents suivant lesquels il souhaite visualiser les résultats. En général, ces attributs correspondent aux termes de la requête. Par exemple dans *Vibe* [Olsen, 1993] et *VR-Vibe* pour sa version 3D (Figure 19), l'utilisateur peut visualiser les documents par rapport à des termes qui l'intéressent.

[4] John Venn 1834-1923

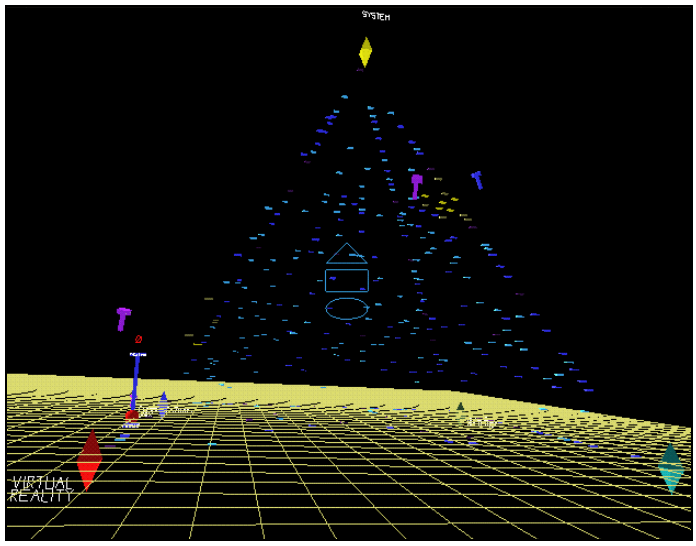


Figure 19* - *VR-Vibe* [Benford, 1995]

Une autre approche intéressante est celle proposée par [Cugini, 1996] au travers de l'interface *Three-Keywords Axes Display* (Figure 20). Cette interface permet de visualiser des termes ou combinaisons de termes sur des axes orthogonaux. Chaque document est représenté par un carré bleu dont la taille est proportionnelle à la taille du document. Dans chaque carré, l'importante d'un critère est représenté au moyen de lignes de couleurs plus ou moins longues proportionnellement à leur importance.

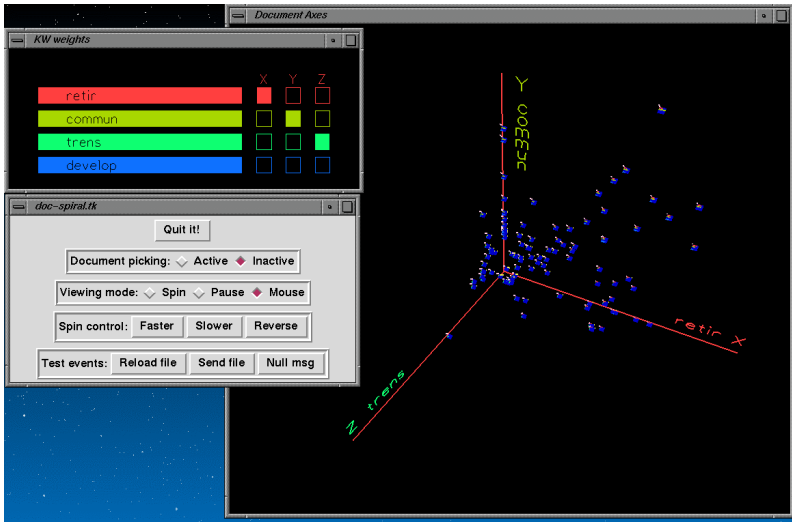


Figure 20* - *Three-Keywords Axes Display* [Cugini, 1996]

Par contre, dans *DocCube* [Mothe, 2002] (Figure 21), la visualisation ne se base plus sur les termes de la requête mais sur des hiérarchies de concepts. Les documents sont visualisés au travers d'une interface en 3D relatives aux différentes entrées des branches des hiérarchies de concepts sélectionnées.

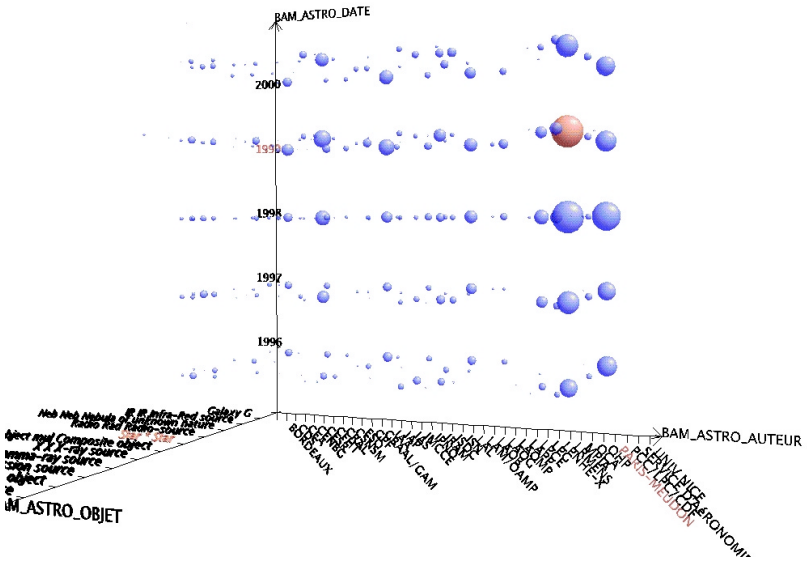


Figure 21 - Interface DocCube [Mothe, 2002]

I.4.2.2.5.2 Visualisation des relations inter-documents

Ces techniques de visualisation visent à mettre en évidence les relations entre les différents documents. Ces projets reposent essentiellement sur les notions de similarité et de classification.

I.4.2.2.5.2.1 Relations document-document

La visualisation des relations entre documents permet d'apprécier les documents similaires pour un document donné.

L'approche la plus commune est la visualisation des documents au travers d'un réseau comme le propose l'outil de recherche *Kartoo*[5] dans lequel les liens entre les documents correspondent aux termes que les documents ont principalement en commun (Figure 22).

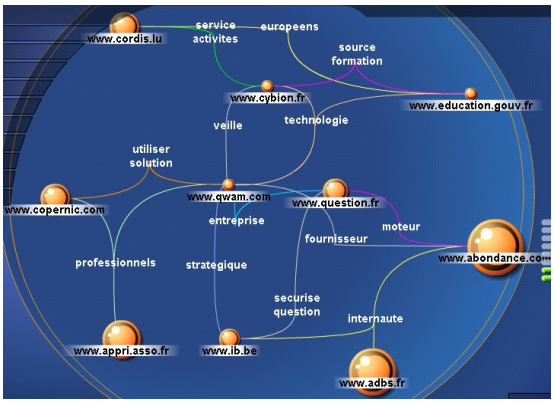


Figure 22* – *Kartoo*

Une alternative à la représentation en réseau est l'utilisation d'un affichage pondéré des documents. Par exemple, l'affichage en « œil de poisson » ou encore l'utilisation d'un procédé de force d'attraction et répulsion permet de réorganiser les documents spatialement

[5] http://www.kartoo.com

selon leur similarité. Grâce à ces affichages, l'utilisateur comprend à partir d'un document quels sont ceux qui sont proches de celui-ci. *Bead* [Chalmers, 1992] est un autre exemple d'interface basée sur ce type de représentation.

I.4.2.2.5.2.2 Relations classes de documents - classes de documents

La classification des documents permet de regrouper les documents similaires permettant ainsi de diminuer le nombre d'éléments visualisés puisque l'on ne représente plus les documents de façon indépendante mais les classes de documents.

Différentes techniques peuvent être utilisées pour représenter les relations entre des classes de documents.

La représentation la plus simple des classes de documents sont les classes elles-mêmes. Par exemple, *Grouper* [Zamir, 1999] (Figure 23) ou encore *Scatter/Gather* [Cutting, 1993] reposent sur une liste de résultats au sein de laquelle apparaissent des classes de documents (définies par un ensemble de groupes de mots caractérisant la classe de documents).

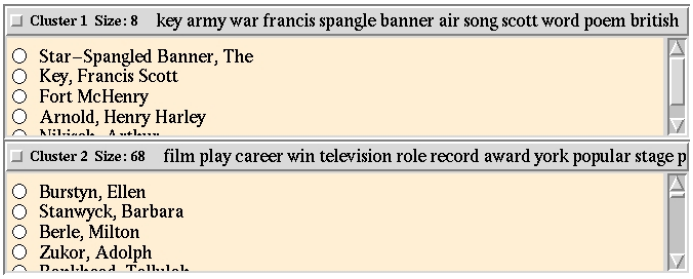


Figure 23 - *Grouper* [Zamir, 1999]

Par ailleurs, la visualisation proposée par l'outil de recherche *Mapstan*[6] (Figure 24) vise à présenter les classes de documents au moyen d'une métaphore d'un réseau urbain. Les quartiers (cercles) représentent des classes de documents similaires tandis que les « rues » représentent la similarité entre les quartiers. Ainsi, plus une rue est courte et épaisse, plus la similarité entre les quartiers sera importante.

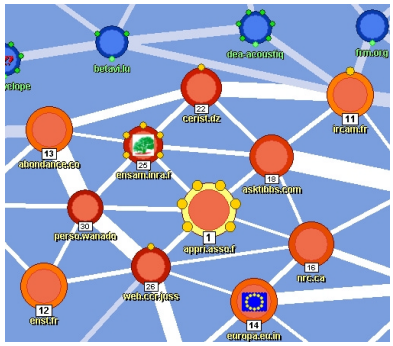


Figure 24* – *MapStan*

Plutôt que de représenter les relations des classes sous la forme d'un réseau, dans les cartes auto-organisatrices [Kohonen, 1982], la carte est une grille où chaque case correspond à une classe de documents similaires. Les classes sont positionnées les unes par rapport aux

[6] http://search.mapstan.net/

autres suivant leur similarité respective. [Lesteven, 1996] propose une utilisation de ces cartes dans le domaine de l'astronomie tandis que *Websom* en propose une application sur le web [Lagus, 1996].

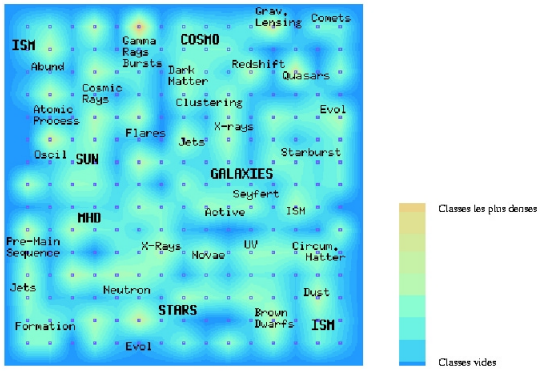


Figure 25* - Carte auto-organisatrice dans le domaine de l'astronomie [Poinçot, 1999]

Une autre approche de visualisation des relations entre classes de documents est celle de *Umap*[7] (Figure 26). Elle repose sur les « Arbres de connaissance® » et propose une carte « maritime » où chaque îlot correspond à un thème décrit par un ensemble de termes issus des documents retrouvés. Chaque élément de la carte représente un terme dont la position est calculée en fonction de sa corrélation avec les autres termes issus des documents. Cette visualisation permet de voir la cohérence du contenu des documents retrouvés au regard des termes qu'ils contiennent.

Figure 26* - *Umap*

Les interfaces précédentes ne représentent qu'une partie des travaux existants. Par exemple le Pacific Northwest National Laboratory (http://www.pnl.gov/infoviz/) propose d'autres interfaces de visualisation mais elles ne sont pas directement destinées à la visualisation de résultats de RI.

I.4.2.2.5.3 Comparaison des interfaces de visualisation

Enumérer les interfaces ne suffit pas pour comprendre l'impact qu'elles ont sur le processus de RI. Cette section ne vise pas à proposer une méthodologie d'évaluation d'une

[7] http://www.trivium.com

interface de visualisation. Des informations précises sur ce thème se trouvent notamment dans [Moore, 1995], [Shneiderman, 1998]. Dans cette section, nous proposons une évaluation comparative des interfaces présentées selon différents axes. En effet, une interface de visualisation utilise une certaine métaphore (représentation) des informations. Cette métaphore peut être construite à partir de plusieurs axes visuels qu'il est nécessaire de prendre en compte pour comparer les interfaces car ils conditionnent l'interprétation de la représentation graphique. Pour cette comparaison, nous avons fait le choix de présenter l'axe « espace de visualisation» et l'axe « couleur » car ils se retrouvent couramment dans la plupart des métaphores utilisées.

I.4.2.2.5.3.1 L'espace de visualisation : Texte, 2D ou 3D ?

Au travers de la catégorisation des interfaces précédentes nous avons pu constater qu'il existe trois catégories de représentation des informations : L'affichage textuel, l'affichage en deux dimensions et l'affichage en trois dimensions[8]. Ces représentations correspondent au type d'affichage utilisé et ce quel que soit le nombre de caractéristiques visualisées des informations. Le choix de la dimension utilisée n'est pas sans conséquence pour l'utilisation générale de l'interface. Ce choix conditionne :

- la mise en œuvre du système,
- le nombre d'éléments visualisés,
- l'expérience requise pour que l'utilisateur puisse s'orienter dans l'espace des informations,
- l'attrait de l'utilisateur.

Intuitivement, la complexité de la mise en œuvre et le coût (en temps ou en ressources) d'une interface augmente avec le nombre de dimensions. Ainsi les interfaces textuelles sont plus simples à mettre en œuvre que le sont les interfaces en 2D. Du fait du nombre de calculs accrus, les interfaces 3D sont encore plus coûteuses que les interfaces 2D. Le coût (temps et ressources) peut être aujourd'hui amorti par les performances actuelles des micro-ordinateurs personnels.

Par contre, le nombre d'éléments visualisés augmente avec le nombre de dimensions. Les interfaces textuelles permettent de visualiser un nombre restreint d'éléments comparativement aux interfaces 2D et 3D. Les interfaces 2D quant à elles permettent de visualiser un nombre d'éléments inférieur aux interfaces 3D.

En ce qui concerne l'expérience requise, elle augmente avec le nombre de dimensions utilisées. En effet, le fait d'utiliser une interface textuelle ne nécessite quasiment aucun apprentissage car elle est naturellement interprétée tandis qu'une interface 3D peut nécessiter un apprentissage relativement important.

L'attrait visuel pour un utilisateur a également son importance dans l'utilisation d'une interface. Ainsi, [Sebrechts, 1999] montre que l'utilisateur est plus sensible aux représentations graphiques qu'à une représentation textuelle et que pour une même tâche la représentation 3D permet d'obtenir globalement de meilleurs résultats.

---

[8] Dans ce document, un affichage en trois dimensions ne désigne pas la représentation stéréoscopique de l'espace d'information mais la représentation en perspective traditionnellement utilisée en visualisation.

Pour résumer, nous pouvons dire que la visualisation en 2D est peut-être la mieux adaptée à tous utilisateur car elle permet d'obtenir un bon compromis entre le nombre d'éléments visualisés et l'apprentissage nécessaire. Cependant, les interfaces 3D, sous réserve d'un apprentissage minimum, peuvent donner d'aussi bons, voire de meilleurs résultats que les autres types de visualisations [Sebrechts, 1999].

I.4.2.2.5.3.2 La couleur

La couleur est un axe exploité dans la plupart des visualisations proposées dans la littérature. En effet, la couleur permet de faire une distinction visuelle entre les différents éléments. [Keim, 1995] explique que la coloration a un fort impact sur l'interprétation des résultats. [Cugini, 2000] souligne d'ailleurs que le balayage visuel des couleurs (qui est un processus automatique) demande moins de temps et d'efforts que le balayage visuel des termes. Par ailleurs, l'utilisation des dégradés de couleurs permet de réaliser visuellement une hiérarchisation des éléments [Hearst, 1995].

Pour comparer l'utilisation des couleurs, nous avons défini deux catégories d'utilisation de la couleur. Elles peuvent être utilisées à des fins :

- *Pragmatiques*. Les couleurs correspondent à des actions réalisées par l'utilisateur ou à des caractéristiques uniquement destinées à distinguer les éléments. Par exemple, le fait de faire passer un document de la couleur verte à la couleur bleue traduit l'action « document visité »,
- *Sémantiques*. Les couleurs traduisent l'importance d'un critère ou d'une caractéristique des métaphores. Par exemple, un document de couleur verte traduit un document pertinent alors qu'un document de couleur rouge traduit un document non pertinent.

Le Tableau 2 synthétise les principaux travaux présentés dans cette section.

Légende du Tableau 2. (« N/d » signifie « non disponible »)

Dimensions	***Couleurs***
T : dimension textuelle	***P*** : Pragmatiques
2 : 2D	***S*** : Sémantiques
3 : 3D	***D*** : Dégradés

Projet	*Dimension*				*Couleur*			
Nom	*T*	*2*	*3*	*Détails*	*P*	*S*	*D*	*Détails*
Cougar	X	X		Chaque catégorie sélectionnée est représentée par un cercle. Les documents sont positionnés selon leur appartenance aux différentes catégories.	X			La couleur des cercles correspond à la couleur de la catégorie sélectionnée.
DocCube			X	A l'intersection des différentes entrées des hiérarchies de concepts, une sphère est affichée dont la taille est proportionnelle au nombre de documents contenus dans toutes ses concepts.	X			Selon les actions réalisées par l'utilisateur, les éléments changent de couleur (pour la sélection notamment).
Grouper	X			Les documents sont rassemblés en classes.				N/d
InfoCrystal		X		Ne présente que le nombre d'éléments par rapport aux combinaisons des différents critères.				N/d
Kartoo		X		Les documents sont représentés par des cercles dont la taille correspond à la pertinence globale. Ils sont reliés par des traits correspondant à des termes qu'ils partagent.	X			Chaque trait reliant les documents possède une couleur qui correspond au terme partagé entre les différents documents.
MapStan		X		Les documents sont répartis en quartiers (cercles) selon leur similarité. Ils sont reliés par des rues dont la taille dépend de la similarité entre les quartiers.	X			Les quartiers retrouvés par l'outil de recherche sont présentés en rouge alors que les documents recommandés sont présentés en bleu.
Three-Keywords Axes Display			X	3 critères maximum sont visualisés de façon orthogonale. Plusieurs mots-clés peuvent être assignés au même axe. Les documents sont visualisés par un carré positionné selon l'importance des différents critères. La taille de l'élément dépend de la taille du document.	X			Chaque document est représenté par un carré bleu. A l'intérieur de ce carré, chaque critère est représenté par une droite de la couleur spécifique au critère et dont la longueur dépend de l'importance du critère.
TileBars	X	X		Chaque *Tile* (petit carré) présente la longueur du document. Le document est découpé en sections où est visualisée l'importance de chaque critère.	X	X	X	L'intensité de la couleur (dégradé du blanc au noir) correspond à l'importance du critère dans la section en cours. Une autre utilisation des couleurs sert à caractériser chaque critère.
Umap		X		Les îlots représentent des termes issus des documents et très corrélés.		X	X	La couleur dépend du nombre de documents dans lesquels apparaît le terme.
Vibe		X		Chaque critère est représenté par un cercle et tous les documents sont positionnés par rapport à ces critères. Plus ils sont proches d'un critère plus celui-ci est important dans les documents.				N/d
VR-Vibe			X	Chaque critère est présenté par un octaèdre. Les documents sont représentés par des cubes qui sont positionnés comme dans *Vibe* selon l'importance des différents critères. La taille des cubes correspond à la pertinence globale des documents.	X		X	Tous les documents ont une même couleur initiale (bleue). Par contre, l'intensité de celle-ci correspond à la pertinence globale du document. La teinte dépend des actions de l'utilisateur : ouverture d'un document, désignation d'un document pertinent.
WebSom		X		Les documents sont présentés sous forme d'une grille.		X	X	La couleur dépend du nombre de documents présents dans l'entrée de la grille.

Tableau 2 - Catégorisation des principales interfaces de visualisation

Outre ces axes de visualisation, un problème important des interfaces pour la RI reste la multitude des tâches (recherche précise d'un document, survol d'un domaine...) et la diversité humaine. [Levialdi, 1994], par exemple, souligne l'intérêt de l'adaptation de l'interface aux utilisateurs. Ainsi, en plus des critères cognitifs « standards », [Vernier, 1997] a introduit le critère « *d'affordance* » qui indique que l'utilisateur doit comprendre qu'il est nécessaire de changer de visualisation selon sa tâche et son niveau d'expérience. Ceci implique donc l'intégration au sein du même système de plusieurs visualisations possibles,

que ce soit pour répondre au grand nombre de tâches ou pour s'adapter à l'utilisateur en lui proposant des outils exploitables et ce quelque soit son niveau d'expérience.

Nous venons de souligner les approches visant à améliorer la recherche grâce à des outils adhoc. D'autres approches ont été développées afin de proposer des solutions non seulement PULL mais également PUSH.

I.4.2.3 Les agents

Les approches précédentes reposent généralement sur la méthode PULL. Or, cette méthode d'accès à l'information nécessite une implication continue et importante de l'utilisateur. Pour donner un peu plus de liberté à l'utilisateur, les agents peuvent jouer un rôle important en proposant à l'internaute un véritable « compagnon » de recherche.
Un agent est communément défini comme « *une personne chargée des affaires et des intérêts d'un individu* » (*Petit Robert*).

L'application informatique du concept d'agents respecte cette définition et est caractérisée par plusieurs aspects [Caglayan, 1998] :

- *de délégation*. L'agent accomplit un ensemble de tâches à la demande d'un utilisateur (ou d'un autre agent), ensemble de tâches explicitement approuvées par l'utilisateur,
- *de communication*. L'agent a besoin d'interagir avec l'utilisateur afin d'en recevoir des instructions pour accomplir sa tâche, de l'informer du statut et de la réalisation de la tâche. Ces interactions s'effectuent au moyen d'une interface utilisateur-agent ou d'un langage de communication agent,
- *d'autonomie*. L'agent travaille sans intervention directe (en tâche de fond) jusqu'à un point défini par l'utilisateur. L'autonomie d'un agent peut aller du simple lancement d'une sauvegarde nocturne, à la négociation de prix d'un produit désigné par l'utilisateur,
- *de contrôle*. L'agent est capable de contrôler son environnement pour effectuer l'ensemble des tâches de manière autonome,
- *de mise en action*. L'agent doit être capable d'interagir sur son environnement par un mécanisme de mise en action pour un fonctionnement autonome,
- *d'intelligence*. L'agent doit interpréter les événements observés afin de prendre les décisions appropriées, de manière autonome.

La Figure 27 présente l'architecture simplifiée d'un agent.

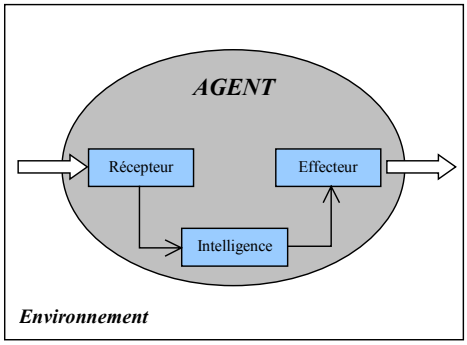


Figure 27 - Structure simplifiée d'un agent intelligent

Il existe cependant de nombreuses catégories d'agents (agent bureautique, agent de notification...) qui sont notamment présentées dans [Caglayan, 1998]. Dans ce document, nous mettons l'accent sur les agents de recherche, les agents de recommandation puisqu'ils interviennent dans le contexte de l'aide à la Recherche d'Information sur le web mais également sur les systèmes multi-agents qui peuvent apporter de la valeur ajoutée dans l'aide à la recherche.

I.4.2.3.1 Les agents de recherche

La catégorie des *agents de recherch*e comprend les méta-moteurs de recherche d'information de dernière génération (listes synthétisées) jusqu'aux outils de recherche off-line (*Copernic* par exemple). Ces derniers permettent à l'utilisateur d'initier des recherches d'informations qui s'exécuteront même lorsque l'utilisateur ne sera plus connecté à Internet.

I.4.2.3.2 Les agents de recommandation

Les *agents de recommandation*, quant à eux, visent à optimiser la recherche d'information de l'internaute en lui proposant automatiquement de nouveaux documents au regard de ses besoins ou de ses actions. Ils reposent essentiellement sur une approche *Push* proposant des informations à l'utilisateur et une caractérisation des besoins au moyen d'un profil utilisateur. Nous présentons différentes catégories d'agents de recommandation en fonction des buts recherchés.

Dans un premier temps, nous soulignons les agents qui proposent les liens qui mènent vers des informations répondant aux besoins de l'utilisateur. Ceci permet notamment à l'utilisateur d'optimiser sa navigation en se consacrant à des documents potentiellement pertinents. Les systèmes *Letizia* [Lieberman, 1995], *WebWatcher* [Armstrong, 1995] ou encore *BroadWay* [Jaczynski, 1997] sont de bons représentants de cette catégorie.

Letizia utilise les documents visités comme base du profil utilisateur. Celui-ci est construit à partir des termes issus des documents visités et des actions de l'utilisateur (ajout dans les signets par exemple). Lors de la visite d'un document, *Letizia* télécharge le contenu de tous les documents liés au document courant pour en évaluer l'appariement avec le profil utilisateur. Les liens menant vers des documents pertinents sont présentés à l'utilisateur au moyen d'une fenêtre annexe.

Webwatcher demande à l'utilisateur de formuler ses besoins explicitement (au travers d'une requête). Chacun des documents visités est annoté par les termes de la requête de l'utilisateur ainsi que par un jugement de pertinence de la part de l'utilisateur. A chaque visite d'un document, *Webwatcher* recherche les liens pointant vers des documents similaires aux besoins de l'utilisateur. Cet appariement est réalisé grâce aux annotations laissées par les utilisateurs ayant précédemment visité le document et la requête de l'utilisateur qui navigue. Les liens pointant vers des documents susceptibles d'intéresser l'utilisateur sont présentés directement au sein du document visité au moyen d'icônes.

Broadway poursuit le même but mais s'appuie sur une démarche légèrement différente car il repose sur l'idée que si plusieurs utilisateurs suivent le même parcours au sein de l'hypertexte, c'est vraisemblablement qu'ils recherchent les mêmes informations. Le profil utilisateur repose cette fois-ci sur une représentation de la navigation en cours. La

représentation de la navigation est construite à partir de séries temporelles correspondant à quatre variables : le vecteur descripteur du document visité, l'évaluation explicite faite par l'utilisateur, le temps de lecture et l'URL du document. L'échelle de temps appliquée à ces séries temporelles correspond à une visite. Pour identifier les documents pertinents, *Broadway* utilise le concept de raisonnements par cas sur les représentations des navigations antérieures. Les documents ainsi retrouvés sont présentés à l'utilisateur au moyen d'une fenêtre annexe.

Du point de vue du résultat, la principale différence entre ces trois agents est la profondeur de recommandation. Dans *Letizia* ou *Webwatcher*, les recommandations ne concernent que les documents directement liés au document courant, tandis que *Broadway* anticipe en quelque sorte la navigation en proposant des documents liés par transitivité au document courant.

Les systèmes précédents n'indiquent à l'utilisateur que les documents potentiellement pertinents pour la navigation en cours. Or, il est tout aussi intéressant d'informer l'utilisateur sur le fait que le document qu'il est en train de visiter est pertinent (ou non pertinent) pour les thèmes qui l'intéressent. Le système *Syskill & Webert* [Pazzani, 1996] vise à remplir cette tâche. L'utilisateur peut créer autant de thèmes que nécessaire avec la restriction que les thèmes sont organisés sous forme de liste non hiérarchique. A chaque fois qu'il trouve un document pertinent (ou non pertinent) pour les thèmes qu'il a créé, l'utilisateur le signale au système qui met à jour un profil correspondant au thème. Lors de la navigation, le système indique à l'utilisateur si le document qu'il visite est pertinent (ou non pertinent) pour les thèmes qu'il possède.

De façon plus générale, certains agents peuvent également prendre en considération l'environnement de recherche de l'utilisateur (ensemble des applications) pour déduire ses besoins et lui recommander des documents qui ne sont pas forcément reliés (par des liens hypertextes) aux documents visités. *WBI* [Barrett, 1997], par exemple, étudie les actions de l'utilisateur et organise les documents visités en classes représentant les besoins des utilisateurs. A partir de ces besoins, *WBI* interroge des outils de recherche pour proposer le résultat de recherche aux utilisateurs. Il propose en outre des fonctionnalités de gestion d'historique, de notification (signale les modifications dans le contenu d'un document) et de génération de raccourcis (un utilisateur qui, chaque jour, utilise la même séquence de liens pour atteindre un document se verra proposer un raccourci vers celui-ci). Cette approche repose sur une architecture multi-agents et l'interface de WBI est intégrée aux documents web.

Fab [Balabanovic, 1997] repose également sur une approche multi-agents qui rapproche les individus par centres d'intérêt. Le système collecte ensuite des documents sur le web grâce à un agent spécifique qui recherche les thèmes intéressant des utilisateurs, pour les leur proposer. Un document est recommandé s'il correspond bien au profil de l'utilisateur ou s'il correspond au profil d'un utilisateur proche. Le profil est construit à partir des jugements exprimés par les utilisateurs concernant les documents recommandés.

Le système *Watson* [Budzik, 1999] apporte une innovation concernant l'identification d'expressions régulières dans les documents visités (comme les adresses par exemple) afin de proposer des services contextuels à l'utilisateur. Pour une adresse identifiée, *Watson* propose

à l'utilisateur la génération d'un plan urbain permettant de localiser celle-ci. L'interface de *Watson* est présentée dans une fenêtre indépendante.

Dans le contexte de l'étude de l'environnement de travail de l'utilisateur pour la construction de son profil, l'approche de *Suitor* [Maglio, 2000] innove par différentes fonctionnalités. Il utilise notamment une fenêtre déroulante permettant d'afficher les informations proposées par *Suitor* en base de l'écran, mais il propose surtout un système permettant de suivre le regard de l'utilisateur afin d'identifier les parties de l'écran qu'il visionne.

Dans une toute autre approche, *SiteSeer* [Rücker, 1997] est un agent permettant de proposer à un utilisateur des signets directement dans la hiérarchie de signets qu'il possède. Ce système repose sur la couverture entre les différents répertoires des hiérarchies de signets des utilisateurs du système. En effet, plus la couverture entre le contenu de deux répertoires de deux utilisateurs respectifs est importante, plus les deux répertoires traitent d'un thème similaire. Ainsi, les documents n'entrant pas dans la couverture composée par le contenu des deux répertoires sont jugés pertinents et sont proposés pour le répertoire de l'utilisateur qui ne les possède pas.

I.4.2.3.3 Approches multi-agents

Les systèmes multi-agents visent à faire coopérer une série d'agents afin d'obtenir un résultat. Un modèle général de coopération peut être vu dans [Cloutier, 1998]. Ainsi, dans le cadre de la recherche d'information, il peut être intéressant de faire coopérer les agents pour répondre à des besoins précis en information d'un utilisateur.

Par exemple, l'approche multi-agents proposée au travers du projet *Marvin* (http://www.hon.ch/Project/Marvin_project.html) permet de construire des collections thématique de documents (Figure 28). Chaque agent parcours le web à la recherche de documents pour un thème donné (exemple la médecine). Ces collections servent à alimenter des outils de recherche adhoc sectoriels.

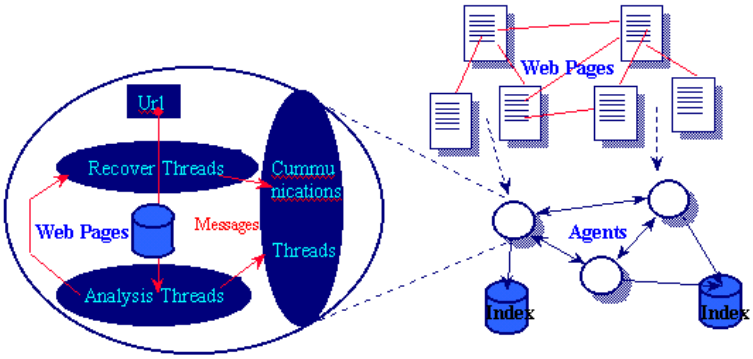


Figure 28 - Architecture multi-agents de Marvin

L'intérêt des approches multi-agents peut également être vu au travers du projet *Abrose* [Carré, 1999] qui permet de construire et de maintenir un profil utilisateur pour le commerce électronique mais qui pourrait être adaptée à la recherche d'information. Ce profil utilisateur contient l'ensemble des préférences d'utilisation d'Internet d'un utilisateur qui peut être utilisé pour répondre à des requêtes de façon plus fine ou proposer à

l'utilisateur de nouveaux documents ou offres plus intéressantes. L'évolution des préférences est adaptative et est caractérisée par un ensemble d'agents.

I.4.2.3.4 Synthèse sur les agents

En conclusion, les agents de recommandation permettent à l'utilisateur d'obtenir une aide précieuse à la RI en proposant pertinents à l'utilisateur de nouveaux documents potentiellement. Ils se différencient par la représentation des centres d'intérêt d'un utilisateur ainsi que par les recommandations qu'il réalise. Nous proposons dans le Tableau 3 une synthèse des agents de recommandation que nous avons présentés.

Légende du Tableau 3.

Profil : précise comment l'agent caractérise les besoins de l'utilisateur. Il peut utiliser un profil à court ou à long terme. Un profil à court terme est un profil créé et utilisé durant quelques minutes voire quelques heures alors qu'un profil à long terme représente un profil persistant sur un grand nombre de recherches.

Recommandation : indique la source d'information utilisée pour mettre en œuvre les recommandations. Elle peut reposer sur des outils de recherche externes tels qu'Altavista ou Google, sur les documents liés au document courant (hypertexte local) ou encore sur une base d'informations interne construite au fur et à mesure des utilisations par exemple.

Nom	*Profil*	*Recommandation*	*Appariement*
Broadway	Court-Terme	Interne (Time Series)	Raisonnement par cas
FAB	Long-Terme	Externe (Outils de recherche)	Similarité contenu
Letizia	Court-Terme	Hypertexte Local	Similarité contenu
SiteSeer	Long-Terme	Interne (Hiérarchies de signets)	Couverture
Suitor	Court-Terme	Interne (News, Aides)	Similarité contenu
Syskill & Webert	Long-Terme	Hypertexte Local	Similarité contenu
Watson	Court-Terme	Externe (Outils de recherche)	Similarité contenu
WBI	Court-Terme	Externe (Outils de recherche)	Similarité contenu
WebWatcher	Court-Terme	Interne (Annotations implicites)	Similarité contenu

Tableau 3 - Synthèse sur les agents de recommandation

I.4.3 Organisation et sauvegarde des informations retrouvées

Nous avons présenté, dans les sections précédentes, les approches visant à aider l'utilisateur durant sa recherche d'information. Or, lors d'une tâche de recherche, l'utilisateur identifie des documents qui l'intéressent et dont il souhaite sauvegarder des informations (URL, remarque...). Pour lui permettre d'organiser et de retrouver aisément les documents qui l'intéressent sur le web, deux grands types d'outils sont disponibles : les systèmes de signets et les systèmes d'annotation.

I.4.3.1 Les systèmes de signets

Les navigateurs actuels proposent des systèmes de signets web (encore appelés « favoris ») qui sont une métaphore des marque-pages.

Un signet est un pointeur (ou un raccourci) vers un document web. Il est constitué généralement d'un nom ainsi que d'une URL correspondant au document web pertinent. Ces signets sont généralement organisés sous la forme d'une hiérarchie de répertoires. L'organisation des signets dépend du nombre de signets [Abrams, 1998]. Ainsi, pour moins de 35 signets, l'utilisateur n'utilise généralement qu'un seul répertoire pour organiser ses signets. Par contre, pour un nombre plus important de signets, il est recommandé de les structurer hiérarchiquement pour en faciliter l'accès.

La plupart des internautes possèdent une hiérarchie de signets. [GVU, 1998] et [Abrams, 1998] montrent qu'en moyenne les internautes possèdent plus de 30 signets.

Cependant ces hiérarchies de signets souffrent des limites suivantes :

- *sous-information*. Les informations contenues dans les signets sont assez limitées et l'utilisateur peut avoir du mal à se souvenir de la raison pour laquelle il a créé un signet vers un document sans avoir à en visualiser le contenu [Maarek, 1996],
- *fréquentation faible*. Contrairement à leur but initial, les signets d'une hiérarchie de signets sont peu visités (un utilisateur parcours la totalité de ses signets en environ 300 jours [Abrams, 1998]),
- *évolution rapide*. Le nombre de signets évolue avec le temps. [Abrams, 1998] montre qu'à chaque session l'utilisateur crée quelques signets (3 ou 4).
- *réorganisation difficile*. Dès lors que l'internaute possède un nombre de signets important, il est difficile et contraignant de les ré-organiser manuellement,
- *disponibilité locale*. Les signets ne sont accessibles, au travers des navigateurs actuels, que de façon locale. Ceci signifie que si l'utilisateur change de lieu de connexion, ses signets ne seront pas accessibles,
- *mise à jour inexistante*. Les signets ne sont pas toujours cohérents avec les documents qu'ils référencent. La rapide évolution des documents du web implique que, compte tenu de la faible fréquentation des signets, certains des signets pointent vers des documents qui n'existent plus ou vers des documents dont le contenu a été modifié.

Pour tenter de pallier à certaines de ces limites, des outils tels que *Oneview* (http://ww.oneview.com) permettent d'obtenir un accès distant à ces signets ainsi qu'une mise à jour relative (vérification des liens morts).

Cependant, malgré ces limites, les signets constituent le moyen privilégié pour créer un espace d'informations personnel [Abrams, 1998] ou encore pour servir de base à un échange d'informations.

I.4.3.2 Les systèmes d'annotations

Grâce à un système d'annotations, un utilisateur peut ajouter à un document visité des informations personnelles telles que des jugements de valeurs, des remarques...

Définition : Une **annotation** est composée :
- d'une *ancre* qui permet de rattacher l'annotation au document,
- d'*attributs* facultatifs comme des commentaires.

Il existe différentes catégories de systèmes d'annotations correspondant aux différentes formes d'annotations proposées.

Par exemple, dans *Pharos* [Bouthors, 1999], l'utilisateur peut annoter les documents web qu'il visite de façon intégrale par rapport à des centres d'intérêt (« channels ») structurés de façon hiérarchique. Une annotation est composée d'une date, d'un jugement de pertinence [-1 ;+1] et d'une série de mots-clés issus d'un thésaurus proposé par le système.

Yawas [Denoue, 2000], quant à lui, propose l'annotation de passages de documents. Une annotation est composée de métadonnées telles que le type et le thème du document, le type du texte annoté, un commentaire libre ainsi que son type et l'auteur de l'annotation. L'ancre utilisée repose sur la norme *DOM* (Document Object Model)[9].

Une alternative est proposée par le *World Wide Web Consortium* (W3C) au travers du système *Annotea* [Kahan, 2001]. Ce projet a comme particularité de permettre d'annoter soit les documents en leur intégralité soit des parties de documents. De plus, les ancres reposent sur le modèle *RDF (Resource Desrciption Framework)*[10]. Les annotations peuvent être typées (conseil, commentaire, explication, question...). Elles sont également composées d'un ensemble de méta-données comme l'auteur, le type...

I.4.3.3 Synthèse Signets-Annotations

Le Tableau 4 synthétise ces différentes approches.

Légende du Tableau 4.

Annotation : indique si le système permet de créer des annotations pour le document intégral ou pour des parties de documents,

Navigateur : précise si le projet utilise les navigateurs classiques ou un navigateur spécifique,

Typée : précise si l'annotation peut avoir différentes formes (conseil, question...) ou s'il ne s'agit que d'un commentaire textuel,

Pointeur : précise comment sont liées les annotations au document. Cette liaison est basée soit sur DOM soit sur RDF.

Nom	*Annotation*	*Navigateur*	*Typée*	*Pointeur*
Pharos	Intégral	Classique	Non	n/a
Yawas	Partie	Classique	Oui	DOM
Annotea	Intégral/Partie	Spécifique	Oui	RDF

Tableau 4 - Synthèse sur les systèmes d'annotations

La grande différence entre les systèmes de signets et d'annotations se situe au niveau de l'implication de l'utilisateur pour créer le signet ou l'annotation. Dans les systèmes de

[9] http://www.w3c.org/DOM

[10] http://www.w3c.org/RDF

signets, l'utilisateur est peu présent dans le processus et se contente d'indiquer dans quel répertoire il souhaite ajouter le signet. Par contre, dans les systèmes d'annotations, l'utilisateur doit avoir une démarche active et prendre le temps de remplir les différentes informations relatives à l'annotation. Du fait de cette implication importante de l'utilisateur, le contenu des annotations est plus riche que le contenu des signets.

Cependant, l'organisation intrinsèque des hiérarchies de signets permet à l'utilisateur de construire son propre espace d'informations organisé selon ses besoins de façon aisée. En effet, les systèmes d'annotations proposent tous un outil de recherche dans les annotations mais peu proposent à l'utilisateur d'organiser ces informations (sous forme de liste, d'arborescence...). De ce fait, l'utilisateur a du mal à obtenir un vision personnelle de ses annotations ainsi qu'un accès rapide aux documents annotés comme le propose *Pharos* par exemple.

Ces deux approches permettent aux utilisateurs de conserver des traces ou des informations concernant les documents qu'il visite. De plus, au regard de l'aide à la RI, ces éléments sont une base intéressante pour le partage d'information. En effet, ces informations sont fiables et validées par un utilisateur, ce qui permet un échange d'information plus pertinent. Nous présentons dans la section suivante cet aspect coopératif dans l'optique de l'aide à la RI.

I.4.4 L'aspect coopératif dans la Recherche d'Information

Grâce aux systèmes visant à aider un internaute dans sa recherche, la tâche d'un utilisateur peut être réalisée dans de meilleures conditions. Cependant, la plupart des approches visent uniquement à répondre aux besoins de l'utilisateur grâce à des informations essentiellement collectées et traitées par le système.

Comment arrive-t-on sur un site Web
CommerceNet/Nielsen Media - Juillet 1997

71,0%	*Par les moteurs de recherche*
9,8%	*Conseillé par amis ou collègues*
8,5%	*Journaux quotidiens ou périodiques*
8,4%	*Lien sur un autre site*
8,1%	*Par hasard, en surfant*
3,6%	*Signalé à la TV*
3,3%	*Guides sur les sites web*

Tableau 5 - Comment arrive-t'on sur un site web

Or, dans la réalité, lorsque nous avons besoin d'aide, nous faisons appel à des collègues, amis... Dans le cadre de la Recherche d'Information, ce n'est généralement pas le cas et nous pouvons constater que moins de 10% des informations retrouvées sur le web proviennent de collègues ou amis (Tableau 5). Cet état de fait ne permet donc pas à un utilisateur de profiter des informations collectées par d'autres utilisateurs et ne lui permet pas de faire évoluer sa connaissance. En effet, l'aspect coopératif dans la RI permet aux internautes de découvrir des documents pertinents, comme par exemple, des outils de recherche spécifiques pour leur domaine de recherche provenant de dialogues avec les autres internautes. De plus,

il est dommage de ne pas profiter des ressources collectées par un utilisateur ayant les mêmes centres d'intérêt.

Cette section se focalise donc sur le partage d'informations entre les utilisateurs dans le cadre de l'aide à la recherche d'information.

La Recherche d'Information coopérative vise à faire profiter les utilisateurs des informations, des jugements de pertinence ou encore de l'expérience d'autres utilisateurs. L'intérêt de cette approche est qu'elle se base sur une pertinence utilisateur plutôt qu'une pertinence système (valeur mesurée par la machine).

Un indicateur de l'intérêt réel que suscite une telle approche peut être constaté au travers de l'essor des communautés basées sur l'échange d'information entre les utilisateurs [Fresse, 2002] ou encore de l'intérêt qu'éprouvent les internautes pour les groupes de discussion (« *newsgroups* ») et les foires aux questions (« *faq* »).

Dans le processus de Recherche d'Information, cette approche coopérative peut être appliquée à tous les niveaux. Cette section vise à présenter les différentes possibilités que cette approche offre durant le processus de RI. Ainsi, nous présentons dans un premier temps l'aspect coopératif afin d'améliorer la connaissance du domaine de l'utilisateur. Dans un second temps, nous détaillons cet aspect coopératif au sein du processus de recherche.

I.4.4.1 La connaissance du domaine

Nous pouvons considérer que des systèmes comme *SiteSeer* [Rücker, 1997] permettent d'améliorer la connaissance du domaine de l'utilisateur de façon coopérative. En effet, il se base sur une représentation des centres d'intérêt de l'utilisateur à partir de sa hiérarchie de signets. Le système propose de nouveaux signets à l'utilisateur qui sont issus des hiérarchies de signets des autres utilisateurs du système. Les recommandations faites peuvent être considérées comme des informations permettant à l'utilisateur d'enrichir sa connaissance du thème représenté par le répertoire dans lequel les recommandations ont été effectuées.

L'approche du système *GroupMark* [Pemberton, 2000] qui fait partie du projet *Select* [Procter, 1999] repose sur ce même principe qui est utilisé pour la définition et le partage d'informations au sein d'un groupe d'utilisateur. Le système propose à l'utilisateur des groupes pertinents par rapport à ses centres d'intérêt. Pour cela, l'utilisateur doit posséder au sein de ses signets certains documents qui définissent ce groupe. Ces documents sont choisis et définis par le créateur du groupe.

