Profil-doc : Evaluation d'un système personnalisé de requête à des bases de données en texte intégral

Christine Michel, Thierry Lafouge

Profil-doc : un système personnalisé de requête à des bases de données en texte intégral.

Christine Michel, Thierry Lafouge

Laboratoire RECODOC,
Bat 721, Université Claude Bernard LYON I
43, Bd du 11 Novembre 1918
69622 VILLEURBANNE CEDEX
Tel : 04 72 43 13 91
michel@recodoc.univ-lyon1.fr
lafouge@enssibhp.enssib.fr

1. Introduction

Grâce aux systèmes documentaires manipulant du texte intégral, l'interrogation des bases s'est révélée être beaucoup plus conviviale et transparente pour l'utilisateur. Mais la production d'information en augmentation régulière, la prolifération des sources étend considérablement le volume d'information à consulter pour obtenir une information pertinente. *« On peut faire un constat simple : si le bruit et le silence sont toujours à peu près les mêmes, par exemple de 50%, un utilisateur qui reçoit dix documents en réponse à une question, en trouvera cinq pertinents. Un utilisateur qui obtiendra cent documents, en trouvera sans doute cinquante pertinents, mais aussi cinquante hors sujet. Le facteur bruit devient une gêne très réelle pour l'utilisateur dès que le volume des réponses dépasse un certain seuil «tolérable » »* [Lain 94]. D'une manière très schématique, dans une opération de recherche documentaire classique, l'utilisateur se contente de formuler une requête, puis le système apparie les mots de la requête avec ceux du dictionnaire qu'il possède et génère ainsi une réponse. Dans les systèmes référentiels, la structure de la base assure un certain tri au niveau de la réponse. Au contraire, dans le cas des systèmes documentaires en texte intégral, il est toujours possible de trouver des documents contenant un des termes de la question, mais cela ne veut pas dire qu'ils seront vraiment pertinents pour l'utilisateur. Si ces systèmes savent presque toujours proposer une réponse à la demande de l'utilisateur, ils ne répondent que partiellement à ses besoins. « *La tâche d'interrogation s'inscrit à l'intérieur d'une activité de recherche d'information. En effet, la particularité de la recherche d'information provient du fait que l'utilisateur collecte des données pour un problème qu'il va résoudre par la suite, en dehors du système. Sur ce problème, sur le contexte dans lequel il effectue sa recherche, sur les buts qu'il poursuit nous avons très peu d'information. Pour qu'un système fournisse des réponses satisfaisantes, il faut qu'il ait une certaine connaissance du problème que l'utilisateur se pose. La recherche d'information ne peut pas être considérée comme une tache d'exécution, indépendante du contexte dans lequel elle se passe.* » [Poli 94]

Si l'on cherche l'information sur un réseau un autre fait vient amplifier le bruit. En effet, si un serveur particulier structure généralement l'information d'une manière cohérente ; sur un réseau

comme Internet où circulent des informations issues de multiples serveurs répartis un peu partout dans le monde, les documents qui sont hétérogènes, sont présentés au même niveau, sans distinction particulière de domaine (la physique , la chimie, l'économie ...), de nature (on retrouve pêle-mêle des images, du texte, du son), de contenu (pages personnelles, catalogues publicitaires, publications scientifiques, ...), ou de format (HTML, Postscript, texte, ...). De plus, la nature transversale de certaines sciences, en particulier les sciences de l'information et de la communication, rend inévitable une recherche étendue à plusieurs champs disciplinaires.

Pour pallier aux limites de l'« indexation » et avoir une meilleure connaissance du fonds, les systèmes documentaires traditionnels et automatiques ont tenté de décrire les documents par des critères externes à leurs contenus. Ainsi en bibliothéconomie classique, la dimension d'un ouvrage, son nombre de pages, ..., sont autant de critères supplémentaires permettant de gérer le fonds, mais il est rare qu'un utilisateur se serve de ces critères pour sélectionner des documents. Grâce aux systèmes de gestion de fichiers ou aux systèmes de gestion de bases de données, la recherche d'une notice par l'ensemble des champs (zones) la décrivant est devenue possible ; des champs définissant des caractéristiques externes au contenu ont ainsi pu être rajoutés : le pays et le champ disciplinaire de l'auteur, le nom du laboratoire, etc.