Une autre approche peut être vue au travers de *Fab* [Balabanovic, 1997] car un utilisateur profite des documents qui sont pertinents pour les utilisateurs ayant des centres d'intérêt similaires. *Fab* recommande un document à un utilisateur si il est pertinent pour ses centres d'intérêt mais aussi s'il est jugé pertinent par un utilisateur ayant des centres d'intérêt similaires.

I.4.4.2 L'aspect coopératif au cours de la recherche d'information

L'aspect coopératif idéal de la RI correspond à un expert à qui l'on demande d'effectuer la recherche désirée. De tels procédés existent comme ceux proposés dans *Cybion*[11] ou encore *Cybervigie*[12]. Des experts en recherche traitent les demandes des utilisateurs pour proposer en retour un résultat de recherche synthétisé, vérifié... Cependant, ce type de service n'est pas gratuit et cette solution n'est réellement envisageable que lorsque les besoins en informations sont très importants voire sensibles du fait du coût élevé du service.

Dans le cadre de la formulation de la requête, un système comme *Cosydor* [Jeribi, 2001] permet de faire profiter l'utilisateur de l'expérience d'autres utilisateurs pour mieux formuler ses besoins.

Il existe, par ailleurs, des outils de recherche par interrogation proposant un aspect coopératif comme *HumanLinks*[13] ou *IronWeb* [Dussaux, 2000]. *HumanLinks* par exemple est un outil de recherche basé sur le protocole *Peer-To-Peer* (P2P). L'utilisateur formule une requête et le système recherche, sur les machines de tous les utilisateurs connectés, les informations susceptibles d'être pertinentes pour cette requête. Une approche plus traditionnelle est proposée dans *IronWeb* qui est un outil de recherche adhoc dont la base d'indexation est construite à partir des hiérarchies de signets d'un groupe d'utilisateurs jugés comme experts du domaine.

En ce qui concerne les interfaces de visualisation, l'interface VR-Vibe [Benford, 1995] permet d'effectuer des recherches de façon coopérative dans un environnement 3D. Les différentes recherches des utilisateurs sont présentées dans l'univers virtuel proposé par l'interface. Chaque utilisateur peut ainsi accéder aux recherches des autres utilisateurs, laisser des informations....

Pour aider l'utilisateur à trouver des documents pertinents lors de sa navigation, certains agents sur le web proposent une approche coopérative. Ces outils mettent en évidence les documents pertinents que l'utilisateur devrait visiter par rapport à la tâche de navigation qu'il est en train de réaliser. Ces recommandations proviennent de l'utilisation d'informations d'autres utilisateurs. Par exemple, *Broadway* repose sur l'utilisation des historiques des navigations antérieures d'autres utilisateurs du système pour anticiper la navigation de l'utilisateur courant. *WebWatcher*, quant à lui, repose sur une démarche coopérative basée sur les mots-clés laissés par les différents utilisateurs du système lors de leurs visites précédentes. A partir de cet ensemble de termes, le système identifie pour un document visité, parmi ceux de l'hypertexte local, les documents qui correspondent aux besoins de l'utilisateur. Le projet *Select* [Procter, 1999], quant à lui, adopte une démarche qui informe l'utilisateur de la pertinence du document visité mais aussi des liens qu'il contient vis-à-vis du groupe auquel est inscrit l'utilisateur. Ces jugements reposent sur des informations explicites ou implicites (temps de lecture...).

L'approche coopérative peut être également constatée au travers des systèmes d'annotations. Les systèmes comme *Yawas* et *Pharos*, par exemple, informent un utilisateur,

[11] http://www.cybion.fr

[12] http://www.cybervigie.com

[13] http://www.human-links.com

qui visite un document, de toutes les annotations émises au préalable par l'ensemble des utilisateurs.

I.4.4.3 Synthèse sur l'aspect coopératif en RI

Afin d'évaluer les différentes approches coopératives en RI, nous présentons, dans le Tableau 6, pour chaque système, quel est le producteur de l'information partagée et qui en est le destinataire, en précisant le type de ces informations.

Nom	*Producteur*	*Consommateur*	*Type information*
Broadway	Tous les utilisateurs	Utilisateur enregistré	Navigation
Cybion	Cybion	Client	Résultats de recherche
Cosydor	Utilisateurs ayant une expérience dans le domaine de recherche	Tous les utilisateurs	Termes pour la formulation de la requête
Fab	Tous les utilisateurs ayant les mêmes centres d'intérêt	Utilisateur enregistré	Documents
GroupMark	Utilisateurs du groupe	Utilisateur enregistré	Groupe pertinent
HumanLinks	Tous les utilisateurs	Utilisateur enregistré	Documents
IronWeb	Utilisateurs experts	Tous les utilisateurs	Signets web
Pharos	Tous les utilisateurs ayant les mêmes centres d'intérêt	Utilisateur enregistré	Annotations
Select	Tous les utilisateurs	Utilisateur enregistré	Documents
Siteseer	Utilisateur enregistré le plus proche du consommateur	Utilisateur enregistré	Signets web
VR-Vibe	Un utilisateur	Tous les autres utilisateurs	Recherches d'informations ainsi que leurs résultat
WebWatcher	Tous les utilisateurs	Utilisateur enregistré	Type d'annotation
Yawas	Tous les utilisateurs	Utilisateur enregistré	Annotations

Tableau 6 - Synthèse des approches coopératives

Au regard de ce tableau, nous pouvons souligner la diversité des approches proposées pour réaliser le partage d'informations. Par exemple, certains des systèmes ne diffusent l'information qu'entre utilisateurs qui ont les mêmes centres d'intérêt (*SiteSeer*, *GroupMark*...). D'autres approches reposent plutôt sur le partage universel des informations comme (*Broadway*, *Select*...).

Cependant, l'efficacité des systèmes reposant sur une approche coopérative réside dans la représentation des utilisateurs et plus particulièrement de leur appariement. Par exemple, *SiteSeer* ou *GroupMark* reposent sur une similarité qui est fonction du taux de couverture entre les signets des deux utilisateurs. Or, dans le cadre du web, il est peu probable que deux

utilisateurs ayant le même centre d'intérêt possèdent les mêmes documents au sein de leur hiérarchie de signets.

I.5 Conclusion

Cette première partie a présenté les fondements et caractéristiques de la RI sur le web.

Nous avons souligné les concepts généraux de Recherche d'Information tels que les modèles de recherche permettant d'apprécier l'appariement entre une requête et les informations disponibles et les stratégies de recherche qui dépendent de l'environnement de l'outil de recherche. La spécificité de la recherche sur le web a également été détaillée.

La RI sur le web souffre, nous l'avons souligné, de divers problèmes qui influent de façon négative sur l'efficacité des recherches de l'utilisateur.

Pour réaliser une bonne recherche, l'utilisateur doit posséder une bonne connaissance pratique pour lui permettre d'utiliser les fonctionnalités offertes par le web (liens hypertextes, manipulation d'un outil de recherche...). Il doit également posséder une bonne connaissance du domaine de recherche relativement importante lui permettant de formuler ses besoins de façon précise (avec des termes bien choisis) mais également avoir un meilleur jugement de pertinence des documents qu'il visite.

Par ailleurs, durant la navigation, il est nécessaire d'apporter à l'utilisateur une aide lui permettant de ne pas avoir à réaliser un effort cognitif trop important au travers de visualisations de navigation ou de cartographies pour éviter qu'il se perde dans l'hypertexte ou qu'il se sente désorienté. La prise en compte de la navigation est actuellement considérée par les systèmes existants (*sauf Broadway*) globalement par rapport aux termes résultant d'une analyse des documents visités. Ce point est important car cela signifie que ces systèmes ne proposent à l'utilisateur que des documents possédant des termes similaires à ceux des documents visités.

Durant la tâche de recherche, l'utilisateur peut également avoir du mal à exploiter le trop grand nombre de résultats retournés par l'outil de recherche interrogé. De ce fait, il n'exploite qu'une portion minime des documents retrouvés en ne se consacrant qu'aux premiers résultats, et ainsi occulte d'éventuels documents pertinents. Pour remédier à cela, l'utilisateur dispose d'interfaces de visualisation.

Enfin, en aval de la recherche, l'utilisateur peut éprouver des difficultés à garder et surtout à organiser les informations qui l'intéressent.

Des solutions existent pour chacun de ces problèmes. Seulement, celles-ci ne tiennent pas compte du processus global de recherche mené par l'utilisateur (tâche de navigation, tâche de recherche...). Le découpage de ce processus ne permet pas aux différents systèmes d'exploiter une représentation unique des utilisateurs (besoins, centres d'intérêt...).

Le dernier aspect négatif du processus de RI classique sur le web provient de son caractère « solitaire ». Alors qu'Internet est a priori un média « ouvert », la RI sur le web reste une tâche essentiellement solitaire pour l'utilisateur. Au regard du nombre de systèmes basés sur une approche coopérative, nous pouvons constater que cet aspect revêt une grande importance car il permet de rapprocher les utilisateurs ayant des centres d'intérêt communs et ainsi de partager les informations qu'ils ont précédemment collectées.

Dans la partie suivante, nous présentons l'approche que nous proposons dans le but d'aider l'utilisateur durant son processus de recherche de façon globale. Cette approche porte sur un aspect coopératif fort, privilégiant le partage d'informations pour offrir une meilleure aide à l'utilisateur.

II

INTERFACE ADAPTATIVE POUR L'AIDE A LA RECHERCHE D'INFORMATION SUR LE WORLD WIDE WEB

APPLICATION DU PARTAGE DE CONNAISSANCES

INTERFACE ADAPTATIVE POUR L'AIDE A LA RI SUR LE WEB

II.1 Introduction

L'internaute, en quête d'informations, nous l'avons souligné, a besoin d'outils dans sa recherche sur le web. Or, un internaute n'est pas une entité solitaire et bon nombre d'autres individus (au sein d'organisations, d'entreprises ou de laboratoires de recherches par exemple) doivent avoir, ou ont déjà eu, des besoins proches voire similaires. Dans ce contexte, l'aspect coopératif prend tout son sens afin que chacun puisse amener sa contribution à aider autrui.

Cette partie développe notre contribution dans le cadre de l'aide à la Recherche d'Information sur le web. Ce travail est issu d'une réflexion sur les problèmes que peut rencontrer un internaute lors d'une recherche sur le web. Les problèmes que peut rencontrer un utilisateur ont été présentés dans la section précédente et sont relatifs à :

- la connaissance du domaine de recherche qui peut induire des difficultés à formuler ses besoins et un jugement incertain des documents visités,
- la difficulté qu'il peut ressentir à apprécier les résultats d'une recherche adhoc,
- l'aspect solitaire de sa navigation pouvant rendre une recherche infructueuse mais également l'effort cognitif qu'elle nécessite pour obtenir des documents pertinents et surtout éviter de se perdre au travers des liens,
- la difficulté de gérer et suivre l'évolution des documents intéressants qu'il a choisi de mémoriser. Cette difficulté est directement liée à l'évolution rapide du web.

Notre réflexion repose donc sur une vision globale du processus de recherche (navigation, recherche...) sur le web, ainsi que sur un aspect coopératif pour que chaque utilisateur puisse profiter de l'expérience de recherche mais également des informations capitalisées par les autres utilisateurs ayant des besoins proches voire similaires.

Nous présentons tout d'abord l'approche générale que nous proposons en mettant en évidence les différents objectifs que nous souhaitons atteindre. Nous détaillons ensuite ces solutions au travers du système *Easy-DOR* puis les résultats obtenus au cours des expérimentations réalisées. Nous terminons enfin ce chapitre par la présentation de l'interface que nous avons conçue et développée.

La présentation qui suit permet de justifier l'intérêt d'une telle interface basée sur le partage d'informations pour l'aide à la RI sur le web.

II.2 Approche générale proposée

Notre approche s'inscrit dans le processus de Recherche d'Information et vise à aider un utilisateur durant cette tâche. Cette approche est intrinsèquement orientée utilisateur et non système car l'utilisateur se situe au centre du processus de recherche.

Comme nous l'avons montré dans le chapitre précédent, les principaux éléments pouvant influencer l'efficacité (temps, qualité des résultats) de la recherche d'un utilisateur sont sa connaissance du domaine, le choix de l'outil de recherche, la façon de naviguer dans les documents, ainsi que la gestion et l'organisation des informations pertinentes retrouvées.

Ces aspects sont pris en compte dans notre démarche qui peut en outre être qualifiée :

- *d'intégrée* puisqu'elle propose au sein d'un même système différents modules permettant d'aider l'utilisateur dans les différentes tâches de la RI,
- *de coopérative* car elle repose sur un partage des informations au sein d'un groupe d'utilisateurs.

Nous avons privilégié une approche *intégrée*, car le fait d'utiliser divers modules de façon indépendante, comme c'est le cas dans les systèmes d'aide à la RI actuels, ne permet pas aux applications d'avoir une représentation unique des besoins et connaissances des utilisateurs permettant de mieux le connaître et donc de mieux le servir. Dans notre cas, l'intégration permet au système d'obtenir une représentation unique des centres d'intérêt de l'utilisateur exploitable au travers des différents modules, chacun d'eux dédié à un type d'aide particulier.

L'aspect *coopératif* est une réponse à la question :

« *Pourquoi ne pas exploiter les informations et connaissances capitalisées par un ensemble d'utilisateurs (ayant ou ayant eu des besoins similaires) pour aider un autre utilisateur dans ses propres recherches ?* ». Ainsi, notre approche consiste à faire profiter un internaute des informations acquises par d'autres utilisateurs afin de l'aider lors de ses recherches sur le web. Les domaines applicatifs privilégiés pour fournir une aide coopérative à la RI sur le web sont les organisations consommatrices d'informations provenant des nouvelles technologies (web, intranets...) telles que les entreprises, les écoles ou encore les laboratoires de recherche... Dans ces structures, les individus possèdent des centres d'intérêt proches voire similaires pour lesquels les informations devraient être partagées. Ces centres d'intérêt peuvent être soit motivés par l'emploi exercé, soit par un besoin personnel. Cette approche a pour but d'éviter aux différents individus de perdre trop de temps à chercher des informations que d'autres individus du groupe auraient déjà collectées. Elle permet ainsi d'optimiser leurs recherches sur le web (temps de recherche minimisé, effort cognitif limité...), en exploitant au mieux les informations partagées qui peuvent être issues du web visible (accessibles par des moteurs par exemple) mais également du web caché.

Nous avons choisi d'illustrer nos propos par l'application de notre approche dans un contexte de laboratoire de recherche. En effet, au sein d'un laboratoire de recherches, l'information revêt une importance capitale. L'intérêt d'une telle approche coopérative basée sur le partage d'informations peut se justifier à plusieurs niveaux :

- au niveau d'une équipe de recherche. Un chercheur intégrant une équipe se doit de faire le point sur les différentes recherches et informations précédemment capitalisées par l'équipe. Notre approche permet d'éviter de perdre du temps à « re-rechercher » certaines informations pour ainsi mieux se concentrer sur les nouvelles informations à prospecter,
- au niveau des différentes équipes présentes au sein du laboratoire. Les équipes d'un laboratoire ont des thématiques qui leur sont propres (intelligence artificielle, systèmes d'information par exemple). Elles peuvent cependant avoir des problématiques qui couvrent plusieurs thématiques et peuvent être partagées avec d'autres équipes. Ainsi, une équipe travaillant en intelligence artificielle et une équipe en systèmes d'information peuvent traiter d'une même problématique qui est la

représentation des besoins d'un utilisateur par exemple. Certes, les approches de ces deux équipes seront traitées d'un point de vue différent mais elles feront tout de même appel à des savoirs communs, par exemple pour constituer un état de l'art. Le partage d'informations prend alors tout son sens pour éviter de laisser chacune des équipes rechercher des informations que l'autre équipe posséderait déjà. Cet aspect permet d'accroître les collaborations et les rapprochements entre les individus.

Dans un tel cadre applicatif, le partage « manuel » des informations demande un investissement en temps trop important pour les chercheurs qui doivent régulièrement se tenir au courant de toutes les évolutions des centres d'intérêt de tous les membres du laboratoire, afin de pouvoir diffuser l'information de manière optimale.

Dans un contexte plus « industriel » ou « grand public », cette approche peut également être exploitée pour rapprocher différents individus par l'intermédiaire d'un partage d'information non anonyme. En effet, un problème majeur que l'on peut rencontrer dans une communauté est de trouver quelqu'un qui puisse nous aider. Si une personne profite d'un grand nombre d'informations provenant d'un même individu, elle peut voir en ce dernier un interlocuteur privilégié pour discuter de ses besoins. Cette démarche peut donc s'inscrire dans le cadre plus important des applications « People-to-People » (Individu à Individu).

Cependant, dans ce mémoire, nous nous sommes limités à un partage d'informations anonyme qui s'applique bien à un laboratoire de recherche.

Par ailleurs, la réflexion que nous avons menée nous a incité à proposer une approche n'introduisant pas un changement radical des habitudes de l'utilisateur afin que cela ne représente pas une charge trop contraignante. En effet, l'utilisateur du web est un utilisateur exigeant qui souhaite obtenir des informations rapidement, sans trop de possibilités d'erreurs ou d'échecs [Shneiderman, 1998], [Pejtersen, 1998]. Il préférera donc des outils simples dont il connaît bien le fonctionnement et qui lui fournissent une réponse rapidement même si celle-ci n'est pas toujours la plus pertinente.

Afin d'aider l'utilisateur au cours de sa tâche de recherche, nous proposons un ensemble de quatre modules applicatifs. Ces modules s'insèrent dans le processus général de recherche d'information sur le web comme le souligne la Figure 29. L'aide que nous proposons à l'utilisateur se situe à différents stades de la Recherche d'Information :

- *en amont de la recherche*. Un module basé sur la diffusion de l'information fournit à l'utilisateur de nouveaux documents relatifs à ses propres centres d'intérêt. Il offre à l'utilisateur la possibilité d'améliorer ou de mettre à jour sa connaissance des différents domaines qui l'intéressent grâce à de nouveaux documents qui lui sont proposés par le système. Ce module repose sur une tâche de filtrage cognitif (I.3.1.2.3.2),
- *au cours de la recherche.* Un premier module propose à l'utilisateur des documents potentiellement pertinents par rapport à sa navigation hypertexte en cours. Ces recommandations proviennent des connaissances capitalisées par l'ensemble des utilisateurs du système. Un second module offre une interface de visualisation des résultats de recherche afin d'aider l'utilisateur à comprendre et à manipuler les documents retrouvés de façon globale. Cette interface est associée à un outil de

recherche par interrogation permettant à l'utilisateur d'effectuer une recherche au travers des informations collectées par l'ensemble des usagers,

- *en aval de la recherche*. Un module permet une meilleure organisation et gestion des informations mémorisées qui intéressent l'utilisateur.

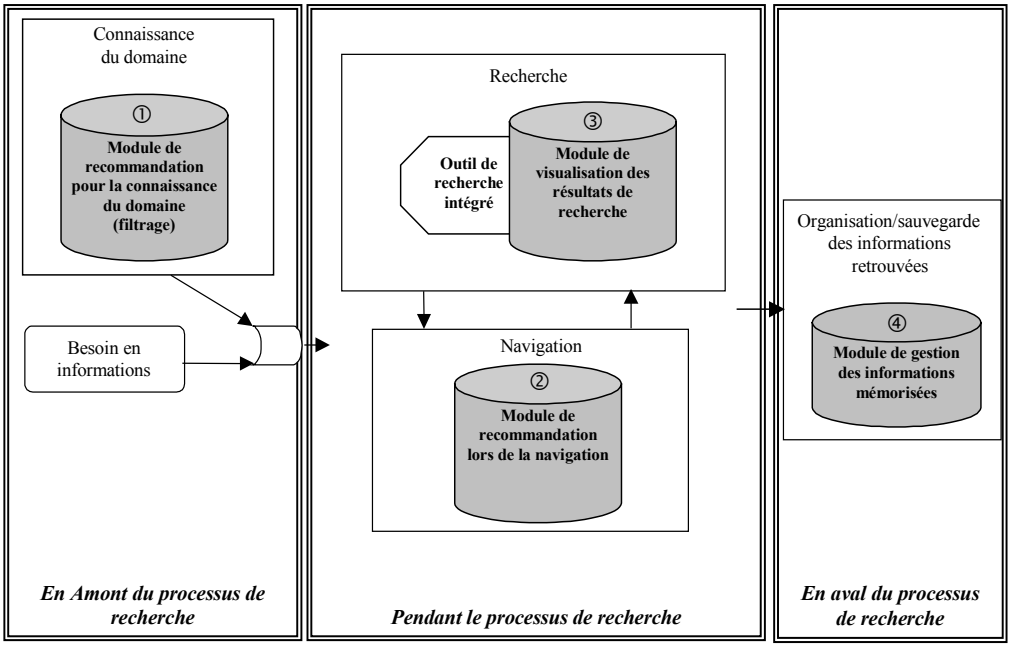


Figure 29 – Modules proposés d'aide à la RI sur le web

Nous détaillons chacun des modules proposés dans le cadre du projet *Easy-DOR* avant de présenter les spécificités du prototype qui en découle.

II.3 Le projet Easy-DOR

Le projet *Easy-DOR* (« Easy DOcument Retrieval ») [Chevalier, 2001b] est un projet visant à aider un utilisateur au cours d'une recherche d'information grâce à une approche coopérative en intégrant l'ensemble des modules décrits dans la section précédente.

Ce projet consiste à définir une *interface adaptative pour l'aide à la Recherche d'Information* sur le web qui résulte de l'étude des problèmes rencontrés par l'internaute lors d'une recherche d'information classique sur le web.

Définition. Interface adaptative pour l'aide à la RI sur le web :
- *Interface*. Application située entre l'utilisateur et le web,
- *Adaptative*. aptitude à s'adapter à une situation, à choisir en fonction des circonstances. Dans notre cas l'application s'adapte aux besoins de l'utilisateur et lui apporte une aide contextuelle,
- *Aide*. Apporte à l'utilisateur des informations lui permettant d'effectuer une recherche d'information plus efficace.

La plupart des modules proposés (module de recommandation pour la connaissance du domaine, module de recommandation lors de la navigation et le module de gestion des informations mémorisées) dans le projet repose sur les centres d'intérêt des utilisateurs.

Nous décrivons tout d'abord le choix réalisé quant au moyen utilisé pour représenter les centres d'intérêt, pour ensuite présenter chacun des modules proposés :

- module de recommandation pour la connaissance du domaine,
- module de recommandation durant la navigation,
- module de visualisation des résultats de recherche,
- module de gestion des informations mémorisées par l'utilisateur.

II.3.1 Représentation des centres d'intérêt d'un utilisateur

La plupart des modules (module de recommandation pour la connaissance du domaine, module de recommandation lors de la navigation et le module de gestion des informations mémorisées) repose sur une représentation des centres d'intérêt de l'utilisateur à partir des informations implicites que le système peut exploiter. Ces informations implicites ne nécessitent pas d'effort supplémentaire de la part de l'utilisateur mais sont exploitables par les modules au dépend d'une pertinence moins importante que les informations explicites [Procter, 1999].

Or, dans le contexte du web, l'utilisateur utilise différents outils tels que les hiérarchies de signets ou les systèmes d'annotations afin de mémoriser des documents intéressants qu'il rencontre au fil de ses recherches. En mémorisant ces documents, l'utilisateur réalise un jugement de pertinence implicite. Une organisation, de préférence hiérarchique, de ces informations pertinentes permet à l'utilisateur de construire un véritable espace d'informations concernant ses centres d'intérêt.

Ainsi, compte tenu du fait que nous ne souhaitons pas changer les habitudes des utilisateurs, nous avions le choix entre utiliser un système d'annotations ou un système de signets comme représentation des centres d'intérêt d'un utilisateur. Dans notre approche, les signets sont de meilleurs représentants des centres d'intérêt d'un utilisateur car :

- les signets sont naturellement utilisés par les internautes au travers de leur navigateur,
- les signets ne nécessitent que peu d'effort de création contrairement aux annotations,
- les signets peuvent intrinsèquement être organisés de façon hiérarchique contrairement aux annotations,
- un signet correspond à un jugement de pertinence positif implicite du document pour la problématique représentée par le répertoire dans lequel il a été inséré.

L'intérêt de l'organisation des signets en hiérarchie réside dans le fait qu'elle permet d'obtenir les relations entre un centre d'intérêt (représenté par un répertoire) et les documents qui le décrivent. Du fait que l'insertion d'un document dans un répertoire spécifique de la hiérarchie résulte d'un effort cognitif de la part de l'utilisateur [Rücker, 1997], les documents issus d'un même répertoire sont similaires, du point de vue de l'utilisateur, par rapport aux centres d'intérêt correspondant à ce répertoire.

Ces relations de similarité, issues d'un effort cognitif, nous sont notamment utiles afin de détecter les documents à recommander durant la navigation de l'utilisateur. Cependant, une nuance doit être apportée concernant cette organisation hiérarchique des signets. En effet, les utilisateurs peuvent avoir des stratégies d'organisation de leurs signets très

différentes [Abrams, 1998], [GVU, 1998]. La Figure 30 présente les résultats obtenus par Abrams et al [Abrams, 1998].

Légende de la Figure 30.

Aucune : les signets sont organisés sous la forme d'une liste au sein de laquelle les signets restent dans l'ordre dans lequel ils ont été créés.

Liste ordonnée : l'utilisateur reclasse manuellement les signets dans la liste.

Liste de répertoires : l'utilisateur utilise une liste simple de répertoires pour rassembler les signets (pas de sous-répertoires).

Hiérarchie de répertoires : l'utilisateur utilise une arborescence de répertoires pour classer ses signets.

Externe : les signets sont exportés vers une autre application.

Page web : l'utilisateur crée une page web à partir des signets.

Autre : autre méthode.

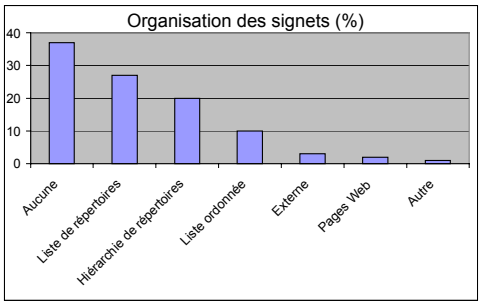


Figure 30 - Organisation des signets [Abrams, 1998]

Au regard de la Figure 30, nous pouvons constater que seulement 50% des utilisateurs organisent leurs signets en listes de répertoires ou en hiérarchie de répertoires. Dans notre approche, seules ces deux organisations nous permettent d'extraire des connaissances de ces hiérarchies de signets. Ce point a été pris en compte et l'incidence de cette dernière remarque est limitée par l'introduction du module de gestion des informations mémorisées.

Les centres d'intérêt d'un utilisateur sont donc issus, dans notre approche, de sa hiérarchie de signets qui permet également d'obtenir facilement des relations implicites entres les divers documents qu'elle contient.

Il est important de mentionner que, dans notre approche, les signets situés dans la racine de la hiérarchie de signets d'un utilisateur ne sont pas utilisés. En effet, les documents s'y trouvant n'ont fait l'objet d'aucun effort cognitif et aucune hypothèse ne peut être formulée par rapport à l'appartenance de ces documents à un quelconque thème. Ces documents sont généralement sauvegardés à cet endroit de façon temporaire pour permettre un rappel d'une navigation antérieure, pour une future réorganisation par exemple.

Dans les sections suivantes, les quatre modules proposés sont présentés au travers d'un rappel de la problématique lié à chaque module ainsi qu'au travers de la présentation de l'approche proposée.

II.3.2 Module de recommandation pour la connaissance du domaine

II.3.2.1 Problématique

La connaissance d'un domaine correspond au niveau d'expertise de l'utilisateur sur ce domaine. De cette connaissance découle, entre autres, la formulation de la requête ou encore l'interprétation des documents visités [Pejtersen, 1998]. Ainsi, un utilisateur qui ne connaît pas bien le domaine de recherche aura du mal à retrouver des documents pertinents sur le web. Il est donc primordial de tenir compte de cet aspect dans une approche d'aide à la RI sur le web car elle conditionne tout le reste du processus de recherche.

L'objectif de ce module est de permettre à l'utilisateur d'acquérir des informations pertinentes pour ses centres d'intérêt de façon automatique et régulière pour ainsi lui permettre soit d'améliorer soit de mettre à jour sa connaissance du domaine. Pour cela, le système exploite les documents visités par l'ensemble des utilisateurs du système afin d'identifier ceux qui peuvent intéresser chaque utilisateur.

II.3.2.2 Approche proposée

La nécessité de faire évoluer la connaissance d'un centre d'intérêt est né d'un constat de la vie réelle : il existe différents moyens pour acquérir une connaissance d'un domaine particulier (enseignements, livres...). Dans notre contexte, nous avons choisi d'augmenter la connaissance d'un domaine au travers des documents web visités par un ensemble d'utilisateurs. Nous avons proposé des mécanismes qui permettent au système de recommander à l'utilisateur des documents pertinents afin qu'il obtienne un grand nombre d'informations sur les sujets qui l'intéressent ; l'objectif étant, pour l'utilisateur, de consolider ou mettre à jour sa connaissance à partir de ces informations. Cette démarche peut être transcrite à la recherche d'information sur le web au travers d'un module de recommandation visant à enrichir de façon automatique et régulière les différents centres d'intérêt de l'utilisateur à partir de nouveaux documents pertinents. Cette approche est basée sur une démarche PUSH pour éviter à l'utilisateur de réaliser manuellement des recherches répétées sur le même thème. Grâce à ces documents recommandés, l'utilisateur a ainsi la possibilité d'acquérir de nouvelles connaissances ou simplement de mettre à jour celles qu'il possédait déjà. Cette meilleure connaissance lui permet d'effectuer de meilleures recherches ultérieures concernant ses centres d'intérêt grâce à des requêtes plus précises et un jugement des documents visités plus objectif.

Il s'avère qu'une approche de recommandation basée sur un aspect coopératif favorise la diffusion de l'information au sein d'un groupe d'individus ou d'une organisation (I.4.2.3.2). Cet aspect permet à un utilisateur d'exploiter les informations provenant d'utilisateurs qui partagent les mêmes centres d'intérêt.

Dans notre exemple, au sein d'un laboratoire de recherche, cette diffusion d'informations est d'autant plus importante qu'elle permet de partager les mêmes informations entre plusieurs chercheurs sans avoir à effectuer d'interminables recherches. De plus, les documents diffusés provenant des autres utilisateurs qui partagent éventuellement les

mêmes centres d'intérêt sont au moins aussi pertinents que ceux issus d'un outil de recherche.

Pour permettre une évolution de cette connaissance, nous proposons un module de recommandation de documents coopératif pour les centres d'intérêt des utilisateurs basée sur le contenu des documents visités par les utilisateurs du système. *SiteSeer*, par exemple, propose une approche intéressante de partage d'informations entre deux utilisateurs. En effet, celle-ci ne se base pas sur le contenu des documents mais sur les URLs des documents référencés par les signets des utilisateurs. Il repose sur le fait que deux individus possédant des signets identiques dans un répertoire ont des centres d'intérêt similaires qui peuvent être partagés (Figure 31).

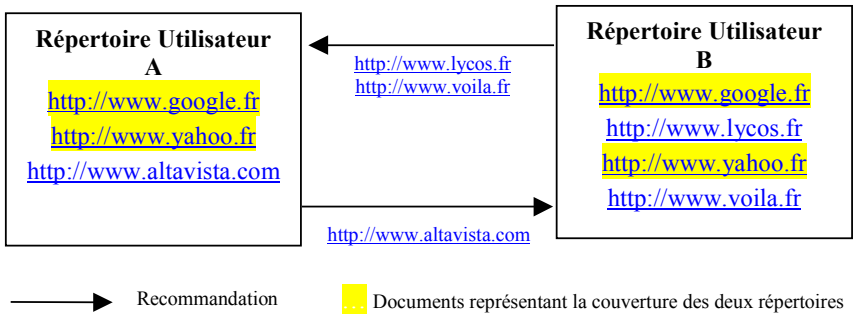


Figure 31 - Principe de recommandation de *SiteSeer*

Cependant, cette solution souffre de certaines limites :

- d'une part, compte tenu du nombre très important de documents disponibles sur le web, nous pouvons légitimement nous interroger sur la probabilité que deux individus, ayant des centres d'intérêt communs, possèdent des signets vers des documents identiques,
- d'autre part, *SiteSeer* ne tient pas compte de la construction arborescente d'une hiérarchie de signets et considère les répertoires individuellement. Or, le niveau de profondeur d'un répertoire correspond à un niveau de granularité d'une problématique initiée dans un répertoire racine.

Cette approche est également restrictive dans le sens où la recommandation ne repose que sur une fraction limitée des documents pertinents pour une problématique puisque l'utilisateur ne sauvegarde au sein de sa hiérarchie de signets que moins de la moitié des informations pertinentes qu'il trouve, les autres étant facilement retrouvées soit par l'URL soit par un moteur de recherche [Rücker, 1997].

Pour toutes ces raisons, nous avons fait le choix d'utiliser tous les documents web visités par l'ensemble des utilisateurs du système comme base de la recommandation plutôt que les seuls documents mémorisés au sein de leur hiérarchie de signets. Ce choix a été motivé par le fait que :

- il est nécessaire d'utiliser un maximum de documents pertinents pour obtenir une représentation plus précise des centres d'intérêt. Or, un utilisateur visite des documents pertinents qu'il n'insère pas nécessairement dans sa hiérarchie de signets,
- dans le contexte du partage d'informations, un individu qui visite un document (qu'il soit pertinent ou non pour ses centres d'intérêt) n'est pas censé savoir si celui-ci

peut intéresser un autre individu du groupe. Dans le cadre d'un laboratoire de recherche, ce dernier point est important car les chercheurs d'une même équipe de recherche ont des centres d'intérêt relativement proches correspondant à la ligne directrice des travaux de l'équipe (par exemple, la *recherche d'information*). Dans ce contexte, la diffusion manuelle d'information est facilitée car tous les membres d'une même équipe savent globalement sur quel thème les autres travaillent. Malheureusement, même si les chercheurs connaissent les centres d'intérêt « professionnels » de leurs collègues, ils peuvent ne pas soupçonner leurs centres d'intérêt personnels (*aquariophile* par exemple). Dans le cas des centres d'intérêt personnels, la diffusion manuelle des informations connaît alors ses limites du fait qu'il est nécessaire de connaître tous les centres d'intérêt de tous les individus de l'équipe pour pouvoir la réaliser. Cette limite s'accroît avec le nombre d'individus dans l'équipe. Ainsi, nous pouvons intuitivement comprendre la raison pour laquelle la diffusion d'informations « manuelle » entre les individus d'un laboratoire de recherche est limitée.

Les documents visités par l'ensemble des utilisateurs du système dont donc exploités afin de détecter ceux qui sont pertinents pour un utilisateur. Ce mécanisme repose sur une tâche de filtrage cognitif. Les centres d'intérêt de l'utilisateur sont extraits de sa hiérarchie de signets qui est construite sous la forme d'une arborescence. Celle-ci traduit une organisation selon les notions de généralisation/spécialisation de thèmes voire selon une certaine utilité (Figure 32).

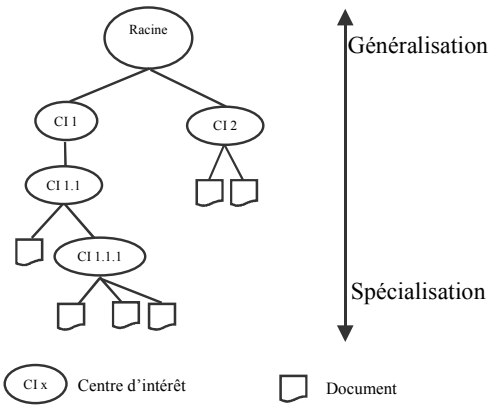


Figure 32 – Relations de spécialisation / généralisation dans une hiérarchie de signets

La profondeur d'un répertoire correspond à un niveau de granularité du centre d'intérêt initié dans le répertoire de niveau le plus faible c'est-à-dire contenu dans la racine de la hiérarchie de signets. Par exemple, dans la Figure 32, le répertoire CI1.1.1 est une spécialisation des répertoires CI1.1 et CI1.

Du fait de cette construction hiérarchique, le but de la tâche de filtrage est non seulement d'identifier si un document visité est pertinent pour un utilisateur, mais également de déterminer quel niveau de granularité d'un centre d'intérêt correspond le mieux au document. Ensuite, le module propose à l'utilisateur le document au travers d'un nouveau signet directement inséré dans sa hiérarchie de signets. L'utilisateur peut, à tout moment, émettre un jugement concernant la pertinence du document recommandé pour le

nœud dans lequel il a été inséré. Ces jugements positifs ou négatifs sont représentés par un triplet *(document, nœud, jugement)*. Ces jugements sont utilisés afin de caractériser le contenu des répertoires d'une hiérarchie de signets afin de vérifier si un document est pertinent ou non pour ce nœud.

Note : Afin d'uniformiser les informations manipulées, l'insertion manuelle d'un signet dans un répertoire donne lieu à un jugement positif entre le document pointé par le signet et le nœud dans lequel il a été inséré.

La tâche de filtrage que nous proposons repose sur la notion de classifieur qui est présentée dans la sous-section suivante.

Ces classifieurs sont ensuite considérés dans une arborescence qui correspond à la hiérarchie de signets d'un utilisateur. La démarche proposée est enfin évaluée dans le contexte de la collection de documents OHSUMED.

II.3.2.2.1 Notion de classifieur

Afin de savoir si un document visité est pertinent pour un utilisateur, nous avons utilisé la notion de « classifieur ». Un classifieur est un outil qui a pour but, à partir d'un document fourni en entrée et d'un ensemble de classes de documents, de fournir, en sortie une ou plusieurs de ces classes pour lesquelles le document est pertinent.

Un classifieur est caractérisé par deux paramètres :

- une fonction nommée *CSV* [0 ; 1] qui correspond à l'appariement entre le document et la représentation d'une classe. Plus la valeur est importante, plus le document est pertinent pour la classe,
- un seuil τ qui permet de passer de la valeur réelle de *CSV* à une valeur binaire (pertinent / non pertinent). Si l'appariement entre le document et le nœud (valeur de *CSV*) est supérieure au seuil τ, le document est accepté par le classifieur et jugé pertinent pour la classe. Dans le cas contraire, il est jugé non pertinent et est rejeté.

Divers classifieurs existent dans la littérature (MégaDocument, kNN, Rocchio, réseaux de neurones...) qui sont comparés dans [Yang, 1999], [Sebastiani, 1999], [Sebastiani, 2002].

Dans notre approche, nous avons envisagé d'utiliser les classifieurs Rocchio et Mégadocument afin de privilégier le lien avec l'utilisateur. En effet, à notre connaissance, seuls ces classifieurs sont basés sur la représentation d'une classe en profil qui est une représentation explicite des informations qu'elle contient. Un profil permet à l'utilisateur d'interroger le système pour comprendre le contenu de l'un de ses répertoires ou formuler une requête par exemple.

Nous présentons, dans cette section, les classifieurs Rocchio et Mégadocument qui seront utilisés et comparés dans nos expérimentations. Nous détaillons, ensuite, l'intérêt de réduire le nombre de termes pris en compte dans ces classifieurs et de la valeur du seuil τ

II.3.2.2.1.1 Rocchio

Le classifieur de Rocchio est une adaptation de la formule de Rocchio utilisée en réinjection de pertinence (I.4.2.2.2.1.2). Le classifieur Rocchio est donné par la formule 8.

$$w_{t,p} = \frac{\beta}{|D_p|} * \sum_{\{d \in D_p\}} w_{t,d} + \frac{\gamma}{|D_n|} * \sum_{\{d \in D_n\}} w_{t,d} \qquad 8$$

Où D_p(respectivement D_n) correspond à l'ensemble des documents ayant fait l'objet d'un jugement positif (respectivement négatif) pour le noeud,

$w_{t,d}$ correspond au poids du terme t dans le document d,

$w_{t,p}$ correspond au poids du terme t dans le profil,

β, γ sont des paramètres tels que $\beta \geq 0$ et $\gamma \leq 0$.

Le vecteur obtenu correspond au vecteur centroïde de la classe.

II.3.2.2.1.2 Mégadocument

Le mégadocument [Klas, 2000], quand à lui, repose sur l'idée de fusion de tous les documents pertinents en un seul document plutôt que de calculer un vecteur moyen comme avec Rocchio. La construction d'un mégadocument est réalisée à partir du nombre d'occurrences de tous les termes des documents pertinents dans la classe comme le montre la Figure 33. Une pondération des termes du mégadocument est ensuite réalisée par la formule classique *tf.idf* (I.3.1.2.1.3) pour obtenir la représentation de la classe.

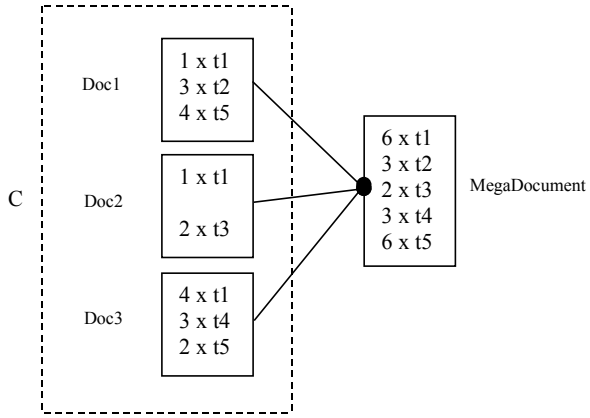


Figure 33 - Construction d'un mégadocument

II.3.2.2.1.3 Réduction de l'espace des termes

Un élément important à prendre en considération lors de la construction d'un classifieur est l'espace des termes utilisé. En effet, plus il y a de documents dans une classe, plus le nombre de termes augmente [Boughanem, 1992].

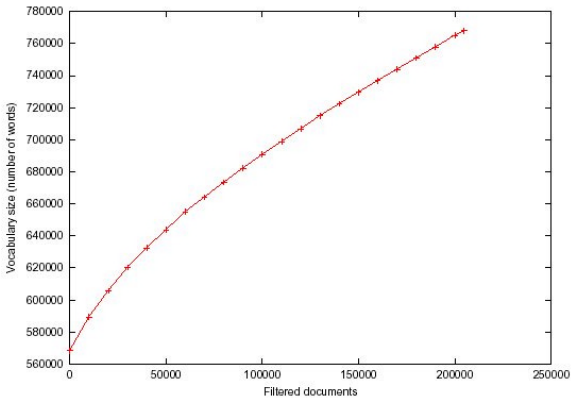


Figure 34 - Nombre de termes par rapport au nombre de documents filtrés (Financial Times 92-94 documents [Johansson, 2000])

Afin d'obtenir un classifieur optimal, que ce soit en terme de performances (limiter les calculs nécessaires) et de qualité (pertinence des termes par rapport au contenu du nœud), il est nécessaire de ne sélectionner que les termes qui sont les plus représentatifs des documents de la classe, c'est-à-dire le centre de gravité de la classe. Il existe un grand nombre de critères afin d'identifier et sélectionner l'ensemble des termes significatifs d'un ensemble de documents : fréquence du terme dans l'ensemble des documents, coefficient de corrélation, information mutuelle... Une étude comparative des principaux critères de sélection peut être trouvée dans [Yang, 1997]. Dans notre approche, nous avons choisi d'utiliser le critère statistique du χ^2 car il est l'un des critères permettant d'obtenir les meilleurs résultats [Yang, 1997]. Cette mesure privilégie les termes étant fréquemment présents dans les documents pertinents et peu présents dans les documents non pertinents. Cette mesure statistique est définie par l'équation 9.

$$\chi^2(t,c) = \frac{N(N_{r+}N_{n-} - N_{r-}N_{n+})^2}{(N_{r+} + N_{r-})(N_{n+} + N_{n-})(N_{r+} + N_{n+})(N_{r-} + N_{n-})} \qquad 9$$

Où t est un terme et c une classe,

N_{r+}(respectivement N_{n+}) correspond au nombre de documents pertinents (respectivement non pertinents) pour la classe c dans lesquels apparaît le terme t,

N_{r-}(respectivement N_{n-}) correspond au nombre de documents pertinents (respectivement non pertinents) pour la classe c dans lesquels n'apparaît pas le terme t,

N représente le nombre total de documents.

Outre la sélection des termes les plus significatifs d'une classe de documents, la valeur du seuil τ est un paramètre essentiel qui doit être pris en compte lors de la construction d'un classifieur. Nous rappelons que ce seuil permet de passer de la valeur réelle de l'appariement entre le document et la classe de documents à une valeur binaire (pertinent/non pertinent). De cette valeur dépend les performances de notre module de recommandation.

II.3.2.2.1.4 *Valeur du seuil* τ

Le seuil τ d'un nœud permet de définir si un document est jugé pertinent ou non par rapport à la valeur de *CSV*. Le document est jugé comme pertinent si la valeur de *CSV* est supérieure au seuil τ.

La valeur de ce seuil est donc primordiale pour obtenir la plus grande proportion de documents pertinents et la proportion la plus faible de documents non pertinents (Figure 35).

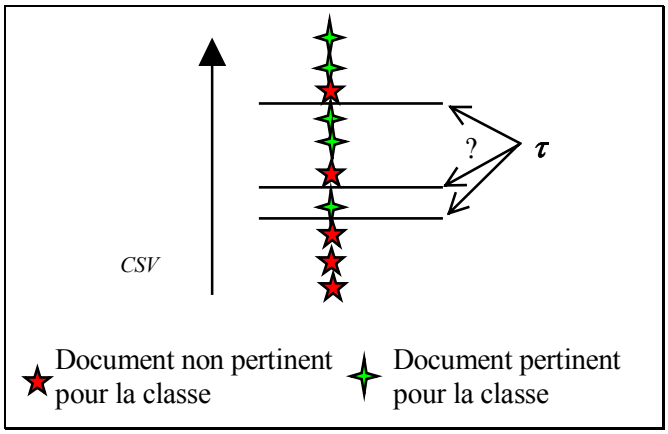


Figure 35 - Intérêt du seuil τ

Du fait que les différentes classes de documents peuvent être relativement hétérogènes, fixer le seuil à une valeur unique n'est pas envisageable. Une approche générale consiste à fixer le seuil d'un nœud à une valeur qui optimise les performances du système qui est également nommé seuil optimal ([Hoashi, 2000], [Wu, 2001], [Ruiz, 2001]). Une approche plus récente [Tmar, 2002] repose sur l'étude de la distribution des scores des documents. Dans notre approche, nous utiliserons un seuil permettant d'optimiser la valeur de la fonction *F1* (I.3.1.2.3.1). Cette approche est notamment utilisée dans [Ruiz, 2001].

II.3.2.2.2 Application des classifieurs aux hiérarchies de signets

Dans notre contexte, le rôle du classifieur est de définir si un document est pertinent pour un centre d'intérêt de l'utilisateur. Or, chaque centre d'intérêt est décomposé au travers de sa hiérarchie de signets. Nous avons suivi la démarche de [Koller, 1997] qui consiste à décomposer le problème général en des problèmes plus « fins » et avons affecté un classifieur à chaque répertoire de la hiérarchie. Le rôle du classifieur est d'indiquer si un document est pertinent ou non pour le répertoire auquel il est affecté (Figure 36).

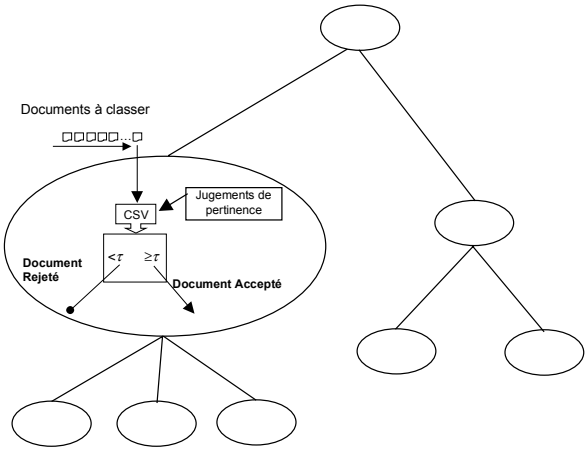


Figure 36 – Prise en compte de la hiérarchie

Par ailleurs, du fait de la structure arborescente, nous avons tenu compte de la distribution des termes permettant de caractériser les nœuds dans la hiérarchie. [Chakrabarti, 1998] souligne le fait qu'un terme a une importance qui dépend de sa profondeur dans la hiérarchie. Par exemple, un terme qui se retrouve dans tous les documents d'une hiérarchie doit avoir une forte importance dans les nœuds des niveaux faibles (proches de la racine) tandis qu'il doit avoir une importance faible dans un nœud ayant une profondeur importante (proche des feuilles). A l'inverse, un terme spécifique à un nœud proche des feuilles doit avoir un poids important à ce niveau et un poids très faible dans les nœuds proches de la racine. Pour obtenir une telle répartition des termes au sein de la hiérarchie, nous avons adopté la démarche utilisée notamment dans [Popescul, 2000], [Mladenic, 1998], [D'alessio, 2000] qui consiste à faire remonter l'ensemble des jugements de pertinence (*document, nœud, pertinence*) des nœuds fils vers les nœuds pères. Ainsi, pour obtenir les meilleurs termes discriminants pour un nœud, nous avons calculé le coefficient χ^2 en tenant compte que :

- l'ensemble des *documents pertinents pour le nœud* est composé des documents pertinents pour le nœud et pour tous les nœuds fils de ce dernier,
- l'ensemble des *documents non pertinents pour le nœud* est composé des documents non pertinents pour ce nœud et pour tous ses nœuds fils ainsi que tous les documents pertinents pour tous les autres nœuds de la hiérarchie (différents du nœud considéré et de ses nœuds fils).

Par ailleurs, pour obtenir la répartition adéquate des termes au sein de la hiérarchie et trouver le niveau le plus adéquat pour le document, nous avons utilisé un parcours descendant de l'arborescence (Figure 37). Cette approche est fréquemment utilisée dans la littérature [Koller, 1997], [Ruiz, 2001]. Elle consiste à parcourir l'arbre à partir de la racine en direction des feuilles. La recherche se poursuit au niveau des nœuds fils si et seulement si le document a été accepté par le classifieur du nœud courant.

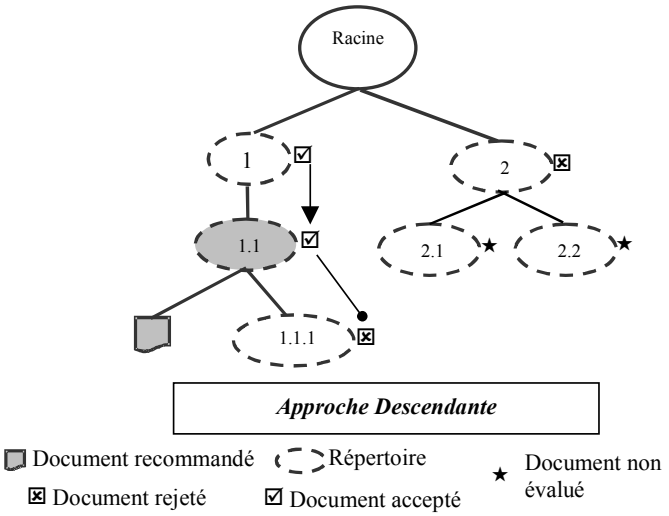


Figure 37 - Principe de recommandation hiérarchique descendant

Dans cette approche, nous avons limité la recommandation au nœud le plus profond dans un chemin pour lequel le document est pertinent. Dans l'exemple de la Figure 37, le document a été affecté seulement au nœud 1.1 et non au nœud 1.

Afin de vérifier l'efficacité d'un tel module, nous avons réalisé une évaluation. Celle-ci est présentée dans la section suivante.

II.3.2.3 Expérimentations

II.3.2.3.1 Collection test de documents

Afin d'évaluer la démarche proposée, à défaut de posséder une hiérarchie de signets « correcte », nous avons utilisé la collection de documents *OHSUMED* qui a notamment été utilisée dans la plate-forme d'évaluation *TREC* [Voorhees, 2001]. Cette collection est composée originalement de 349 566 documents issus de la base de documents en ligne *MEDLINE*. Les documents sont issus de 270 revues médicales sur une période de cinq ans (1987-1991).

Les documents sont indexés manuellement en utilisant les catégories de la hiérarchie *MEdical Subject Headings* (*MESH*) de la *National Library of Medicine*. Les documents sont fournis en langue anglaise et sont accompagnés de diverses informations complémentaires. La Figure 38 présente un document tel qu'il est donné dans la collection.

Les champs qui nous ont particulièrement intéressés sont les suivants :

- **.U** désigne l'identifiant du document,
- **.M** désigne les termes d'indexation correspondant à la hiérarchie de MESH,
- **.T** désigne le titre du document,
- **.W** désigne le contenu du document.

.I 16819
.U
87197590
.S
J Neurosurg 8708; 66(6):830-4
.M
Adolescence; Adult; Arteriovenous Malformations/*DI; Case Report; Female; Human; ….
.T
Magnetic resonance imaging of spinal arteriovenous malformations.
.P
JOURNAL ARTICLE.
.W
Magnetic resonance imaging (MRI) was performed on 12 patients with spinal arteriovenous malformations (AVM's). Six lesions were intramedullary, five were dural, and one was in a posterior extramedullary location. Serpentine filling defects similar to the classic myelographic findings were demonstrated within the high-signal cerebrospinal fluid on T2-weighted coronal scans. The intramedullary nidus was identified by MRI as an area of low-signal intensity within the cord in all six intramedullary AVM's. Neither the dural nor the posterior extramedullary lesions showed intramedullary components. It is concluded that MRI may noninvasively provide the initial diagnosis of a spinal AVM and distinguish intramedullary from dural and extramedullary lesions.
.A
Doppman JL; Di Chiro G; Dwyer AJ; Frank JL; Oldfield EH.

Figure 38 - Exemple d'un document issu de la collection OHSUMED

Les documents ont un contenu limité à 250 mots mais certains autres ne possèdent pas de contenu (.W).

II.3.2.3.2 Cadre expérimental

Afin de réaliser une évaluation la plus proche possible de notre contexte applicatif nous n'avons utilisé qu'un sous-ensemble de la hiérarchie de *MESH*. En effet, les hiérarchies de signets ne contiennent pas un grand nombre de nœuds et de documents dans chaque nœud.

Pour réduire le nombre de nœuds dans les expérimentations, nous avons exploité une sous-partie de la hiérarchie initiale de MESH qui correspond à la hiérarchie des maladies cardio-vasculaires. Cette arborescence est composée de 146 nœuds (seulement 119 d'entre eux contiennent des documents). Elle possède un niveau de profondeur maximum égal à 6. Cette arborescence est présentée en annexe H.

De plus, nous avons réalisé les expérimentations à partir d'un ensemble restreint de documents issus de la collection de test pour permettre de limiter le nombre de documents par nœuds dans les expérimentations. Or, la collection de documents issue de la plate-forme TREC est composée de deux ensembles de documents :

- une collection d'entraînement, qui consiste en un ensemble de documents, associée à un ensemble de jugements de pertinence permettant au système de débuter le processus,
- une collection de test permettant d'évaluer les performances du système.

Dans nos expérimentations, nous nous sommes donc limités aux 4974 documents de la collection d'entraînement utilisée dans *TREC* qui s'avèrent pertinents pour les nœuds de la hiérarchie des maladies cardio-vasculaires. Les nœuds de la hiérarchie ainsi obtenue possèdent entre 4 et 302 documents et près de 70% d'entre eux possèdent moins de 54 documents (qui constitue la moyenne). Nous pouvons également constater le fait que les documents utilisés sont assez courts car, après radicalisation et épuration, la longueur moyenne des documents est d'environ 90 termes.

Pour réaliser nos expérimentations, nous avons scindé la collection de 4974 documents en une collection d'entraînement et une collection de test. La proportion utilisée est 2/3 de documents d'entraînement et 1/3 de documents de test.

L'indexation des documents a été réalisée en utilisant le titre ainsi que le contenu des documents de manière traditionnelle (§ I.3.1.2.1.2) avec analyse morpho-syntaxique, suppression des mots vides de la langue anglaise et radicalisation par les règles de Porter.

Les expérimentations conduites dans cette section se basent sur deux aspects :

- la comparaison des classifieurs dans notre contexte,
- l'impact de la hiérarchie sur les performances du système (incidence de la migration des documents des fils vers le nœud père, parcours descendant de l'arborescence).

A des fins de comparaison entre les différentes expérimentations, nous avons utilisé les mesures dites *Micro*, *Macro* et *MacroAvg*. Ces mesures s'appliquent sur la précision, le rappel ou encore sur la fonction *F1* (§ I.3.1.2.3.1). Ces mesures sont utilisées pour combiner les performances obtenues par les différents classifieurs (Tableau 7). La mesure *Micro* tend à accentuer la performance des classes de grande taille. A l'inverse, la mesure *Macro* tend à accentuer les performances des petites classes [Lewis, 1991].

	Précision	***Rappel***	***F_1***
Micro	$microP = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} (TP_i + FN_i)}$	$microR = \frac{\sum_{i=1}^{\|C\|} TP_i}{\sum_{i=1}^{\|C\|} (TP_i + FP_i)}$	$microF_1 = \frac{2 * microP * microR}{microP + microR}$
Macro	$macroP = \frac{\sum_{i=1}^{\|C\|} \frac{TP_i}{(TP_i + FN_i)}}{\|C\|}$	$macroR = \frac{\sum_{i=1}^{\|C\|} \frac{TP_i}{(TP_i + FP_i)}}{\|C\|}$	$macroF_1 = \frac{2 * macroP * macroR}{macroP + macroR}$
MacroAvg			$macroAvgF_1 = \frac{\sum_{i=1}^{\|C\|} F_{1i}}{\|C\|}$

Tableau 7 - Mesures d'évaluation des classifieurs

Où TP_i = ensemble des documents réellement pertinents reconnus comme pertinents par le classifieur *i*,

FP_i = ensemble des documents réellement pertinents reconnus comme non pertinents par le classifieur *i*,

FN_i = ensemble des documents réellement non pertinents reconnus comme pertinents par le classifieur *i*,

$F1_i$ = valeur de *F1* pour le classifieur *i*,

C = nombre de classifieurs.

Notre but étant de maximiser les performances pour des classes ayant peu de documents, dans nos expérimentations, nous avons privilégié les mesures *macro.*

II.3.2.3.2.1 Comparaison des classifieurs : Rocchio vs Mégadocument

Dans notre cadre expérimental, nous avons souhaité, dans un premier temps, mettre en relief les différences entre les deux classifieurs sélectionnés, c'est-à-dire Rocchio et Mégadocument appliqués à notre contexte (cf. II.3.2.3.2). Le classifieur Rocchio utilise deux paramètres α et β qui permettent de pondérer l'importance d'un terme dans les documents pertinents (α) et dans les documents non pertinents (β).

Pour tenir compte de ces paramètres, nous avons comparé le classifieur mégadocument au classifieur Rocchio avec les valeurs les plus répandues dans la littérature, à savoir :

- $\alpha = 1$; $\beta = 0$
- $\alpha = 1$; $\beta = 1$
- $\alpha = 2$; $\beta = 1$
- $\alpha = 16$; $\beta = 4$

Dans un premier temps, afin d'éviter toute incidence de la hiérarchie sur la comparaison des classifieurs, nous avons limité cette première expérimentation aux seuls nœuds feuilles de la hiérarchie utilisée. Ils sont au nombre de 95 et contiennent entre 2 et 105 documents pertinents (la moyenne étant 24 documents par nœud).

Dans un premier temps, nous avons construit un classifieur optimal pour chacun de ces nœuds à partir des documents d'entraînement. Pour cela, nous avons utilisé les termes issus des documents d'entraînement présents dans les nœuds.

A partir de ces termes, nous avons construit le classifieur et l'avons appliqué sur l'ensemble des documents d'entraînement afin d'en déduire le seuil qui optimise la valeur *F1*. Ce seuil correspond à la valeur *CSV* (cf. II.3.2.2.1) du document pour laquelle la valeur de *F1* du classifieur est maximum.

Cependant, étant donné que les performances du classifieur dépendent des termes utilisés, nous avons reproduit ce processus de calcul du seuil pour différents sous-ensembles de termes. Nous avons calculé le seuil ainsi que la valeur de *F1* pour les *n*% meilleurs termes du nœud par rapport à leur valeur χ^2 (*n* variant de 0 à 100% par pas de 5%).

Un classifieur que nous qualifions d'*optimal* est ainsi obtenu pour une valeur de *n* et un seuil qui optimisent la valeur *F1* du classifieur.

Pour évaluer les performances réelles du système, nous avons appliqué l'ensemble de test aux différents classifieurs. Le résultat de cette première expérimentation est présenté dans la Figure 39. Cette figure montre, pour chaque classifieur, la valeur de la MacroF1.

Dans cette figure, la lettre *R* signifie le classifieur Rocchio et *M* le mégadocument. En ce qui concerne les classifieurs Rocchio, les valeurs de α et β sont également précisées. Par exemple, $R_{1/0}$ correspond au classifieur Rocchio pour lequel α vaut 1 et β vaut 0.

Méthode	MacroF1
$R_{1/0}$	0.699
$R_{1/1}$	0.613
$R_{2/1}$	0.682
$R_{16/4}$	0.658
M	0.623

Figure 39 - Comparatif des différents classifieurs

Nous pouvons constater qu'il y a une grande différence entre les performances des classifieurs. Le classifieur de Rocchio $R_{1/0}$ obtient les meilleurs résultats en terme de MacroF1 (0.699). Cependant, nous avons remarqué que les différences se situent également au niveau du nombre moyen de termes utilisés (*n*) pour construire chaque classifieur. Au travers de la Figure 40, nous pouvons constater que le classifieur Rocchio $R_{1/0}$ utilise, en moyenne, plus de termes que les autres classifieurs.

	$R_{1/0}$	$R_{1/1}$	$R_{2/1}$	$R_{16/4}$	MegaDoc
%Termes utilisés	21.53%	18.10%	14.53%	5.95%	5.68%

Figure 40 - Pourcentage moyen de termes utilisés dans les classifieurs

Nous avons souhaité aller plus loin dans cette expérimentation en appréciant la sensibilité des classifieurs au nombre de documents présents dans les nœuds. Nous avons alors évalué la sensibilité des classifieurs en fonction du nombre de termes utilisés par rapport au nombre de documents dans le nœud. En effet, plus il y a de documents dans un nœud, plus il y a de termes permettant de décrire ce nœud. Pour cela, nous avons calculé le coefficient de corrélation entre le nombre de termes utilisés par le classifieur et le nombre de documents pertinents pour le nœud (Figure 41).

	$R_{1/0}$	$R_{1/1}$	$R_{2/1}$	$R_{16/4}$	MegaDoc
Coéfficient Corrélation	0.30	0.36	0.39	0.83	0.89

Figure 41 - Corrélation entre le nombre de termes utilisés par le classifieur et le nombre de documents pertinents dans le nœud

Cette figure permet de souligner la différence de sensibilité des classifieurs au nombre de documents. Nous pouvons ainsi constater que pour les classifieurs Rocchio $R_{1/0}$, $R_{1/1}$ et $R_{2/1}$, la corrélation entre le nombre de termes à utiliser *(n)* pour construire le classifieur optimal et le nombre de documents est relativement faible (moyenne 0.35). Ainsi, il est difficile de généraliser le nombre de termes à prendre en compte dans chaque classifieur. Cet aspect est une limite importante pour obtenir un ensemble de classifieurs optimaux car, pour cela, il est nécessaire d'évaluer le nombre de termes optimal pour chacun des nœuds de la hiérarchie, ce qui nécessite un gros investissement notamment en temps de calcul.

A l'inverse, nous pouvons constater que les classifieurs Rocchio $R_{16/4}$ et Mégadocument sont sensibles au nombre de documents. Le nombre de termes utilisés est très lié au nombre de documents dans le nœud (moyenne 0.86). Il est donc plus facile de généraliser le nombre de termes à prendre en compte pour obtenir un classifieur optimal.

Nous avons rapproché ces résultats avec la proportion de termes à prendre en compte pour obtenir les classifieurs optimaux (Figure 42). Cette figure montre que la proportion des termes est assez homogène (écart-type faible) pour les classifieurs sensibles au nombre de documents dans le nœud. A l'inverse, les autres classifieurs possèdent un écart-type important signifiant que la proportion des termes est assez hétérogène sur l'ensemble des classifieurs et ne permettant pas une généralisation.

	R1/0	R1/1	R2/1	R16/4	MegaDoc
Proportion Moy.	21.53%	18.05%	14.53%	5.95%	5.68%
Ecart-Type	**17.98%**	**18.92%**	**14.48%**	**3.95%**	**2.01%**

Figure 42 - Proportion moyenne de termes utilisés

Pour résumer cette première expérimentation, nous pouvons dire que le classifieur Rocchio $R_{1/0}$ optimal permet d'obtenir les meilleurs résultats au prix d'une recherche individuelle du nombre de termes à utiliser pour chaque nœud gourmande en ressources. A l'inverse, les classifieurs Rocchio $R_{16/4}$ et Mégadocument obtiennent des performances moindres (respectivement -5,9% et -10,94%) mais sont, à l'opposé, généralisables en ce qui concerne la proportion des termes à prendre en compte pour caractériser le nœud. Le

classifieur Rocchio $R_{16/4}$ est donc une alternative au classifieur Rocchio $R_{1/0}$ car les performances restent acceptables (-5.9% par rapport à Rocchio $R_{1/0}$) ce qui n'est pas le cas pour le Mégadocument (-10.94% par rapport à Rocchio $R_{1/0}$).

Dans la suite des expérimentations, nous nous sommes limités au meilleur classifieur évalué c'est-à-dire le classifieur Rocchio $R_{1/0}$ optimal sur l'ensemble de la hiérarchie des maladies cardio-vasculaires.

II.3.2.3.2.2 Impact de la hiérarchie sur la performance des classifieurs

Afin de mettre en évidence l'impact de la hiérarchie sur les performances du système, nous avons réalisé une expérimentation « à plat ». Cette expérimentation est utilisée comme référence pour apprécier l'apport de l'aspect hiérarchique.

Dans l'expérimentation « à plat », nous avons considéré que chaque nœud est indépendant. Pour ce faire, nous avons calculé le classifieur optimal de chaque nœud en suivant la même démarche que précédemment (recherche du seuil et de la proportion de termes permettant d'optimiser le fonction *F1* du nœud) en ne prenant en compte que l'ensemble des documents contenus dans chaque nœud.

Dans un premier temps, dans la suite des expérimentations, nous avons mis en évidence l'intérêt de faire remonter les jugements de pertinence des fils vers le père afin d'obtenir une répartition des termes dans l'arborescence répondant à la notion de généralisation/spécialisation. Dans un second temps, nous avons mis en évidence l'impact du parcours descendant de la hiérarchie de signets sur les performances des classifieurs.

Le premier aspect que nous avons souhaité mettre en évidence est l'aspect hiérarchique des nœuds. Contrairement à l'expérimentation précédente, dans cette expérimentation « hiérarchique », les documents d'un nœud fils migrent vers les nœuds pères. Ainsi, un document pertinent pour un nœud fils l'est également pour tous ses pères. Par contre, un document pertinent pour le nœud père n'est pas pertinent pour ses nœuds fils.

La comparaison entre l'expérimentation « à plat » (sans remontée des documents d'un nœud fils aux nœuds pères) et l'expérimentation « hiérarchique » (remontée des documents d'un nœud fils aux nœuds pères) est présentée dans le Tableau 8.

MacroF1 hiérarchique	0.524
MacroF1 « à plat »	0.498

Tableau 8 - Résultats de l'expérimentation hiérarchique

Nous pouvons constater que le fait de faire remonter les documents des nœuds fils vers les nœuds pères permet d'obtenir une représentation plus précise des nœuds. Cette amélioration de la représentation d'un nœud favorise une meilleure affectation des documents pertinents (+5% de la *macroF1*) principalement au niveau des nœuds non feuilles. En effet, les nœuds étant des nœuds feuilles de l'arborescence ne profitent pas de l'aspect hiérarchique.

Cependant, cette amélioration n'est pas systématique et pour certains nœuds, l'aspect hiérarchique est considéré comme préjudiciable. L'exemple le plus significatif est le nœud 1195 nommé « Cerebrovascular Disorders » (cf annexe H) pour lequel la valeur F1 chute de

près de 42% si nous prenons en compte l'aspect hiérarchique. D'autres nœuds connaissent également des chutes de performances tels que les nœuds (999 « Heart Block », 1066 « Heart Rupture », 1343 « Ischemia ») mais cette baisse est moins significative.

Après vérification, il s'avère que la chute de performance provient essentiellement d'une chute du taux de précision dans ces nœuds (Figure 43).

Nœud	*Précision Expérimentation à plat*	*Précision Expérimentation hiérarchique*	*Evolution*
999	*0.35*	*0.31*	**-11.11%**
1066	*0.83*	*0.33*	**-60.00%**
1195	*0.46*	*0.18*	**-59.69%**
1343	*0.42*	*0.40*	**-5.00%**

Figure 43 - Baisse constatée de la précision dans les nœuds pour lesquels la valeur de F1 chute

Nous interprétons cette baisse de la précision par une hétérogénéité des documents contenus dans ces nœuds ainsi que dans leurs arborescences filles. Nous avons utilisé la méthode du χ^2 pour identifier les termes qui se retrouvent fréquemment dans les documents pertinents et peu fréquemment dans les documents non pertinents. Le χ^2 permet d'identifier les termes qui se retrouvent le plus souvent dans les documents pertinents et le moins souvent dans les documents non pertinents. Or, si les documents du nœud sont très hétérogènes, les termes issus du χ^2 peuvent d'être très généraux. De ce fait, des documents à priori non pertinents risquent d'être jugés pertinents par le classifieur.

Pour conclure cette série d'expérimentations, nous avons évalué l'impact d'un parcours descendant spécifique sur les performances du système.

Les résultats obtenus pour ce parcours sont proposés dans la Figure 44. Ce parcours descendant permet d'obtenir de meilleurs résultats du fait que l'on n'assigne un document qu'à un seul nœud de l'arborescence ce qui limite les erreurs à un seul nœud. Ainsi, les erreurs possibles ne sont pas diffusées dans différents nœuds comme c'est le cas sans parcours spécifique de l'arborescence (expérimentation précédente). Nous pouvons constater que l'approche descendante permet d'obtenir des résultats meilleurs de plus de 43% par rapport à une expérimentation sans parcours spécifique.

	MacroF1
Sans parcours hiérarchique	**0.524**
Parcours Descendant	**0.752**

Figure 44 – Résultats du parcours descendant

Nous avons également étudié le temps nécessaire pour classer les documents grâce à ce parcours descendant (Figure 45). Dans cette figure, les temps de calculs sont donnés en milli-secondes hors temps de construction des classifieurs. Le parcours descendant permet de réduire le temps de calcul de plus de 69% du fait que le nombre d'appariements est limité aux seules branches pour lesquelles un document est pertinent.

	Temps (ms)
Sans parcours hiérarchique	**1179125**
Parcours Descendant	**365953**

Figure 45 - Temps de classification des documents selon le mode de parcours de l'arborescence

Pour résumer cette série d'expérimentations, nous pouvons souligner l'impact de la hiérarchie, que ce soit au niveau de la migration des documents des nœuds fils vers les nœuds pères ou des parcours de la hiérarchie. Cet aspect permet d'accroître les performances du système de plus de 43% et le temps de calcul de plus de 68%. Cependant, cette méthode semble trouver ses limites dans un contexte où le contenu des différents nœuds fils sont hétérogènes (peu de termes en commun notamment).

II.3.2.4 Bilan sur le module de recommandation pour la connaissance du domaine

Le module de recommandation pour la connaissance du domaine offre la possibilité à un utilisateur de profiter automatiquement des documents visités par l'ensemble des utilisateurs du système et qui sont pertinents pour ses propres centres d'intérêt. Contrairement aux approches existantes, la nôtre repose uniquement sur l'utilisation des documents visités par les utilisateurs du système et non pas uniquement sur les documents qu'ils mémorisent (comme *SiteSeer*) ou sur les documents issus d'une interrogation d'un moteur de recherche (*Watson* ou *WBI* par exemple). Cette approche permet une meilleure circulation des documents visités du fait qu'un document n'intéressant pas un utilisateur peut en intéresser un autre ayant le même centre d'intérêt mais ayant un point de vue différent. D'autre part, cette approche permet à tout utilisateur de bénéficier de l'expérience en recherche qu'ont acquis les autres utilisateurs du système. Le système proposé permet donc de mettre l'utilisateur à l'écoute de nouveaux documents au travers de recommandations. A partir de ces recommandations, l'utilisateur peut mettre à jour, voire faire évoluer sa connaissance, dans les domaines qui l'intéressent. Ce module permet également, du point de vue du partage des informations, de construire une représentation plus complète des centres d'intérêt de chaque utilisateur au travers d'un plus grand nombre de jugements de pertinence qui peuvent être exploités dans les autres modules d'aide que nous proposons.

Au travers des expérimentations, nous avons, dans un premier temps, souligné la sensibilité des classifieurs Rocchio et mégadocument au nombre de documents dans une classe lorsque l'on les associe à la mesure du χ^2. Il s'agit d'une caractéristique importante puisque le nombre de termes à prendre en compte en dépend. Nous avons également pu mettre en évidence l'intérêt mais aussi la limite de tenir compte de la hiérarchie des nœuds pour obtenir de meilleures recommandations. Le fait de remonter les documents des nœuds fils vers les nœuds pères permet globalement d'améliorer l'affectation des documents aux nœuds (visible au travers de la mesure macroF1) entre la racine et les feuilles. Mais cet aspect hiérarchique est également une limite. Dans le cas où des nœuds fils sont assez hétérogènes, le fait de remonter les documents au nœud père implique une baisse des performances du fait de la dispersion du centre de gravité des nœuds. Cette limite peut être attribuée à la linéarité des classifieurs utilisés.

Nous avons également démontré l'intérêt de pratiquer un parcours descendant de l'arborescence afin d'optimiser les performances du système (affectation des documents aux nœuds ainsi que temps de calcul) et limiter les erreurs de recommandations.

En conclusion, la qualité des recommandations faites par les classifieurs proposés dépend de l'organisation de la hiérarchie initiale de documents et plus particulièrement de l'homogénéité du contenu des documents au sein d'une même branche d'un arbre. Pour vérifier nos conclusions, il serait intéressant de pratiquer des expérimentations à partir d'autres classifieurs (non linéaires tel que les k-plus-proches-voisins) et d'autres hiérarchies de documents. Une étude en profondeur des caractéristiques des classifieurs serait également intéressante à mener comme leur sensibilité au nombre de documents ou encore l'impact de l'aspect hiérarchique sur leurs performances. Le module proposé permet d'atteindre les objectifs escomptés c'est-à-dire la recommandation de documents pertinents pour les centres d'intérêt d'un utilisateur.

Ce module permet de tenir compte des centres d'intérêt à long terme de l'utilisateur en amont de la Recherche d'Information. Cependant, il est nécessaire de tenir également compte des besoins qui pourraient être ponctuels lors d'une recherche d'information. Les sections suivantes s'intéressent à l'aide que le système offre durant le processus, c'est-à-dire durant sa navigation ou lors de l'exploitation des résultats d'une recherche adhoc au travers d'une interface de visualisation.

II.3.3 Module de recommandation lors de la navigation

II.3.3.1 Problématique

La Recherche d'Information fait appel alternativement à une tâche de recherche ainsi qu'à une tâche de navigation. Cette dernière offre la possibilité de parcourir, à partir d'un document, les autres documents de l'hypertexte en suivant les liens. Elle peut correspondre à des besoins récurrents relatifs aux centres d'intérêt de l'utilisateur ou à des besoins ponctuels. Au cours de cette navigation, l'utilisateur visite des documents qui correspondent à ses besoins. Les outils actuels tendent à optimiser cette recherche par navigation en proposant des documents pertinents à l'utilisateur sous la forme de recommandations.

Dans la littérature, deux grands types d'approches existent (cf I.4.2.3.2) :

- les approches basées sur des recommandations issues de l'hypertexte local du document visité,
- les approches basées sur des recommandations issues de l'interrogation d'un outil de recherche extérieur.

Cependant, la principale limite des approches actuelles comme celles utilisées dans *Webwatcher* ou encore *Letizia*, réside dans le fait que les documents recommandés restent situés dans l'hypertexte local du document visité. Ainsi, un document recommandé est forcément accessible par le biais de la navigation au travers d'un nombre de liens plus ou moins important (pour *Webwatcher* ou *Letizia* ce nombre est égal à 1). Ces approches peuvent être qualifiées comme étant des « accélérateurs » de navigation car les documents auraient pu être visités sans leur assistance en suivant les liens hypertextes.

D'autres outils tels que *WBI* et *Watson* recommandent, quant à eux, des documents qui ne sont pas forcément dans l'hypertexte local des documents visités. Ces documents sont

issus d'une interrogation d'outils de recherche externes. Dans ce cas, l'aspect coopératif est limité voire inexistant car les documents proviennent d'une interrogation de moteurs de recherche traditionnels basés sur le contenu. Cet aspect coopératif est plus développé dans *FAB* car un utilisateur profite de documents issus d'une interrogation d'un outil de recherche externe qui sont soit pertinent pour ses propres centres d'intérêt soit pertinents pour un utilisateur ayant les même centres d'intérêt (similarité entre profils utilisateurs).

La navigation est représentée, dans la plupart des approches de recommandation, durant la navigation (cf. I.4.2.3.2), par un profil (liste de termes pondérés) construit à partir des représentations des documents visités. Cette démarche limite ainsi la recommandation aux seuls documents similaires au profil de navigation. Or, un thème ou une problématique peuvent être décrits par un ensemble de termes différents d'un document à un autre selon le point de vue envisagé (utilisation de synonymes par exemple...).

L'objectif de ce module est de proposer à un internaute, au cours de sa navigation, des documents ayant une bonne valeur ajoutée qui correspondent à son thème de recherche. En effet, ces documents sont issus des informations collectées et organisées par l'ensemble des utilisateurs du système. De ce fait, les documents ainsi recommandés ne sont pas nécessairement présents dans l'hypertexte local des documents visités.

II.3.3.2 Approche proposée

La démarche coopérative que nous proposons au travers du module de recommandation durant la navigation peut être qualifiée de complémentaire aux approches traditionnelles. Elle se traduit par une approche exploitant les informations collectées par les utilisateurs d'un groupe afin d'identifier les documents pertinents pour une navigation en cours [Chevalier, 2002]. Cette navigation est appréciée de façon globale, c'est-à-dire que les recommandations évoluent au fur et à mesure des documents visités pour adapter au mieux les documents proposés au contexte de la navigation. Cette adaptation permet également au module de recommandation, durant la navigation, de suivre l'évolution des besoins de l'utilisateur à partir des documents visités.

Par exemple, dans le cadre d'un laboratoire de recherche, cette démarche de recommandation permet à des individus d'obtenir des documents pertinents complémentaires correspondant au thème de leur navigation. Un doctorant qui effectue une recherche par navigation concernant le filtrage d'informations obtiendra automatiquement des documents en rapport avec ce thème provenant des membres du laboratoire partageant ce même centre d'intérêt.

Contrairement aux projets existants, la pertinence des documents recommandés ne repose pas sur une représentation statistique du contenu des documents et de la navigation. Afin de caractériser la navigation, nous avons défini un profil de navigation qui contient les informations relatives à la navigation en cours c'est-à-dire les URLs des documents visités par l'utilisateur ainsi que la liste des recommandations que le système va proposer à l'utilisateur.

La démarche pour construire et mettre à jour la liste de recommandations repose sur un ensemble de documents pertinents pour chaque document visité. Cet ensemble de

documents pertinents est issu des informations collectées et organisées par les utilisateurs du système. Puisque dans le processus proposé, les informations collectées par les utilisateurs sont organisées au sein des hiérarchies de signets, nous avons considéré qu'il serait intéressant d'exploiter cette connaissance pour obtenir une similarité implicite entre les documents situés dans une même hiérarchie de signets. En effet, un utilisateur utilise sa hiérarchie de signets pour organiser les informations pertinentes qu'il retrouve sur le web. L'action d'insérer un document dans le même répertoire qu'un autre document résulte d'un effort cognitif [Rücker, 1997] (cf. I.3.2.2.2) et peut être interprétée comme une similarité entre deux documents par rapport à la thématique exprimée au travers du nœud. Par ailleurs, la relation hiérarchique de spécialisation/généralisation permet également de définir un niveau de similarité entre les différents documents contenus dans ces hiérarchies.

Nous rappelons que, dans notre approche, nous ne tenons pas compte des signets insérés dans la racine de la hiérarchie de signets afin de construire la liste des recommandations car ils ne font l'objet d'aucune organisation hiérarchique.

L'identification des documents pertinents pour un document visité *d* repose donc sur le principe que si un répertoire d'une hiérarchie de signets d'un utilisateur fait référence à *d*, les documents associés aux signets proches de celui-ci peuvent être considérés comme pertinents par rapport à *d*, et ce du point de vue de l'utilisateur. Chacun de ces documents, est sauvegardé dans une liste et est pondéré pour refléter l'importance de celui-ci par rapport au document visité *d* au travers de la hiérarchie de signets étudiée.

Du fait que le système exploite les hiérarchies de signets de plusieurs utilisateurs, il se peut que, pour un document visité *d* nous obtenions plusieurs listes de documents pertinents par rapport à *d*. Nous avons choisi de fusionner l'ensemble de ces listes en une seule qui sert de synthèse des points de vue des différents utilisateurs du système. Cette dernière permet en outre de mettre à jour la liste des recommandations du profil de navigation de l'utilisateur du document visité.

Dans un premier temps, nous présentons, le profil de navigation qui renferme les informations concernant la navigation de l'utilisateur ainsi que la liste de recommandations proposées par le module. Nous détaillons ensuite la démarche proposée afin de détecter les documents pertinents pour un document visité qui sont destinés à mettre à jour la liste de recommandations pour la navigation en cours. Enfin, la démarche proposée est évaluée de façon à vérifier nos hypothèses.

Pour illustrer nos propos, nous avons utilisé, dans ce qui suit, l'exemple de deux hiérarchies de signets de deux utilisateurs différents. Ces hiérarchies sont présentées dans la Figure 46.

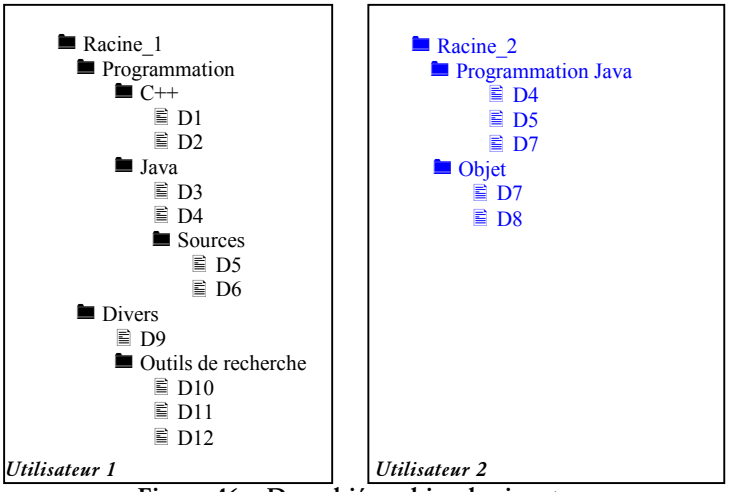


Figure 46 – Deux hiérarchies de signets

II.3.3.2.1 Profil de navigation

Le profil de navigation est utilisé afin de caractériser la navigation en cours d'un utilisateur. Ce profil contient :

- une liste des documents visités,
- une liste de recommandations qui évoluent au fur et à mesure des documents visités lors de la navigation.

La liste des documents visités est utilisée afin que le système ne propose pas un document précédemment visité par l'utilisateur.

La liste de recommandations qui est vide au début de la navigation. Sa construction repose sur une approche incrémentale pour refléter la nécessité d'adapter les recommandations à l'ensemble des documents visités. En effet, le système vise à proposer à l'utilisateur des recommandations qui correspondent au mieux à ses besoins déduits de sa navigation.

Le profil de navigation *PN* d'un utilisateur contient donc :

- l'ensemble *VISITES* = $\{D_1, D_2,..., D_n\}$ qui correspond à l'ensemble des documents visités par l'utilisateur,
- l'ensemble $RECOMMANDATIONS_d$ = $\{(D_4, w_4), (D_5, w_5),..., (D_n, w_n)\}$ qui contient l'ensemble des documents pertinents pour la navigation, associés à un poids *w* correspondant au degré d'adéquation entre un document et la navigation. Il est mis à jour à chaque fois que l'utilisateur visite un nouveau document.

A partir de la liste de recommandations, qui est mise à jour à chaque document visité, le système propose périodiquement à l'utilisateur les *k* meilleures recommandations.

II.3.3.2.2 Recherche des recommandations pour un document visité

L'approche proposée pour la recommandation pour un document visité, repose sur l'identification des documents pertinents qui permettent ensuite de mettre à jour la liste des recommandations. Or, afin de détecter ces documents, il est nécessaire de vérifier si le document visité existe dans chaque hiérarchie de signets et ce pour chaque utilisateur du système. C'est pour cela que nous avons choisi de simplifier ce processus et d'exploiter une structure unique construite à partir des hiérarchies de signets de l'ensemble des utilisateurs.

Nous utilisons pour cela le concept de multi-arbres. Un multi-arbres permet de représenter l'ensemble des hiérarchies de signets des utilisateurs au sein de la même structure.

Nous présentons ce concept de multi-arbres dans la section suivante. Ensuite, nos détaillons la démarche permettant d'identifier les documents pertinents pour un document visité au travers de ce multi-arbres.

II.3.3.2.2.1 Représentation des hiérarchies de signets sous la forme de Multi-arbres

Afin d'obtenir une représentation unique de l'ensemble des hiérarchies de signets des utilisateurs du système, nous avons utilisé la représentation en *multi-arbres* dont le but est, initialement, la conception coopérative de documents pédagogiques [Furnas, 1994].

Dans notre contexte, un multi-arbres caractérise les liens entre les documents et tous les nœuds des hiérarchies de signets de l'ensemble des utilisateurs. Il permet d'exploiter la structure des différentes hiérarchies de signets directement à partir des documents qu'elles contiennent.

Un exemple de représentation des deux hiérarchies de signets (cf Figure 46) en une structure de multi-arbres est présentée dans la Figure 47. Dans cette figure, chacune des hiérarchies initiales est représentée par un style de trait différent.

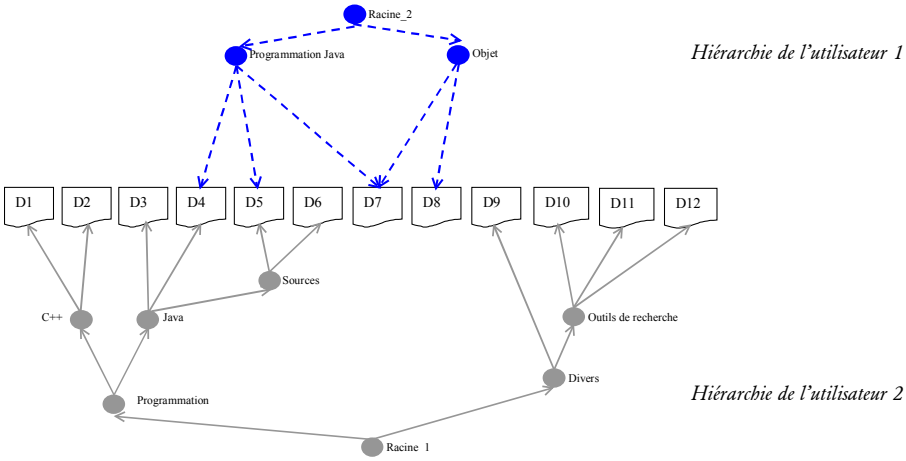


Figure 47 - Représentation multi-arbres des hiérarchies de signets de la Figure 46

Dans notre application, la construction du multi-arbres est réalisée à partir de l'ensemble des jugements de pertinence positifs pour chaque nœud issus du module précédent. Une application de multi-arbres aux hiérarchies de signets a été présentée au travers du projet *GAB* [Wittenburg, 1995]. Cependant, celui-ci est principalement orienté vers la navigation au travers du multi-arbres ainsi que sa visualisation graphique (sous forme de Tree-Map). Dans notre contexte, les objectifs visés ne sont pas les mêmes ; ils consistent à identifier tous les documents pertinents pour un document visité. C'est pour cette raison qu'il est nécessaire de définir certains des concepts auxquels nous ferons fréquemment référence.

Un *arbre* est une arborescence dont la racine correspond à la racine de la hiérarchie de signets d'un utilisateur.

Dans l'exemple de la Figure 47, il y a 2 arbres dont les racines sont *Racine_1* et *Racine_2*.

Une *branche* d'un arbre, dans notre approche, correspond à un sous-arbre dont la racine est un répertoire contenu dans la racine de l'arbre.

Dans la Figure 47, l'arbre défini par *Racine_1*, contient deux branches nommées *Programmation* et *Divers*. L'arbre Racine_2 contient deux branches nommées *Programmation Java* et *Objet*.

Afin d'identifier les relations de similarité entre les documents des hiérarchies de signets, nous avons utilisé les concepts de documents « frères » d'un document *d* [Wittenburg, 1995] :

Un document *frère* de *d* est un document contenu dans un nœud contenant *d*.

Dans notre exemple, les documents frères du document D5 sont {D4, D6, D7}. En effet au travers du nœud *Sources*, nous obtenons {D6} et au travers du nœud *Programmation Java* {D4, D7}.

Cependant, cette notion de *document frère* se limite aux documents se trouvant au plus près du document visité. Or, la construction d'une hiérarchie de signets dépend des besoins de l'utilisateur à qui elle appartient (subjectivité). Ainsi, deux utilisateurs peuvent construire deux hiérarchies de signets totalement différentes à partir d'un même ensemble de documents. Pour ne pas se limiter à ces seuls documents, nous avons proposé une notion de *document filié*[14] qui étend cette notion de document frère.

Un document *filié* à *d* est un document contenu dans une branche contenant *d*.

Cette notion repose sur l'idée qu'une branche reflète une spécialisation successive d'un thème générique initié dans le nœud racine de la branche.

Ainsi, dans l'exemple de la Figure 47, l'ensemble des documents {D1, D2, D3, D4, D6} sont des documents filiés à D5 au travers de l'arbre *Programmation* car ils traitent de ce même thème.