Une étude approfondie[1] sur un certain nombre de textes, livres, thèses, articles de revues scientifiques, a montré qu'on pouvait trouver, pour chacun d'eux, une structure générique facilement identifiable. En effet, dans la majorité des cas, un texte (article, conférence, rapport, ouvrage, etc.) a une *structure générale*, il forme une unité car il est construit pour faire passer un message : résultats de synthèse, nouvelles pistes de recherche, etc. Cette unité matérielle et intellectuelle est le résultat d'un lien parfaitement établi entre ses différentes parties, celles-ci pouvant former à leur tour des unités indépendantes remplissant une fonction bien déterminée. Ainsi, par exemple, la bibliographie est utilisée généralement pour étayer les propos cités dans les différentes parties du texte et pour donner au lecteur une idée plus ou moins exhaustive de tout ce qui a été écrit sur le sujet traité, ce qui représente d'une certaine manière le contexte du texte. Cette constatation, nous a conduit à admettre que «l'éclatement » du document en unités documentaires nous permet, tout en préservant l'unité globale du document (le lien entre l'unité documentaire et le document auquel elle appartient), de présenter à l'usager une information plus affinée et plus facile à saisir.

Mais cette structuration de document n'est pas unique ; en effet, on peut aussi considérer *les différents types de textes* (publicitaires, scientifiques), *le mode d'organisation du discours* (narratif, argumentatif, etc.) ou même encore *la structure physique* (attributs typographiques, polices, espaces, etc.) comme des caractéristiques propres à discriminer une fraction du document.

Le projet Profil-doc [Lain 96] utilise ces différentes structures pour décrire les documents en unités documentaires, au sein d'un système documentaire en texte intégral. Chacune des unités est alors accessible par des index bien sûr, mais aussi par ses propriétés. Le découpage est basé sur la fonction remplie par ces parties du document et non sur leur contenu. Au niveau de l'utilisateur, ces propriétés seront autant d'outils supplémentaires utilisables lors de la requête, pour sélectionner

[1] Norme 5963 : Méthode d'analyse des documents - Norme ISO 2145 : Numérotation des divisions et subdivisions dans les documents écrits - Norme ISO 8613 : Architecture du document.
Norme ISO 7144 : Présentation des thèses et des documents assimilés.
Norme ISO 5966 : Présentation des documents scientifiques et techniques.

l'information « pertinente ». En effet, on peut remarquer que l'utilisateur, face à un système en texte intégral qui lui fournit généralement trop d'information, va développer une stratégie de recherche empirique. Il va par exemple se limiter à certaines bases de données, selon la discipline ou le type de revues répertoriées, ou encore, selon la langue, pays ou année. Toutes ces stratégies ont deux caractéristiques : elles portent sur des critères (la forme, le support, le style, ...) autres que le contenu du document, elles sont très fortement individualisées et permettent une personnalisation de la recherche [Lain 96].

Dans cette optique, le système Profil-doc veut aller plus loin que l'utilisation simple de ces critères pour la description et sélection des documents. En effet, ces propriétés nous permettront de sélectionner un corpus « personnalisé » suivant les caractéristiques de l'utilisateur, corpus sur lequel portera la question. En d'autres termes, ces propriétés, appariées avec le profil de l'utilisateur, nous permettent de présélectionner un ensemble de documents.

2. Le principe

Ce système de sélection s'appuie sur trois composantes fondamentales :

• un découpage des documents en unités documentaires, en se basant sur des fonctions d'usage, c'est à dire, sur l'utilité qu'elles peuvent avoir, en d'autres termes sur le type d'utilisateur qui pourrait en avoir besoin

• une caractérisation du profil de l'utilisateur, de ses compétences et habitudes, mais aussi de l'objectif de sa recherche

• un système d'aiguillage qui à partir du profil de l'utilisateur extrait un sous-corpus personnalisé d'unités documentaires.

Nous venons de voir l'utilité du découpage et de la caractérisation des documents.

Le « profil » de l'utilisateur est défini par diverses caractéristiques : son niveau éducationnel, son champ disciplinaire, le type de recherche souhaitée (recherche exhaustive, pointue, etc.), la situation de la recherche (réalisation d'un projet, mise à jour des connaissances, etc.). Cette caractérisation nous permet de cerner ses besoins informationnels.

Le système d'aiguillage est le coeur du processus, en effet, c'est cette fonction qui va nous permettre de définir l'ensemble des propriétés des unités documentaires souhaitables en fonction d'un profil donné. Nous n'expliciterons pas en détail dans ce travail le processus d'aiguillage, une thèse [Bena 97] est en cours de réalisation sur le sujet.

3. La phase d'interrogation

Nous allons à présent présenter le processus d'interrogation du système profil-doc. Reportons nous à la figure 1 et explicitons les cinq étapes qui composent l'interrogation.