Cependant, l'éloignement entre des documents au travers d'une hiérarchie de signets peut être interprétée comme un degré de similarité. Ce degré de similarité est également nécessaire pour rendre compte de la différence de profondeur entre le document visité *d* et ses documents filiés dans les différentes branches. La similarité entre le document d et un document filié utilise la fonction de similarité entre deux URLs présentée dans [Jaczynski, 1997].

[14] Nous avons utilisé le barbarisme « filié » pour représenter un lien de filiation entre des documents.

La formule 10 présente la mesure de similarité que nous avons utilisée entre le document visité *d* et un document frère *f* de *d*. Cette formule est une adaptation de celle de [Jaczynski, 1997].

$$Sim(b, n_d, n_f) = 1 - \frac{c(n_d, MSCA(n_d, n_f)) + c(n_f, MSCA(n_d, n_f))}{c(n_d, racine(b)) + c(n_f, racine(b)) + 2} \qquad \textbf{10}$$

Où *b* est une branche contenant le document filié *f* et le document visité *d*,
n_d(respectivement n_f) est un nœud contenant le document visité *d* (respectivement le document filié *f*),
c(x, y) calcule la distance entre un nœud *x* et un nœud *y* au travers de l'arborescence, Cette distance tient compte du nombre de nœuds dans la branche entre *x* et *y*,
MSCA(x, y) retourne le plus grand chemin commun entre les chemins des nœuds *x* et y. Par exemple, si *x* = *Programmation/Java* et *y* = *Programmation/Java/Sources*, la valeur de *MSCA(x, y)* vaut *Programmation/Java*.
Racine(b) retourne le nœud racine de l'arbre contenant la branche *b*.

A partir du multi-arbres, le module identifie, pour chaque document visité *d*, les documents pertinents. Pour cela, nous avons exploité le concept de documents filiés. Nous recherchons le document *d* dans le multi-arbres, puis, au travers de ce concept, le système identifie les documents pertinents pour *d*.

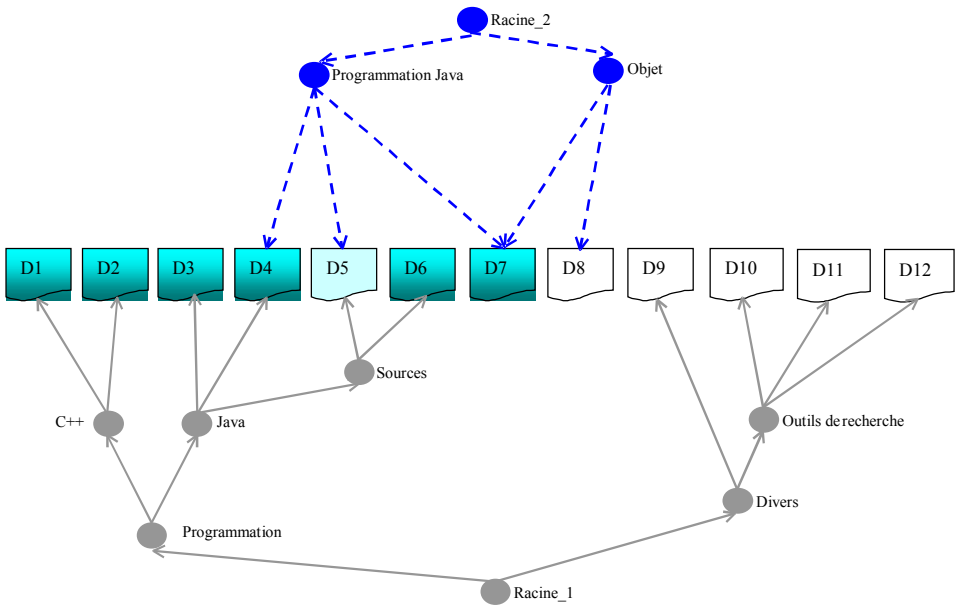


Figure 48 – Documents pertinents pour le document visité d=D5

Par exemple, la Figure 48 présente les documents filiés au document visité D5. Les documents filiés sont {D1, D2, D3, D4, D6, D7}.

Pour obtenir ce résultat, le système recherche d'abord toutes les branches *b* qui pointent vers *d*. Ensuite, le système identifie l'ensemble des documents de *d* au travers de ces branches *(étape 1)*. A la suite de cela, ces documents sont pondérés pour refléter leur l'intérêt par rapport au document visité *d (étape 2)*.

Etape 1 – Sélection des documents filiés à d.

Pour identifier les documents filiés au document visité *d*, le système identifie tout d'abord les branches pointant vers *d*. Dans l'exemple de la Figure 47, si le document visité est D5, deux branches pointent vers *d* c'est-à-dire les branches *Programmation Java* et *Programmation*.

Pour chacune de ces branches *b*, le système crée la liste de documents filiés $documents\_filiés_{b,d}$ qui correspond à l'ensemble des documents filiés à *d* au travers de *b*.

Etape 2 - Pondération des documents frères.

A partir de la liste des documents frères contenus dans l'ensemble $documents\_filiés_{b,d}$,, le système pondère chacun d'entre eux pour refléter leur similarité avec *d*. Chaque document filié f_j est pondéré par rapport à *d* en fonction de deux paramètres :

- la proportion d'utilisateurs possédant des branches contenant f_j et *d*. En effet, plus le nombre d'utilisateurs ayant construit leur hiérarchie de signets de façon à ce que le document f_j soit un document filié à *d*, plus f_j est similaire à *d*,
- la distance moyenne entre f_j et *d* au travers des différentes branches identifiées. Plus la distance entre f_j et *d* est grande au travers des hiérarchies de signets, moins le document f_j est similaire à *d*.

Ainsi, pour chaque document présent dans au moins un des ensembles $documents\_filiés_{x,y}$, la pondération est effectuée par la fonction 11 :

$$Poids_{f_j,d} = \exp(\frac{n_e}{N}) * Similarité\_moyenne(f_j, d) \qquad \textbf{11}$$

Où n_e est le nombre d'utilisateurs possédant au moins une branche pointant vers f_j et *d*,

N est le nombre d'utilisateurs possédant une hiérarchie de signets au sein du système,

Similarité_moyenne(f_j, d) correspond à la similarité moyenne entre le document f_j et *d* dans la totalité des occurrences des ensembles $documents\_filiés_{x,y}$. Cette similarité utilise la formule 10.

Cette fonction privilégie les documents filiés dont la proportion d'utilisateurs possédant le document *d* et *f* au sein d'une même branche est importante. Elle privilégie également les documents filiés situés, en moyenne, à une faible distance de *d* au sein de ces branches car ces documents filiés correspondent à la même profondeur que *d*.

La Figure 49 présente un exemple de pondération des documents filiés à un document visité D5. Dans cet exemple, nous avons uniquement proposé le calcul du poids des documents filiés D1, D4, D7. Cette liste est volontairement limitée pour permettre d'apprécier la pondération de documents situés à différents endroits de l'arborescence.

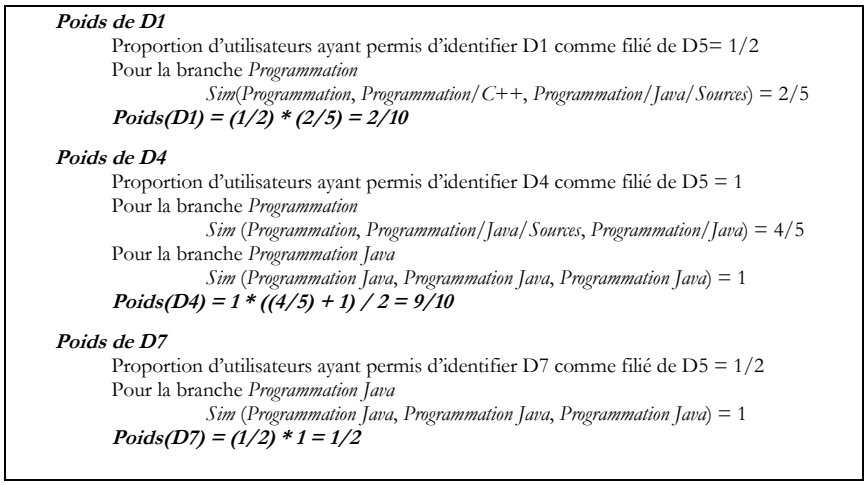


Figure 49 - Exemples de pondération des documents recommandés pour le document visité D5

L'ensemble des documents ainsi pondérés permet de construire la liste des documents pertinents pour le document visité *d* nommé $Liste\_Recommandations_d$. Cette liste de documents permet de mettre à jour la liste des recommandations du profil de navigation afin d'adapter celles-ci par rapport à l'ensemble des documents visités.

II.3.3.2.3 Mise à jour du profil de navigation

La mise à jour du profil de navigation passe par la mise à jour des différents éléments le composant.

Dans un premier temps, le document visité est ajouté à la liste des documents visités *VISITES* s'il n'y figurait pas. Cette mise à jour permet au système de garder en mémoire l'ensemble des documents visités (l'ordre importe peu dans notre approche) afin d'éviter de recommander à l'utilisateur des documents qu'il aurait déjà visité au cours de sa navigation. Ainsi tous les documents de $RECOMMANDATIONS_d$ qui sont également présents dans l'ensemble VISITES (historique de navigation) sont supprimés de la liste des recommandations.

Dans un second temps, la liste de recommandations $RECOMMANDATION_d$ est mise à jour. Afin de favoriser les documents qui correspondent le mieux à la problématique de la navigation, nous avons privilégié les documents qui ont été pertinents pour le plus grand nombre de documents visités. A l'opposé, nous avons exclu les documents qui ne sont pertinents que pour un nombre limité de documents visités.

L'ensemble des recommandations est mis à jour à partir de la liste des documents pertinents $Liste\_Recommandations_d$ du dernier document visité. Tous les documents existant dans $Liste\_Recommandations_d$ mais n'existant pas dans $RECOMMANDATIONS_d$ sont simplement ajoutés avec leur poids d'origine. Pour tous les autres documents, leur poids dans l'ensemble $RECOMMANDATIONS_d$ est augmenté de la valeur du poids de la nouvelle recommandation.

Pour tenir compte de l'évolution des besoins de l'utilisateur au cours de la navigation (il peut éventuellement changer radicalement de thème de recherche) et pour exclure les documents peu pertinents pour l'ensemble des documents visités, tous les documents de $RECOMMANDATIONS_d$ qui n'ont pas été ajoutés ou modifiés voient leur poids décroître.

Grâce à une adaptation du profil, le processus construit une liste de documents pertinents pour la navigation de l'utilisateur qui évolue au fur et à mesure des documents visités.

II.3.3.2.4 Recommandation des documents à l'utilisateur

A partir de cette liste de recommandations, le processus peut régulièrement proposer à l'utilisateur (après chaque visite d'un document par exemple) une liste de documents pertinents pour sa navigation globale et non seulement pour le dernier document visité. Afin de présenter les recommandations à l'utilisateur, le système ne sélectionne que les k recommandations les plus importantes issues des ensembles $RECOMMANDATIONS_d$. Pour cela, les recommandations sont ordonnées selon leur poids et seules les k premières sont présentées à l'utilisateur.

II.3.3.3 Expérimentations

Afin de réaliser l'évaluation de ce module, nous avons collecté, au sein de notre équipe de recherche ainsi que sur Internet, 21 hiérarchies de signets. Ces hiérarchies de signets partagent au moins un centre d'intérêt. Les hiérarchies de signets sont utilisées en l'état sans aucune réorganisation préalable. Elles contiennent au total 5098 signets, soit près de 200 signets par hiérarchie en moyenne. L'ensemble de ces signets correspond à 4566 documents HTML existants (soit plus de 11% de signets qui pointent vers des documents similaires). Chaque nœud de ces hiérarchies contient en moyenne 6 documents.

Un multi-arbres a été construit à partir de ces 21 hiérarchies de signets afin d'évaluer notre approche.

Pour les différentes expérimentations, nous avons sélectionné 24 documents faisant partie des thèmes partagés au sein des hiérarchies de signets comme « la recherche d'information », « XML », « la programmation java » etc. Plusieurs d'entre eux traitent du même thème mais ne sont pas forcément présents dans les nœuds des hiérarchies de signets. Nous avons simulé la visite d'un ou de plusieurs documents successivement afin d'évaluer la qualité des recommandations faites par le système. Pour cela, nous avons jugé la relation entre un document restitué et le(s) document(s) visité(s), et ce, pour les 100 premiers documents restitués par le système. Si un document restitué correspond au thème ou est « utile » pour le(s) document(s) visité(s), un jugement positif est émis et le document recommandé est jugé pertinent. Par exemple, un document en rapport avec les « dictionnaires bilingues » est jugé pertinent pour un document visité concernant la « RI multilingue » au même titre qu'un document traitant exactement de ce thème.

Le but de ces expérimentations est, dans un premier temps, de démontrer l'influence de la distance entre les documents frères et les instances du document visité sur la qualité et la quantité des recommandations faites par le système. Ces expérimentations visent également à souligner l'évolution des recommandations au fil des documents visités lors de la navigation de l'utilisateur.

La première expérimentation que nous avons menée a pour objectif de montrer l'influence de la distance de recommandation sur les résultats obtenus. En effet, nous avons

identifié comme documents pertinents pour un document visité *d*, tous les documents filiés à *d*. Cependant, il est intéressant de voir si la prise en compte des nœuds pères et nœuds frères du nœud pointant vers *d* a une influence sur les résultats. Ainsi, nous avons réalisé trois séries de mesures sur les 24 documents sélectionnés (Figure 50). Dans un premier temps, nous avons pris en compte uniquement les documents situés dans les nœuds contenant le document visité *d* ainsi que leurs arborescences filles. Cette expérimentation est dite avec une *Distance de 0*. Nous avons conservé les nœuds fils du nœud contenant *d* en raison de l'aspect généralisation/spécialisation.

Ensuite, nous avons évalué l'influence de la distance au travers des *Distance de 1* et *Distance de 2.* A chaque niveau, un père supplémentaire du nœud contenant *d* est pris en compte dans l'évaluation. Nous nous sommes limités à une expérimentation avec une distance maximale de 2 car la profondeur moyenne des hiérarchies de signets étudiées est de 1.19 nœuds (hors racine).

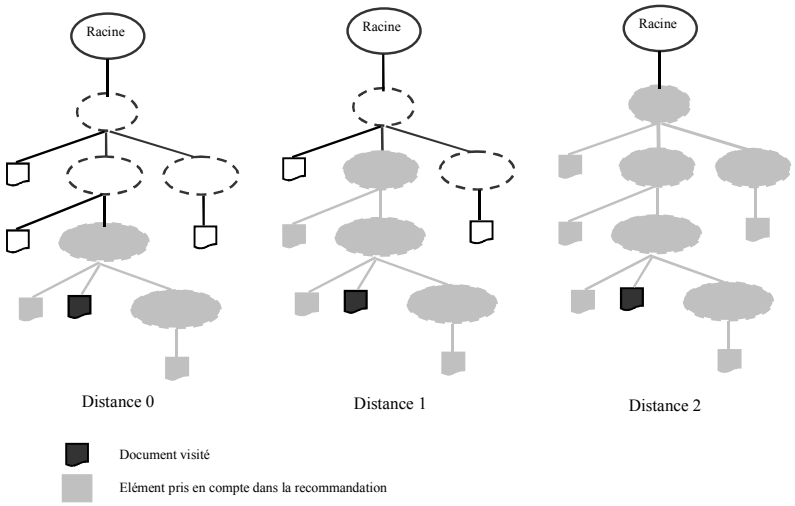


Figure 50 - Expérimentations concernant la distance de recommandation

Le premier résultat concerne l'influence de cette distance sur le nombre de documents dans la liste de recommandations. La Figure 51 présente les résultats obtenus. En moyenne, la distance 1 permet d'obtenir 4 fois plus de documents qu'avec une distance de 0. Par contre, la distance 2 ne permet d'augmenter le nombre de documents que de façon sporadique et, pour la plupart des cas, elle n'apporte rien en terme de nombre de recommandations.

Document	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12
Distance 0	6	8	18	8	57	19	12	8	11	71	1	16
Distance 1	66	29	43	8	57	43	43	65	61	73	24	69
Evolut° entre Distance 1 et Distance 0	1000.00%	262.50%	138.89%	0.00%	0.00%	126.32%	258.33%	712.50%	454.55%	2.82%	2300.00%	331.25%
Distance 2	66	29	43	8	57	43	56	65	61	73	24	69
Evolut° entre Distance 2 et Distance 1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	30.23%	0.00%	0.00%	0.00%	0.00%	0.00%

Document	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d23	d24
Distance 0	18	8	35	69	19	4	71	11	7	8	20	8
Distance 1	23	43	67	76	22	27	76	18	21	8	76	67
Evolut° entre Distance 1 et Distance 0	27.78%	437.50%	91.43%	10.14%	15.79%	575.00%	7.04%	63.64%	200.00%	0.00%	280.00%	737.50%
Distance 2	23	43	67	76	22	27	76	18	52	8	76	67
Evolut° entre Distance 2 et Distance 1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	147.62%	0.00%	0.00%	0.00%

Figure 51 - Influence de la distance sur le nombre de documents retrouvés

Le second résultat intéressant est la pertinence des résultats retournés par le système. Nous utilisons la notion de pertinence pour indiquer qu'un document recommandé est en relation avec le thème du document visité. La Figure 52 présente les résultats obtenus. Cette figure présente la proportion de documents pertinents au sein des recommandations pour chacune des distances étudiées.

Document	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12
Distance 0	0.33	0.50	0.94	0.75	0.72	0.95	0.92	0.88	0.82	0.70	1.00	0.88
Distance 1	0.67	0.10	0.42	0.75	0.72	0.44	0.42	0.15	0.13	0.70	0.13	0.26
Evolut° entre Distance 1 et Distance 0	100.00%	-79.31%	-55.68%	0.00%	0.00%	-53.36%	-54.33%	-82.42%	-83.97%	-0.79%	-87.50%	-70.19%
Distance 2	0.67	0.10	0.42	0.75	0.72	0.44	0.55	0.15	0.13	0.70	0.13	0.26
Evolut° entre Distance 2 et Distance 1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	32.24%	0.00%	0.00%	0.00%	0.00%	0.00%

Document	d13	d14	d15	d16	d17	d18	d19	d20	d21	d22	d23	d24
Distance 0	1.00	0.75	0.91	0.84	1.00	1.00	0.15	1.00	0.86	0.63	0.75	1.00
Distance 1	0.83	0.56	0.64	0.67	1.00	0.11	0.21	0.72	0.57	0.63	0.21	0.22
Evolut° entre Distance 1 et Distance 0	-17.39%	-25.58%	-29.80%	-20.17%	0.00%	-88.89%	35.89%	-27.78%	-33.33%	0.00%	-71.93%	-77.61%
Distance 2	0.83	0.56	0.64	0.67	1.00	0.11	0.21	0.72	0.46	0.63	0.21	0.22
Evolut° entre Distance 2 et Distance 1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-19.23%	0.00%	0.00%	0.00%

Figure 52 - Influence de la distance sur la pertinence des recommandations

Nous pouvons constater que pour une distance égale à 0, la pertinence des recommandations est relativement élevée. La moyenne de pertinence sur l'ensemble des documents est de plus de 80%. Seuls les résultats pour les documents 1 et 19 sont inférieurs à 0.5. Après vérification, il s'avère que ces documents sont présents dans des répertoires « fourre-tout » qui contiennent des documents ayant peu de rapport entre eux. En ce qui concerne la distance 1, nous pouvons souligner le fait qu'à l'exception de deux documents (1 et 19) tous les résultats sont inférieurs à ceux obtenus avec une distance de 0.

Le résultat obtenu avec une distance de 2 n'est pas étonnant car il est lié au résultat de la Figure 51 c'est-à-dire que la distance 2 ne permet pas d'obtenir globalement beaucoup plus de recommandations. Par ailleurs, lorsque la distance 2 permet d'en obtenir plus (documents 7 et 21) la proportion des documents pertinents n'est pas forcément plus importante (document 21 par exemple). Pour résumer, nous pouvons constater que la distance 1 permet d'obtenir de plus nombreux résultats mais au prix d'une pertinence

moindre. Par ailleurs, la distance 2 ne permet pas d'obtenir globalement de meilleurs résultats que la distance de 1.

Pour toutes ces raisons, pour la suite des expérimentations, nous avons utilisé uniquement la distance 0.

La deuxième expérimentation que nous avons menée concerne l'évolution des recommandations au fil de la navigation de l'utilisateur. Pour cela, nous avons étudié le poids moyen des documents pertinents dans l'ensemble des recommandations, ainsi que le poids moyen des documents non pertinents. Nous rappelons que le poids est calculé à l'aide de la formule 11 et que plus le poids est important, plus le système qualifie un document de pertinent pour la navigation en cours. Ceci signifie qu'il a été jugé pertinent pour plusieurs des documents visités.

Nous avons réalisé 5 études à partir de 14 documents individuels. Nous avons simulé une succession composée de 2 à 4 documents. La Figure 53 présente les résultats obtenus avec 2 documents visités successivement. Ceux obtenus avec 3 documents sont présentés dans la Figure 54, tandis que la Figure 55 montre l'incidence d'une succession de 4 documents visités sur les poids moyens des documents pertinents et des documents non pertinents. *Poids+* (respectivement *Poids-*) représente le poids moyen des documents pertinents (respectivement non pertinents). Dans chacun des cas, les documents utilisés traitent du même thème.

2 Docs	d14	d19	d14 & d19
Poids+	1.05	1.02	**1.31**
Poids-	1.05	0.83	**0.87**

2 Docs	d22	d8	d22 & d8
Poids+	1.05	1.05	**1.01**
Poids-	1.05	1.05	**0.97**

Figure 53 – Poids moyen des documents pertinents/non pertinents pour une succession de 2 documents

3 Docs	d13	d16	d20	d13 & d16	d13 & d16 & d20
Poids+	1.06	1.06	1.05	1.19	**1.24**
Poids-	0.00	1.05	0.00	1.05	**1.05**

3 Docs	d1	d17	d24	d1 & d17	d1 & d17 & d24
Poids+	1.05	1.05	1.05	1.04	**1.17**
Poids-	1.05	0.00	0.00	0.94	**1.05**

Figure 54 - Poids moyen des documents pertinents/non pertinents pour une succession de 3 documents

4 Docs	d2	d7	d21	d12	d2 & d7	d2 & d7 & d21	d2 & d7 & d21 & d12
Poids+	1.05	1.05	1.05	1.05	1.03	1.19	**1.36**
Poids-	1.05	1.05	1.05	1.05	0.96	0.90	**1.00**

Figure 55 - Poids moyen des documents pertinents/non pertinents pour une succession de 4 documents

A partir de ces figures, nous pouvons constater que, globalement, l'écart entre le poids des documents pertinents et le poids des documents non pertinents augmente lorsque le nombre des documents visités s'accroît. Cette augmentation de poids résulte du fait que

certains documents ont un poids élevé du fait qu'ils sont pertinents pour plusieurs des documents visités.

II.3.3.4 Bilan du module de recommandation durant la navigation

Nous avons utilisé, dans ce module, le concept de documents frères issus des multi-arbres que nous avons étendu au concept de document filié afin d'identifier les documents pertinents pour un document visité. Par ailleurs, une approche incrémentale permettant d'adapter une liste de recommandations par rapport à la succession de documents visités a été proposée.

L'utilisateur profite ainsi des connaissances capitalisées par l'ensemble des utilisateurs du système correspondant à ses besoins au travers de recommandations réalisées par le système.

La dynamique de la navigation permet de faire émerger les recommandations qui sont en rapport avec le thème de celle-ci au fil des documents visités. Le système permet d'obtenir des résultats satisfaisants à partir des hiérarchies de signets utilisées pour effectuer les expérimentations.

Par le biais de ce module de recommandation lors de la navigation, l'utilisateur peut obtenir non seulement des documents traitant du même thème que sa navigation mais également des documents lui permettant d'élargir son champ de recherche ou au contraire l'aiguiller plus rapidement vers des documents relatifs à ses besoins.

La démarche proposée s'affranchit de l'étude du contenu des documents. Les documents pertinents pour un document visité reposent sur une similarité implicite déduite de l'organisation des hiérarchies de signets des utilisateurs. Ainsi, l'intérêt majeur d'utiliser la manière dont sont construites les hiérarchies de signets est qu'elle repose sur l'aptitude naturelle qu'ont les utilisateurs à organiser des documents et par extension sur leur connaissance du domaine ou leurs besoins.

Les expérimentations menées montrent que, globalement, les utilisateurs organisent leurs signets de façon thématique. Cependant, de la qualité de l'organisation des signets au sein des hiérarchies de signets dépend la qualité des recommandations faites par le système. Notre approche trouve cependant ses limites lorsqu'un internaute recherche des documents par navigation dont le thème n'est pas possédé par un utilisateur du système. En effet, le système n'est utilisable que si au moins un utilisateur du système est intéressé dans la thématique de la navigation d'un internaute tiers. Quoi qu'il en soit, ce module représente un complément aux approches traditionnelles qui reposent soit sur l'interrogation de moteurs de recherche soit sur la recommandation de documents situés dans l'hypertexte local du document visité. En effet, les recommandations réalisées au travers de ce module repose sur les informations collectées et organisées par les utilisateurs du système.

La navigation est utilisée alternativement avec une recherche adhoc au sein du processus de recherche. La principale limite de cette recherche est qu'elle retourne généralement un nombre très important de documents qu'il est difficile de gérer globalement. Or, ces outils de recherche adhoc ne proposent pas à l'utilisateur une représentation des résultats qui

permettrait de l'apprécier de façon globale. Pour cette raison, nous proposons un module de visualisation de résultats de recherche associée à un outil de recherche adhoc coopératif.

II.3.4 Module de visualisation des résultats de recherche

Afin de permettre à l'utilisateur d'initier une recherche d'information, nous offrons la possibilité à l'utilisateur d'interroger un outil de recherche adhoc coopératif. La base d'indexation de cet outil de recherche est construite à partir des documents issus des hiérarchies de signets des utilisateurs du système comme le proposent *IronWeb* [Dussaux, 2000] ou encore *Yawas* [Denoue, 2000]. Cet outil permet donc à un utilisateur d'interroger les informations que possèdent les autres individus du système de façon transparente.

Par exemple, les chercheurs ayant de nouveaux centres d'intérêt ou encore les doctorants ont intérêt à retrouver les informations préalablement collectées par les individus du laboratoire pour réaliser des états de l'art. Un tel outil de recherche permet de réaliser cette tâche en réduisant les efforts de recherche.

Nous avons associé cet outil de recherche à une interface de visualisation des résultats de recherche pour en permettre une appréciation globale.

II.3.4.1 Problématique

Les outils de recherche adhoc génèrent un grand nombre de résultats que l'utilisateur doit parcourir un à un au travers de la liste de résultats pour vérifier leur pertinence par rapport à ses besoins. Dans la plupart des cas, l'utilisateur se limite aux premiers documents pertinents occultant ainsi de nombreux documents potentiellement pertinents (cf. I.4.2.2.5). D'autre part, cette liste de résultat ne permet pas d'apprécier la similarité des documents par rapport à la requête.

Il est donc nécessaire d'avoir une vision globale du résultat de recherche permettant à l'utilisateur d'identifier, dans l'ensemble des résultats, les documents les plus pertinents pour ses besoins.

II.3.4.2 Approche proposée

La solution découle des interfaces de visualisation des résultats de recherche (Visual Information Retrieval Interface ou VIRI) qui permettent une interprétation et une manipulation globale des résultats de recherche.

Nous avons précédemment souligné le grand nombre de visualisations possibles (cf. I.4.2.2.5) pour les résultats de recherche (relations inter-documents, répartition des termes de la requête...).

Afin de permettre à l'utilisateur d'apprécier l'importance des documents retrouvés par rapport à la requête, présentée sous la forme d'une liste de mots-clés, nous avons estimé, dans notre approche, que la présentation des résultats de recherche par rapport aux termes de la requête devait être privilégiée. En effet, si l'on écarte les problèmes de formulation de requête, l'utilisateur emploie des termes qu'il souhaite généralement retrouver dans les

documents. Une visualisation en liste de résultats ne permet pas facilement d'identifier l'importance des termes de la requête dans les documents retrouvés. Certes, la plupart des outils de recherche fait apparaître quelques occurrences des termes de la requête dans le résumé des documents, mais il est toujours difficile d'évaluer leur importance globale.

Cette interface doit également permettre une bonne interprétation des résultats visualisés. Dans les projets existants, un critère de visualisation (auteur du document, taille du document, importance d'un mots-clé dans le document...) est associé à un seul axe d'interprétation. Cependant, du fait de la diversité humaine, un utilisateur pourra être plus sensible aux couleurs alors qu'un autre sera plus sensible à la localisation spatiale des documents. De ce fait, nous proposons une combinaison de plusieurs axes d'interprétation pour un même critère de visualisation. Dans notre approche, nous avons choisi de combiner les axes couleur et répartition spatiale qui sont des axes couramment utilisés en visualisation. Cette combinaison permet de faciliter l'interprétation des documents visualisés pour un plus grand nombre d'utilisateurs, tout en minimisant l'effort cognitif nécessaire.

L'utilisation des couleurs est très importante dans cette interface graphique car elle permet une distinction naturelle entre les documents. Les nuances de couleurs sont également très intéressantes pour l'interprétation et doivent être privilégiées car elles permettent d'effectuer intuitivement une hiérarchie visuelle entre les documents de même teinte (cf. I.4.2.2.5.3.2). Contrairement à la plupart des interfaces actuelles, la couleur, dans notre approche, est utilisée de façon *sémantique* permettant ainsi à l'utilisateur d'interpréter l'importance des termes de la requête à partir des nuances de couleurs.

L'organisation spatiale permet de représenter l'importance des critères de visualisation au sein des documents selon une répartition spécifique des documents dans l'espace.

Cette section propose une interface de visualisation visant à aider l'utilisateur dans l'appréciation de la globalité des résultats de recherche par rapport aux termes de la requête grâce aux couleurs et à une répartition spatiale 3D spécifique [Chevalier, 2000], [Chevalier, 2001]. Nous avons utilisé un espace en 3D car il permet d'afficher un plus grand nombre d'éléments que les modèles 2D ou textuels et parce qu'un apprentissage minimal permet d'estomper les problèmes de manipulation.

Nous présentons, dans un premier temps, les aspects cognitifs qu'il est nécessaire de prendre en compte lors de la conception d'une telle interface. Dans un second temps, nous justifions nos choix concernant les axes d'interprétation qui sont liés à la couleur et à l'espace 3D. Nous détaillons, ensuite, la démarche proposée afin de visualiser les résultats de recherche d'informations suivie d'une méthode d'interprétation de celle-ci. Enfin, nous présentons les fonctionnalités liées à cette interfaces afin d'exploiter la visualisation (rotation, sélection...) avant de terminer par des expérimentations qui ont permis de valider nos propositions.

II.3.4.2.1 Aspects cognitifs

Nous avons employé les deux axes couleurs, répartition spatiale car nous souhaitons privilégier les aspects cognitifs proposés par l'interface. En effet, l'interface doit respecter au mieux les quatre aspects cognitifs suivants [Wiss, 1998], [Vernier, 1997] :

- *l'attention*. L'utilisateur doit avoir la possibilité d'identifier les documents pertinents d'un simple coup d'œil,
- *l'abstraction*. L'utilisateur doit pouvoir se concentrer sur une partie spécifique de l'espace des informations,
- *l'intuition*. L'utilisateur doit comprendre intuitivement quel va être le résultat d'une action qu'il peut réaliser,
- *« l'affordance »*. L'utilisateur doit comprendre quels outils correspondent à ses besoins. C'est pour respecter ce critère qu'il est nécessaire de proposer plusieurs visualisations différentes.

Pour que l'interface respecte ces aspects cognitifs, nous nous sommes basés sur l'étude de Lohse [Lohse, 1994] qui a mis en évidence l'intérêt des différentes représentations de l'information. Cette étude est basée sur des évaluation réalisées par 16 personnes d'horizons différents, avec comme sujets d'étude des représentations visuelles de l'information numérotées de 1 à 60 (voir annexe A). Le but de cette étude était de classifier les différentes représentations possibles de l'information. Elle a permis de faire émerger 11 catégories de base de représentations de l'information qui sont évaluées selon 10 critères (spatialité, temporalité, simplicité de compréhension...). Le bilan de cette étude est présenté dans la Figure 56. Dans cette figure, les valeurs représentent la position des catégories par rapport aux différents critères et sont comprises entre 1 (critère du haut) et 9 (critère opposé). Par exemple, pour les testeurs, une représentation en carte est spatiale (1.9) et est relativement facile à comprendre (7.5). Les numéros d'icônes précisés en exemple font référence aux dessins présentés en annexe A. Afin de bien interpréter les catégories utilisées, il est important de comprendre ce que chacune d'elle représente :

- les *graphiques* représentent l'information quantitative en utilisant le positionnement et l'importance d'objets géométriques. Les données numériques à 1, 2 ou 3 dimensions sont affichées dans un repère cartésien ou polaire. Les représentants de cette catégorie sont les histogrammes, les camemberts, les nuages de points...,
- les *tables* sont des arrangements de mots, nombres, symboles ou une combinaison de ceux-ci pour exprimer un ensemble de faits ou de relations dans un format compact. Cette catégorie peut être divisée en deux sous-rubriques selon le format des informations.
  - les tables numériques (icône 7),
  - les tables graphiques (icône 21),
- les *graphes temporels* mettent en avant des données temporelles (icônes 14 & 55),
- les *graphes en réseau* permettent de visualiser les relations entre les informations (par une proximité, des flèches...),
- les *diagrammes* représentant des données spatiales. Il en existe deux types.
  - les diagrammes de structure décrivant de façon statique un objet physique (icônes 59 & 11),
  - les diagrammes de processus décrivant les relations et les processus relatifs à un objet physique. Les informations spatiales représentent des relations dynamiques, continues ou temporelles entre les objets,

- les *cartes* représentent des informations relatives à la géographie physique. Cette représentation visualise les informations à l'aide de symboles et de légendes (cartes topographiques, marines...),
- les *cartogrammes* sont des cartes sur lesquelles on superpose des informations quantitatives (indications de flux, courbes de niveaux...),
- les *icônes* sont des représentations visuelles revêtant une seule interprétation, étant destinées à une audience spécifique (icône 30 & 58),
- les *images* sont des visions réalistes d'un objet ou d'une scène (voir icônes 44 & 49).

Critère : 1 → 9 / Catégorie	Spatiale – Non spatiale	Non Temporel – Temporel	Difficile – Facile à comprendre	Concret – Abstrait	Type continu – Discret	Attractif – Non attractif	Non synthétique – Synthétique	Non numérique – Numérique	Statique – Dynamique	Apporte beaucoup – Peu d'informations
Diagramme De structure	3.0	2.9	6.8	3.3	4.8	4.1	3.8	1.7	3.8	3.7
Cartogramme	2.9	4.7	5.6	4.7	4.4	4.2	4.7	3.4	4.9	4.3
Carte	1.9	2.0	7.5	3.6	4.3	3.9	4.5	3.8	2.4	2.7
Table graphique	6.3	4.7	6.2	4.6	5.3	4.5	3.0	3.7	4.3	2.5
Diagramme de processus	4.7	5.3	5.9	4.9	4.1	4.5	4.1	2.8	6.2	3.9
Icône	6.4	2.0	7.3	5.8	4.2	3.9	2.9	2.5	3.4	4.9
Graphe temporel	5.9	7.8	6.1	4.9	4.9	3.9	3.7	4.5	4.8	3.7
Graphe en réseau	5.2	4.0	5.6	5.3	4.3	5.9	3.9	2.6	5.3	4.3
Image	**3.2**	**1.9**	**6.7**	**4.9**	**5.3**	**3.1**	**6.7**	**1.7**	**3.1**	**7.2**
Table	7.1	1.8	5.1	5.1	5.4	5.2	2.5	8.0	2.4	2.8
Graphique	4.7	4.0	6.3	4.5	4.8	4.6	4.9	7.0	3.7	4.1

Figure 56 - Bilan de l'étude de Lohse [Lohse, 1994]

D'après l'étude de Lohse, nous pouvons souligner que les graphiques sont des représentations attractives qui synthétisent bien les informations. De plus, ils permettent de représenter des données numériques faciles à appréhender. Ce dernier point est d'autant plus important que nous avons souhaité visualiser l'importance des mots-clés de la requête dans les documents qui se base sur une information numérique (poids du terme dans le document).

La visualisation que nous proposons repose ainsi sur une représentation des informations sous la forme d'un graphique et plus particulièrement d'un nuage de points colorés dans un espace à trois dimensions afin de représenter l'ensemble des documents de façon globale.

II.3.4.2.2 Utilisation des couleurs

Dans notre approche, les couleurs sont utilisées de façon sémantique afin de représenter l'importance des différents mots-clés de la requête dans les documents. En moyenne, peu de mots-clés composent une requête (< 3) [Silverstein, 1998], [Jansen, 2000], [Spink, 2002]. Pour cette raison, nous avons utilisé les trois couleurs que sont le rouge, le vert et le bleu qui sont utilisées dans la plupart des modèles de couleurs. Chacune de ces couleurs représente chaque terme de la requête et chaque couleur est ainsi associée à un critère de visualisation.

Cependant, soucieux du fait que le nombre de mots-clés dépend des requêtes et des utilisateurs, nous offrons la possibilité de combiner plusieurs termes dans un critère de visualisation qui est associé à une des couleurs rouge, vert ou bleu à l'aide d'opérateurs booléens tels que *et* ou *ou*. Un critère de visualisation peut, par exemple, correspondre à la combinaison des termes « Système » *et* « Information ».

Pour rendre compte de l'importance de chaque critère de visualisation pour un document, nous avons utilisé la corrélation entre l'intensité de la couleur et l'importance du critère (Figure 57). Plus le critère est faible dans le document, plus l'intensité de la couleur du point est faible (sombre) alors qu'à l'inverse, plus le critère est important dans le document, plus la couleur est vive.

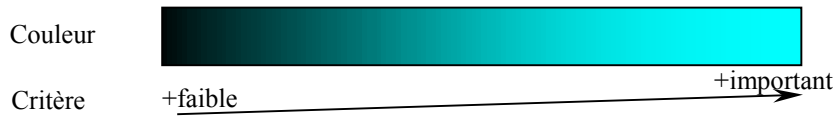


Figure 57* - Correspondance entre l'importance d'un critère et l'intensité de la couleur

Pour combiner l'importance des différents critères au niveau d'un même point correspondant à un document, nous avons utilisé intuitivement la synthèse additive. La couleur d'un point correspondant à une combinaison de l'importance des différents critères résulte donc d'un mélange des couleurs des critères (rouge, vert et bleu). La Figure 58 présente cette synthèse additive.

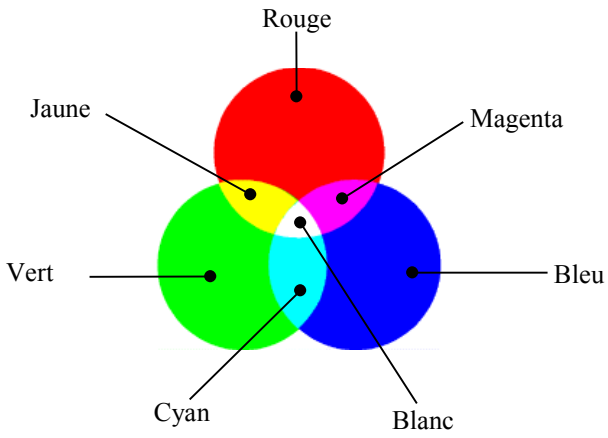


Figure 58* - Synthèse additive des couleurs

Cette synthèse additive permet, à partir des couleurs des différents critères de visualisation (associés aux couleurs rouge, vert ou bleu) d'obtenir une couleur unique après addition de celles-ci. Ainsi, d'après la Figure 58, si un document a une couleur jaune vif, le critère de visualisation affecté au rouge et celui affecté au vert sont importants alors que le troisième critère (associé au bleu) est inexistant.

La représentation de la synthèse additive peut faire penser à l'interface *Cougar* [Hearst, 1994]. En effet, les couleurs de bases (rouge, vert ou bleu) correspondent exactement aux zones utilisées dans cette interface. Cependant, *Cougar* ne permet pas une interprétation

relative des différents documents se trouvant dans la même zone les uns par rapport aux autres. Notre approche, en revanche, permet de réaliser cette hiérarchisation de façon visuelle grâce notamment aux dégradés de couleurs [Hearst, 1995] (cf. I.4.2.2.5.3.2).

II.3.4.2.3 Espace 3D

L'organisation spatiale devant permettre une interprétation identique à celle des couleurs, nous avons étudié intuitivement les espaces de couleurs qui permettent de représenter les couleurs dans l'espace.

Parmi ces espaces, nous pouvons citer les espaces de couleurs RVB et HSV (Figure 59). Ces deux modèles sont en 3D et se différencient par l'organisation spatiale des couleurs proposées.

Figure 59* - Modèles de couleurs RVB (à gauche) et HSV (à droite)

Le modèle RVB est représenté sous la forme d'un cube. Les couleurs sont organisées selon la synthèse additive à partir des 3 axes de couleurs Rouge, Vert et Bleu perpendiculaires. Ce modèle est orienté technologie et est proche de la répartition spatiale utilisée dans l'interface *Three Keywords Display* [Cugini, 1996].

Le modèle HSV [Smith, 1978] est, quant à lui, représenté sous la forme d'un cône. Les couleurs ne sont plus réparties selon 3 axes perpendiculaires mais par rapport à la caractéristique des couleurs (Teinte, Saturation, Intensité). Contrairement au modèle RVB, ce modèle est orienté utilisateur et fait appel, intuitivement, aux concepts de teinte, de nuance et de ton connus des artistes [Foley, 1995]. [Ware, 1985] qui a étudié l'implication de la couleur dans l'analyse de données souligne par ailleurs le fait que le modèle HSV est un modèle basé sur la perception humaine contrairement au modèle RVB. Il précise cependant que les modèles basés sur la perception sont limités lorsque les éléments visualisés ont une intensité faible car la résolution des autres axes est faible en raison du rétrécissement du cône. Il souligne également le fait qu'au travers de ses expérimentations l'interprétation de la corrélation entre les différents critères est optimum contrairement au modèle RVB.

Pour toutes ces raisons, nous avons utilisé le modèle HSV qui est un modèle orienté utilisateur et qui permet à l'utilisateur de mieux appréhender la corrélation entre les différents critères visualisés. En ce qui concerne le problème de faible résolution pour des valeurs des critères faibles soulignée dans [Ware, 1985], il n'est pas préjudiciable dans notre approche puisque les documents représentés par des points sombres (intensité faible) sont des documents peu pertinents pour l'utilisateur. De plus, cette agglutination de documents non pertinents vers le sommet du cône permet de faciliter la manipulation de ces informations (suppression par exemple). Cependant, pour tenir compte de cette limite qui

peut éventuellement poser des problèmes d'interprétation, nous avons proposé une visualisation alternative à la représentation en cône qui est une visualisation en forme de cylindre. Cette dernière permet de ne plus avoir la résolution qui faiblit avec la diminution de l'importance des critères.

II.3.4.2.4 Visualisation des résultats

La Figure 60 présente, dans sa partie gauche, le cône de l'espace HSV. La *teinte* (*h*) correspond à un angle indiquant la teinte principale de la couleur (rouge, jaune...). La saturation (*s*) correspond à la « pureté » de la teinte (degré de mélange des couleurs). L'*intensité* de la couleur (*v*) correspond à l'importance de la teinte principale. Plus l'intensité est élevée, plus la couleur est vive.

Dans sa partie droite, cette figure présente la visualisation alternative en cylindre que l'on propose à partir de l'espace de couleurs HSV.

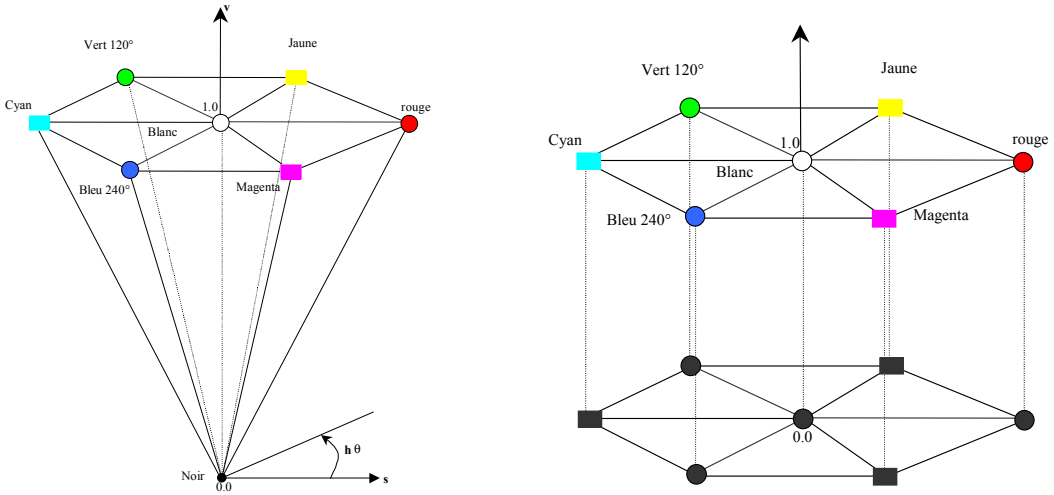


Figure 60 - Le cône du modèle HSV (Hue, Saturation, Value) et notre visualisation en cylindre

Remarque : un point à l'écran peut matérialiser un ou plusieurs documents si ceux-ci possèdent la même combinaison de valeurs pour les différents critères.