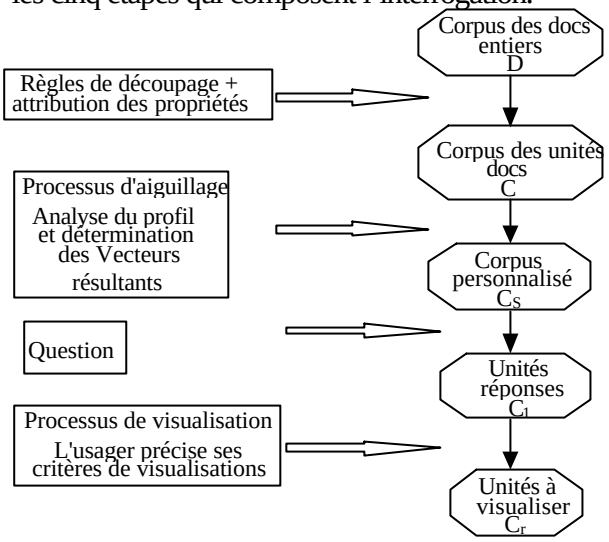


Figure 1 : Processus d'interrogation

3.1. Le découpage des documents

Les unités appartenant au même document *(des unités soeurs)* héritent des propriétés du document père et se distinguent par des propriétés qui leur sont propres. Nous ne passons pas par une étape de compréhension du contenu pour attribuer les propriétés. Elles peuvent être facilement repérées à l'intérieur du document ; comme le champ disciplinaire de l'auteur ou le style de l'unité documentaire ; ou alors repérées automatiquement par certains marqueurs (linguistiques ou autres), c'est le cas de la propriété *forme discursive* de l'unité documentaire.

Une partie des propriétés sont **propres au document entier**, le document père :

- les caractéristiques usuelles (titre du document, titre de la revue, auteur du document, coauteurs, affiliation de l'auteur, pays, année, ...)
- l'environnement de production (champ disciplinaire, communauté et profession de l'auteur).
- le support de diffusion (environnement éditorial, type d'article).

La seconde partie des propriétés est elle, **propre aux unités documentaires**.

- le type de l'unité logique (résumé, introduction, description de contexte, description de thème, description de la méthode, environnement, expérimentation, résultat, discussion, développement, conclusion, bibliographie, table des matières, annexes).
- la forme discursive du document (argumentatif, descriptif, narratif).
- le style du document (littéraire, littéraire contenant des données numériques, équations et formules de calculs schémas ou figures).

A partir d'un profil donné, ces propriétés[2] sont utilisées comme nous l'avons dit précédemment pour présélectionner automatiquement certains types d'unités documentaires. Cependant, elles peuvent être utilisées directement par l'utilisateur lors de sa requête.

Prenons un exemple simple : un utilisateur veut savoir comment les travaux de Chomsky ont été exploités par les chercheurs. Il va questionner sur les unités documentaires ayant le type logique « bibliographie » avec la requête "Chomsky". Le corpus C_1 renvoyé est donc uniquement composé des bibliographies comportant des références à Chomsky. L'utilisateur choisit de voir les unités documentaires de type « introduction ». Le système remonte aux documents pères des unités sélectionnées et en extrait les unités « bibliographies » qu'il présente à l'utilisateur.

3.2. Description du profil de l'usager

Dans un contexte documentaire, déterminer la connaissance d'un utilisateur sur un sujet permet d'améliorer considérablement sa recherche. Nous n'irons pas aussi loin dans l'étude de l'usager dans le cas du système profil-doc, seules nous intéressent les caractéristiques qui influencent la réalisation de la recherche d'information.

M.-F. Blanquet [Blan 94] a présenté comment les intermédiaires en information développent et utilisent les modèles usagers. Pour elle, le dialogue professionnel/usager comprend deux parties : l'une porte sur le sujet de la recherche en soi, la question ; l'autre sur l'environnement spécifique de cette dernière, étant entendu qu'une même question à formulation identique peut entraîner des réponses fort différentes suivant le profil du demandeur. L'environnement de la question a précisément pour objectif de prendre la mesure des besoins de l'usager. **Une même question peut induire des réponses différentes en fonction d'un grand nombre de facteurs objectifs ou subjectifs : le niveau de l'utilisateur, sa profession, les langues comprises, l'utilisation précise de l'information trouvée et tout un contexte personnel dont, d'une façon implicite, le professionnel de l'information tient compte pour effectuer sa recherche.**

En nous basant sur les études de P. J. DANIELS [Dani 86], nous avons choisi les quatre caractéristiques suivantes : Niveau éducationnel, Champ disciplinaire, Etapes de recherche, Type de recherche. Lorsque l'usager « entre » sur le système il renseigne donc à partir des listes suivantes, chaque caractéristique[3].

Niveau éducationnel

Maîtrise / DEA / Recherche[4]

Champ disciplinaire

SIC / Informatique / Agronomie / Pharmacie ...

Etapes de recherche

Constitution d'une bibliographie

Définition du sujet

[2] Ces propriétés sont fixées pour les besoins de l'étude, œtte liste n'est cependant pas exhaustive, elle pourra être complétée dans le futur.

[3] La définition des modalités s'est fait en suivant les résultats d'un questionnaire que nous avons effectué auprès de chercheurs en SIC, sciences pharmaceutiques, et sciences physiques

[4] Comprend les doctorants, chercheurs, enseignants chercheurs, ... considérés comme spécialistes dans un domaine.

Faisabilité
Expérimentation
Interprétation des données
Rédaction
Repérage des approches expérimentales
Plan de travail
Compréhension de la problématique
Etat de l'art
Synthèse bibliographique
Dégagement des nouveaux axes de recherche
Mise à jour des connaissances

Type de recherche

Recherche pointue
Recherche généraliste

3.3. Mise en place d'une fonction d'aiguillage

La fonction d'aiguillage est le coeur du système, c'est elle qui va extraire les unités documentaires du corpus, en fonction du profil donc de l'usage fait par l'utilisateur. Brièvement, nous nous sommes basés sur la littérature ainsi que sur une enquête [Bena 97] que nous avons effectuée sur les usages et habitudes des chercheurs en SIC, sciences pharmaceutiques et sciences physiques, pour construire une matrice « profil-utilisateurs ». La sélection, dans cette matrice, des colonnes décrivant le profil de l'utilisateur permet d'obtenir un ensemble de propriétés, utilisées pour présélectionner un ensemble d'unités documentaires, ensemble sur lequel portera la requête de l'utilisateur [Bena 97].

Par exemple ; considérons deux utilisateurs ayant les profils distincts P1 et P2 qui sont :

P1	**P2**
Etudiant en maîtrise en Sciences de l'Information et de la Communication, voulant approfondir la problématique d'un sujet.	Chercheur en Sciences de l'Information et de la Communication, effectuant l'état de l'art sur un sujet où il n'est pas spécialiste.

Tableau 1

Nous aurons donc comme caractéristiques :

	P1	P2
Niveau éducationnel **Champ disciplinaire**	Maîtrise Sciences de l'Information et de la Communication	Chercheur Sciences de l'Information et de la Communication
Etapes de recherche	Compréhension de la problématique	Constitution de bibliographie
Type de recherche	Recherche pointue	Recherche généraliste

Tableau 2

Ces caractéristiques nous permettent de sélectionner certaines propriétés que doivent valider les unités documentaires, elles sont répertoriées dans le tableau ci-dessus. Nous appellerons ce tableau le **vecteur résultant**.

	P1	P2
Type d'unité logique	• Résumé • Introduction • Description de méthode • Discussion • Conclusion	• Résumé • Introduction • Description de thème • Conclusion
Forme discursive du document	• Argumentatif • Descriptif	• Argumentatif
Style	• Littéraire avec données numériques • Schémas • Formalisation	• Littéraire avec données numériques • Littéraire pur
Type d'environnement éditorial	• Thèse / Mémoire • Revue primaire	nul*
Champs disciplinaire de l'auteur	• SIC	• SIC
Profession de l'auteur	• Etudiant • Enseignant chercheur	• Etudiant • Enseignant chercheur
Communauté de l'auteur	• Etudiant • Universitaire • Industriel	• Etudiant • Universitaire • Industriel

Tableau 3

* Le type d'environnement éditorial n'est pas renseigné pour le profil 2, cette caractéristique n'intervient pas dans la discrimination des unités documentaires présentées.

3.4. Sélection d'un sous corpus de la base

Nous voyons bien dans l'exemple précédent que les propriétés induites par un profil sont très variées, le problème de la sélection d'une unité dans la base se pose alors. Une unité validant le type d'environnement éditorial mais pas le type de l'unité logique est elle à sélectionner ou à laisser ? C'est justement ce type d'ambiguï té que la fonction d'aiguillage permet de résoudre. Une série de fonctions booléennes appliquées au vecteur résultant nous permet de combiner ces propriétés, pour effectivement extraire les unités qui constitueront le premier sous-corpus Cs. C'est sur ce sous corpus que la requête de l'usager va porter.