Pour effectuer le positionnement d'un document dans l'espace, nous utilisons l'espace de couleurs RVB pour passer ensuite à l'espace de couleur HSV. En effet, nous souhaitons que chaque couleur de base (rouge, vert et bleu) corresponde à un critère de visualisation. Ainsi, pour obtenir la visualisation dans l'espace HSV, chaque document est représenté par un point dans l'espace RVB *(r, g, b)* dans lequel chacune des composantes r, g ou b correspond à l'importance effective du critère dans le document.

Le passage du point de l'espace RVB à l'espace HSV *(h, s, v)* s'effectue par un algorithme qui est présenté en annexe B. L'algorithme modifié permettant d'obtenir un cylindre est également présenté dans cette annexe.

Ainsi, les points dans l'espace RVB sont obtenus à partir de l'ensemble des documents retrouvés par l'outil de recherche. Cet ensemble *Résultat* = {[d_1, p_1, (t_1, $w_{1,d1}$), (t_2, $w_{2,d1}$), ..., (t_n, $w_{n,d1}$)], ..., [d_m, p_m, (t_1, $w_{1,dm}$), (t_2, $w_{2,dm}$), ..., (t_n, $w_{n,dm}$)]} est composé de la liste des

documents retrouvés pour lesquels sont donnés la pertinence système (p_i) par rapport à la requête ainsi que le poids (w_i) normalisé [0 ; 1] de chacun des termes de la requête.

Les Figure 61 et Figure 62 présentent un exemple de visualisation en cône et en cylindre obtenus en réponse à une requête sur un outil de recherche.

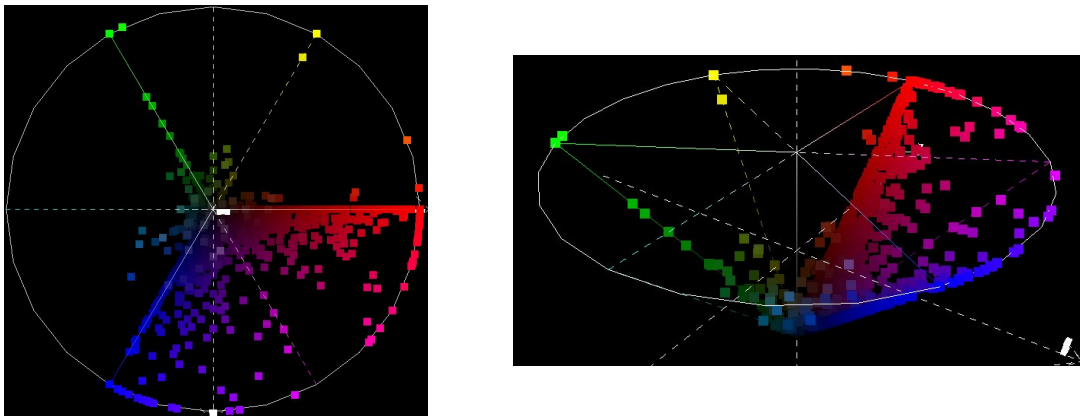


Figure 61* - Visualisation en cône (à gauche une vue de dessus, à droite une vue de ¾)

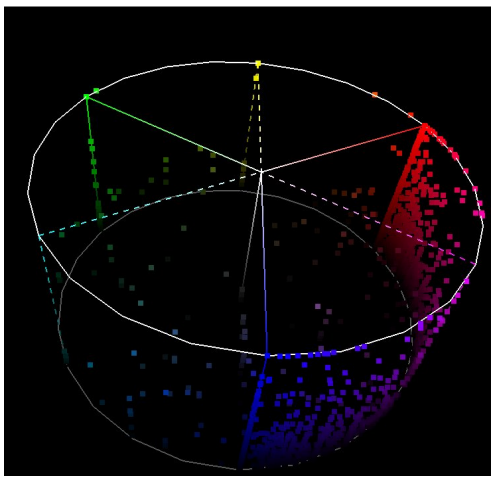


Figure 62* - Visualisation en cylindre

A partir de ces figures, nous pouvons remarquer que la distinction visuelle entre les différents documents est grandement facilitée par les couleurs.

II.3.4.2.5 Interprétation de la visualisation

Les éléments visualisés sont représentés par des points colorés placés à proximité du(des) critère(s) prédominant(s). Les critères sont représentés par les axes extérieurs de l'espace de documents : rouge, vert, bleu, cyan, magenta ou jaune. Les trois axes cyan, magenta, jaune correspondent aux combinaisons de deux des critères de visualisation ayant une importance égale (cf synthèse additive). L'axe central (du blanc au noir) représente les documents dont les trois critères de visualisation ont une importance égale.

L'interprétation de visualisation peut se faire tout d'abord au niveau de la détection du/des critère(s) dominant(s) dans les documents. Ce niveau est le plus simple car il ne demande pas trop d'effort. Cette interprétation est purement visuelle et il suffit d'étudier la couleur (teinte et intensité) ou la position du point interprété. En effet, il suffit de

rapprocher la couleur ou la position du point d'un des critères de base ou de l'une des combinaisons des critères pour comprendre quels sont ceux qui prédominent dans le document.

L'interprétation peut également être réalisée au niveau de la valeur réelle des critères. Cette interprétation fait appel à un effort cognitif plus important que la précédente car elle nécessite un recoupement de plusieurs informations telles que l'intensité de la couleur avec la teinte ou encore la hauteur du point et sa distance par rapport au centre du modèle.

Une méthode d'interprétation complète est fournie en annexe C.

Afin de permettre à l'utilisateur de manipuler les résultats visualisés, nous offrons au travers de cette interface différentes fonctionnalités.

II.3.4.2.6 Fonctionnalités liées à l'interface

Tout d'abord, une fonctionnalité de rotation de l'espace de documents est proposée pour permettre à l'utilisateur de visualiser l'espace des documents selon le point de vue qu'il souhaite. Cette rotation est effectuée selon les 3 axes de rotation de l'espace (X, Y et Z).

Soucieux du fait que l'utilisateur n'a pas toujours en mémoire la synthèse additive des couleurs, une légende est présentée. Cette légende permet à l'utilisateur de se remémorer à la fois les termes assignés à chaque critère ainsi que la correspondance entre chaque critère et la couleur correspondante.

Dans la représentation proposée, les documents pouvant intéresser l'utilisateur peuvent se situer au milieu du cône. Pour faciliter l'identification des documents, nous proposons un outil de sélection fine permettant de définir une zone de choix à l'intérieur de l'espace.

De plus, pour gérer les documents visualisés, l'utilisateur a la possibilité de sélectionner et supprimer tout ou partie des documents. Cette sélection lui permet, par exemple, d'accéder aux informations relatives aux documents comme l'importance des mots-clés de la requête ou tout simplement leur contenu.

Une fonctionnalité annexe est également proposée dans l'optique de l'évolution de l'interface notamment pour favoriser le lien entre l'interface et d'autre outils ou d'autres visualisations. Nous avons nommé cette fonctionnalité « *panier* » car son but est de sauvegarder les documents qui intéressent l'utilisateur pour faire le lien avec les applications annexes. A partir de ce panier, l'utilisateur a la possibilité, par exemple, d'exporter les documents au sein de sa propre hiérarchie de signets, de visualiser cet ensemble de documents au travers d'une autre représentation comme une interface de classification ([Chevalier, 2000]) ou encore de formuler une requête automatiquement à partir de ces documents.

II.3.4.3 Expérimentations

Afin d'évaluer l'intérêt de la visualisation que nous proposons, nous avons mis en place une application de test. Les aspects que nous avons souhaité souligner dans cette évaluation sont l'intérêt de la combinaison de deux axes d'interprétation (couleurs, localisation spatiale), ainsi que les aspects cognitifs de la visualisation proposée.

II.3.4.3.1 Détail de la tâche d'évaluation

L'application de test a pu soit être exécutée au sein de notre laboratoire de recherche (*évaluation locale*), soit être téléchargée et exécutée à domicile par exemple (*évaluation distante*). Dans ce dernier cas, le résultat du test était retourné par courrier électronique.

Durant la tâche d'évaluation, nous avons assisté chaque participant local. Un descriptif de la méthode d'évaluation ainsi que l'ensemble des questionnaires à rendre (annexes E & F & G) étaient également fournis comme le questionnaire personnel qui nous permet de mieux connaître les participants (annexe D).

La phase d'évaluation se décompose en trois parties :

- une première partie visant à mettre en évidence les aspects cognitifs de l'interface (en particulier l'aspect intuitif). Cette partie confronte les participants à l'interface sans aucune information préalable. Le résultat de cette phase est un questionnaire ouvert au travers duquel les utilisateurs doivent indiquer le but de l'interface, les outils disponibles etc,
- une deuxième partie visant à mettre en évidence l'intérêt de combiner plusieurs axes de visualisation pour favoriser l'interprétation exacte des documents représentés. Un participant est confronté à une série de 20 points correspondant à une combinaison des critères. Ces points sont présentés au participant selon différents points de vue (avec un point de couleur uniquement, avec un point blanc dans l'espace 3D HSV, avec un point de couleur dans l'espace 3D HSV et les même tests dans l'espace 3D HSV*). Pour chacun des points présentés, le participant doit indiquer la valeur exacte de chacun des critères,
- une troisième partie est destinée à évaluer la satisfaction de l'utilisateur et à vérifier l'utilisation réelle de l'interface grâce à un cas concret.

II.3.4.3.2 Détail des participants à l'évaluation

12 personnes, n'ayant pas participé au développement de l'interface ont accepté de participer à cette évaluation durant le mois de juillet 2001. Ce nombre est relativement faible mais il permet tout de même d'obtenir une première appréciation de l'interface.

Le panel de testeurs est mixte bien que majoritairement composé d'hommes (2/3 hommes et 1/3 femmes). L'âge moyen des participants est de 28 ans qui correspond globalement à l'âge moyen des internautes (21-35 ans) [GVU, 1998]. L'âge des participants varie cependant de 20 à 52 ans (cf Tableau 9).

Chaque participant est identifié par un numéro. Les numéros des participants à une évaluation distante sont précédés du symbole « X ».

ID	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***X1***	***X2***	***X3***	***X4***
Sexe	*M*	*F*	*M*	*F*	*F*	*M*	*M*	*M*	*M*	*M*	*F*	*M*
Age	*20*	*25*	*23*	*38*	*25*	*29*	*34*	*23*	*24*	*21*	*52*	*24*

Tableau 9 – Panel des participants

En ce qui concerne les connaissances des participants concernant l'outil informatique et le monde de la recherche d'information, chacun d'entre eux avait déjà utilisé l'outil informatique ainsi que des outils de recherche.

II.3.4.3.3 Résultats

II.3.4.3.3.1 Partie 1 : aspects cognitifs

Cette première partie nous a permis de mettre en évidence si les utilisateurs comprenaient l'intérêt de l'interface ainsi que les fonctionnalités offertes après une phase d'utilisation limitée (intuitivité). Le dépouillement du questionnaire souligne que tous les utilisateurs ont globalement compris l'intérêt de l'interface. De plus, ils ont globalement identifié les différentes fonctionnalités offertes.

II.3.4.3.3.2 Partie 2 : combinaison des axes d'interprétation

Le dépouillement des résultats de cette phase nous a permis d'apprécier l'impact de la combinaison des deux axes d'interprétation.

Pour chaque échantillon présenté, le système a sauvegardé les valeurs des trois critères saisies par l'utilisateur. A partir de ces informations enregistrées par le système, nous avons évalué si un point dans l'interface était correctement interprété c'est à dire si un utilisateur pouvait, sans aucune aide extérieure, indiquer l'importance des critères correspondant à un échantillon. Pour chaque échantillon évalué, nous avons calculé la distance euclidienne entre les valeurs réelles des critères dans l'échantillon et les valeurs saisies par l'utilisateur. Cette distance correspond à l'erreur commise par l'utilisateur.

Les distances euclidiennes calculées au cours des différentes phases, pour l'ensemble des utilisateurs, est présentée dans la Figure 63. Les résultats individuels sont présentés en annexe H.

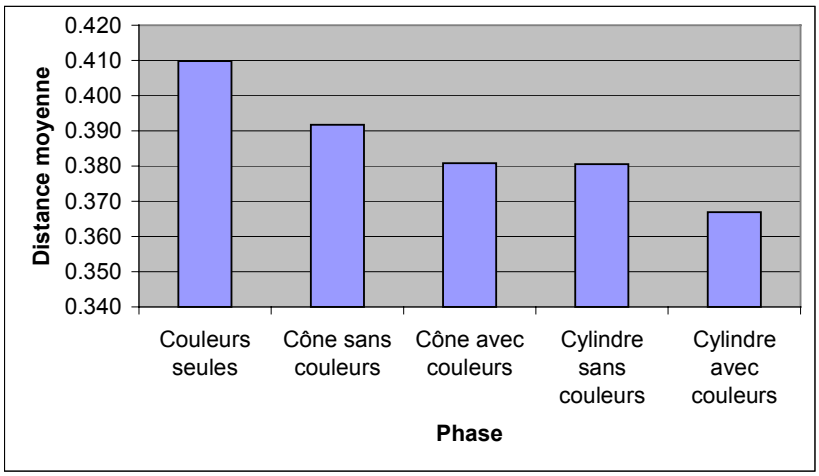


Figure 63 - Distance moyenne entre les échantillons réels et les valeurs saisies par les participants

Cette étude a mis en évidence que la combinaison des deux axes d'interprétation (couleur et localisation spatiale) permet d'obtenir de meilleurs résultats qu'avec seulement un seul axe. En effet, les visualisations en cône ou en cylindre avec couleurs permettent de réaliser une erreur plus faible de 3% que les visualisations en cône et en cylindre sans couleur. En utilisant le cylindre avec les couleurs, les participants ont réalisé une erreur inférieure de 10.5% par rapport à l'utilisation des couleurs seules. Compte tenu qu'aucune explication ou résultat intermédiaire n'étaient fournis à l'utilisateur, la baisse de l'erreur résulte donc

uniquement des visualisations proposées et non d'une expérience plus grande des utilisateurs.

II.3.4.3.3.3 Partie 3 : satisfaction de l'utilisateur

Le résultat du dépouillement des questionnaires de satisfaction subjective est présenté dans la Figure 64. Les réponses des différentes questions son données sur une échelle de valeur de 1 à 9, indiquant le degré de satisfaction de l'utilisateur.

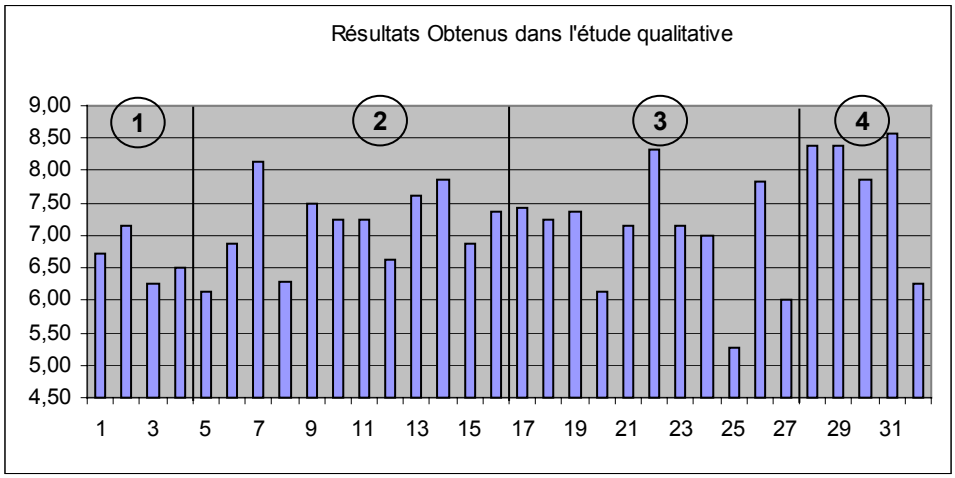


Figure 64 - Résultats de l'étude qualitative

Les réponses sont présentées par rapport au numéro de la question. Les quatre zones correspondent aux différentes parties du questionnaire (annexe F) :

1 – les réactions générales vis-à-vis du système,

2 – l'écran,

3 – l'apprentissage,

4 – le système du point de vue général.

Dans un premier temps, nous pouvons souligner le fait que les réponses sont toutes supérieures à la moyenne ce qui indique que les participants étaient globalement satisfaits de l'interface. Cependant, nous pouvons remarquer que :

- les utilisateurs déplorent le nombre trop important d'étapes à réaliser pour obtenir le résultat escompté (question 25),
- l'achèvement de la tâche n'est pas réellement compris par l'utilisateur (question 27).

Ces constats doivent être nuancés car le questionnaire a été, nous l'avons remarqué, rempli avant de réaliser l'étude de cas dans la plupart des cas. Les valeurs fournies par les participants se basent donc principalement sur une intuition et non par sur une réelle utilisation de l'interface.

Par contre, nous pouvons souligner que l'aspect intuitif de l'interface est assez bien apprécié (questions 12 et 13). De plus, l'utilisation des couleurs sur laquelle repose l'interface est assez bien perçue (questions 14, 15 et 16).

Par ailleurs, cette étude nous montre que l'appréciation de l'apprentissage est bonne (questions 20) même si certains trouvent qu'il est trop long.

Enfin, nous pouvons constater que les performances de l'interface, du point de vue des participants, sont satisfaisantes (questions 28 et 29).

Les remarques libres fournies par les participants nous ont également permis de faire évoluer la visualisation proposée. Par exemple, l'aspect abstraction a été amélioré grâce au clignotement des points correspondant aux documents sélectionnés.

Concernant l'étude de cas, 75% des participants ont réussi à identifier les documents correspondant aux critères dans le premier exercice de repérage. Ces résultats sont encourageants car cela permet de mettre en évidence que les participants ont globalement compris l'interprétation des résultats. Cependant, le deuxième exercice de repérage relatif à des documents situés dans le sommet du cône confirme la remarque faite dans [Ware, 1985]. Elle concerne le problème de la baisse de la résolution lorsque les valeurs sont faibles dans le modèle HSV. En effet, la majorité des utilisateurs a utilisé la visualisation en cône et seulement 12.5% des participants ont réussi à identifier des documents répondant aux critères. Malgré tout, la plupart des participants savaient où se trouvaient les documents mais n'ont pas réussi à les identifier du fait du grand nombre de documents situés dans cette zone. Ce point nous suggère que l'aspect « affordance » (l'utilisateur doit sélectionner l'interface la plus adaptée à ses besoins) de l'interface n'est pas optimale car ces même documents apparaissaient de façon plus significative dans la visualisation en cylindre.

II.3.4.4 Bilan sur l'interface de visualisation

Du fait de la répartition des documents (grâce aux couleurs et à leur localisation spatiale), l'interface permet d'apprécier visuellement l'importance des différents mots-clés de la requête au sein de l'ensemble des documents retrouvés.

Cette représentation offre également à l'utilisateur la possibilité d'effectuer une re-pondération des termes de la requête. En effet, si l'utilisateur souhaite donner une plus grande importance à un critère précis, il lui suffit de se concentrer sur la partie de l'espace de documents ou sur les couleurs proches du critère désiré.

Contrairement à la plupart des approches de la littérature, nous avons fait le choix d'utiliser les couleurs de façon sémantique afin d'interpréter les critères de visualisation.

Cependant, certaines fonctionnalités comme la sélection fine (pour répondre au problème de résolution faible soulevé par [Ware, 1985]) ont été introduits suite à la phase d'évaluation. Cette évaluation nous a également permis de mettre en évidence que les visualisations proposées permettent d'interpréter les résultats présentés de façon plus précise. En effet, la combinaison de deux axes d'interprétation apporte une meilleure interprétation qu'avec un axe unique. Ces résultats sont encourageants puisque cela nous laisse penser que les performances des utilisateurs (temps, qualité d'interprétation) progresseront avec l'expérience liée à l'utilisation de l'interface. Par ailleurs, la satisfaction des participants nous encourage dans le développement de l'interface proposée. Du point de vue des utilisateurs, la phase d'évaluation nous permet de confirmer, de façon pragmatique, que chaque utilisateur est vraiment unique. En effet, l'allure des courbes d'erreurs individuelles est différente d'un utilisateur à l'autre, ce qui prouve que chaque utilisateur réagit de façon singulière face à la même interface.

L'interface respecte les aspects cognitifs dans leur ensemble, même si certains d'entre eux peuvent être améliorés comme l'affordance, pour permettre une manipulation des documents résultant d'une recherche d'information encore plus aisée.

Néanmoins, quelle que soit la visualisation proposée, celle-ci ne représente qu'une partie d'une solution générale de visualisation du résultat de recherche. En effet, nous avons pu vérifier que, compte tenu des nombreuses tâches ainsi que de l'expérience variable des utilisateurs, la visualisation doit proposer plusieurs façons de représenter le résultat de recherche. Ceci nous permet de confirmer ce que soulignaient [Vernier, 1997], [Shneiderman, 1998] à savoir le fait qu'une interface développée pour une communauté d'utilisateurs ou pour une tâche précise pourra ne pas être appropriée à une autre communauté d'utilisateurs ou une autre tâche. Le fait de proposer différentes visualisations combinées doit permettre à chaque utilisateur de trouver l'outil le plus adéquat pour ses besoins et son niveau d'expérience. La combinaison de plusieurs visualisations doit également permettre à l'utilisateur d'apprécier différents types de relations entre les documents retrouvés [Hetzler, 1998] (relation de similarité, liaison par un lien hypertexte...).

L'interface de visualisation proposée montre cependant ses faiblesses du fait de l'interprétation exacte des informations représentées qui n'est pas forcément évidente pour un utilisateur néophyte. Cependant, son apprentissage semble relativement facile et à force d'essais, les utilisateurs semblent pouvoir la maîtriser. Cette visualisation se positionne donc comme un outil plutôt destiné à des utilisateurs experts en recherche d'information ce qui implique que des visualisations plus « simples » à interpréter doivent être proposées.

Outre la possibilité de retrouver des informations intéressantes en tirant partie de l'expérience des autres utilisateurs, l'utilisateur a besoin d'un outil lui permettant d'organiser les documents pertinents mémorisés qu'il trouve au gré de ses recherches pour pouvoir, par exemple, y accéder à nouveau.

II.3.5 Module de gestion et d'organisation des documents mémorisés

II.3.5.1 Problématique

L'utilisateur, nous l'avons précédemment souligné, mémorise les documents qui l'intéressent au travers de sa hiérarchie de signets. Cependant, la plupart des outils existants ne permettent que la création et l'organisation manuelle des informations mémorisées en ne permettant pas un suivi de ces informations (modification du contenu, déplacement des documents mémorisés sur le web). Or, l'utilisation d'un tel outil peut rapidement devenir difficile si l'utilisateur n'y prend pas garde. En effet, divers constats doivent être faits [Abrams, 1998] :

- les utilisateurs créent, en moyenne, environ 4 signets à chaque session de recherche,
- les utilisateurs organisent principalement leur hiérarchie de signets a posteriori (> 45%) et près de 25% ne la réorganise tout simplement pas (Figure 65).

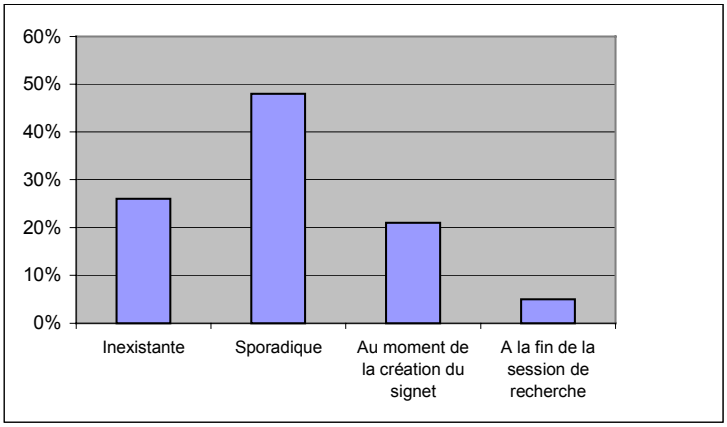


Figure 65 - Fréquence de réorganisation des signets [Abrams, 1998]

Ces constats nous permettent de souligner le problème de la gestion de l'évolution de la hiérarchie de signets. Ainsi, au fil du temps, le nombre de signets augmente tandis que la réorganisation est généralement réalisée de façon sporadique, lorsque celle-ci intervient.

Le fait que les utilisateurs ne réorganisent pas leur hiérarchie de signets peut venir du fait que l'effort cognitif que cela demande est trop important [Rücker, 1997]. Par ailleurs, cet effort à réaliser est accru par le fait qu'un simple signet ne contient pas suffisamment d'information pour rendre compte du contenu du document. Par exemple, il n'est pas facile de se remémorer le contenu d'un document ayant pour titre « Please title (Homepage) » et dont l'URL est « http://www.qsl.net/n4upb/ » ? Cette sous-information conditionne la réorganisation des signets surtout si celle-ci est réalisée a posteriori (ou sporadiquement) car elle nécessite que l'utilisateur visite à nouveau les documents pointés par les signets qu'il souhaite réorganiser.

Pourtant, plus le nombre de document augmente au sein de la hiérarchie de signets, plus la nécessité d'avoir une organisation arborescente se fait ressentir. [Abrams, 1998] souligne, en particulier, qu'il est souhaitable d'utiliser une organisation en répertoires pour plus de 30 signets.

Par ailleurs, une caractéristique fondamentale du web est sa fréquence élevée de mise à jour, c'est-à-dire que les informations naissent, disparaissent ou sont fréquemment déplacées. Or, l'utilisateur ne peut pas être constamment à l'affût des modifications qui pourraient subvenir à l'un des documents qui l'intéresse. Il existe des outils en ligne comme *Mind It* (http://www.mymindit.com/) qui permettent d'informer un internaute des modifications apportées à un certain document web, mais cela demande à l'utilisateur d'enregistrer manuellement tous les documents dont il souhaite connaître l'évolution. De plus, ce type d'outil n'indique à l'utilisateur que les modifications du contenu qui ont été effectuées mais n'informe pas l'utilisateur des déplacements éventuels des documents (changement d'URL par exemple).

L'objectif de notre module de gestion des signets est de proposer à l'utilisateur un véritable suivi de ses signets ainsi qu'une aide à leur réorganisation sous la forme d'une classification hiérarchique.

II.3.5.2 Approche proposée

Afin de permettre un véritable suivi des informations mémorisées par l'utilisateur, il est envisagé d'informer automatiquement cet utilisateur sur l'état des documents issus de ses signets : modification du contenu, changement d'URL ou simple disparition.

Nous proposons également à l'utilisateur une véritable gestion des signets grâce à une réorganisation automatique hiérarchique.

Nous présentons, dans ce qui suit, les deux aspects de ce module, c'est-à-dire :

- la mise à jour des signets,
- l'aide à la réorganisation des signets.

II.3.5.2.1 Mise à jour des signets

La mise à jour des signets vise à renseigner l'utilisateur sur les modifications éventuelles apportées aux documents qu'il possède dans sa hiérarchie de signets.

La démarche proposée consiste à obtenir les informations sur le web relatives aux documents mémorisés (date de dernière modification, nouvelle URL d'un document...). Le système peut ainsi vérifier si le document a été modifié ou déplacé par exemple.

II.3.5.2.2 Aide à la réorganisation des signets

Nous avons souligné le fait que peu d'utilisateurs réorganisent leurs signets. Cependant, pour utiliser de façon optimale ce type d'information, il est nécessaire que les signets soient organisés de préférence au sein d'une hiérarchie de répertoires. Celle-ci permet à l'utilisateur de retrouver facilement un document et ainsi de limiter l'effort cognitif nécessaire pour y accéder [Maarek, 1996].

Pour aider l'utilisateur à construire cette hiérarchie de répertoires, nous proposons un outil de réorganisation de tout ou partie de ses signets. En effet, la réorganisation ne s'impose pas forcément sur la hiérarchie de signets complète. Cet outil permet donc à l'utilisateur de sélectionner tout ou partie de ses signets afin que le système lui en propose une réorganisation thématique hiérarchique. Pour cela, nous avons utilisé la *Classification Hiérarchique Ascendante* [Rijsbergen, 1971] qui permet d'obtenir à partir d'un ensemble de documents une classification arborescente. Cette approche utilise la notion de *cluster* qui correspond à un nœud de l'arborescence et qui peut contenir soit un document soit d'autres clusters. Un cluster vise à rassembler des éléments afin que la similarité intra-cluster soit plus importante que la similarité inter-cluster.

La méthode de classification est la suivante :

> Chaque document est inséré dans un cluster.
> Les clusters les plus similaires sont fusionnés 2 à 2 dans un nouveau cluster jusqu'à ce qu'il n'y ait plus qu'un seul cluster.

Le résultat obtenu est une arborescence de clusters qui est communément appelée *dendogramme* dépendant de la similarité inter-clusters (Figure 66).

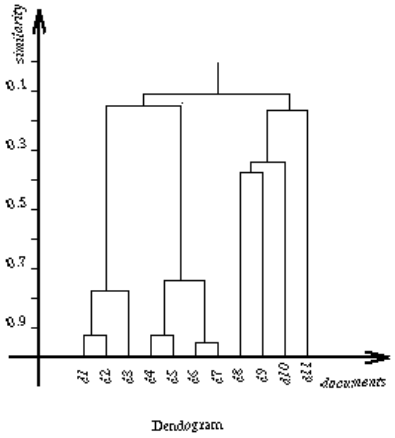


Figure 66 - Un dendogramme

Pour estimer la similarité inter-cluster, différentes mesures ont été proposées telles que le saut minimal, le saut maximal. Dans notre approche, le choix de la méthode est laissé à l'utilisateur qui peut préférer une organisation en classes générales (saut minimal) ou en classes spécifiques (saut maximal). Pour calculer la similarité entre deux clusters selon ces deux méthodes, le système calcule la similarité entre les documents d'un des clusters et ceux de l'autre cluster. Le saut minimal consiste à utiliser la similarité des documents les plus similaires (MAX). A l'inverse, le saut maximal consiste à utiliser la similarité entre les documents les moins similaires (MIN).

Pour présenter à l'utilisateur la réorganisation que le système propose, chaque cluster est transformé en un répertoire.

Cependant, cette méthode souffre du fait que la fusion des clusters se fait de façon binaire. En effet, la profondeur maximale est fonction de l'hétérogénéité des documents ainsi que de leur nombre. Il est, en effet, inconcevable de proposer à un utilisateur une hiérarchie ayant une profondeur maximale de 10 niveaux au sein de laquelle chaque nœud ne contient que deux éléments ! Pour laisser l'utilisateur maître de la réorganisation, nous avons fait appel à une fonction de seuillage comme proposée dans [Maarek, 1996]. Elle permet de contrôler la profondeur maximale de la hiérarchie proposée. Ce seuillage permet de fusionner tous les clusters compris entre deux paliers de similarité en un seul cluster (Figure 67).

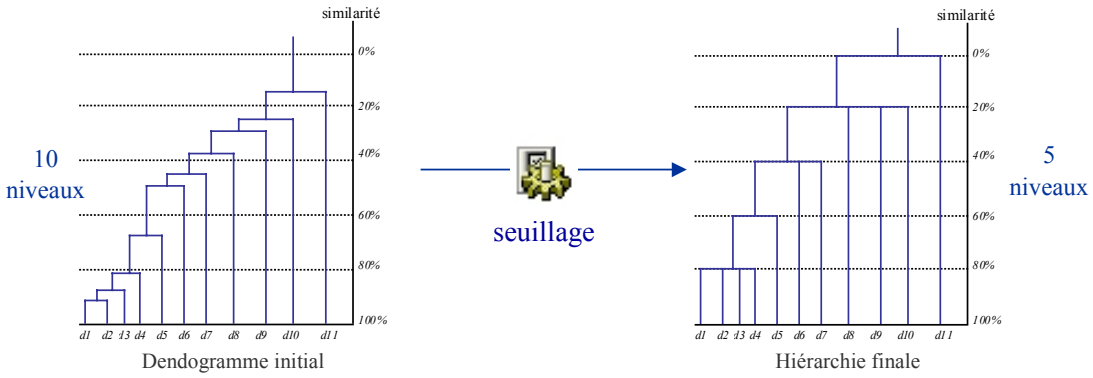


Figure 67 - Contrôle de la profondeur de la classification par seuillage

Chacun des clusters correspond à un ensemble anonyme de documents. Le but étant de présenter la réorganisation à l'utilisateur, nous avons ajouté un processus qui permet de nommer chaque répertoire correspondant à un cluster pour aider l'utilisateur à apprécier la classification obtenue.

Notre proposition repose sur le fait que les termes représentatifs d'un cluster sont les termes qui se retrouvent dans la plupart des documents qu'il contient.

Pour identifier les termes pertinents d'un répertoire, nous utilisons à nouveau la mesure du χ^2. En effet, elle permet d'obtenir les termes qui se retrouvent fréquemment dans les documents d'un nœud et peu fréquemment dans les documents extérieurs au nœud.

L'étiquette de chaque répertoire est construite à partir des k termes ayant la meilleure valeur du χ^2. Cette étiquette a pour seul but d'aider l'utilisateur à comprendre le contenu du répertoire (et de son arborescence fille). L'utilisateur a tout loisir de modifier ce nom de répertoire selon ses besoins.

A partir de la réorganisation proposée par le système, l'utilisateur peut accepter tout ou partie de la proposition du système pour l'appliquer à sa propre hiérarchie de signets.

II.3.5.3 Bilan du module de gestion et organisation des documents mémorisés

La gestion de la hiérarchie de signets des utilisateurs permet à chacun d'avoir un espace d'information personnel à jour et bien organisé, permettant ainsi un accès efficace aux documents mémorisés. Grâce à ce module, l'utilisateur est informé de toute modification apportée aux documents qui l'intéressent. Cette solution offre la possibilité à l'utilisateur d'obtenir une organisation, à jour, des documents mémorisés sans trop d'investissement, que ce soit en temps ou en effort cognitif.

De plus, par le biais de ce module, le système peut exploiter des connaissances mieux organisées et les partager efficacement (notamment lors de la navigation). En effet, les signets sont organisés de manière cohérente et non laissés anarchiquement à la racine de la hiérarchie de signets.

Cependant, cette approche trouve ses limites dans le cas où l'utilisateur ne souhaite pas organiser ses signets de façon thématique. En effet, le module actuel repose sur le contenu des documents pointés par les signets, ce qui peut ne pas satisfaire tous les besoins d'un utilisateur qui pourrait souhaiter organiser ses signets par auteur par exemple.

II.3.6 Respect de l'utilisateur

L'approche que nous proposons au travers du processus global du projet *Easy-DOR* utilise les informations collectées et visitées par les utilisateurs. Cependant, cette démarche peut être critiquable du point de vue de la protection de la vie privée. Bon nombre d'internautes utilisent une messagerie en ligne ou consultent par exemple leurs comptes bancaires qu'il ne serait pas judicieux de diffuser au sein du groupe d'utilisateurs.

Pour toutes ces raisons, nous avons intégré un moyen permettant à l'utilisateur de protéger les informations personnelles qu'il détient ou visite. La confidentialité doit être une caractéristique forte d'un tel système de partage d'informations et c'est la raison pour laquelle nous avons laissé l'utilisateur décider de ce qui est personnel ou non et donc de ce

qui est partageable ou non. Cette confidentialité intervient dans un premier temps au niveau des documents visités. Pour offrir à l'utilisateur une certaine « intimité » vis-à-vis du système, il peut à tout moment activer ou désactiver le processus de partage d'informations à partir des documents qu'il visite. Ainsi, avant de consulter son courrier électronique sur une messagerie en ligne par exemple, l'utilisateur désactive le processus pour annuler tout traitement de partage des documents qu'il va visiter. L'utilisateur pourra, à tout moment, réactiver ce processus pour indiquer au système qu'il peut traiter à nouveau les documents visités et permettre aux autres utilisateurs de profiter de sa navigation.

Dans un second temps, la confidentialité doit intervenir au niveau de la hiérarchie de signets d'un utilisateur qui est utilisée comme base de recommandation durant la navigation. Cependant, l'utilisateur peut souhaiter que certains documents qu'il possède ne soient pas diffusés aux autres utilisateurs comme des documents personnels par exemple. Pour cela, l'utilisateur peut indiquer au système que certains documents ou répertoires de sa hiérarchie de signets sont confidentiels, lui interdisant ainsi l'exploitation de leur contenu pour un quelconque partage d'informations.

Par ailleurs, pour éviter que l'utilisateur soit submergé d'informations qu'il ne désire pas, il a la possibilité d'activer ou de désactiver l'ensemble des modules de recommandation (recommandations pour la connaissance du domaine et durant la navigation). Concernant la recommandation pour la connaissance du domaine, cette activation/désactivation peut également être réalisée pour tout ou partie des nœuds de la hiérarchie de signets.

II.3.7 Bilan sur l'aspect coopératif

Ce projet permet d'aider un utilisateur à effectuer une recherche d'information avec comme leitmotiv le partage de connaissances. Un utilisateur au sein d'un groupe d'utilisateurs ou au sein d'une organisation telle qu'un laboratoire de recherche ou une entreprise, a intérêt à tirer partie des informations collectées par les autres utilisateurs pour obtenir des informations jugées et validées par des utilisateurs ayant les mêmes centres d'intérêt.

Il est évident que le partage de connaissances réalisé à grande échelle de façon manuelle demande un investissement personnel important de la part de chacun des acteurs du groupe ou de l'organisation. C'est pour cette raison qu'*Easy-DOR* propose une solution permettant d'effectuer ce partage de façon transparente dans le but d'aider à la recherche d'information. Cette aide intervient à différents niveaux de la recherche d'informations, ce qui permet à l'utilisateur de profiter des informations collectées par le groupe d'utilisateurs à tout moment de sa recherche. Elle est fournie par l'intégration de différents outils partageant une représentation unique des centres d'intérêt des utilisateurs reposant sur leur hiérarchie de signets. Ce partage de connaissances est basé sur différentes exploitations des informations collectées par les utilisateurs (selon leur contenu, selon leur organisation au sein des hiérarchies de signets...), ce qui permet de recommander des documents selon différents points de vue. Ainsi, les documents visités par les utilisateurs sont proposés aux différents utilisateurs potentiellement intéressés. Par ailleurs, l'organisation des documents mémorisés au sein des hiérarchies de signets des utilisateurs est exploitée afin de proposer des

documents pertinents au cours de la navigation d'un internaute tiers. Ces approches reposent sur le principe que les utilisateurs qui possèdent des centres d'intérêt similaires sont intéressés par les mêmes documents.

De plus, au delà de l'aide à la RI sur le web pour laquelle le partage d'informations a été conçu, il peut servir de support à d'autres processus. Par exemple, afin de rendre les documents persistants, une mémoire documentaire du groupe d'utilisateurs pourrait être imaginée [Mothe, 2000], [Chevalier, 2001c]. Un tel entrepôt a pour but de rendre persistantes les informations pertinentes pour le groupe d'utilisateurs. De plus, cette mémoire documentaire peut être exploitée au travers de différentes analyses (multidimensionnelle par exemple) ou au travers de recherche par mots-clés. Dans cette approche, l'aspect coopératif peut être notamment utilisé au niveau de l'alimentation de l'entrepôt pour filtrer les documents afin qu'il ne contienne que des documents pertinents pour le groupe d'utilisateurs.

II.4 Prototype Easy-DOR

L'ensemble des modules proposés a donné lieu au développement d'un prototype permettant de montrer l'efficacité d'un telle approche.

L'architecture est de type client-serveur. Le prototype est réalisé entièrement en langage Java (1.2) en utilisant les API telles que JDBC pour l'accès aux bases de données. Ce prototype a fait l'objet de plus de 7000 lignes de code.

Du côté serveur, plusieurs applications (filtrage, proxy, indexation) s'exécutent de façon parallèles sur une ou plusieurs machines différentes. Le dialogue avec le client est réalisé par l'intermédiaire d'un port serveur spécifique pour chacune des tâches.

L'architecture du prototype est présentée dans la Figure 68.

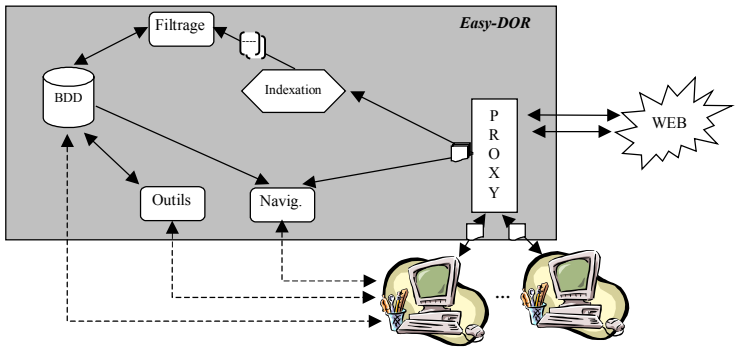


Figure 68 - Architecture générale du système

Cette architecture comprend :

a) L'élément *Proxy* correspond à un composant Proxy qui permet au système d'avoir la liste des documents visités par les utilisateur ainsi qu'à leur contenu,

b) L'élément *BDD* correspond à la base de données relationnelles permettant de stocker, entre autres, les hiérarchies de signets ainsi que toutes les informations s'y référant.

c) Les modules sont représentés par les 3 éléments *Filtrage*, *Navig* et *Outils* :

- *Filtrage* correspond au module de recommandation pour la connaissance du domaine de l'utilisateur,
- *Navig.* est le module permettant les recommandations durant la navigation de l'utilisateur,
- *Outils* intègre le module de recherche, ainsi que le module de réorganisation des signets web.

d) L'interface de visualisation est, quant à elle, intégrée chez le client.

Dans cette section, nous détaillons, dans un premier temps, l'architecture proxy utilisée suivie de la présentation de la modélisation du système, puis dans un second temps, l'implantation des différents modules définis dans le projet *Easy-DOR*. Dans cette présentation des modules, nous avons mis l'accent sur l'interaction entre l'utilisateur et le système.

II.4.1 Architecture Proxy

Afin de détecter les documents web visités par l'ensemble des utilisateurs, le système repose sur une application de type *proxy* [Mignot, 2000] entre les utilisateurs et le web. Habituellement, un proxy joue le rôle de cache permettant d'accélérer les connexions web. Dans notre cas, l'architecture proxy utilisée se limite à la détection et à la sauvegarde du contenu des documents visités par les utilisateurs (Figure 69). Il représente le composant principal du partage de connaissances et son rôle est celui « d'un ami qui regarde par dessus l'épaule de l'internaute pour lui donner des informations visant à l'aider dans ses recherches ». Ce composant est appelé « *sniffer* » pour refléter sa fonction d'écoute des actions de l'utilisateur.

Le principe est le suivant. Le contenu de chaque document visité par un utilisateur du système est sauvegardé pour être indexé et utilisé dans les différents modules (Figure 69). Ce contenu est représenté par un vecteur descripteur du modèle vectoriel. Cette représentation des documents est sauvegardée dans une base de données pour être exploitée notamment lors des recommandations. Pour les documents HTML, par exemple, les balises sont supprimées et l'indexation ne porte que sur le contenu textuel brut. Cependant, les informations contenues dans les balises de méta-données comme les mots-clés ou la description du contenu du document sont incorporées au texte brut pour être utilisées lors de l'indexation. Dans notre approche, le système repose sur une indexation classique comme décrite dans la section cf. I.3.1.2.1.

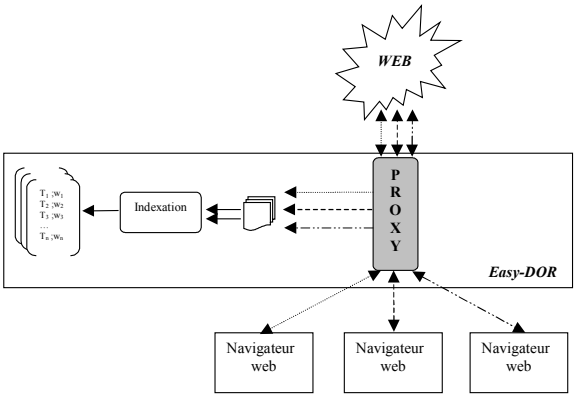


Figure 69 - Architecture proxy utilisée dans le prototype *Easy-DOR*

Par ailleurs, le proxy que nous avons utilisé peut être configuré pour n'écouter que certaines des connexions. En effet, le sniffer a pour but de sauvegarder les entêtes et les contenus de documents pouvant servir au partage d'informations. Dans notre processus, nous n'avons pour l'instant considéré que le partage d'informations textuelles. Afin de se concentrer sur de telles informations, le sniffer a été conçu de façon à ne sélectionner que les connexions dont le contenu est d'un type *MIME* prédéfini (dans notre cas textuel). Le type *MIME* d'un document visité est accessible au travers de l'entête des réponses provenant du serveur web. Le type *MIME* des documents acceptés et dont on souhaite exploiter le contenu peut être fixé par l'administrateur du système. Dans notre application, nous avons utilisé les types *MIME* tels que *text/html*, *text/plain*.