3.5. Requête de l'usager

La requête de l'usager (ainsi que l'alimentation de la base) se fait à partir du logiciel documentaire SPIRIT[5]. Il permet la génération automatique de bases de données textuelles, et leurs interrogations en langage naturel. La réponse du système est présentée sous forme d'une liste triée de documents [Rada 88]. Cette valeur de pertinence est évaluée par un poids qui est calculé pour chacune de ces classes. Elles sont présentées à l'utilisateur selon leur pertinence, et représentent la meilleure équation booléenne élaborée à partir des mots de la question, qui permet d'obtenir ces documents. Le poids de la classe est l'addition des poids de tous ces mots, composant la question. La particularité de SPIRIT est de coupler deux analyses, linguistique et statistique, qui sélectionnent les concepts et les pondèrent.

3.6. Visualisation et navigation de l'usager

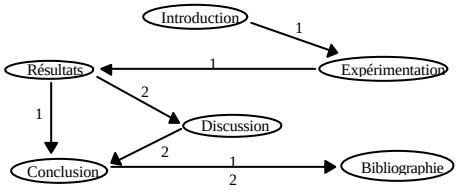
La requête de l'usager a permis de sélectionner une partie de Cs, C1 (cf. figure 1) qui contient les unités de sens « pertinent ». Les parties du document à présenter à l'utilisateur ne sont pas nécessairement celles sur lesquelles la requête a été effectuée. C'est l'utilisateur, par le processus de navigation, qui va choisir de visualiser telle ou telle partie du document entier.

Nous pensons choisir de proposer une lecture de type navigationnelle car la lecture de l'article scientifique s'y prête du fait de sa structure et de «l'indépendance sémantique» de ses différentes parties. En termes grossièrement simplifiés, la lecture navigationnelle traduit une démarche «naturelle» de compréhension. Ainsi un chemin donné de lecture reflète, généralement, un sens voulu par l'utilisateur (le lecteur).

Nous proposons ici trois scénarios de navigation qui ne sont pas limitatifs.

Dans le premier scénario nous proposons à l'utilisateur de naviguer dans les unités d'un même document suivant un ordre prédéfini par les habitudes de lecture des chercheurs mise en évidence dans le questionnaire.

Exemple :



Le second scénario consisterait à définir des chemins de navigation en fonction des propriétés de l'unité d'appel, en présentant en premier par exemple les unités validant des propriétés type de l'unité logique, forme

[5] Développé par la société T-GID - 84-88 Bd de la mission Marchand 92411 Courbevoie Cedex.

discursive et/ou style, puis les unités validant l'environnement de production puis celle validant le support de diffusion.

Le dernier type de scénario suit la lecture séquentielle de l'article, en présentant les unités, lorsqu'elles sont extraites du même document, selon leur ordre dans le document initial.

4. Expérimentation

4.1. Constitution d'une base

En appliquant les règles de découpage mentionnées ci-dessus, nous avons constitué une base de données d'unités documentaires.

A l'heure actuelle la base comprend 505 unités documentaires extraites de 55 documents. Ils proviennent :

- pour le champ des **sciences de l'information** de revues de presse professionnelle (Bulletin des bibliothèques de France, Documentaliste Sciences de l'information), de revues de presse fondamentale (La revue française de bibliométrie, Laforia, Cahiers du Lerass) et d'un ouvrage (IDT 96 Paris 21-23 mai 1996)
- de revues en presse fondamentale et actes de congrès en **biomécanique** (AUTOMEDICA, Archives de physiologie et de biochimie, 21ème congrès de la société de biomécanique, ITBM, Fifth European Conference of medical and Health Libraries)
- d'un ouvrage en **biologie** (C. R. Acad. Sc. Paris, t. 272, p. 1391-1393 (8 mars 1971) Série D). Nous avons inséré de plus, des mémoires et articles de recherche disponibles sur Internet (l'origine des publications n'a pu être retrouvée), ainsi que des écrits didactiques
- de revues de presse fondamentale en **pharmacie** (Journal de Pharmacie de Belgique)

Les auteurs sont soit des étudiants, soit des spécialistes du domaine, industriels ou universitaires du domaine public ou parapublic.

4.2. Utilisation directe des propriétés

Un utilisateur, par exemple un industriel, souhaite faire une recherche uniquement sur les réalisations techniques d'un domaine, prenons par exemple les méthodes qui permettent de faire de l'indexation automatique. Dans un système classique en texte intégral il formulera sa requête en langage naturel sous la forme Q = « Les méthodes d'indexation automatique ». Nous pouvons considérer que l'utilisateur a besoin de documents de nature **P** = « méthode », sur le sujet **S** = « L'indexation automatique ». Nous pouvons donc décomposer cette requête Q en **S+P**. Avec notre système, l'utilisateur peut faire *une utilisation directe des propriétés* en faisant un tri préalable sur les unités de type logique P = « méthode » avec comme clé d'interrogation S = « indexation automatique ». Simulons sa requête sans filtrage sur la base que nous avons constituée. Nous obtenons **130** unités documentaires extraites de 38 documents, donc en moyenne 3,30 unités extraites par document.

Observons à présent la distribution du type d'unité logique qui lui est présenté.

annexe	4
bibliographie	7
conclusion	12
contexte	18
développement	13
discussion	13
expérimentation	5
introduction	14
méthode	**25**
résultat	6
résumé	9
thème	4
Total	130

Si l'on considère que l'utilisateur n'est intéressé que par les unités de type « méthode », il ne retiendra que 25 unités sur les 130 unités initiales. Le terme « méthode » est indexé dans la question et repéré dans le texte. Il rapatriera donc des documents décrivant effectivement des méthodes d'indexation automatiques, mais aussi des documents analysant ou discutant ces méthodes (unités de type résultats, discussion, contexte, développement, ...), ainsi que des unités décrivant des méthodes d'autres sujets.

L'utilisateur n'a aucun moyen de faire le tri rapidement entre les unes et les autres.

Effectuons à présent la même requête avec un filtre préalable sur les unités de type logique « méthode ». Nous obtenons uniquement **5** unités documentaires.

Cet exemple montre bien l'utilité pour l'utilisateur de préciser si les mots de sa requête portent sur des éléments structurels (les propriétés) des unités documentaires, ou bien effectivement, s'ils concernent le contenu du document.

De multiples exemples peuvent être construits sur ce modèle, « Je voudrais des textes descriptifs *(Style du document)* sur ... » « Je voudrais des résultats de recherche *(Type de l'unité logique)* sur ... ». Nous pouvons augmenter l'effet en construisant des requêtes composant les divers attributs « Je voudrais des résultats de recherche *(Type de l'unité logique)* descriptifs *(Style du document)* des travaux des laboratoires de Lyon *(Affiliation)* sur la circulation de l'information ».

4.3. Utilisation du profil

Nous allons observer comment le système réagit avec l'application des 2 profils précédents.

1er cas : L'utilisateur ne dispose d'aucun outil de filtrage, il pose une question en langue naturelle du type « Je voudrais connaître l'ensemble des travaux sur la circulation de l'information effectués dans les laboratoires de Lyon ou de Villeurbanne ».

2eme cas : L'utilisateur dispose de l'outil de filtrage il renseigne donc son profil. Nous allons procéder aux deux types d'interrogation, dans le cas du profil 1 et 2 décrit précédemment. La question de l'utilisateur sera alors, «la circulation de l'information », il aura précisé préalablement dans le champ *affiliation de l'auteur*, qu'il souhaite avoir des travaux de laboratoires de Lyon ou de Villeurbanne.

Nous appellerons **R** le corpus d'unités documentaires présenté à l'utilisateur lorsqu'il pose sa question sans aucun filtrage et **R1** (respectivement **R2**), le corpus d'unités documentaires lorsqu'il renseigne son profil P1 (respectivement P2).

Rappelons que SPIRIT présente des classes de documents, ordonnées suivant le degré de pertinence des mots informationnels qui ont servi à les créer. Par exemple pour R nous avons la classe 1 caractérisée par « information, circulation, Lyon, Villeurbanne », et la classe 2 par «information, circulation, travaux ». La constitution des classes est entièrement liée au logiciel SPIRIT, l'utilisation d'un autre système de recherche nous donnerait des résultats différents et donc le système influe sur

notre évaluation. Cependant, nous avons construit notre protocole pour qu'il puisse être réutilisable sur un autre système.

Nous avons choisi de comparer les classes de même niveau, par exemple la classe 1 de R, R1 et R2. Ces trois classes ne sont pas forcément définies par les mêmes mots informationnels, cependant, nous partons du principe que l'utilisateur va visionner les documents en respectant ce classement, nous pouvons donc les comparer.

Nous obtenons comme résultats :

	Nombre total d'unités documentaires rapatriées	Nombre total de classes effectuées par spirit
Sans filtrage	308	18
avec P1	5	4
avec P2	26	9

Tableau 5

Nous observons qu'effectivement le profil filtre en volume d'une manière conséquente les unités documentaires. Nous passons de 308 unités à 5 pour P1 et 26 pour P2. De plus, si l'on considère que la question est analysée d'une manière similaire dans tous les cas, le sens contenu dans chaque document sera identiquement évalué. Les différences que nous observons tiennent donc d'une part à l'absence de présélection, et d'autre part à l'ajout de classes supplémentaires ; celles constituées de mots non informationnels pour l'utilisateur comme « travaux » ou « laboratoires ».