Le fait de ne s'intéresser qu'aux connexions de types textuels permet de ne traiter qu'un sous-ensemble restreint des connexions web. Le Tableau 10 présente les types des connexions web réalisées, au travers de 345 postes clients, lors d'une journée, au sein de notre laboratoire de recherche en décembre 2000. A partir de ce tableau, nous pouvons constater que les connexions de type textuel (HTML, SHTML, Text) ne représentent en réalité qu'un petit nombre de connexions (< 4%) permettant d'économiser énormément de ressources ce qui ne pénalise pas les utilisateurs du système. *Remarque : Si, par exemple, une page web contient 1 image, deux connexions sont crées : une pour le document source HTML et une pour le fichier image. La page web est donc décomposée en granules du point de vue des connexions.*

Type de connexion	***Nombre de connexions***	***Pourcentage***
Images	92823 connexions	(65,2 %)
HTML	4487 connexions	(3 %)
Directory	3282 connexions	(2,3 %)
SHTML	306 connexions	(0,21 %)
Bundle	201 connexions	(0,15 %)
Text	112 connexions	(0,08 %)
Movie	59 connexions	(0,05 %)
Audio	12 connexions	(0,01 %).
Autres	41092 connexions	(29 %)

Tableau 10 - Types de connexions web

Afin de ne sauvegarder que le contenu des documents ayant un type *MIME* prédéfini, le fonctionnement du sniffer se base sur une comparaison simple du type *MIME* du document avec ceux acceptés par l'application (Figure 70).

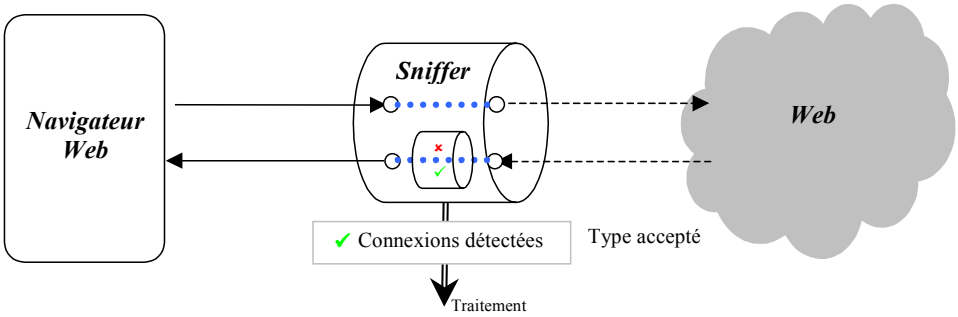


Figure 70 - Fonctionnement du sniffer

II.4.2 Modèle sous-jacent au système

Le système *Easy-DOR* a été représenté au travers d'un diagramme de classes UML (Figure 71). Ce diagramme représente les éléments essentiels utilisés dans les différents modules proposés (principalement les utilisateurs et leur hiérarchie de signets).

La classe *Utilisateur* représente l'ensemble des utilisateurs enregistrés dans le système. Chaque utilisateur possède une hiérarchie de signets caractérisée par un répertoire racine. Chaque signet correspond à un document et est caractérisé par un attribut nommé *Attributs* qui représente l'état du document lié à ce signet (supprimé, modifié, non visité...).

Pour chaque répertoire, le système sauvegarde un ensemble de jugements correspondant aux jugements de pertinences correspondant à la phase de recommandation pour la connaissance de l'utilisateur.

Le type de jugement peut avoir la valeur 0 si le document n'est pas accepté pour le répertoire, 1 s'il est accepté et jugé pertinent, 2 s'il est accepté et jugé non pertinent, 3 s'il est accepté et en attente de jugement de la part de l'utilisateur. Ainsi, grâce à ces valeurs le système peut distinguer les documents pertinents pour les nœuds des documents non pertinents. Les documents pertinents mais non jugés ne sont pas pris en compte dans la phase de recommandation pour la connaissance du domaine car ils n'ont pas fait l'objet d'un jugement explicite de la part de l'utilisateur.

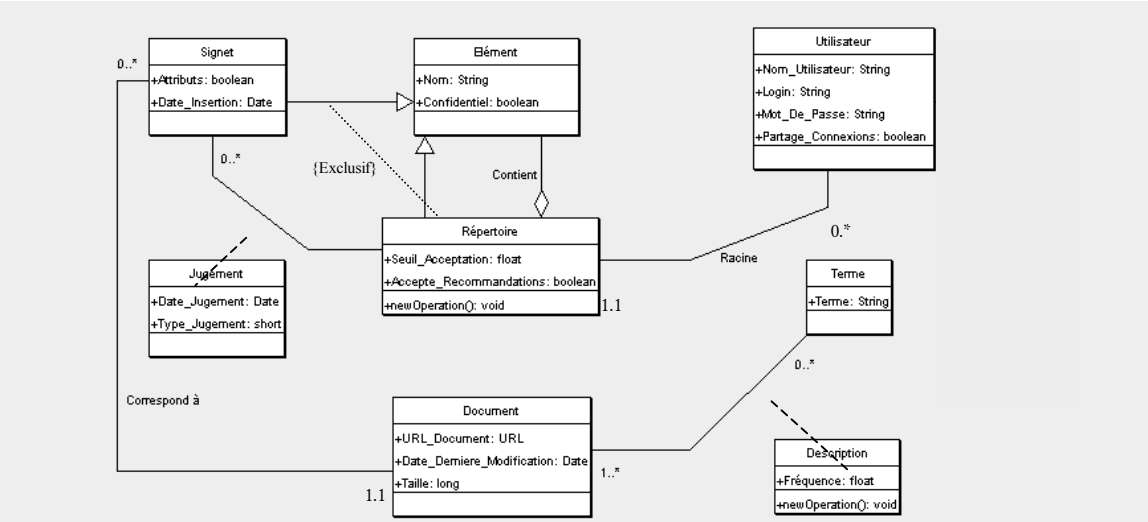


Figure 71 – Diagramme UML des classes de base du système

II.4.3 Module de recommandation pour la connaissance du domaine

Dans ce module, nous avons adopté un classifieur de type Rocchio $R_{16/4}$ car il permet de généraliser la proportion de termes à prendre en compte lors de sa construction. Au vue des expérimentations, nous avons choisi de prendre en compte 5.5% des termes issus des documents pour construire le classifieur. Par ailleurs, nous avons opté pour une mise à jour du seuil de chaque classifieur de façon périodique par la méthode du seuil optimal.

Les recommandations pour la connaissance du domaine de l'utilisateur sont directement proposées dans sa hiérarchie de signets. Pour permettre à l'utilisateur de faire une distinction entre sa hiérarchie de signets initiale et les signets qui lui sont proposés pour un nœud, le système emploie un répertoire spécifique. Ce répertoire (*ED_Nouveaux_Signets*) est ajouté à chaque répertoire de la hiérarchie de signets de l'utilisateur et ne peut contenir que les documents recommandés par le système. Pour un meilleur confort d'utilisation, ce répertoire n'est visible que s'il contient des documents recommandés (Figure 72). Dans cette figure, la hiérarchie de signets initiale de l'utilisateur est présentée à gauche alors que la même hiérarchie dans laquelle un document a été recommandé est représentée à droite.

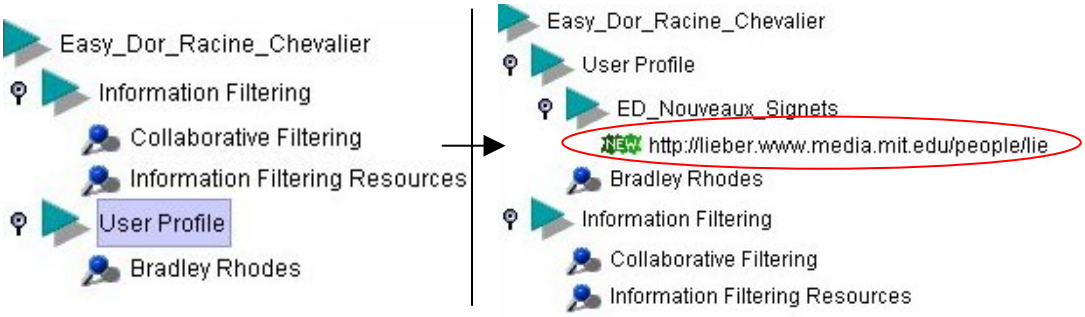


Figure 72 - Recommandation dans la hiérarchie de signets

Les icônes jouent un rôle important dans l'application permettant à l'utilisateur d'accéder d'un coup d'œil à des informations complémentaires relatives aux éléments de la hiérarchie de signets.

Les icônes qui précèdent les éléments de la hiérarchie utilisés dans la hiérarchie sont les suivants (Tableau 11) :

Répertoire partagé (bleu)	Répertoire Confidentiel (rouge)	Signet partagé (bleu)	Signet Confidentiel (rouge)	Document recommandé jamais visité
				

Tableau 11* - Icônes de la hiérarchie

II.4.4 Module de recommandation lors de la navigation

L'approche générale de ce module est présentée dans la Figure 73.

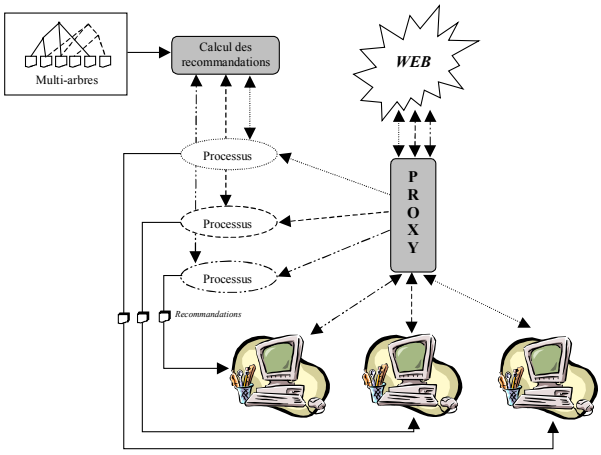


Figure 73 - Approche de la recommandation durant la navigation

Chaque utilisateur connecté au système est associé à un processus. Chacun de ces processus à pour rôle de gérer les relations avec le client et de construire la liste des recommandations au fur et à mesure que le proxy lui indique les documents visités par l'utilisateur.

Les recommandations sont proposées à l'utilisateur de façon périodique. Dans notre système, le client interroge le serveur pour obtenir la nouvelle liste de recommandations toutes les minutes.

Afin que l'utilisateur comprenne pourquoi le système lui recommande tel ou tel document, la présentation des informations est réalisée sous la forme d'une liste ordonnée des recommandations. Cette présentation a été particulièrement soignée de façon à en faciliter l'interprétation. Pour chacune des recommandations, un composant graphique (Figure 74) est inséré dans la fenêtre du client. Ce composant présente :

- le titre du document,
- son importance par rapport à la navigation en cours sous la forme d'une barre de couleur (du point de vue système),
- la liste des répertoires des utilisateurs du système qui contiennent un signet pointant vers ce document. Un répertoire est représenté par son chemin complet dans l'arborescence de l'utilisateur initial (chaque nœud de l'arborescence étant séparé par un « / »). Cette présentation des répertoires dans lesquels se trouve le document

recommandé permet à l'utilisateur de connaître quelles sont les problématiques auxquelles le document a été rattaché par les utilisateurs du système. Cette représentation permet de comprendre la relation entre le document et la navigation en cours. A partir de ces chemins, à l'utilisateur a la possibilité d'accéder directement au contenu d'un nœud en cliquant sur son nom.

Figure 74 – Deux composants d'une liste de recommandations

La Figure 75 présente une copie d'écran relative à la recommandation de type documents avec la visite des répertoires relatifs aux documents proposés. Nous pouvons voir, au travers de cette figure, que l'utilisateur a choisi de visualiser le contenu de deux répertoires des hiérarchies de signets dans lesquels se trouvent des documents recommandés. Ces répertoires sont présentés dans la fenêtre du bas qui fait apparaître tous les documents qu'ils contiennent. Par ce biais, l'utilisateur a la possibilité de parcourir les hiérarchies de signets des utilisateurs.

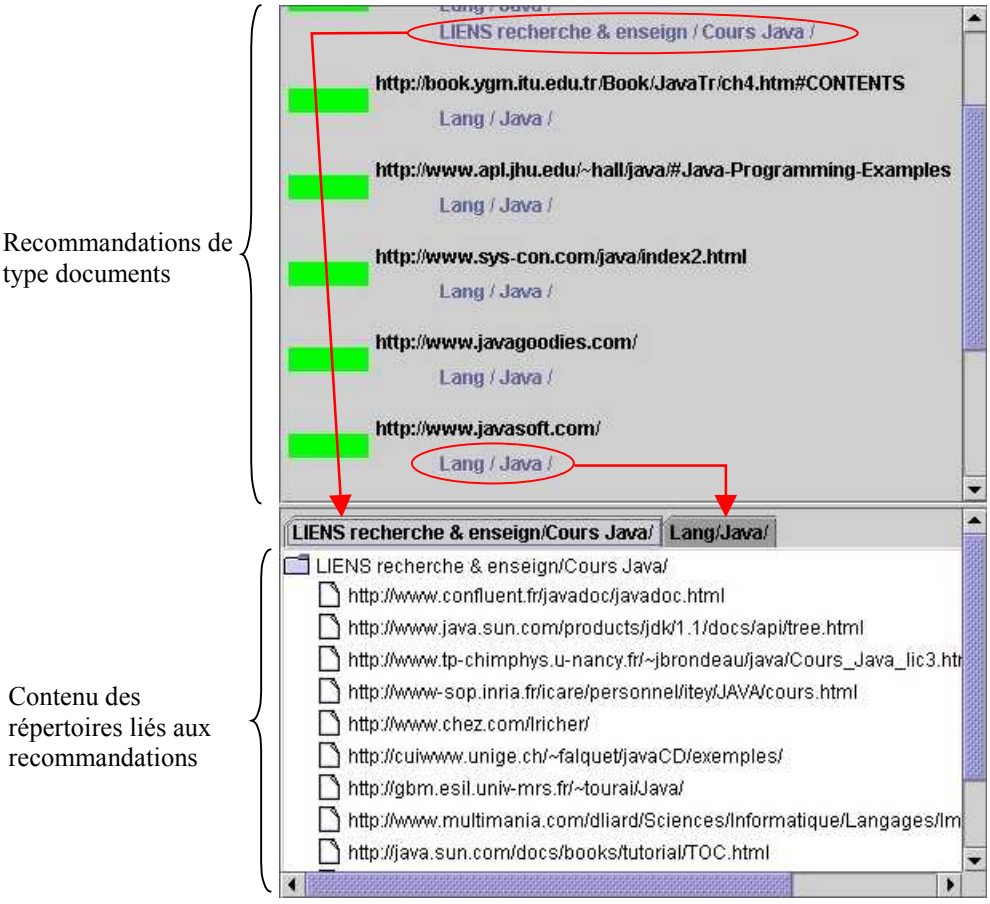


Figure 75 - Recommandations durant la navigation

II.4.5 Module de visualisation des résultats de recherche

L'interface de visualisation permet de représenter de façon graphique les résultats de recherches d'informations afin de permettre à l'utilisateur de comprendre ce résultat de façon globale (Figure 76).

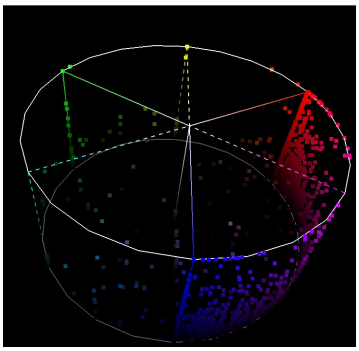


Figure 76 - Un exemple de visualisation en cylindre

Pour faciliter la compréhension des éléments représentés, différents outils et fonctionnalités ont été proposés en plus de la visualisation 3D. En premier lieu, pour aider l'utilisateur à garder à l'esprit les différents critères utilisés, une fenêtre de légende est proposée. Cette légende reprend les couleurs et les critères auxquels elles se réfèrent.

Différents outils sont également proposés dans cette fenêtre. On y retrouve les rotations, les sélections document par document ou par zone...

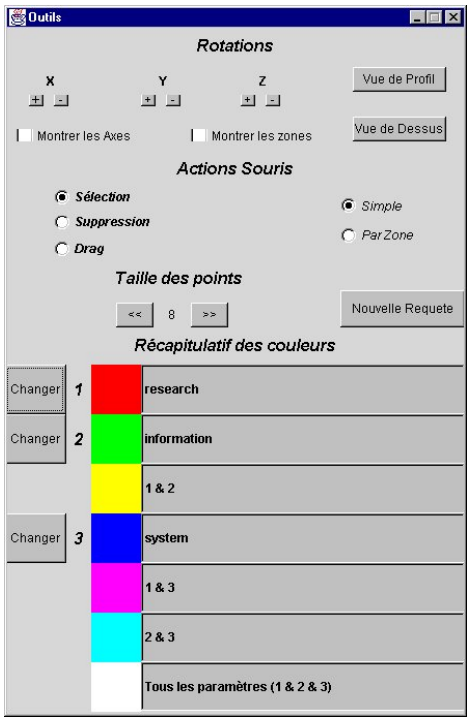


Figure 77 - Fenêtre d'outils de l'interface de visualisation

Dans la figure précédente, nous pouvons voir des boutons nommés « changer » qui permettent à l'utilisateur de choisir ou combiner les termes de la requête qu'il souhaite affecter à un critère.

Afin de permettre une meilleure sélection des documents, nous avons introduit une fonctionnalité de « sélection fine » au travers de la visualisation. Celle-ci est accessible via les carrés de couleurs correspondant aux différents critères. Nous proposons deux types de sélection :

- la sélection pour les critères de base (rouge, vert, bleu). Elle permet de sélectionner les documents ne répondant qu'à un critère de base. Par exemple, si l'utilisateur souhaite sélectionner les documents qui ne possèdent que le critère *system* (critère bleu), il suffit de cliquer sur le carré bleu correspondant au critère 3,
- la sélection pour les critères combinés (cyan, magenta, jaune ou blanc). Elle permet de sélectionner une zone du cône autour de ces critères. Cette sélection permet de sélectionner des zones géographiques de l'espace 3D. Par ce biais, l'usager peut préciser l'étendue de sélection autour du critère sélectionné mais aussi la hauteur minimum (importance des documents). La Figure 78 présente un exemple de sélection dans le cône pour la combinaison des 3 critères (critère blanc). La forme rouge correspond à la zone de sélection définie par l'utilisateur.

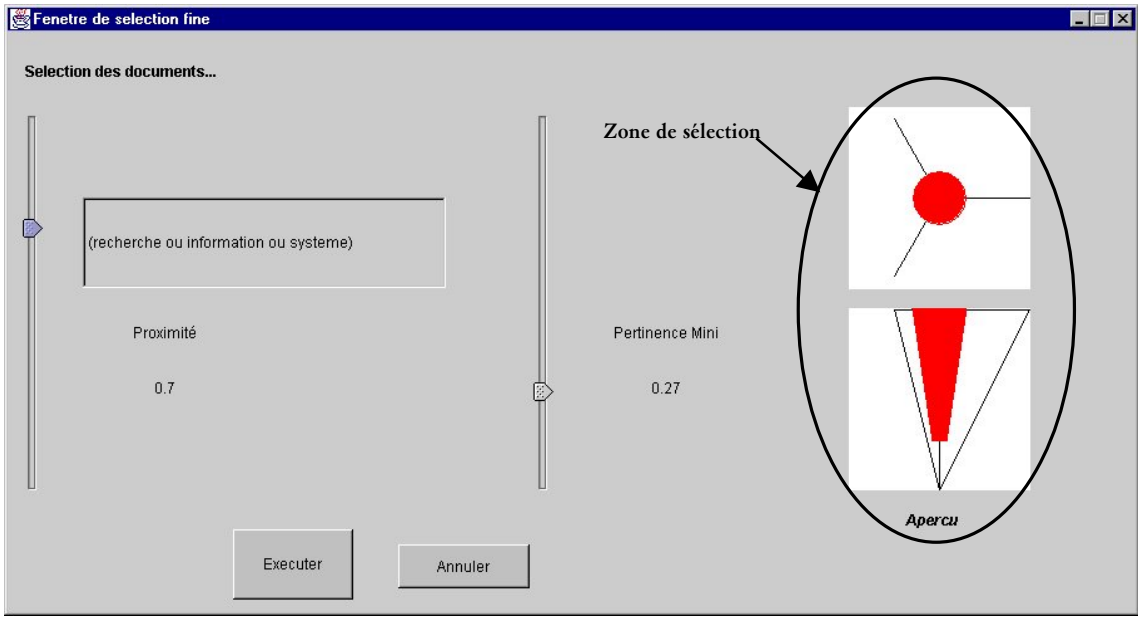


Figure 78 - Outil de sélection fine

La plupart des commandes disponible est également accessible par des touches du clavier permettant une multi-modalité en entrée.

II.4.6 Module de gestion des signets

II.4.6.1 Mise à Jour des signets

Afin de détecter les modifications apportées aux différents documents qui intéressent les utilisateurs, nous avons fait le choix d'utiliser :

- la navigation des utilisateurs,
- une tâche périodique de mise à jour.

Quelle que soit l'approche utilisée, le système utilise l'entête des réponses provenant des serveurs web pour vérifier la persistance de l'information. D'après le contenu de ces entêtes, le système accède aux informations concernant le document : code de retour ou encore date de dernière modification du document. Par exemple, le code de retour (404) pour un document, permet au système d'informer tous les utilisateurs possédant un signet vers ce document que celui-ci n'existe plus. A partir de ce code de retour, le système peut informer l'utilisateur du changement d'URL d'un document (grâce aux codes de retour 3xx).

Dans la même optique, le système utilise le champs *Last-Modified* pour vérifier la date de dernière modification d'un document. Si celle-ci est différente de celle conservée par le système (depuis la dernière visite), le système informe tous les utilisateurs ayant un signet pointant vers le document que celui-ci a été modifié.

La navigation des différents utilisateurs du système peut être utilisée afin de participer à la mise à jour des documents qui intéressent les usagers. Ainsi, si un utilisateur visite un document qui n'existe plus, le système, au travers du proxy, le détecte et avertit tous les utilisateurs possédant un signet pointant vers ce document. Cette approche trouve tout son sens dans le contexte d'un groupe d'utilisateur car les documents se rapportant à des thèmes communs ont une probabilité importante d'être visité par l'un des utilisateurs du système.

La tâche périodique de mise à jour est envisagée pour vérifier si les documents qui intéressent les utilisateurs n'ont pas subi de modification. Cette vérification est réalisée grâce à la commande *Head* du protocole *HTTP* qui permet d'accéder uniquement à l'entête du document.

Pour faire comprendre à l'utilisateur que certains des documents qu'il possède dans sa hiérarchie de signets ont été supprimés, déplacés ou modifiés, des icônes spécifiques sont utilisés. Ces icônes précèdent le noms des signets et remplacent les signets traditionnels jusqu'à ce que ceux-ci soient visités par l'utilisateur (Tableau 12).

Document déplacé Document modifié Document supprimé

Tableau 12* - Icônes utilisés pour indiquer la mise à jour des signets

II.4.6.2 Réorganisation des signets

En ce qui concerne la réorganisation des signets, l'utilisateur doit sélectionner l'ensemble des documents dont il souhaite voir la réorganisation à partir d'un arbre à cocher. La Figure 79 présente la fenêtre qui succède à cette sélection et qui présente la réorganisation des signets.

La partie gauche de la fenêtre présente l'arborescence initiale de l'utilisateur avec les documents qui ont été sélectionnés pour la réorganisation. La partie de droite, présente la réorganisation des signets sélectionnés réalisée par le système. Par souci de compréhension, nous utilisons les couleurs de façon pragmatique pour permettre à l'utilisateur de repérer les modifications qu'il apporte à sa propre hiérarchie initiale. Ainsi, les signets provenant de la réorganisation sont signalés avec une couleur bleue, tandis que les signets initiaux sont en noir.

A partir de cette réorganisation, nous proposons à l'utilisateur de sélectionner les parties de la réorganisation qui l'intéressent et de les réinsérer au sein de sa hiérarchie de signets.

Pour accepter une partie de la réorganisation (un répertoire ou simplement un signet) et pour privilégier l'interaction, l'utilisateur utilise le principe du « glisser-déposer ». Par ailleurs, pour éviter tout problème de manipulation, toutes les modifications réalisées par l'utilisateur concernant sa hiérarchie de signets ne seront effectives qu'après validation.

Afin d'aider l'utilisateur dans sa démarche, les couleurs sont utilisées pour rendre compte des actions qu'il a réalisées (Figure 79) :

- les éléments acceptés par l'utilisateur sont grisés dans la fenêtre de la classification proposée par le système et sont signalés en bleu dans la hiérarchie de signets initiale de l'utilisateur,
- les éléments supprimés par le système sont barrés et signalés en rouge dans la hiérarchie initiale de l'utilisateur. En effet, suite à une acceptation, nous avons considéré que tous les signets réorganisés qui pointent vers les documents réintroduits n'ont plus lieu d'être puisque l'utilisateur a accepté la nouvelle classification. Il peut toutefois annuler cette suppression à tout moment.

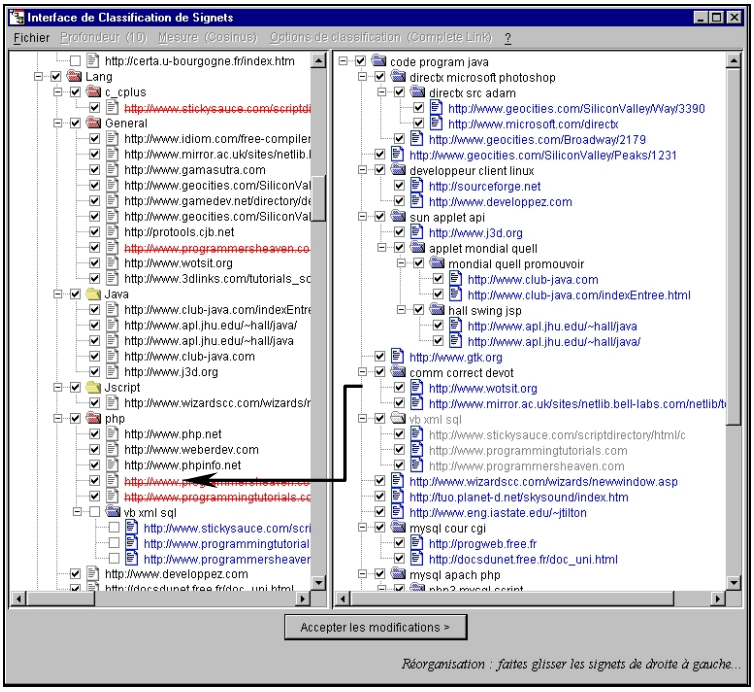


Figure 79 - Acceptation d'une partie de la réorganisation

II.5 Conclusion

L'utilisateur en quête d'informations a besoin d'effectuer une recherche efficace. Dans ce contexte, les outils de recherche d'information sont donc indispensables pour permettre à l'utilisateur d'identifier les informations pertinentes dans l'amas de documents disponibles sur le web.

Ce chapitre a porté sur notre contribution dans l'aide à la recherche d'information sur le web. Elle a conduit à la conception du système *Easy-DOR*. L'innovation de ce système repose

sur la dynamique de la RI. En effet, l'utilisateur dispose d'une aide à la recherche d'information durant tout le processus de recherche d'information au travers d'une application unique. Cette aide repose sur un aspect coopératif qui permet à un utilisateur de tirer partie des informations collectées par d'autres utilisateurs offrant ainsi une meilleur circulation des informations au sein du groupe.

Conclusion & Perspectives

Les travaux développés dans ce mémoire s'inscrivent dans le cadre de la conception d'un système d'aide à la recherche d'information sur le web. Nous avons proposé un système qui apporte à l'utilisateur une aide personnalisée à différents stades de la recherche d'information. Cette aide vise à lui permettre d'obtenir plus efficacement des informations pertinentes pour compléter régulièrement les connaissances relatives à ses centres d'intérêt mais également pour répondre à ses besoins ponctuels.

L'approche repose sur le principe que le processus de recherche d'information est plus vaste que la simple interrogation d'un outil de recherche. En effet, une recherche d'information débute par l'utilisation de ses connaissances pour formuler ses besoins en information, se poursuit par la recherche de documents pertinents et l'interprétation des résultats pour s'achever éventuellement par la mémorisation de certains d'entre eux.

Le système *Easy-DOR* tient compte de cette démarche. Il est basé sur l'utilisation et le partage des signets des utilisateurs ainsi que sur les documents qu'ils visitent. Il propose des solutions permettant d'aider l'utilisateur dans ses recherches :

- l'utilisateur construit au travers de sa hiérarchie de signets un véritable espace d'informations personnel. Cet espace d'informations permet au système de caractériser les centres d'intérêt de l'utilisateur afin de proposer de nouveaux documents pertinents pour chacun d'entre eux. Ce processus offrant ainsi la possibilité à l'utilisateur de mettre à jour voire de faire évoluer sa connaissance du domaine,
- durant la navigation, l'utilisateur suit un certain cheminement pour atteindre des documents répondant à ses besoins. Au cours de cette tâche, le système lui propose des documents pertinents à partir des informations collectées par les autres utilisateurs au travers de leurs signets,
- lors d'une recherche, le système propose à l'utilisateur une interface de visualisation des résultats issus de cette recherche. Cette visualisation permet d'appréhender de manière globale l'ensemble des résultats obtenus,
- lors de la réorganisation et du suivi des signets, le système indique les modifications apportées aux documents pertinents pour l'utilisateur. Par ailleurs, il lui propose une aide à la classification thématique des signets pour obtenir une arborescence plus facile à exploiter.

Easy-DOR repose sur des approches issues de systèmes existants (utilisation des signets, outils de recommandations...) mais aussi sur des aspects innovants tels que la recommandation durant la navigation et la visualisation des résultats de recherche proposées. *Easy-DOR* est décomposé en quatre modules, chacun d'entre eux dédié à une aide particulière :

- l'outil de recommandation pour la connaissance de l'utilisateur permet de faire évoluer la connaissance à long terme de ses centres d'intérêt grâce à de nouveaux documents pertinents. Ces centres d'intérêt sont déduits de sa hiérarchie de signets. Du fait de cette meilleure connaissance, l'utilisateur peut formuler de manière plus précise les requêtes et ainsi obtenir de meilleurs résultats,
- la recommandation durant la navigation repose sur une représentation des centres d'intérêts des utilisateurs sous la forme d'un multi-arbres afin d'obtenir une

représentation unique de tous les centres d'intérêt des utilisateurs. Ce multi-arbres est exploité pour identifier des documents pertinents pour la navigation d'un utilisateur,

- l'interface de visualisation des résultats de recherche d'information permet d'appréhender les résultats de recherche de façon globale au travers d'une représentation des documents dans un espace 3D en utilisant les couleurs. La combinaison de deux axes d'interprétation (organisation spatiale et couleurs) permet à l'utilisateur une meilleure appréhension de l'importance des différents critères de recherche au sein des documents,
- l'outil de gestion de la hiérarchie de signets permet à un usager de s'affranchir du travail fastidieux qu'est le suivi des documents mémorisés ainsi que de la réorganisation des signets. Cet outil offre à l'utilisateur la possibilité de connaître les modifications apportées aux documents mémorisés mais également une aide pour organiser ces documents.

Les différentes approches proposées ont été validées au travers des expérimentations menées. Ces expérimentations nous ont permis d'apprécier les performances mais aussi les limites des différents modules proposés :

- le module de recommandation pour la connaissance du domaine, au travers de l'aspect hiérarchique (parcours spécifique de l'arborescence, remontée des documents aux nœuds pères) permet d'obtenir de bons résultats. Cependant, ce module semble dépendre de l'organisation des nœuds,
- le module de recommandation pour la navigation permet de proposer des documents pertinents pour la navigation de l'utilisateur. Ces recommandations sont déduites de l'organisation des documents au sein des hiérarchies de signets des utilisateurs du système. La limite à ce système est qu'il ne permet pas de répondre à des besoins inexistants au sein du groupe. En ce sens, ce module représente un complément aux méthodes traditionnelles de recommandations lors de la navigation qui reposent sur l'interrogation d'un outil de recherche externe ou sur l'étude des documents de l'hypertexte local du document visité,
- le module de visualisation des résultats de recherche d'information proposé, combine deux axes d'interprétation qui permet d'améliorer la compréhension des éléments représentés. Toutefois, cette visualisation est plutôt destinée à des personnes expérimentées car son apprentissage (pour une interprétation précise) ne semble pas adapté à des personnes néophytes du fait de l'aspect 3 Dimensions, des couleurs notamment,
- le module de gestion des signets permet d'obtenir une hiérarchie de signets à jour mais également organisée de façon thématique.

L'intégration de tous ces modules au sein d'un même système permet non seulement d'éviter d'utiliser des outils indépendants mais aussi permet de centraliser une représentation unique des centres d'intérêt des utilisateurs. Ainsi, les différents modules proposés œuvrent intimement pour améliorer leur qualité. Par exemple, de l'organisation des signets va dépendre la qualité des recommandations lors de la navigation mais également la qualité de la représentation des centres d'intérêt de l'utilisateur.

L'aspect coopératif du système semble bien adapté au domaine de l'aide à la recherche d'information car il repose sur le fait que les utilisateurs peuvent avoir des centres d'intérêt communs. Il est donc important de pouvoir partager des informations jugées et validées. Il reste toutefois à expérimenter le système proposé en grandeur nature, au sein d'une entreprise : notre laboratoire de recherche par exemple.

Les travaux que nous avons menés nous ont permis de mettre en évidence certaines limites du système ce qui nous a conduit à envisager bon nombre de perspectives. La première limite vient du fait que la circulation de l'information au sein d'un groupe d'individus semble être cloisonnée dans notre système. En effet, les informations utilisées sont générées uniquement par les utilisateurs du système. Pour remédier à cela, nous envisageons d'intégrer des informations provenant de sources extérieures tels que des moteurs de recherche.

Par ailleurs, le système suppose que les hiérarchies de signets des utilisateurs sont construites selon des thématiques propres à chaque utilisateur. Afin de respecter l'organisation initiale de ces hiérarchies, il est nécessaire de proposer des classifications additionnelles de ces signets selon des critères différents du contenu des documents liés (par auteur, par date par exemple). L'organisation spécifique des signets de l'utilisateur doit également être prise en compte afin que le système caractérise les centres d'intérêt de l'utilisateur au mieux dans le module de recommandation pour le connaissance du domaine. En effet, à l'heure actuelle, le système suppose que l'utilisateur organise ses signets de façon thématique. Une approche moins stricte pourrait être mise en œuvre (détection du type de classification utilisée) afin de coller au mieux à l'organisation spécifique des signets.

De plus, les signets souffrent notamment d'une sous-information induite par quelques caractéristiques (URL, titre) qui ne permettent pas de se remémorer aisément le contenu du document ne facilitant pas leur réorganisation. Une approche intermédiaire entre les signets et les annotations permettant de mémoriser les documents pertinents pour l'utilisateur doit être envisagée. Cette approche pourrait reposer sur une extension des signets par l'intermédiaire de méta-données, par exemple, qui permettraient de qualifier l'intérêt que l'utilisateur porte au document mémorisé. Toutefois, l'organisation hiérarchique de ces informations doit être conservée pour permettre une recherche aisée parmi les documents mémorisés.

La phase de recommandation durant la navigation permet à l'utilisateur d'obtenir des documents pertinents en rapport avec les documents visités. Cependant, de la qualité de l'organisation des hiérarchies de signets utilisées dépend la qualité des recommandations. Une solution alternative est envisagée afin d'homogénéiser l'ensemble des hiérarchies de signets au travers d'un arbre conceptuel unique par exemple composé de l'ensemble des documents mémorisés par les utilisateurs.

L'approche proposée pour la visualisation des résultats de recherche se base sur une représentation permettant d'apprécier l'importance des termes de la requête au sein des différents documents retrouvés. Cette représentation repose sur deux axes d'interprétation favorisant cette interprétation qui sont les couleurs et l'espace en 3D. Cependant, à elle seule, cette interface ne permet pas de répondre aux différents niveaux d'expérience ainsi

qu'aux différents besoins des utilisateurs. Une combinaison de plusieurs visualisations mérite d'être étudiée. Par ailleurs, il serait intéressant d'appliquer cette interface à un méta-moteur de recherche plutôt qu'à un outil de recherche unique.

L'outil d'aide à la réorganisation des signets de l'utilisateur repose sur la profondeur maximale de la hiérarchie que l'utilisateur souhaite obtenir. Cependant, il serait également intéressant d'appliquer un seuil de coupe optimal qui permettrait de proposer à l'utilisateur une hiérarchie de signets plus homogène.

Enfin, nous envisageons de développer l'aspect coopératif au sein du système. Par exemple, la réorganisation des signets pourrait tenir compte de la façon dont les autres utilisateurs ont organisé les mêmes documents. En effet, si la plupart des utilisateurs organisent certains documents d'une manière identique, il serait intéressant que le système tienne compte de cela et propose cette même organisation de ces mêmes documents à un utilisateur tiers. La réorganisation pourrait ainsi reposer à la fois sur le contenu des documents mais également sur la manière dont ils sont insérés dans les hiérarchies de signets de tout ou partie des autres utilisateurs du système. D'autre part, l'aspect coopératif peut intervenir au niveau de la présentation des résultats de recherche en précisant pour chaque document le chemin des hiérarchies de signets dont il est issu. Cette démarche permettrait ainsi à l'utilisateur de mieux apprécier les différents thèmes auxquels les autres utilisateurs ont rattaché un document retrouvé.

BIBLIOGRAPHIE

- A -

(Abrams, 1998) ABRAMS D., BAECKER R., CHIGNELL M., « Information archiving with bookmarks: personal web space construction and organisation », International ACM Conference on CHI, Los Angeles, CA-USA, pp 41-48, April 18-23, 1998.

(Agosti, 1996) AGOSTI M., SMEATON A., « Information retrieval and hypertext », Kluwer Academic Publisher, ISBN 0-7923-9710-X, 1996.

(Agosti, 2000) AGOSTI M., MELUCCI M., « Information retrieval on the web », Lectures Notes on Information Retrieval (1980), Springer Verlag ed., ISBN 3-540-41933-0, pp 243-285, 2000.

(Andrieu, 1998) ANDRIEU O., « Trouver l'info sur internet », Eyrolles éd., ISBN 2212089929, 1998.

(Armstrong, 1995) ARMSTRONG R., FREITAG D., JOACHIMS T., « Webwatcher: machine learning and hypertext », Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford University, California, USA, March 27-29, 1995.

- B -

(Baeza-Yates, 1999) BAEZA-YATES R., RIBEIRO-NETO B., « Modern information retrieval », ACM Press, Addison Wesley ed., ISBN 0-201-39829-X, 1999.

(Balabanovic, 1997) BALABANOVIC M., SHOHAM Y., « Fab : content-based, coopérative recommandation », Communications of the ACM, 40(3) : 66-72, March, 1997.

(Barrett, 1997) BARRETT R., MAGLIO P.P., KELLEM D.C., « How to personalize the web », International ACM Conference on Human Factors and Computing Systems (CHI), Atlanta Georgia, pp 75-82, March 22-27, 1997.

(Belkin, 1992) BELKIN N., Croft B., « Information filtering and information retrieval: two sides of a same coin ? », Communications of the ACM, 35(12) :29-38, December, 1992.

(Benford, 1995) BENFORD S., SNOWDON D, GREENHALGH C., KNOX I., BROWN C., « VR-Vibe: a virtual environment for co-operative information retrieval », EuroGraphics, vol. 14(3), pp 349-360, 1995.

(Bergman, 2000) BERGMAN M.K., « The deep web: Surfacing the hidden value », BrightPlanet ed., July 2000. http://www.brightplanet.com

(Berners-Lee, 1994) BERNERS-LEE T., CAILLIAU R., NIELSEN H.F., SECRET A., « The World Wide Web », Communications of the ACM, vol. 37, n°8, août 1994.

(Boughanem, 1992) BOUGHANEM M., « Les systèmes de recherche d'informations d'un modèle classique à un modèle connexionniste », Thèse de Doctorat de l'Université Paul Sabatier spécialité Informatique, 1992.

(Boughanem, 2000) BOUGHANEM M., CHRISMENT C., SOULÉ-DUPUY C., TAMINE L., « Connectionnist and genetic approaches to achieve IR », Soft Computing in Information Retrieval Techniques and Applications, F. Crestani & G. Pasi ed., Springer Verlag, ISBN 3-7908-1299-4, pp 173-198, 2000.

(Bouthors, 1999) BOUTHORS V., DEDIEU O., « Pharos, a coopérative infrastructure for web knowledge sharing », Technical report num. 3679, ISSN 0249-6399, Institut de Recherche en Informatique et en Automatique (INRIA), Mai 1999.

(Bray, 1996) BRAY T., SPERBERG-MCQUEEN C.M., « Extensible markup language (XML) », W3C Working Draft, WD-xml-961114, 1996. http://www.w3c.org/TR/WD-xml-961114.html

(Budzik, 1999) BUDZIK J., HAMMOND K., « Watson : anticipating and contextualizing information needs », Annual Meeting of the American Society for Information Science (ASIS), Washington, Oct. 31- Nov. 4, 1999.

- C -

(Caglayan, 1998) CAGLAYAN A., HARRISON C., « Les agents », InterEditions ed., ISBN 2225831467, 1998.

(Carré, 1999) CARRE J., MACHONIN A., GLIZE P., « Un système multi-agent auto-organisateur pour l'apprentissage d'un profil utilisateur », Ingénierie des systèmes multi-agents, actes des 7ème journées francophones d'Intelligence Artificielle et Systèmes Multi-Agents (JFIADSMA'99), ISBN 2-7462-0077-5, pp 207-221, 1999.

(Chakrabarti, 1998) CHAKRABARTI S., DOM B., AGRAWAL R., RAGHAVAN P., « Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies », The VLDB Journal, Springer-Verlag ed., vol. 7, pp 163-178, 1998.

(Chalmers, 1992) CHALMERS M., CHITSON P., « Bead: explorations in information visualization », 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, pp 330-337, 1992.

(Chevalier, 2000) CHEVALIER M., VERLHAC M., « ISIDOR: a visualisation interface for advanced information retrieval », 2nd International Conference on Enterprise Information Systems (ICEIS), B. Sharp J. Cordeiro J. Filipe (eds.), ISBN 972-98050-1-6, Stafford (England), pp 414 – 418, July 4-7, 2000.

(Chevalier, 2001) CHEVALIER M., JULIEN C., « ISIDORView : une interface de visualisation des résultats de recherche d'informations », Revue Extraction des Connaissances et Apprentissage (ECA), Hermès (ed.) , ISBN 2-7462-0216-6, 1(1-2), pp 135-140, Janvier 2001.

(Chevalier, 2001b) CHEVALIER M., JULIEN C., CHRISMENT C., « Easy-DOR : un système de gestion des besoins web au sein d'un groupe d'utilisateurs », 19ème Congrès Inforamtique des organisations et Systèmes d'Information et de Décision (INFORSID), ISBN 2-906855-170, Université de Genève, Martigny, pp 217-236, 29 mai – 1er juin, 2001.

(Chevalier, 2001c) CHEVALIER M., JULIEN C., KHROUF K., SOULE-DUPUY C., « Vers une mémoire documentaire », 2ème conférence internationale sur la maîtrise des systèmes complexes et la relation homme-système (NimesTIC), EMA/Site EERIE, Parc scientifique G.BESSE, Nîmes, pp 121-126, 12-14 décembre, 2001.

(Chevalier, 2002) CHEVALIER M., JULIEN C., « Aide à la navigation sur le web », Revue Extraction des Connaissances et Apprentissage (ECA), Hermès (ed.) , ISBN 2-7462-0406-1, 1(4), pp 399-406, 2002.

(Chi, 2000) CHI Ed. H., « A taxonomy of visualization techniques using the data state reference model », INFOVIS'2000, Salt Lake City, pp 69-75, October, 2000.

(Cloutier, 1998) CLOUTIER L., ESPINASSE B., LEFRANÇOIS P., « CAT: a general coordination framework for multi-agent systems », Document de travail 1998-016, Centre de Service, d'Orientation et de Recherche sur la Compétitivité Internationale et l'Ingénierire de l'Entreprise Réseau (SORCIIER), ISBN 2-89524-047-7, 1998.

(Conklin, 1986) CONKLIN J., « A survey of HYPERTEXT », MCC Technical Report, Num STP-356-86, Rev. 2, December 1986.

(Cugini, 1996) CUGINI J.V., PIATKO C., LASKOWSKI S., « Interactive 3D visualization for document retrival », *Actes CIKM*, Rockville MD, 1996.

(Cugini, 2000) CUGINI J.V., LASKOWSKI S., SEBRECHTS M., « Design of 3-D visualisation of search results: evolution and evaluation », 12th International Symposium on Electronic Imaging: Visual Data Exploration & Analysis (SPIE 2000), San Jose, CA, pp 198-210, January 23-28, 2000.

(Cutting, 1993) CUTTING D.R., KARGER D.R., PEDERSON J.O., « Constant interaction-time scatter/gather browsing of very large document collections », 14th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburg, pp 121-131, 1993.

- D -

(D'allessio, 2000) D'ALLESSIO S., MURRAY S., SCHIAFFINO R., KERSHENBAUM A., « The effect of using hierarchical classifiers in text categorization », 6ème Conférence Recherche d'Information Assistée par Ordinateur (RIAO), 2000.