Nous présenterons des résultats nous permettant de saisir les différences entre les sous corpus d'unités documentaires présentés à l'utilisateur avec et sans pré-filtrage. Nous utiliserons deux indicateurs évaluant d'une part les recoupements d'éléments entre R, R1 et R2 (comme le feraient des calculs de distance usuels entre ensembles) mais aussi l'ordre de présentation des unités documentaires.

4.3.1. Indicateur d'« éparpillement » du système

Nous pouvons calculer un indice qui quantifie l'éparpillement des réponses, avec et sans profil. Nous considérons les ensembles R1 et R2 comme des référentiels pour les profils P1 et P2, nous pouvons calculer l'éparpillement des réponses du système lorsqu'on fait une interrogation sans présélection.

Nous noterons $Br_{R1}(x)$ le ratio d'éparpillement de la classe x de R par rapport au référentiel R1.

$$Br_{R1}(x) = \frac{Card\left(R_x \cap \overline{R1}\right)}{Card\left(R_x\right)}$$

Le volume de documents étant faible, nous effectuerons les calculs sur des regroupements de 5 classes[6]. Pour l'ensemble des classes de 1 à 5 nous aurons par exemple :

[6] Sur une base plus large nous effectuerons les calculs pour chacune des classes.

$$Br_{R1}(1-5) = \frac{Card\left(R_{1-5} \cap \overline{R1}\right)}{Card\left(R_{1-5}\right)}$$

R_{1-5} étant l'ensemble des unités documentaires des classes 1 à 5 de R.

Le ratio est maximum (Br_{R1} = 1) si l'ensemble des classes sélectionnées ne contient aucun des documents de R1 et est minimum (Br_{R1} =0) s'il les contient exactement tous. Cet indicateur nous permet de savoir dans quelle classe de R nous allons retrouver les éléments considérés comme pertinents dans le sous corpus référentiel R1 ou R2.

Groupement des classes	R	R_1	R_2	Br_{R1}	Br_{R2}
R_1-R_5	16	5	9	1	11/16=0.68
R_6-R_{10}	14	0	17	1	12/14=0.85
R_{11}-R_{15}	29	0	0	28/29=0.96	28/29=0.96
R_{16}-R_{18}	249	0	0	245/249=0.98	231/249=0.92

Tableau 6

On remarque un ratio maximum pour Br_{R1}(1-5) et Br_{R1}(6-10) qui vient du fait que les éléments de R1 se dispersent dans les dernières classes de R (R_{11} à R_{18}) comme nous le verrons dans la suite. Le ratio le plus faible pour la comparaison avec R2 se situe bien dans la première classe, il est cependant de 0.68, soit 11 documents non « pertinents » pour 16 présentés à l'utilisateur. Nous constatons cependant, comme nous nous pouvions le prévoir, que le ratio augmente au fur et à mesure que l'on descend dans les classes. La présélection a donc un effet certain sur la recherche.

4.3.2. Indicateur de proximité entre ensembles

Observons plus précisément comment les documents se répartissent dans chacun des ensembles de classes. Les intersections de chacun des ensembles nous permettent de calculer des indices de proximité. Cet indice de proximité compare deux à deux les classes d'ordre équivalent, par exemple nous comparons R_{1-5} avec $R1_{1-5}$ ou $R2_{1-5}$. Cet indice, nous permet de savoir si les classes et les éléments qui les composent changent du fait de l'utilisation du pré-filtrage.

L'indice de proximité est calculé en se basant sur l'indice de Jaccard :

$$Prox\,(R_{1-5},\, R1_{1-5}) \;=\; \frac{card\left(R_{1-5} \cap R1_{1-5}\right)}{card\left(R_{1-5} \cup R1_{1-5}\right)}$$

Groupement des classes	R	R_1	R_2	$card(R \cap R1)$	$card(R \cup R1)$	$card(R \cap R2)$	$card(R \cup R2)$	**Prox. (R,R_1)**	**Prox. (R,R_2)**
R_1-R_5	16	5	9	0	21	3	22	**0**	**0.13**
R_6-R_{10}	14	0	17	0	14	1	30	**0**	**0.033**
R_{11}-R_{15}	29	0	0	0	29	0	29	**0**	**0**
R_{16}-R_{18}	249	0	0	0	249	0	249	**0**	**0**

Tableau 7

Nous observons qu'aucune unité documentaire de R1 n'est identiquement classée dans R, et que la proximité est toujours nulle.

La proximité de R par rapport à R2 est maximum pour les premières classes, bien qu'elle ne soit que de 0.13. Le profil est peut être ici moins spécifique, donc plus proche d'une interrogation sans présélection.