(Denoue, 2000) DENOUE L., « De la création à la capitalisation des annotations dans un espace personnel d'informations », Thèse (Informatique) de l'Université de Savoie, 26 octobre 2000.

(Dömel, 1994) DÖMEL P., « Webmap - a graphical hypertext navigation tool », 2nd International World Wide Web Conference, Chicago, September 1, 1994.

(Dreilinger, 1997) DREILINGER D., HOWE A.E., « Experiences with selecting search engines using meta-search », ACM Transactions on Information Systems, 15(3), pp 195-222, 1997.

(Dubin, 1995) DUBIN D., « Document analysis for visualization », 18th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 199-204, 1995.

(Dussaux, 2000) G. DUSSAUX, J.P. PECUCHET, « Création collective de bases de connaissances sur le web : Indexation par l'usage des documents », Colloque International sur le Document Electronique (CIDE), 4-6 juillet 2000, Lyon.

- E -

(Eick, 1994) EICK S.G., « Graphically display text », Journal of the Computational and Graphical Statistics, vol. 3(2), pp 127-142, 1994.

- F -

(Foley, 1995) FOLEY J., FEINER S., VAN DAM A., HUGHES J., « Computer graphics : principles & practice », ISBN 0201848406, Editions Addison-Wesley, 1995.

(Fresse, 2002) FRESSE A., « Quel avenir pour les communautés virtuelles ? », mensuel I-entreprise, n°15, pp 33-36, mars 2002.

(Furnas, 1994) FURNAS G.W., ZACKS J., « Multitrees: enriching and reusing hierarchical structure », ACM International Conference on Human Factors in Comuting Systems (CHI'94), Boston, Massachusetts, pp 330-336, 1994.

- G -

(Gery, 1999) GERY M., « Smartweb : recherche de zones de pertinence sur le world wide web », Actes du 17ème Congrès INFORSID, La Garde, 2-4 juin, pp 133-147, 1999.

(Gravano, 1999) GRAVANO L., GARCIA-MOLINA H., TOMASIC A., « GloSS: text-source discovery over the internet », ACM Transactions on Database Systems, vol. 24(2), pp 229-264, 1999.

(GVU, 1998) « 10th WWW User Survey », Graphic, visualisation & usability center (GVU), 1998. http://www.gvu.gatech.edu/user_surveys/survey-1998-10/

- H -

(Hascoët, 2000) HASCOËT M., « A user interface combining navigation aids », 11th International ACM Hypertext Conference, San Antonio, TX, pp 224-225, 2000.

(Hascoët, 2001) HASCOËT M., BEAUDOUIN-LAFON M., « Visualisation interactive d'information », Revue Information-Interaction-Intelligence (I3), « A Journal in Information Engineering Sciences », 1(1), 2001.

(Hearst, 1994) HEARST M.A., « Using categories to provide context for full-text retrieval results », Conférence Recherche d'Informations Assistée par Ordinateur (RIAO) : Intelligent Multimedia Information Retrieval Systems and Management, Rockefeller University, NY, October, 1994.

(Hearst, 1995) HEARST M.A., « TileBars: visualization of term distribution information in full text information access », ACM Conference on Human Factors in Computing Systems (SIGCHI), Denver CO, pp 59-66, May, 1995.

(Hearst, 1997) HEARST M.A., KARADI C., « Cat-a-cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy », 20th International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, pp 246-255, 1997.

(Hetzler, 1998) HETZLER B., HARRIS W. M., HAVRE S., WHITNEY P., « Visualizing the full spectrum of document relationships », 5th International Conference of the International Society for Knowledge Organization (ISKO), pp 168-175, Lille, August 25-29, 1998.

(Hoashi, 1999) HOASHI K., MATSUMOTO K., INOUE I., HASHIMOTO K., « Experiments on the TREC-8 filtering task », TREC-8, 1999.

(Hölscher, 2000) HÖLSCHER C., STRUBE G., « Web search bahavior of internet experts and newbies », 9th International Conference on the World Wide Web (www9), Amsterdam, May 15-19 2000.

- J -

(Jaczynski, 1997) JACZYNSKI M., TROUSSE B., « Broadway: a world wide web browsing advisor reusing past navigations from a group of users », Proceedings of the 3rd UK Case-Based Reasoning Workshop (UKCBR'97), Manchester, UK, September 9th, 1997.

(Jain, 1994) JAIN V., SHNEIDERMAN B., « Data structures for dynamic queries: an analytical and experimental evaluation », ACM International Workshop on Advanced Visual Interfaces (AVI), Bari, Italy, pp 1-11, June 1-4, 1994.

(Jansen, 2000) JANSEN B.J., SPINK A., SARACEVIC T., « Real life, real users, and real needs: a study and analysis of user queries on the web », Information Processing and Management, 36, pp 207-227, 2000.

(Jéribi, 2001) JERIBI L., RUMPLER B., PINON J.M., « Système d'aide à la recherche et à l'interrogation, fondé sur la réutilisation d'expériences », 19ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID), ISBN 2-906855-170, Université de Genève, Martigny, pp 443-463, 29 mai - 1er juin, 2001.

(Johansson, 2000) JOHANSSON R., « Information filtering using threshold adjustement », Computing Science Master's thesis, ISSN 1100-1836, Uppsala University, Sweden, February 2nd, 2000.

(Julien, 1988) JULIEN C., « Bases d'informations généralisées : contribution à l'étude des mécanismes de consultation d'objets multimédia », Thèse de Doctorat en Informatique de l'Université Paul Sabatier, Toulouse III, octobre, 1988.

- K -

(Kahan, 2001) KAHAN J., KOIVUNEN M-R., PRUD'HOMMEAUX E., SWICK R.R., « Annotea: an open RDF infrastructure for shared web annotations », 10th International World Wide Web Conference (WWW10), Hong Kong, May 2001.

(Kahle, 1996) KAHLE B., « Archiving the internet », submitted to Scientific American, March issue 1997, April 11th 1996. http://www.archive.org/sciam_article.html

(Kamba, 1995) KAMBA T., BHARAT K., ALBERS M.C., « The Krakatoa chronicle - An interactive, personalized, newspaper on the web », 4th International World Wide Web Conference Journal, O'Reilly & Associates, pp 159 - 170, November 1995.

(Keim, 1995) KEIM D.A., KRIEGEL H.P., « Possibilities and limits in visualizing large database », Visual Database Systems (VDB), pp 203-214, 1995.

(Klas, 2000) KLAS C-P., FÜHR N., « A new effective approach for categorizing web document », 22th Colloquium on Information Retrieval (BCS-IRSG), Cambridge, England, April 5-7, 2000.

(Kohonen, 1982) KOHONEN T., « Self-organised formation of topologically correct feature maps », Biological Cybernetics, vol. 43, pp 59-69, 1982.

(Koller, 1997) KOLLER D., SAHAMI M, « Hierarchically classifying documents using very few words », 14th International Conference on Machine Learning (ICML), pp 170-178, 1997.

(Korfhage, 1997) KORFHAGE R.R., « Information storage and retrieval », Wiley Computer Publishing, ISBN 0-471-14-338-3, 1997.

- L -

(Lagus, 1996) LAGUS K., KASKI S., HONKELA T., KOHONEN T., « Browsing digital libraries with the aid of self-organizing maps », 5th International World Wide Web Conference (WWW5), May 6-10, Paris, France, vol. Poster Proceedings, pp 71-79, 1996.

(Lainé-Cruzel, 1999) LAINE-CRUZEL S., « Profildoc – Filtrer une information exploitable », Bulletin des Bibliothèques de France (BBF), 44(5), pp. 60-64, juin, 1999.

(Lawrence, 1999) LAWRENCE S., GILES L., « Accessibility of information on the web », Revue Nature, vol. 400, pp 107-109, 1999.

(Lesteven, 1996) LESTEVEN S., POINÇOT P., MURTHAG F., « Neural networks and information extraction In astronomical information retrieval », Strategies and Techniques of Information for Astonomy, F.MURTHAG & A. HECK éd., Vitas in Astron, 1996.

(Levialdi, 1994) LEVIALDI S., BADRE A.N., CHALMERS M., COPELAND P., MUSSIO P., SALOMON C., « The interface of the future », ACM International Conference on Advanced Visual Interface (AVI), pp 200-205, June, 1994.

(Lewis, 1991) LEWIS D.D., « Evaluating text categorization », Proc. of Speech and Natural Language Workshop, Morgan-Kaufmann pub., pp 312-318, 1991.

(Li, 1997) LI Y., RAFSKI L., "Beyond relevance ranking: hyperlink vector voting", 5th International Conference on Computer Assisted Information Retrieval, RIAO'97, Montréal (Canada), pp 648-651, 25-27 juin 1997.

(Lieberman, 1995) LIEBERMAN H., « Letizia: an agent that assists web browsing », proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95), Montreal, August, 1995.

(Lohse, 1994) LOHSE G.L., BIOLSI K., WALKER N., RUETER H.H., « Classification of visual representations », Communication of the ACM, 12, pp 36-49, December, 1994.

(Luhn, 1958) LUHN H.P., « The automatic creation of literature abstracts », IBM Journal of Research and Development, 2(2), pp 159-165, 1958.

- M -

(Maarek, 1996) MAAREK Y.S., BEN SHAUL I.Z., « Automatically organizing bookmarks per contents », 5th International World Wide Web Conference (www5), May 6-10, Paris, France, 1996.

(Maglio, 2000) MAGLIO P.P., BARRETT R., CAMPBELL C.S., SELKER T., « SUITOR : an attentive information system », International ACM Conference on Intelligent User Interfaces (IUI), New Orleans, January 9-12, pp 169-176, 2000.

(Malone, 1987) MALONE T.W., GRANT K.R., TURBAK A., BROBST S.A., COHEN M.D., « Intelligent information sharing systems », Communication of the ACM (CACM), 30(5), pp 390-402, 1987.

(Mechkour, 1998) MECHKOUR M., HARPER D.J., MURESAN G., « The webcluster project: using clustering for mediating access to the world wide web », Proceedings of the 21st International ACM SIGIR Conference on Research and development in Information Retrieval, Melbourne, Australia, pp 357-358, August 24-28, 1998.

(Mignot, 2000) MIGNOT J-C., « Web caches: state-of-the-art techniques and prototypes », Institut National de Recherche en Informatioque et en Automatique (INRIA), ENS Lyon, rapport de recherche, n°3854, ISSN 0249-6399, Janvier, 2000.

(Mladenic, 1998) MLADENIC D., GROBELNIK M., « Feature selection for classification based on text hierarchy », Working notes of learning from text and the web, Conference on Automatd Learning and Discovery (CONALD), Carnegie Mellon University, June 11-13, 1998.

(Moore, 1995) MOORE A., REDMOND-PYLE D., « Graphical user interface design and evaluation », Prentice Hall ed., ISBN 013315193X, 1995.

(Mothe, 2000) MOTHE J., RAVAT F., RIAHI F., Zurfluh G., « Structuration and enrichment of HTML documents in order to build a specific information warehouse », Proc. European Conference on Information System (ECIS), Vienne, pp 386-395, Jully 3rd-5th, 2000.

(Mothe, 2002) MOTHE J., CHRISMENT C., ALAUX J., « Visualisation globale de collections de documents sous forme d'hypercube », Revue Extraction des Connaissances et Apprentissage (ECA), vol. 4/2001, ed. Hermes Science, ISBN 2746204061, pp 131-142, 2002.

(Murray, 2000) MURRAY B.H., MOORE A., « Sizing the Internet: A white paper », Cyveillance ed., July 10th 2000. http://www.cyveillance.com

- N -

(Notess, 2000) NOTESS G.R., « Search engine statistics: database overlap », SearchEngineShowdown, 2000. http://www.searchengineshowdown.com/stats/overlap.shtml

(Notess, 2002) NOTESS G.R., « Search engine statistics: database total size estimates », SearchEngineShowdown, 2002.
http://www.searchengineshowdown.com/stats/sizeest.shtml

- O -

(Olsen, 1993) OLSEN K.A., KORFHAGE R.R., SOCHATS K.M., SPRING M.B., WILLIAMS J.G., « Visualization of a document collection: the VIBE system », Information Processing and Management Journal (IPM), 29(1), pp 69-81, 1993.

(O'Neill, 1998) O'NEILL E., LAVOIE B., MCCLAIN P., « Characterizing the web and web-accessible information », Report of the web characterization workshop, W3C web characterization group conference, Cambridge, Massachussets, November 5, 1998.
http://www.w3.org/1998/11/05/WC-workshop/Papers/oneill.htm

- P -

(Pazzani, 1996) PAZZANI M., MURAMATSU J., BILLSUS D., « Syskill & Webert: Identifying interesting web sites », National Conference on Artificial Intelligence, Portland, pp 54-61, 1996.

(Pejtersen, 1998) PEJTERSEN A.M., FIDEL R., « A framework for centered evaluation and design: a case study of information retrieval on the web », Working Paper for MIRA Workshop, Grenoble, France, March, 1998.

(Pemberton, 2000) PEMBERTON D., RODDEN T., PROCTER R., « GroupMark: a www recommander system combining coopérative and information filtering », 6th ERCIM Workshop « User Interfaces for All », Florence, Italy, October 25-26, 2000.

(Peters, 2000) PETERS C., « Introduction », workshop of the Cross-Language Evaluation Forum (CLEF), Lecture Notes in Computer Science (2069), Springer Verlag publisher, Lisbon, Portugal, pp 1-6, September 21-22, 2000.

(Poinçot, 1999) POINÇOT P., « Classification et recherche d'information bibliographique par l'utilisation des cartes auto-organisatrices, applications en astronomie », Thèse de Doctorat de l'Université Louis Pasteur, soutenue le 15 décembre 1999.

(Popescul, 2000) POPESCUL A., UNGAR L.H., « Automatic labeling of document clusters », non publié, 2000.
http://www.cis.upenn.edu/~popescul/Publications/labeling_KDD00.pdf

(Porter, 1980) PORTER M.F., « An algorithm for suffix stripping », Program, Vol. 1(3), pp130-137, 1980.

(Procter, 1999) THE SELECT PROJECT TEAM, « SELECT: social and coopérative filtering of web documents and news », 5th International European Research Consortium For Informatics and Mathmatics (ERCIM) Workshop on User Interface For All: User-Tailored Information Environments, Dagstuhl, Germany, pp 23-37, November 28-December 1st, 1999.

- R -

(Rijsbergen, 1971) RIJSBERGEN C.J., JARDINE N., « The use of hierarchical clustering in information retrieval », Information Storage and Retrieval, 7, pp 217-240, 1971.

(Rocchio, 1971) ROCCHIO J., « Relevance feedback in information retrieval », Prentice Hall Inc., 1971.

(Rücker, 1997) RÜCKER J., POLANCO M.J., « Siteseer: personalized navigation for the web », Communications of the ACM, 40(3), pp.73-75, March, 1997.

(Ruiz, 2001) RUIZ M E, « Combining machine learning and hierarchical structures for text categorization », Thèse de l'Université de l'Iowa, décembre 2001.

- S -

(Salton, 1983) SALTON G., MACGILL M.J., « Introduction to modern information retrieval », McGraw Hill International Book Company, ISBN 0-07-Y66526-5, 1983.

(Salton, 1990) SALTON G., BUCKLEY C., « Improving retrieval performance by relevance feedback », Journal of the American Society for Information Science, Vol. 41, n°4, pp 288-297, 1990.

(Sebastiani, 1999) SEBASTIANI F., « A Tutorial on Automated Text Categorisation », Proceedings of {ASAI}-99, 1st Argentinian Symposium on Artificial Intelligence, pp 7-35, 1999.

(Sebastiani, 2002) SEBASTIANI, F., « Machine learning in automated text categorization », ACM Computing Survey, 34(1), pp 1-47, 2002.

(Sebrechts, 1999) SEBRECHTS M., CUGINI J.V, VASILAKIS J., MILLER M.S., LASKOWSKI S.J., « Visualization of search results: a comparative evaluation of text, 2D and 3D interfaces », ACM SIGIR, Berkley, pp 3-10, 1999.

(Seltzer, 1997) SELTZER R., « Altavista, Understanding the limits of accuracy », Internet Search Advantage, Cobb group publishing, ZD Journal, Novembre 1997.

(Shivakumar, 1998) SHIVAKUMAR N., GARCIA-MOLINA H., « Finding near-replicas of documents on the web », International workshop on the web and databases (WebDB), (GVU), Valencia, Spain, March 27-28,1998.

(Shneiderman, 1998) SHNEIDERMAN B., « Designing the user interface », Addison-Wesley Editeur, 3ème édition, ISBN 0-201-69497-2, 1998.

(Silverstein, 1998) SILVERSTEIN C., HENZINGER M., MARAIS H., MORICZ M., « Analysis of a very large web search engine query log », SRC technical note #1998-014, October 26, 1998.

http://gatekeeper.dec.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html

(Singhal, 1997) SINGHAL A. K., « Term weighting revisited », Ph. D. of Cornell University, 1997.

(Smith, 1978) SMITH A.R., « Color gamut transform pairs », Computer Graphics, (12), pp 12-19, 1978.

(Soulé-Dupuy, 1990) SOULE-DUPUY C., « Systèmes de recherche d'informations : mécanismes d'indexation et d'interrogation », Thèse de doctorat de l'Université Paul Sabatier, n°612, Toulouse III, février 1990.

(Soulé-Dupuy, 2001) SOULE-DUPUY C., « Bases d'informations textuelles : des modèles aux applications », Habilitation à Diriger des Recherches, Spécialités Informatique, Université Paul Sabatier, Toulouse III, 2001.

(Spark Jones, 1972) SPARK-JONES K., « A statistical interpretation of term specificity and its application in retrieval », Journal of Documentation, Vol. 28(1), pp 11-20, 1972.

(Spink, 2002) SPINK A., JANSEN B.J., WOLFRAM D., SARACEVIC T., « From e-sex to e-commerce: web search changes », revue IEEE Computer, vol. 35 (3), pp 107-109, March, 2002.

(Spoerri, 1993) SPOERRI A., « InfoCrystal: a visual tool for information retrieval and management », International Conference on Information Knowledge and Management (CIKM), Washington, USA, ISBN 0897916263, pp 11-20, November 1-5, 1993.

(Sullivan, 2001) SULLIVAN D., « Search engines sizes », The search engine report, 18 déc. 2001. http://searchenginewatch.com/reports/sizes.html

- T -

(Tamine, 2000) TAMINE L., « Optimisation de requêtes dans un système de recherche d'information », Thèse de l'Université Paul Sabatier spécialité Infomatique, soutenue le 12 décembre 2000.

(Tmar, 2002) TMAR M., « Calibrage du seuil par linéarisation des scores par intervalles dans un système de filtrage adaptatif », XXème Congrès INFORSID, ISBN 2906855189, pp 55-71, Nantes, 4-7 juin, 2002.

- V -

(Vernier, 1997) VERNIER F., NIGAY L., « Représentation multiples d'une grande quantité d'information », 9ème journées Interaction Homme-Machine (IHM), pp 183-190, Futuroscope Poitiers, France, 10-12 Septembre, 1997.

(Voohrees, 1999) VOOHREES E.M., HARMAN D., « Overview of TREC 8 », eighth Text REtrieval Conference (TREC-8), 1999.

(Voohrees, 2001) VOOHREES E.M., HARMAN D., « Overview of TREC 2001 », tenth Text REtrieval Conference (TREC-2001), Gaithersburg, Maryland, November 13-16, 2001.

- W -

(Ware, 1985) WARE C., BEATTY J.C., « Using colour as a tool in discrete data analysis », CS-85-21, Computer Graphics Laboratory, University of Waterloo, August, 1985.

(Wittenburg, 1995) WITTENBURG K, DAS D., HILL W, STEAD L, « Group asynchronous browsing on the world wide web », 4th World Wide Web Conference (WWW4), Boston, pp 51-62, December 11-14, 1995.

(Wiss, 1998) WISS U., CARR D., « A cognitive classification framework for 3-dimensonnal information visualisation », Technical Report, ISSN 1402-1536 / ISRN LTU-TR-98/04-SE / NR 1998:04, Université de Lulea, 1998.

(Wu, 2001) WU L., HUANG X., GUO J., XIA Y., FENG Z., « FDU at TREC-10, filtering, QA, web and video tasks », TREC-10, 2001.

(Wood, 1995) WOOD A., DREW N., BEALE R., HENDLEY B., « Hyperspace: web browsing with visualisation », 3rd International World Wide Web Conference (WWW3), Darmstadt, Germany, April 10-14, 1995.

- Y -

(Yang, 1997) YANG Y., PEDERSEN J.O., « A comparative study on feature selection in text categorization », 14th International Conference on Machine Learning (ICML), pp 412-420, 1997.

(Yang, 1999) YANG Y., « An evaluation of statistical approaches to text categorization », Journal of Information Retrieval, 1(1/2), pp 67-88, 1999.

- Z -

(Zamir, 1998) ZAMIR, O., « Visualisation of search results in document retrieval systems », General Examination, University of Washington, 1998.

(Zamir, 1999) ZAMIR O., ETZIONI O., « Grouper: a dynamic clustering interface to web search results », Computer Networks, vol. 31(11-16), pp 1361-1374, May, 1999.

(Zipf, 1949) ZIPF G.K., « Human behavior and principles of feast effort », Addison Wesley ed., 1949.

Index des publications citées

Cloutier, 1998 55
Conklin, 1986 11
Cugini, 1996 ..... 45, 110
Cugini, 2000 ..... 41, 50
Cutting, 1993 47

D

D'alessio, 2000 82
Denoue, 2000 ..... 58, 105
Dömel, 1994 35
Dreilinger, 1997 40
Dubin, 1995 41
Dussaux, 2000 ..... 61, 105

E

Eick, 1994 44

F

Foley, 1995 110
Fresse, 2002 60
Furnas, 1994 95

G

Gery, 1999 30
Gravano, 1999 39
GVU, 1998 ..... 29, 30, 57, 74, 114

H

Hascoët, 2000 34
Hascoët, 2001 41
Hearst, 1994 ..... 44, 109
Hearst, 1995 ..... 43, 50, 110
Hearst, 1997 37
Hetzler, 1998 118
Hoashi, 2000 81
Höschler, 2000 29

J

Jaczynski, 1997 ..... 53, 96, 97
Jain, 1994 37
Jansen, 2000 ..... 36, 108
Jeribi, 2001 61
Johansson, 2000 80
Julien, 1988 15

K

Kahan, 2001 58
Kahle, 1996 13
Kamba, 1995 33
Keim, 1995 50
Klas, 2000 79
Kohonen, 1982 47
Koller, 1997 ..... 81, 82
Korfhage, 1997 26

L

Lagus, 1996 48
Lainé-Cruzel, 1999 40
Lawrence, 1999 32
Lesteven, 1996 48
Levialdi, 1994 51
Lewis, 1991 85
LI, 1997 30
Lieberman, 1995 53
Lohse, 1994 107
Luhn, 1958 18

M

Maarek, 1996 ..... 57, 120, 121
Maglio, 2000 55
Malone, 1987 27
Mechkour, 1998 36
Mignot, 2000 125
Mladenic, 1998 82
Moore, 1995 49
Mothe, 2000 124
Mothe, 2002 45
Murray, 2000 12

N

Notess, 2000 32
Notess, 2002 32

O

O'Neill, 1998 12
Olsen, 1993 44

P

Pazzani, 1996 54

Pejtersen, 1998 ..... 29, 71, 75
Pemberton, 2000 60
Peters, 2000 28
Poinçot, 1999 48
Popescul, 2000 82
Porter, 1980 19
Procter, 1999 ..... 60, 61, 73

R

Rijsbergen, 1971 120
Rocchio, 1971 38
Rücker, 1997 ..... 55, 60, 73, 76, 93, 119
Ruiz, 2001 ..... 81, 82

S

Salton, 1983 ..... 19, 21, 24, 38
Salton, 1990 20
Sebastiani, 1999 78
Sebastiani, 2002 78
Sebrechts, 1999 ..... 49, 50
Seltzer, 1997 32
Shivakumar, 1998 13
Shneiderman, 1998 ..... 29, 49, 71, 118
Silverstein, 1998 ..... 36, 40, 108
Singhal, 1997 20
Smith, 1978 110
Soulé-Dupuy, 1990 19
Soulé-Dupuy, 2001 21
Sparck Jones, 1972 20
Spink, 2002 ..... 36, 40, 108
Spoerri, 1993 44
Sullivan, 2001 32

T

Tamine, 2000 39
Tmar, 2002 81

V

Vernier, 1997 ..... 51, 106, 118
Voorhees, 1999 37
Voorhees, 2001 ..... 28, 83

W

Ware, 1985 ..... 110, 117
Wiss, 1998 106
Wittenburg, 1995 ..... 95, 96
Wood, 1995 36
Wu, 2001 81

Y

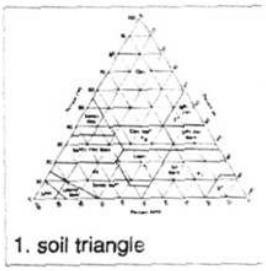
Yang, 1997 80
Yang, 1999 78

Z

Zamir, 1998 41
Zamir, 1999 47
Zipf, 1949 ..... 18, 19

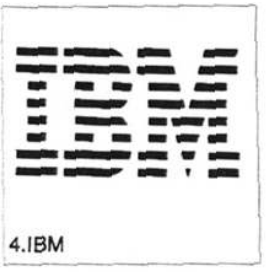
Annexes

Annexe A
Planches de la classification de Lohse


1. soil triangle

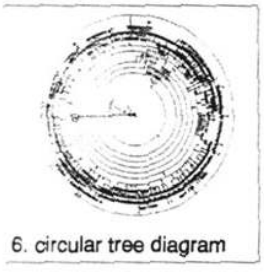

2. missile crisis


3. USA tornadoes



4.IBM

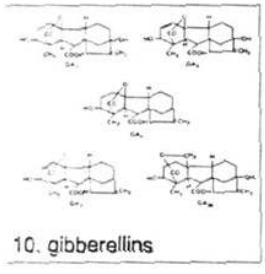

5. microscope


6. circular tree diagram


7. auto repair records

8. pie chart


9. Crater Lake


10. gibberellins


11. heart

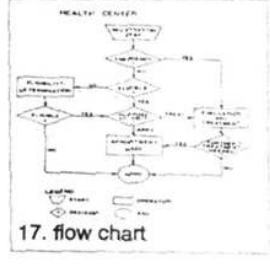

12. microbiology

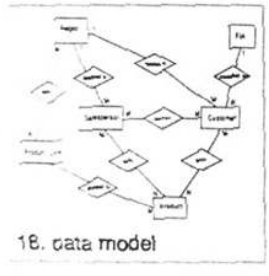

13. spreadsheet budget

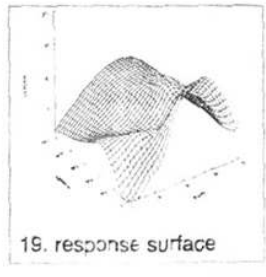

14. Gantt chart


15. Pittsburgh map

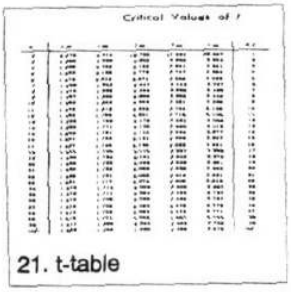

16. list of integrals


17. flow chart


18. data model


19. response surface

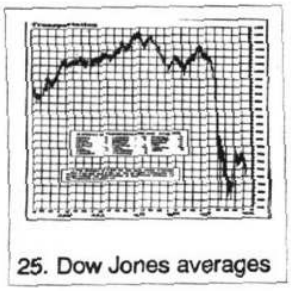

20. wheelbarrow

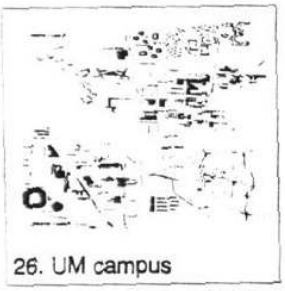

21. t-table

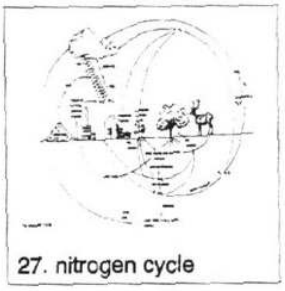

22. USA murder rates

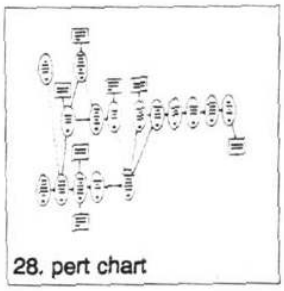

23. dollar bar chart

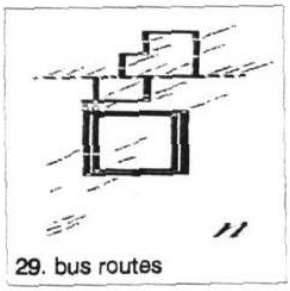

24. yen vs. dollar


25. Dow Jones averages

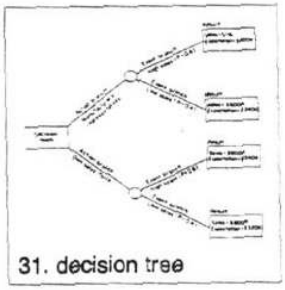

26. UM campus

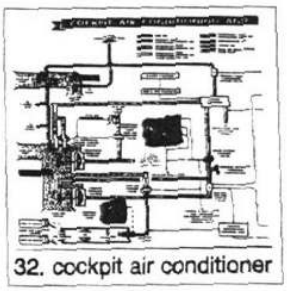

27. nitrogen cycle


28. pert chart

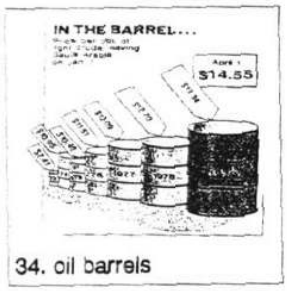

29. bus routes

30. highway signs

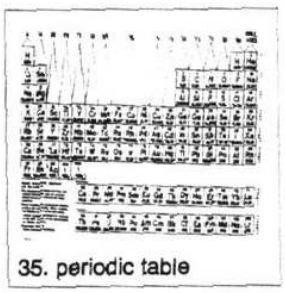

31. decision tree


32. cockpit air conditioner


33. Tale of two cities



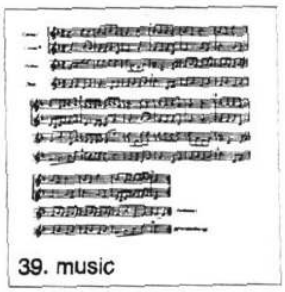
34. oil barrels

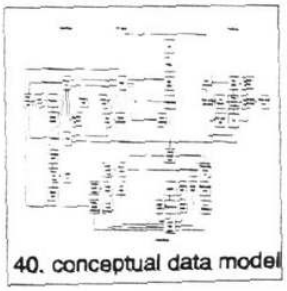

35. periodic table


36. organizational chart


37. floorplan

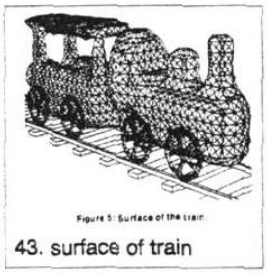

38. dot chart


39. music


40. conceptual data model


41. nervous system

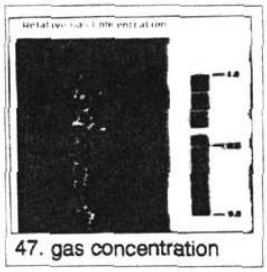

42. software features

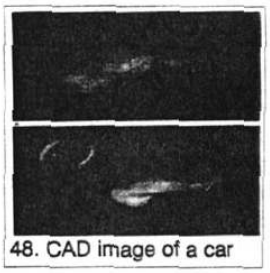

43. surface of train

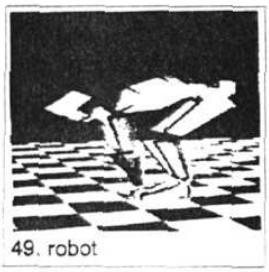

44. city buildings


45. MRI brain images


46. world erosion

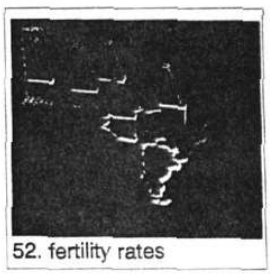

47. gas concentration

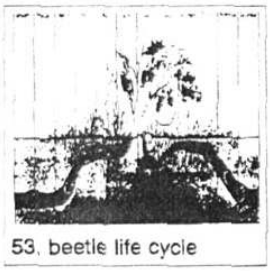

48. CAD image of a car

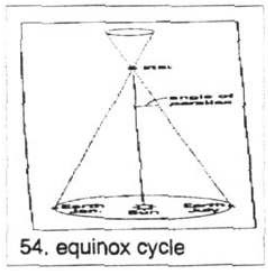

49. robot


50. map of Hong Kong

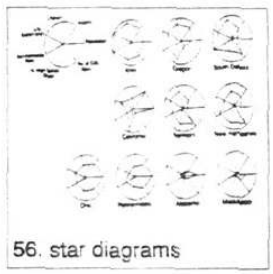

51. chess board


52. fertility rates

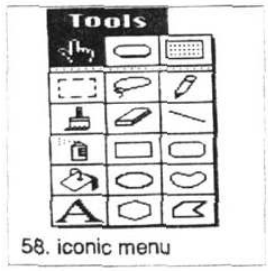

53. beetle life cycle


54. equinox cycle

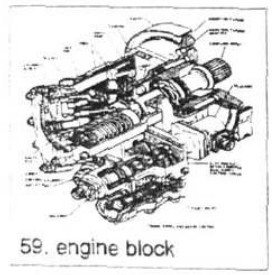

55. Olympics schedule

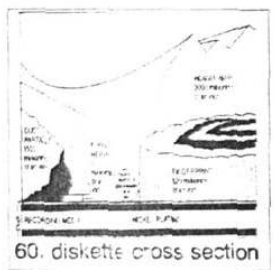

56. star diagrams


57. DNA double helix



58. iconic menu


59. engine block


60. diskette cross section

Annexe B
*Algorithme RVB → HSV et HSV**

Passage d'un point RVB à un point dans l'espace HSV (cône).

Glossaire :

Paramètres en entrée :

Rouge	Valeur réelle [0..1] correspondant au poids du critère 1
Vert	Valeur réelle [0..1] correspondant au poids du critère 2
Bleu	Valeur réelle [0..1] correspondant au poids du critère 3

Paramètres en sortie :

Hue	Valeur réelle [0..360°] correspondant à la Teinte (Hue)
Saturation	Valeur réelle [0..1] correspondant à la Saturation
Value	Valeur réelle [0..1] correspondant à l'Intensité

Début

```
MAXIMUM ← MAX (Rouge, Vert, Bleu)
MINIMUM ← MIN (Rouge, Vert, Bleu)
// On affecte la valeur maximum des 3 composantes Rouge, Vert, Bleu à Value
Value ← MAXIMUM                 // Intensité de la couleur
// Calcul de la saturation. Elle est égale à 0 si Rouge, Vert, Bleu sont nulles.
Si (MAXIMUM = 0) alors
        Saturation ← 0
Sinon
        Saturation ← (MAXIMUM - MINIMUM) / MAXIMUM
FinSi
Si (Saturation = 0) alors
        Hue = INDEFINIE
Sinon   // Cas chromatique, la Saturation est différente de 0
        // Delta valeur réelle locale au Si
        Delta ← MAXIMUM – MINIMUM
        Si (Rouge = MAXIMUM) alors
                Hue ← ((Vert - Bleu) / Delta)*60°
                // Couleur résultante entre le jaune et le magenta
        Sinon   Si (Vert = MAXIMUM) alors
                Hue ← 120° + ((Bleu – Rouge) / Delta)*60°
                // Couleur résultante entre cyan et jaune
                Sinon   Si (Bleu = MAXIMUM) alors
                                Hue ← 240° + ((Rouge – Vert) / Delta)*60°
                                // Couleur résultante entre magenta et cyan
                        FinSi
                FinSi
        FinSi
        Si (Hue < 0) alors
                Hue ← Hue + 360         // Vérifie que Hue toujours positif
        FinSi
FinSi
```

Fin.

Passage d'un point RVB à un point dans l'espace HSV* modifié (cylindre).

Pour transformer ce cône en cylindre nous ne modifions pas la Saturation en fonction de la valeur MAXIMUM. La valeur de la Saturation correspond donc à (MAXIMUM - MINIMUM).

Annexe C
Interprétation précise de la visualisation

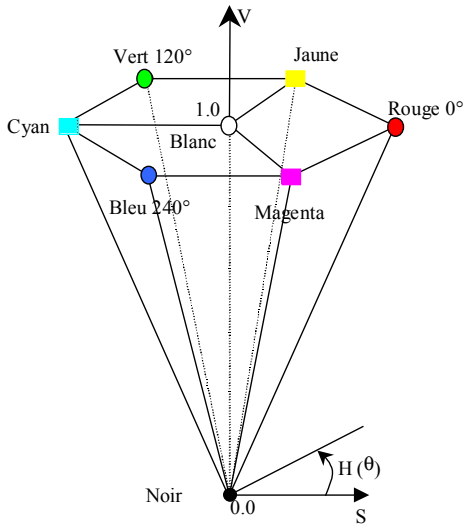


Figure 80 - Cône HSV

Glossaire :

Critère : Nous appelons " critère " les axes sur la périphérie du cône. Ainsi les critères correspondent aux axes de couleur noir-rouge, noir-vert, noir-bleu, noir-cyan, noir-magenta, noir-jaune, noir-blanc.

Interprétation de l'éloignement par rapport à l'axe central (axe Y) :

L'importance de chaque critère les uns par rapport aux autres est donné par la position du point par rapport à l'axe central (axe V). La figure 2 présente une tranche du cône vue de dessus.

Les points sont soumis à l'attraction des critères :

RAPPELS SUR LES COULEURS :

Critère N°1 Rouge : Si le point se trouve sur l'axe noir-rouge, le document ne possède que ce critère.

Critère N°2 Vert : Si le point se trouve sur l'axe noir-vert, le document ne possède que ce critère.

Critère N°3 Bleu : Si le point se trouve sur l'axe noir-bleu, le document ne possède que ce critère.

Critère Jaune : Si le point se trouve sur l'axe noir-jaune, le document possède uniquement les critères 1 (rouge) & 2 (vert). *Ces critères sont d'égale importance.*

Critère Cyan : Si le point se trouve sur l'axe noir-cyan, le document possède uniquement les critères 3 (bleu) & 2 (vert). *Ces critères sont d'égale importance.*

Critère Magenta : Si le point se trouve sur l'axe noir-magenta, le document possède uniquement les critères 3 (bleu) & 1 (rouge). *Ces critères sont d'égale importance.*

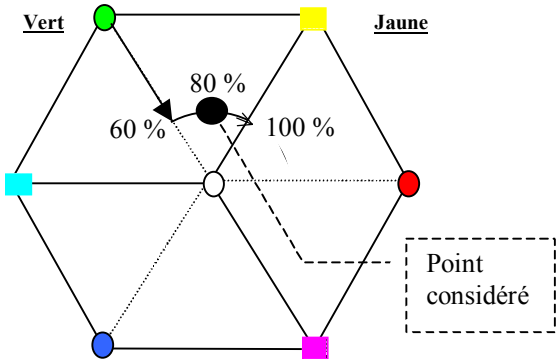
Critère Blanc : Si le point se trouve sur l'axe noir-blanc (axe V), le document possède les 3 critères (rouge, vert, bleu). *Ces critères sont d'égale importance.*

Critère Noir : Si le point se trouve sur le point noir, le document ne possède aucun des critères.

Pour interpréter les couleurs et la position spatiale il est nécessaire de comprendre comment est construit le modèle. Ne pas oublier que les points sont proches des critères dominants.

Pour interpréter la position d'un point il est nécessaire :

- d'identifier le ou les critères dominants : Si le point se situe au centre du cône les 3 critères sont dominants. Pour le point n°4 de la figure 2, le vert est la couleur dominante de par sa proximité du point vert. **L'importance du/des critères dominant(s) sera 100%**
- d'évaluer la valeur des autres critères si le point n'est pas au centre du cône c'est à dire si les trois critères ne sont pas dominants. Deux cas se présentent :
    - Si le point se situe sur l'axe Blanc-Couleurs dominantes : les deux critères restant sont d'une importance égale. L'importance relative de ces critères par rapport au critère dominant correspond à la distance entre le point des couleurs dominantes et le point blanc. Le 0% se situant au point des couleurs dominantes et le 100% au niveau u point blanc.
    - Si le point ne se situe pas sur l'axe Blanc-Couleurs dominantes : on projette le point sur l'axe Blanc-Couleurs dominantes. La distance entre le point des couleurs dominantes et le point considéré indique l'importance du critère minimum (Le 0% se situant au point des couleurs dominantes et le 100% au niveau u point blanc). *Dans la figure ci-dessous, le critère dominant est le vert. Le point se situe entre le vert et le rouge, donc l'ordre relatif des critères est le suivant : Vert > Rouge > Bleu. Le critère minimum, ici le bleu, à pour importance, la distance entre le point vert et le point blanc (60%).*



Pour le critère intermédiaire : on étudie pour cela la distance du point projeté sur l'axe Blanc-Couleurs dominantes et la projection du point considéré sur ***l'axe médian*** entre la couleur dominante et la couleur intermédiaire (cf tableau ci-dessous). L'importance du critère intermédiaire aura une valeur entre la valeur du critère minimum et 100%. *Dans la figure ci-dessus, la valeur du critère minimum est 60%. Le critère rouge aura donc une valeur comprise entre 60% et 100%). Pour la déterminer, on observe la position du point considéré entre l'axe Blanc-couleurs dominantes (Vert-Blanc) et l'axe médian(Jaune-Blanc). On observe que le point se situe à égale distance et donc la valeur du critère Rouge sera de 80%.*

Couleur dominante	Couleur intermédiaire	*Axe médian*
Rouge	Vert	**Jaune**
Vert	Rouge	**Jaune**
Vert	Bleu	**Cyan**
Bleu	Vert	**Cyan**
Rouge	Bleu	**Magenta**
Bleu	Rouge	**Magenta**

Exemples d'interprétation(Figure 81***) :***

1 - Le point se situe entre le critère rouge et le critère blanc.

Explications : Le critère Rouge a pour importance 100% car il est le critère dominant. Les car il ne contient que du critère Rouge. Les critères Bleu et Vert sont d'égale importance car ils se situent sur l'axe Rouge-Blanc. L'importance relative de ces critères est 50% car le point se situe à la moitié de l'axe Rouge-Blanc.

Les importances des critères seront donc : R = 100%, V = 50%, B = 50%

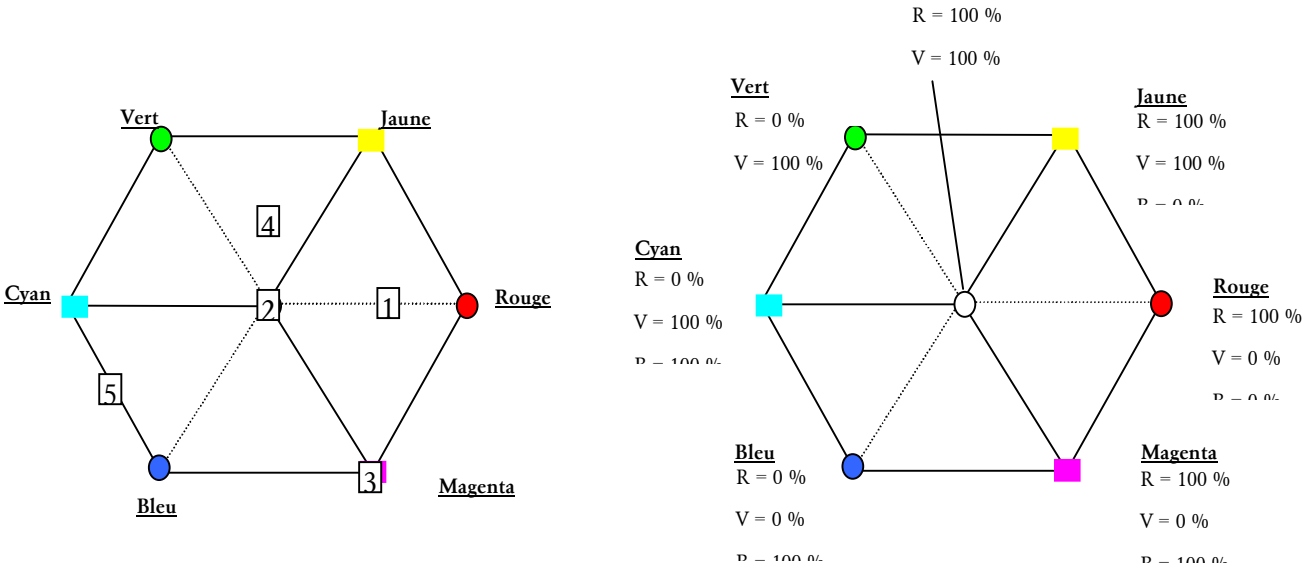


Figure 81 - Exemples d'interprétation (à droite)
Importance relative des critères (Rouge, Vert, Bleu) (à gauche)

2 - Le point se situe sur le point blanc.

Explications : Se situant sur l'axe central (Blanc), le point se situe à égale distance entre les 3 Critères et donc les 3 critères sont dominants.

Les importances des critères seront donc : R = 100%, V = 100%, B = 100%

3 - Le point se situe sur le point magenta.

Explications : Les Critères Rouge et Bleu sont dominant car le point est placé sur l'axe Blanc-Magenta. Le troisième critère Vert est nul car le distance entre le point 3 et le point Magenta est nul.

Les importances des critères seront donc : R = 100%, V = 0%, B = 100%

4 - Le point se situe au centre de la zone définie par le critère blanc, jaune et vert.

Explications : Le critère dominant est le vert car la distance entre le point et le point Vert est minimum. Par ordre d'importance le Rouge est le critère intermédiaire, et le Bleu le critère minimum. Le bleu a pour importance 50% car le point 4 projeté sur l'axe Vert-Blanc se situe à la moitié de cet axe. Le critère intermédiaire aura donc une importance comprise entre 50% et 100%. Après projection du point 4 sur l'axe Blanc-Jaune on s'aperçoit que le point se situe à au milieu des axes Blanc-Jaune et Blanc-Vert (la distance entre le point et ces 2 axes sont égales). L'importance du critère Rouge est donc la moitié de [50% ; 100%] = 75%.

Les importances des critères seront donc : R = 75%, V = 100%, B = 50%

5 - Le point se situe au centre de l'axe Bleu-Cyan..

Explications : Le point se situe sur l'axe Bleu-Cyan. Le critère dominant est le Bleu, le critère intermédiaire est le Vert et le critère minimum est le Rouge. Si l'on projette le point sur l'axe Bleu-Blanc on s'aperçoit qu'il coïncide avec le point Bleu. Le critère minimum est donc nul. Le critère vert a donc un importance comprise entre 0% et 100%. Se situant à la moitié de l'axe Bleu-Cyan, l'importance du critère intermédiaire est donc de 50%.

Les importances des critères seront donc : R = 0%, V = 50%, B = 100%

Interprétation de l'éloignement par rapport au sommet du cône :

Jusque là nous avons seulement pu identifier les importances relatives des critères les uns par rapport aux autres mais nous ne pouvons dire l'importance réelle de ces critères. L'éloignement par rapport au sommet du cône (couleurs sombres) va nous permettre de le dire. L'éloignement par rapport au sommet du cône (ou de la couleur noir) indique l'importance du/des critère(s) dominant(s). L'importance est comprise entre 0.0 (critère noir) et 1.0 (base du cône).

Pour interpréter l'importance des différents critères il suffit d'interpréter l'éloignement par rapport à l'axe central pour connaître l'importance relative de chaque critère, puis de déduire l'importance du/des critère(s) dominant(s) par de l'éloignement du point du critère noir.

Exemples d'interprétation (Figure 82***) :***

Pour connaître la valeur réelle des critères il suffit de constater la valeur correspondante au(x) critère(s) dominant(s) puis appliquer cette valeur à la représentation réalisée au paragraphe précédent.

1 - Le point se situant sur et à mi-hauteur de l'axe central V, les critères dominants ont pour valeur 0.5. Comme le point se situe sur l'axe central Blanc il possède les 3 critères de manière homogène et l'importance de ces critères est 0.5. Explication Rouge = Vert = Bleu = 100%. Or la valeur des critères dominant est 0.5 et donc Rouge = Bleu = Vert = 0.5 * 100% = **0.5**.

2 - Le point se situe sur la base du cône (importance des critères dominants à 1.0). Il correspond au point 1 de la figure 2 pour lequel R = 100 % ; V = 50 % ; B = 50 %. La valeur

réelle des critères est donc : Rouge = 1.0 x 100 % = **1.0** ; Vert = 1.0 x 50 % = **0.5** ; Bleu = 1.0 x 50 % = **0.5**.

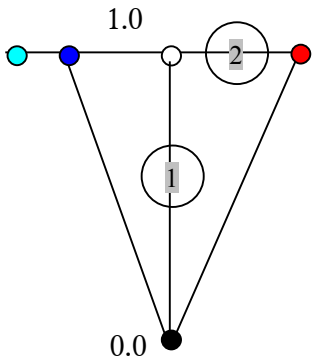


Figure 82 - Exemples d'interprétation (vue de profil du cône)

Particularités de la visualisation en Cylindre

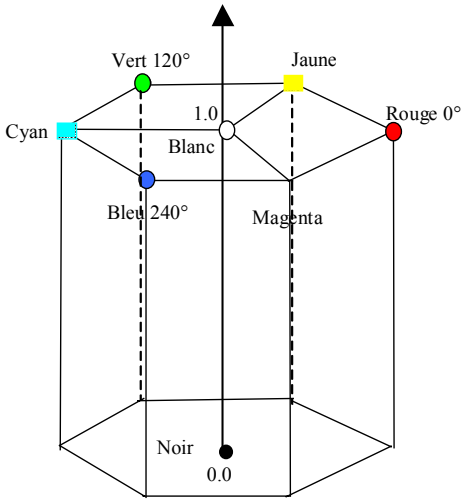


Figure 83 - Modèle en cylindre

Ce modèle est directement inspiré du modèle en cône (HSV). La différence est que la distance entre le point et l'axe central reste constant contrairement à la visualisation en cône. En effet, dans le cône, plus on se rapproche des couleurs sombres et du noir, plus la distance entre le point et l'axe central du cône est faible.

Tout ce qui a été présenté pour le cône reste valable pour la visualisation en cylindre

Annexe D
Déroulement de la phase d'évaluation de l'interface

Partie 1

Cette première partie consiste en un questionnaire ouvert concernant l'interface que nous présentons à l'utilisateur. Cette partie permet de mettre en évidence si elle respecte les critères cognitifs comme l'intuition par exemple. Nous voulions savoir, au travers du questionnaire, si les utilisateurs comprenaient à quoi sert l'interface, quelles en sont les fonctionnalités et comment peut-on les exploiter.

Pour cela, nous avons laissé l'utilisateur utiliser l'interface comme il le souhaitait sans aucune information. Cette manipulation est limitée dans le temps (10 minutes) pour laisser à chacun le soin de s'imprégner librement de l'interface et pour permettre une équité entre les participants. C'est pour cette dernière raison que nous n'avons pas proposé cette partie aux participants distants car aucun contrôle ne pouvait être fait. A la suite de cette manipulation chacun des participants doit remplir un questionnaire visant à mettre en évidence s'ils visualisent les fonctionnalités offertes et s'ils apprécient le résultat de chacune d'entre elles.

Partie 2

La partie 2 est commune aux différentes évaluations (locales ou distantes). Elle vise à mettre en évidence l'efficacité de chacun des axes de représentations, c'est à dire les couleurs et l'espace 3D.

Pour cela, cette partie est divisée en 5 phases. Pour l'ensemble de ces phases, 20 documents ayant 3 critères avec une importance aléatoire sont générés. A chacune des phases successives, les mêmes documents étaient présentés dans un ordre aléatoire pour éviter l'accoutumance de l'utilisateur aux échantillons. Ces échantillons ainsi que les réponses des utilisateurs sont représentés dans un triplet de valeurs correspondant à l'importance de chacun des critères compris entre 0 et 1.

L'interface utilisée dans cette partie est une version bridée, c'est à dire que seules les fonctions de rotation dans l'espace sont disponibles pour que l'utilisateur n'ait à se concentrer que sur l'interprétation des échantillons présentés.

Le déroulement de chacune des phases est relativement simple. Pour chaque phase, les documents sont présentés un à un à l'utilisateur pour que celui-ci indique l'importance de chacun des critères dans le document. A des fins d'analyse, le temps de réponse ainsi que les valeurs saisies par l'utilisateur sont pris en compte pour l'ensemble des 5 phases pour permettre une comparaison des résultats.

La première phase consiste à évaluer la connaissance que possède l'utilisateur concernant la synthèse additive des couleurs. Sur un écran noir, un carré dont la couleur correspond à la couleur d'un document est présenté. Pour chacun d'entre eux, l'utilisateur doit indiquer l'importance de chacun des critères associés à une couleur de base (rouge, vert et bleu).

A la suite de cette phase, chacun des participants avaient accès au modèle d'interprétation (annexe C) qui peut être consulté pendant un laps de temps indicatif de 15mn. Pour les participants distants, les documents sont fournis au format électronique (PDF).

La deuxième phase consiste à utiliser la visualisation en cône mais dans laquelle les documents sont représentés par un carré blanc dans l'espace pour apprécier uniquement l'espace 3D.

La troisième phase repose sur une visualisation similaire sauf que le carré représentant le document possède la couleur adéquat du document. Cette phase nous permet d'évaluer l'apport des deux axes d'interprétation combinés.

Les deux dernières phases reprend les phases 2 et 3 mais avec la visualisation en cylindre.

Partie 3

Cette dernière partie, réservée aux évaluations locales du fait de l'utilisation d'une base de données locale, consiste à utiliser l'interface complète (sauf les outils de sélection fines qui ont fait suite cette évaluation) pour répondre à un questionnaire de satisfaction (cf. annexe F) ainsi qu'à un cas pratique.

Le cas pratique consiste à repérer les documents les plus pertinents pour deux combinaisons spécifiques des critères (ex : les documents qui correspondent à deux ou trois termes de la requête).

Annexe E
Questionnaire individuel

Numéro ID. : **Age :** **Sexe :** ☐ Masculin ☐ Féminin.

0. Avez-vous déjà utilisé un outil informatique (micro-ordinateur) ? Oui / Non / NSP *(Ne Sait Pas)*

1. Avec combien de Systèmes d'exploitation avez-vous déjà travaillé ? (encerclez votre choix)

0	3-4
1	5-6
2	+ de 6

2. Si vous en connaissez, pourriez-vous les citer et indiquer pour chacun d'entre eux votre niveau (cochez la case correspondante, D – Débutant, I – Intermédiaire, A – Avancé) ?

1	_____	D	I	A	4	_____	D	I	A
2	_____	D	I	A	5	_____	D	I	A
3	_____	D	I	A	6	_____	D	I	A

3. Cochez, dans la liste suivante, les périphériques, logiciels et systèmes que vous connaissez ou dont vous avez entendu parlé.

☐ Terminal
☐ Moniteur couleur
☐ Lecteur CD-ROM
☐ Track-ball
☐ Tablette Graphique
☐ Scanner
☐ Tableur
☐ Outil de reconnaissance vocale
☐ Outil de CAO
☐ Ordinateur Personnel (micro-ordinateur)
☐ Ecran Tactile
☐ Clavier
☐ Souris
☐ Traitement de texte
☐ Système de Gestion de Base de Données
☐ Outils de montage vidéo
☐ Atelier de Génie Logiciel
☐ Ordinateur portable
☐ Lecteur de Disquettes
☐ Joystick (manette de jeu)
☐ Modem
☐ Jeux vidéo
☐ Internet
☐ World Wide Web
☐ Message Electronique E-Mail.

4. Cochez, dans la liste suivante, les périphériques, logiciels et systèmes, les cases de ceux que vous avez personnellement utilisé.

☐ Terminal
☐ Moniteur couleur
☐ Lecteur CD-ROM
☐ Track-ball
☐ Tablette Graphique
☐ Scanner
☐ Tableur
☐ Outil de reconnaissance vocale
☐ Outil de CAO
☐ Ordinateur Personnel (micro-ordinateur)
☐ Ecran Tactile
☐ Clavier
☐ Souris
☐ Traitement de texte
☐ Système de Gestion de Base de Données
☐ Outils de montage vidéo
☐ Atelier de Génie Logiciel
☐ Ordinateur portable
☐ Lecteur de Disquettes
☐ Joystick (manette de jeu)
☐ Modem
☐ Jeux vidéo
☐ Internet
☐ World Wide Web

- Message Electronique E-Mail.

5. Savez-vous ce qu'est un Système de Recherche d'Information (SRI) ?

❑ Oui ❑ Non

6. Si vous en utilisez le World Wide Web, connaissez-vous des outils de recherche d'informations sur le web (moteurs de recherche, annuaire, méta-moteur...) ?

❑ Oui ❑ Non

Si OUI, pourriez-vous citer les outils que vous utilisez le plus souvent :

_____
_____
_____
_____

7. ***Si vous avez répondu OUI à la question 5 et/ou 6*** : A quelle fréquence, en moyenne, utilisez-vous ces systèmes ? (indiquez un nombre moyen pour ***UNE SEULE*** des rubriques suivantes)

_____ fois par jour
ou _____ fois par semaine
ou _____ fois par mois
ou _____ fois par trimestre

Diriez-vous que vous êtes : *Pleinement satisfait*, *satisfait*, *moyennement satisfait*, *peu satisfait*, *pas du tout satisfait* des résultats de ces recherches d'informations ? (***rayez les mentions inutiles***)

Annexe F
Questionnaire pour l'évaluation qualitative de l'interface

1. Vos réactions générales vis-à-vis du Système

Terrible — Très Bien

1 2 3 4 5 6 7 8 9 N/A

2.Frustration — Satisfaction

1 2 3 4 5 6 7 8 9 N/A

3.Ennuyant — Stimulant

1 2 3 4 5 6 7 8 9 N/A

4.Puissance Inadéquate — Puissance Adéquate

1 2 3 4 5 6 7 8 9 N/A

5. Rigide — Flexible

1 2 3 4 5 6 7 8 9 N/A

L'écran

6. Les caractères à l'écran

Difficiles à voir — Faciles à voir

1 2 3 4 5 6 7 8 9 N/A

7. Les polices de caractères

Illisibles — Lisibles

1 2 3 4 5 6 7 8 9 N/A

8 Mise en valeur d'une partie de l'écran

Ne sert à rien — Aide précieuse

1 2 3 4 5 6 7 8 9 N/A

9. Organisation Spatiale est une aide

Jamais — Toujours

1 2 3 4 5 6 7 8 9 N/A

10. Quantité d'informations affichées

Inadéquate — Adéquate

1 2 3 4 5 6 7 8 9 N/A

11. Arrangement des infos à l'écran

Illogique — Logique

1 2 3 4 5 6 7 8 9 N/A

12. Les actions et opérations possibles

Imprévisibles — Prévisibles

1 2 3 4 5 6 7 8 9 N/A

13. Icônes et outils graphiques

Incompréhensibles — Compréhensibles

1 2 3 4 5 6 7 8 9 N/A

14 .Couleurs

Non naturelles — Naturelles

1 2 3 4 5 6 7 8 9 N/A

15. Choix des couleurs

Incompréhensible — Compréhensible

1 2 3 4 5 6 7 8 9 N/A

16. Les couleurs aident à l'interprétation

Faux — Vrai

1 2 3 4 5 6 7 8 9 N/A

Veuillez noter vos commentaires sur l'écran ici :

Apprentissage

17. Apprendre à utiliser le Système semble

Difficile — Facile

1 2 3 4 5 6 7 8 9 N/A

18. La prise en main

Difficile — Facile

1 2 3 4 5 6 7 8 9 N/A

19. Apprentissage des commandes

Difficile — Facile

1 2 3 4 5 6 7 8 9 N/A

20. Temps d'Apprentissage

Long — Court

1 2 3 4 5 6 7 8 9 N/A

21. Exploration des fonctionnalités par tâtonnement

Décourageant — Encourageant

1 2 3 4 5 6 7 8 9 N/A

22. Exploration des outils semble

Risquée — Sûre

1 2 3 4 5 6 7 8 9 N/A

23. Souvenir des noms & utilisation des commandes

Difficile — Facile

1 2 3 4 5 6 7 8 9 N/A

24. Les tâches peuvent être accomplies de manière linéaire

Jamais — Toujours

1 2 3 4 5 6 7 8 9 N/A

25. Nombre d'étapes par tâches

Trop élevé — Juste ce qu'il faut

1 2 3 4 5 6 7 8 9 N/A

26. Les étapes se déroulent selon un ordre logique

Jamais — Toujours

1 2 3 4 5 6 7 8 9 N/A

27. Compréhension de l'achèvement d'une tâche

Pas clair — Clair

1 2 3 4 5 6 7 8 9 N/A

Veuillez noter vos commentaires sur l'apprentissage ici :

Le Système

28. Rapidité, Vitesse

Trop lent — Vitesse adéquate

1 2 3 4 5 6 7 8 9 N/A

29. Réponse pour la majorité des opérations

Trop lent — Vitesse adéquate

1 2 3 4 5 6 7 8 9 N/A

30. Fiabilité du Système

Peu fiable — Très fiable

1 2 3 4 5 6 7 8 9 N/A

31. Erreurs dans le fonctionnement apparaissent

Souvent — Presque jamais

1 2 3 4 5 6 7 8 9 N/A

32. Le système prévient d'éventuels problèmes

Jamais — Tout le temps

1 2 3 4 5 6 7 8 9 N/A

Veuillez noter vos commentaires sur le système ici :

Utilisez l'interface et répondre aux questions suivantes :

* Donnez deux noms, s'il y en a, de documents qui traitent le plus de « information retrieval » (S'il n'y en a pas inscrire « pas de réponse ») :

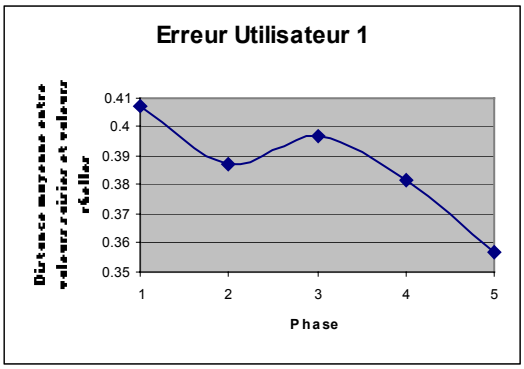
* Donnez deux noms, s'il y en a, de documents qui traitent le plus de « information retrieval system » (S'il n'y en a pas inscrire « pas de réponse ») :

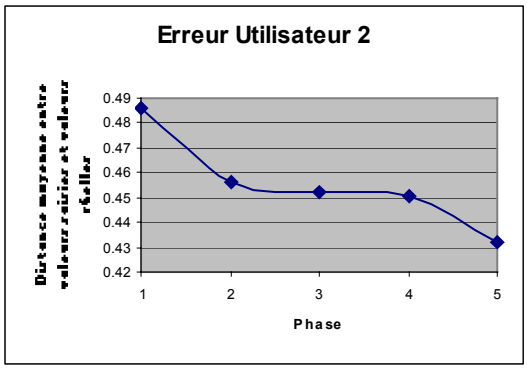
Annexe G
Résultats individuels de l'évaluation de la visualisation

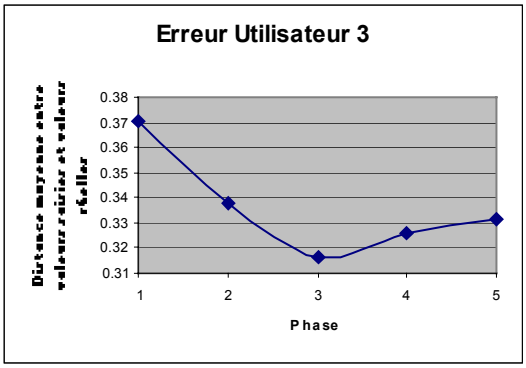
Légende

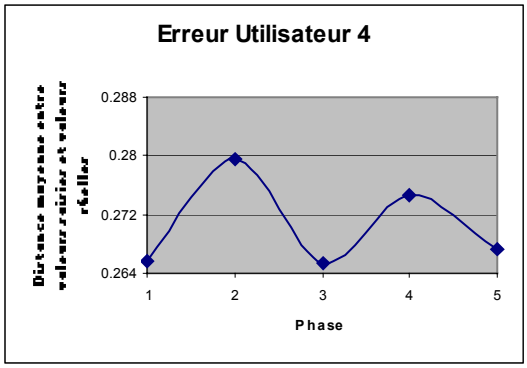
Phase 1 : Couleur seule
Phase 2 : Cône sans couleur
Phase 3 : Cône avec couleurs
Phase 4 : Cylindre sans couleur
Phase 5 : Cylindre avec couleurs

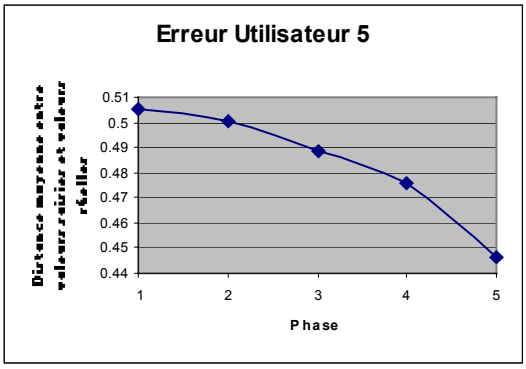
Erreur commise par les participants.

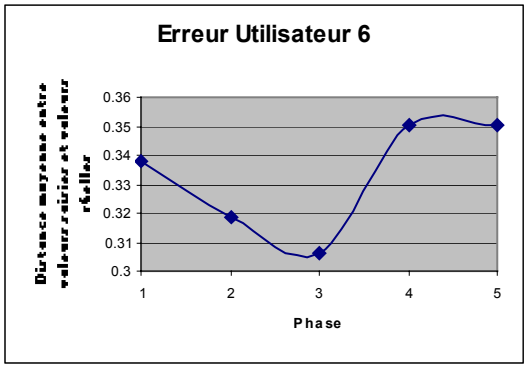


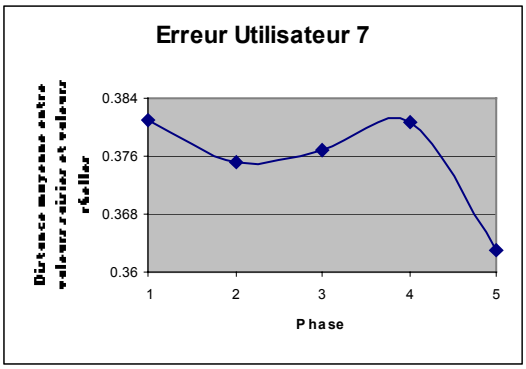


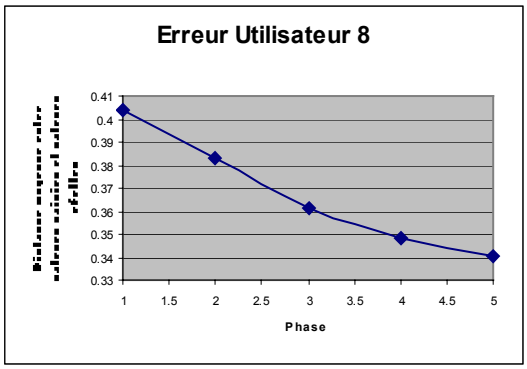


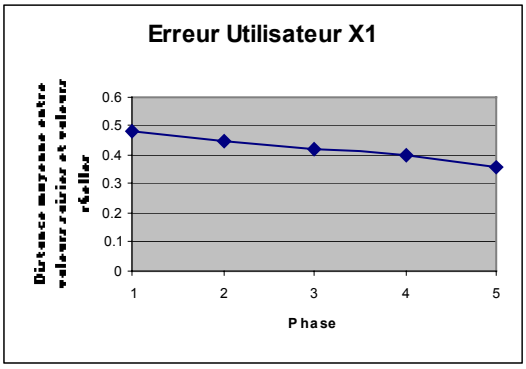


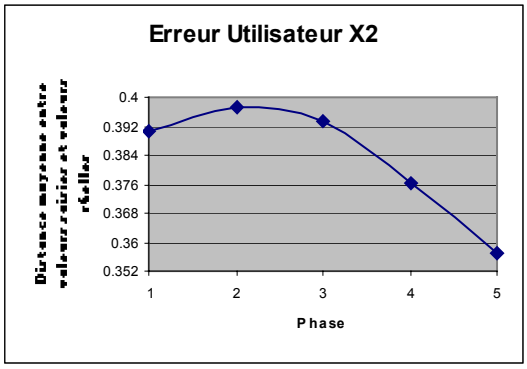


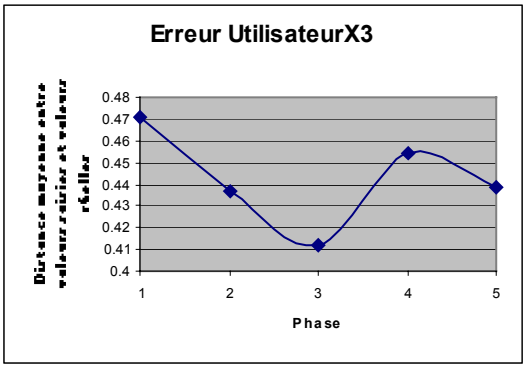


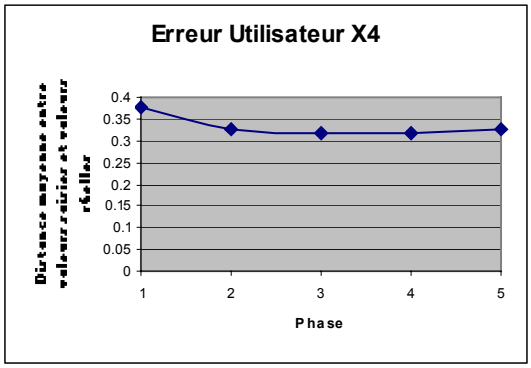




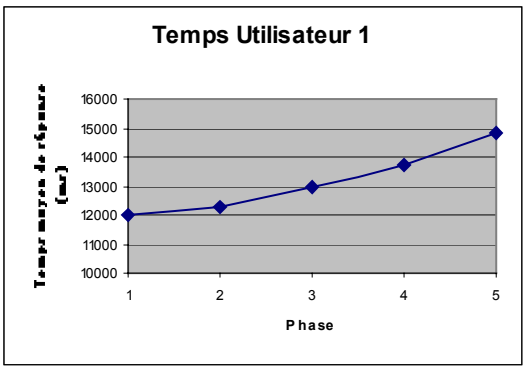


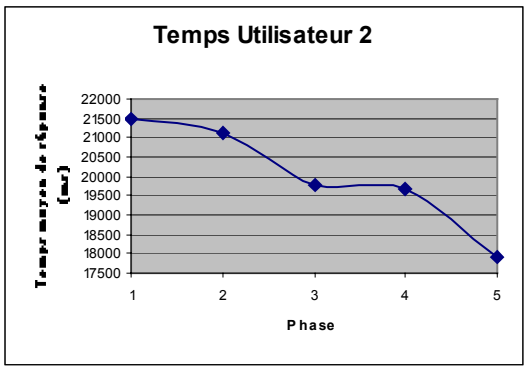


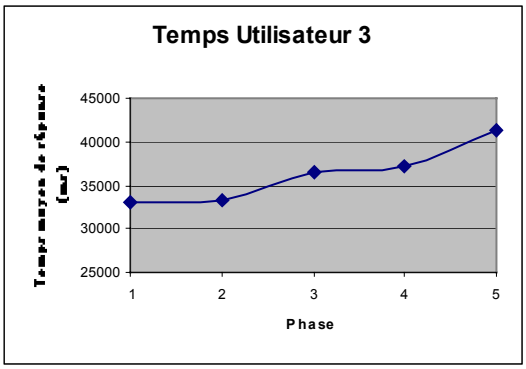


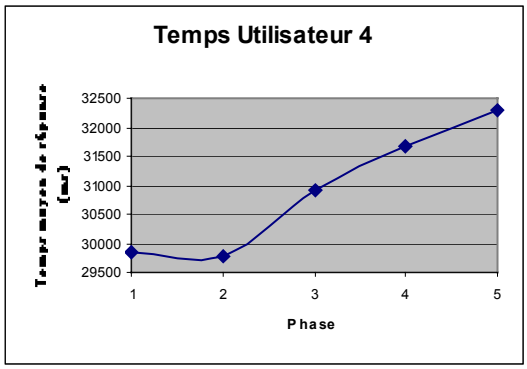


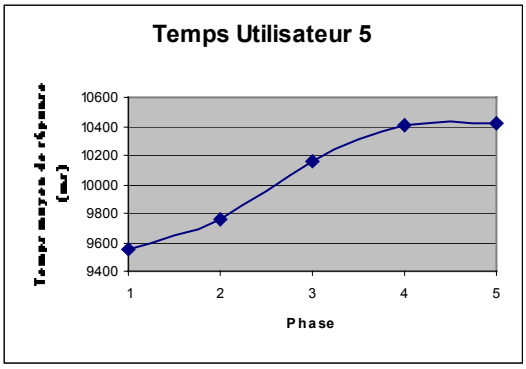
Temps moyen de réponse des participants.

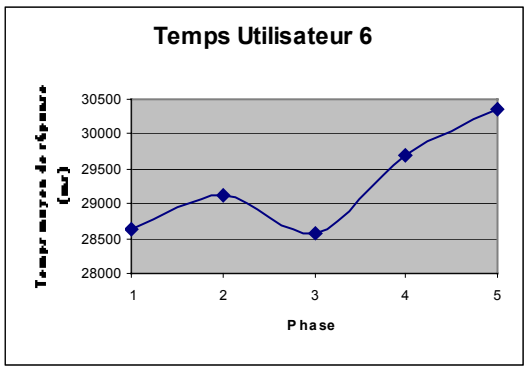


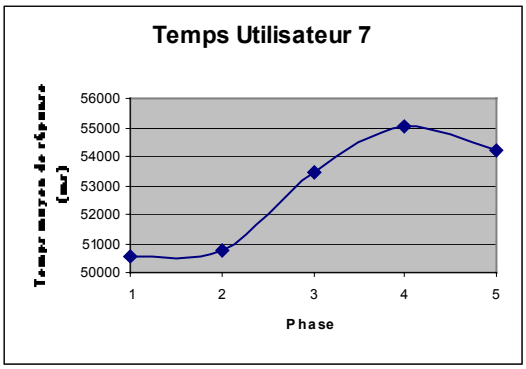


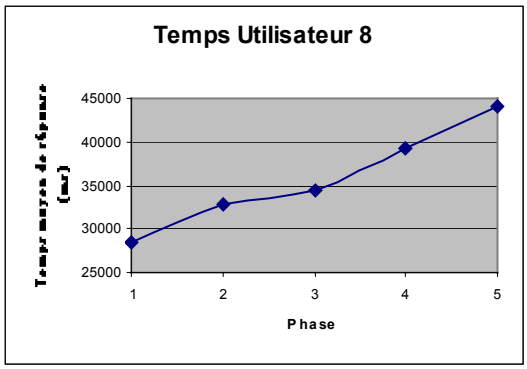

Temps Utilisateur 3
Temps moyen de réponse (ms)
45000
40000
35000
30000
25000
1
2
3
4
5
Phase

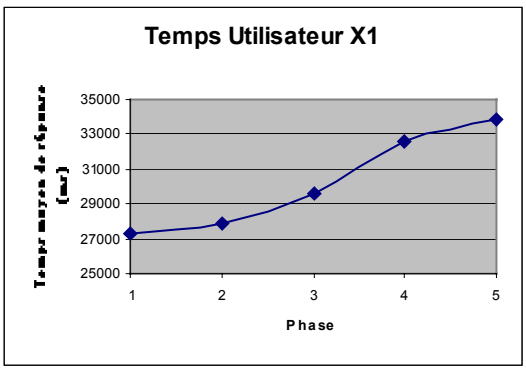

Temps Utilisateur 4
Temps moyen de réponse (ms)
32500
32000
31500
31000
30500
30000
29500
1
2
3
4
5
Phase

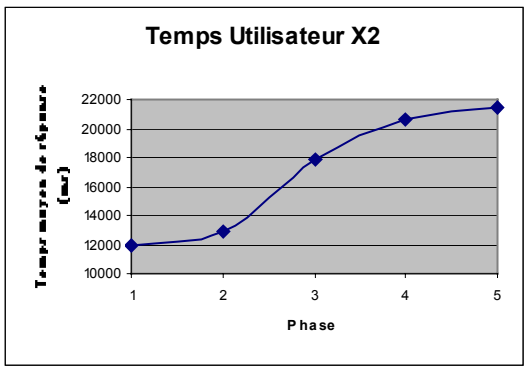

Temps Utilisateur 5
Temps moyen de réponse (ms)
10600
10400
10200
10000
9800
9600
9400
1
2
3
4
5
Phase

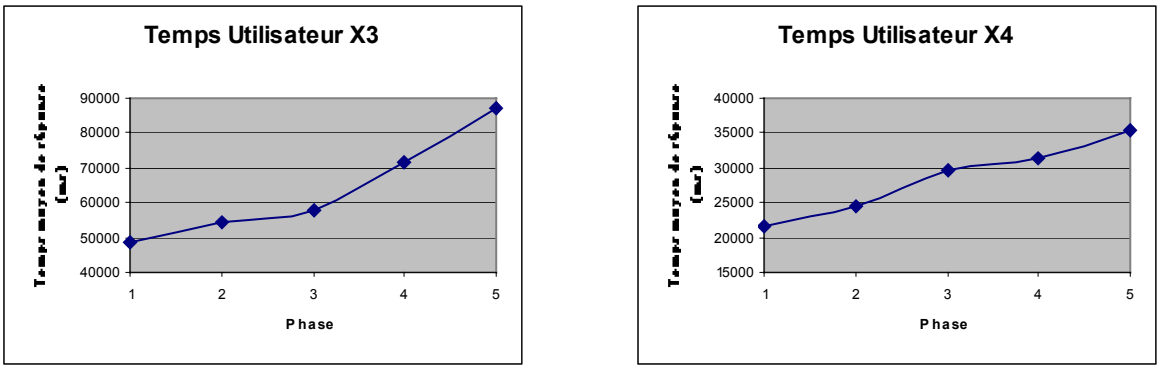

Temps Utilisateur 6
Temps moyen de réponse (ms)
30500
30000
29500
29000
28500
28000
1
2
3
4
5
Phase


Temps Utilisateur 7
Temps moyen de réponse (ms)
56000
55000
54000
53000
52000
51000
50000
1
2
3
4
5
Phase


Temps Utilisateur 8
Temps moyen de réponse (ms)
45000
40000
35000
30000
25000
1
2
3
4
5
Phase


Temps Utilisateur X1
Temps moyen de réponse (ms)
35000
33000
31000
29000
27000
25000
1
2
3
4
5
Phase


Temps Utilisateur X2
Temps moyen de réponse (ms)
22000
20000
18000
16000
14000
12000
10000
1
2
3
4
5
Phase



Temps moyen de réponse selon les différentes phases.

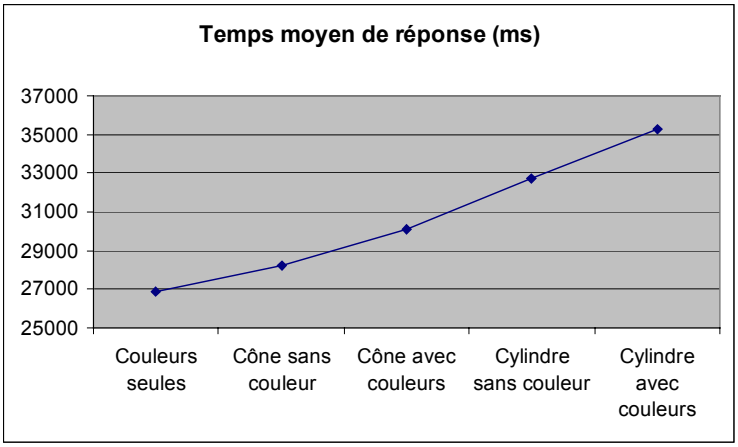


Figure 84 - Temps de réponse moyen (en millisecondes)

Le temps de réponse correspond à la durée entre l'instant où est affiché le point à l'écran et le moment où le participant valide son choix. Il comprend donc toutes les manipulations réalisées par l'utilisateur qui font que le temps de réponse semble très important (à l'optimum le participant 5 met 9 secondes en moyenne pour répondre).

Cette figure est intéressante car elle met en évidence que plus l'utilisateur avance dans la phase d'évaluation plus il met de temps à répondre. En effet, le temps moyen de réponse au cours de la dernière phase est supérieur de 30% du temps moyen de la première phase. Cependant, ce critère ne peut pas être pris en compte dans l'évaluation car, d'une part, quasiment l'unanimité des utilisateurs nous ont fait part de l'aspect répétitif et rébarbatif de l'évaluation. D'autre part, nous pouvons constater qu'il n'y a pas de corrélation entre la courbe individuelle d'erreur moyenne et la courbe de temps.

En effet, la phase dévaluation durait environ 1h30-2h00 avec une bonne heure pour ces cinq phases (5*20 échantillons).

Annexe H
Arborescence MESH

Le nombre de nœuds fils est donné entre <>. A la suite de cette valeur est renseigné le nombre de documents pertinents pour le nœud issus de la collection utilisée.

```
@@ 957 Cardiovascular Diseases <5> 0
-@@ 958 Cardiovascular Abnormalities <2> 0
--@@ 960 Arteriovenous Malformations <1> 32
---@@ 961 Arteriovenous Fistula <0> 47
--@@ 964 Heart Defects, Congenital <6> 148
---@@ 965 Aortic Coarctation <0> 44
---@@ 968 Coronary Vessel Anomalies <0> 29
---@@ 971 Ductus Arteriosus, Patent <0> 23
---@@ 974 Heart Septal Defects <2> 19
----@@ 977 Heart Septal Defects, Atrial <0> 30
----@@ 980 Heart Septal Defects, Ventricular <0> 29
---@@ 983 Tetralogy of Fallot <0> 25
---@@ 984 Transposition of Great Vessels <0> 39
-@@ 990 Heart Diseases <13> 0
--@@ 991 Arrhythmia <8> 299
---@@ 993 Atrial Fibrillation <0> 62
---@@ 994 Atrial Flutter <0> 13
---@@ 995 Bradycardia <0> 40
---@@ 999 Heart Block <1> 85
----@@ 1001 Bundle-Branch Block <0> 20
---@@ 1003 Long QT Syndrome <0> 12
---@@ 1005 Pre-Excitation Syndromes <1> 0
----@@ 1008 Wolff-Parkinson-White Syndrome <0> 39
---@@ 1009 Sick Sinus Syndrome <0> 9
---@@ 1010 Tachycardia <2> 217
----@@ 1011 Tachycardia, Paroxysmal <0> 14
----@@ 1012 Tachycardia, Supraventricular <1> 71
-----@@ 1014 Tachycardia, Atrioventricular Nodal Reentry <0> 15
--@@ 1025 Cardiac Output, Low <0> 10
--@@ 1027 Cardiomegaly <1> 0
---@@ 1028 Cardiomyopathy, Congestive <0> 64
--@@ 1031 Endocarditis <0> 25
--@@ 1034 Heart Aneurysm <0> 27-
--@@ 1035 Heart Arrest <0> 134
--@@ 1066 Heart Rupture <1> 26
---@@ 1067 Heart Rupture, Post-Infarction <0> 12
--@@ 1069 Heart Valve Diseases <7> 0
---@@ 1070 Aortic Valve Insufficiency <0> 60
---@@ 1071 Aortic Valve Stenosis <1> 76
----@@ 1074 Aortic Stenosis, Subvalvular <1> 0
-----@@ 1075 Cardiomyopathy, Hypertrophic <0> 59
---@@ 1077 Heart Murmurs <0> 9
---@@ 1078 Heart Valve Prolapse <1> 0
----@@ 1080 Mitral Valve Prolapse <0> 65
---@@ 1082 Mitral Valve Insufficiency <0> 67
---@@ 1083 Mitral Valve Stenosis <0> 49
---@@ 1089 Tricuspid Valve Insufficiency <0> 19
--@@ 1091 Myocardial Diseases <1> 93
---@@ 1101 Myocarditis <0> 38
--@@ 1102 Myocardial Ischemia <2> 0
---@@ 1103 Coronary Disease <5> 0
----@@ 1104 Angina Pectoris <2> 226
-----@@ 1105 Angina Pectoris, Variant <0> 25
-----@@ 1106 Angina, Unstable <0> 43
----@@ 1108 Coronary Aneurysm <0> 19
----@@ 1109 Coronary Arteriosclerosis <0> 51
----@@ 1110 Coronary Thrombosis <0> 40
----@@ 1111 Coronary Vasospasm <0> 44
---@@ 1113 Myocardial Infarction <1> 0
----@@ 1115 Shock, Cardiogenic <0> 20
--@@ 1117 Pericardial Effusion <0> 40
--@@ 1118 Pericarditis <0> 30
--@@ 1124 Rheumatic Heart Disease <0> 21
-@@ 1129 Hyperemia <0> 15
-@@ 1130 Pregnancy Complications, Cardiovascular <0> 88
-@@ 1136 Vascular Diseases <22> 0
--@@ 1137 Aneurysm <4> 96
---@@ 1138 Aneurysm, Dissecting <0> 50
---@@ 1142 Aneurysm, Infected <0> 13
---@@ 1143 Aneurysm, Ruptured <1> 0
----@@ 1144 Aortic Rupture <0> 41
---@@ 1145 Aortic Aneurysm <0> 107
```

```
--@@ 1155 Angiomatosis <1> 0
---@@ 1157 Hippel-Lindau Disease <0> 5
--@@ 1160 Angioneurotic Edema <0> 32
--@@ 1161 Aortic Diseases <1> 0
---@@ 1166 Aortic Arch Syndromes <0> 14
--@@ 1170 Arterial Occlusive Diseases <4> 180
---@@ 1171 Arteriosclerosis <1> 0
----@@ 1176 Intermittent Claudication <0> 25
---@@ 1178 Fibromuscular Dysplasia <0> 16
---@@ 1180 Mesenteric Vascular Occlusion <0> 25
---@@ 1182 Renal Artery Obstruction <0> 45
--@@ 1188 Arteritis <1> 13
---@@ 1193 Temporal Arteritis <0> 26
--@@ 1195 Cerebrovascular Disorders <6> 302
---@@ 1199 Carotid Artery Diseases <1> 155
----@@ 1200 Carotid Artery Thrombosis <0> 10
---@@ 1222 Cerebral Hemorrhage <0> 160
---@@ 1227 Cerebrovascular Accident <1> 0
----@@ 1228 Brain Infarction <1> 0
-----@@ 1231 Cerebral Infarction <0> 124
---@@ 1255 Intracranial Hemorrhages <2> 0
----@@ 1261 Intracranial Hemorrhage, Traumatic <2> 0
-----@@ 1265 Hematoma, Epidural <0> 13
-----@@ 1266 Hematoma, Subdural <0> 26
----@@ 1271 Subarachnoid Hemorrhage <0> 88
---@@ 1275 Vascular Headaches <2> 0
----@@ 1276 Cluster Headache <0> 12
----@@ 1277 Migraine <0> 110
---@@ 1287 Vertebrobasilar Insufficiency <0> 17
--@@ 1289 Diabetic Angiopathies <1> 0
---@@ 1291 Diabetic Retinopathy <0> 71
--@@ 1292 Embolism and Thrombosis <2> 0
---@@ 1293 Embolism <4> 44
----@@ 1294 Embolism, Air <0> 26
----@@ 1298 Embolism, Fat <0> 10
----@@ 1299 Pulmonary Embolism <0> 126
----@@ 1300 Thromboembolism <0> 76
---@@ 1310 Thrombosis <2> 254
----@@ 1312 Purpura, Thrombotic Thrombocytopenic <0> 10
----@@ 1323 Venous Thrombosis <3> 0
-----@@ 1324 Hepatic Vein Thrombosis <0> 20
-----@@ 1325 Retinal Vein Occlusion <0> 13
-----@@ 1326 Thrombophlebitis <0> 90
--@@ 1328 Hemorrhoids <0> 13
--@@ 1330 Hypertension <3> 0
---@@ 1333 Hypertension, Portal <1> 53
----@@ 1334 Esophageal and Gastric Varices <0> 75
---@@ 1335 Hypertension, Pulmonary <0> 83
---@@ 1337 Hypertension, Renal <1> 26
----@@ 1338 Hypertension, Renovascular <0> 69
--@@ 1340 Hypotension <1> 97
---@@ 1341 Hypotension, Orthostatic <0> 34
--@@ 1343 Ischemia <1> 202
---@@ 1346 Compartment Syndromes <0> 32
--@@ 1373 Raynaud's Disease <0> 31
--@@ 1381 Superior Vena Cava Syndrome <0> 11
--@@ 1382 Telangiectasis <1> 14
---@@ 1385 Telangiectasia, Hereditary Hemorrhagic <0> 9
--@@ 1386 Thoracic Outlet Syndrome <0> 4
--@@ 1388 Varicocele <0> 28
--@@ 1389 Varicose Veins <1> 0
---@@ 1390 Varicose Ulcer <0> 11
--@@ 1394 Vascular Hemostatic Disorders <6> 0
---@@ 1395 Cryoglobulinemia <0> 11
---@@ 1397 Ehlers-Danlos Syndrome <0> 11
---@@ 1398 Hemangioma, Cavernous <0> 24
---@@ 1400 Multiple Myeloma <0> 76
---@@ 1401 Osteogenesis Imperfecta <0> 13
---@@ 1404 Purpura, Schoenlein-Henoch <0> 15
--@@ 1410 Vasculitis <3> 75
---@@ 1418 Behcet's Syndrome <0> 16
---@@ 1420 Mucocutaneous Lymph Node Syndrome <0> 32
---@@ 1432 Wegener's Granulomatosis <0> 31
--@@ 1433 Venous Insufficiency <0> 20
```