4.3.3. Ordre de présentation

Considérons à présent l'ordre de présentation des unités documentaires. En effet, la qualité d'un ensemble de réponses dépend du classement de ses différents éléments. Les tableaux 8 et 9 nous permettent de savoir comment l'utilisation d'un profil modifie l'ordre de présentation des unités documentaires.

Le tableau 8 (respectivement tableau 9) présente les unités documentaires[7] de R1 (respectivement R2) selon leur ordre de présentation (rang), le numéro de leur classe dans R1 (respectivement R2) et dans R.

[7] Les numéros sont seulement des références à nos unités documentaires dans la base.

Rang dans R1	R1	No des classes dans R1	No des classes dans R
1	docu494	1	15
2	docu449	2	13
3	docu424	3	nul[8]
4	docu450	3	nul
5	docu95	4	15

Tableau 8

Rang dans R2	R2	No des classes dans R2	No des classes dans R
1	docu30	1	5
2	docu28	1	2
3	docu465	2	2
4	docu446	3	6
5	docu51	4	18
6	docu45	4	15
7	docu44	4	18
8	docu32	4	18
9	docu247	5	3
10	docu208	6	6
11	docu119	6	11
12	docu117	6	2
13	docu464	7	15
14	docu485	7	18
15	docu495	7	18
16	docu231	8	15
17	docu232	8	15
18	docu407	8	18
19	docu150	9	18
20	docu145	9	15
21	docu143	9	15
22	docu139	9	15
23	docu136	9	18
24	docu132	9	18
25	docu60	9	18
26	docu59	9	18

Tableau 9

Remarquons dans le tableau 8 qu'aucun document de R1 n'apparaît dans les 20 premières unités de R, la première unités documentaire de R1 « docu494 » n'apparaît que dans la 15ème classe. Cela veut dire que l'utilisateur de profil P1 aura à visionner 20 unités documentaires s'il effectue dans requête sans pré-filtrage avant de voir apparaître cette unité.

Dans le tableau 9 nous voyons que le 1er document de R2 n'apparaît qu'en 16eme position dans la 5ème classe R, le 2ème en 3ème position, le 3ème en 9ème position ...

Nous pouvons remarquer qu'aucun document de la classe 1 de R n'apparaît dans R1 et R2, cette classe est caractérisée par les mots informationnels « information, circulation, Lyon, Villeurbanne », et ne contient en fait qu'un seul document. Celui ci n'apparaît pas, en dépit de sa proximité avec la requête car il est défini par des propriétés non conformes à P1 et P2.

Nous voyons bien que l'ordre de présentation des unités documentaires est complètement modifié par l'utilisation des profils. Il nous reste à déterminer un indicateur nous permettant de quantifier le « retard » de lecture, c'est à dire prenant en compte la différence d'ordre des éléments de R et R1 ou R2. Pour cela nous travaillons avec les indicateurs de Tague-Sutcliffe [Tagu95] que nous sommes en train d'adapter à notre système.

5. Conclusion

[8] Ces deux éléments de R1 ne se retrouvent pas dans R car nous avons, en plus du profil, fait un pré-filtrage sur l'affiliation de l'auteur. Si aucun pré-filtrage direct n'est effectué et que seul les propriétés renseignées dans le profil trient l'information alors on a $R1 \subset R$ et $R2 \subset R$.

Cette étude nous montre bien que la recherche d'information est très largement modifiée par les critères que nous y ajoutons. Le découpage diminue le volume d'information présenté en sélectionnant uniquement les parties de document traitant du sujet de la requête. De plus, l'interrogation directe par les caractéristiques décrivant les documents ou bien par le profil de l'utilisateur permet de préciser la nature de l'information à présenter.

Notre problème actuellement est de construire des outils pour effectivement mesurer l'impact de ces différents processus sur la recherche d'information. Nous aurons à mesurer les différences de volume, d'enregistrements rapatriés, et d'ordre de présentation. Les deux indicateurs se rapprochant du bruit et de la proximité que nous avons présentés, ne sont qu'un exemple et doivent être accompagnés pour une étude plus approfondie. C'est la tâche à laquelle nous nous employons maintenant.

6. Bibliographie

[Bena 97] N. Ben Abdallah. Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information utile : vers un système d'information évolué. (Thèse à soutenir le 7 juillet 97 à l'université Lyon 1)

[Blan 94] Blanquet Marie-France. Intelligence artificielle et système d'information. ESF. 1994. 269 p.

[Dani 86] Daniels P.J. Cognitive models in information retrieval an evaluative review. Journal of documentation , Vol. 42, N°4, Décembre 1986, pp. 272-304.

[Lain 94] Lainé-Cruzel Sylvie. Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs. Documentaliste. Sciences de l'information - 1994 - 31 (3) - pp. 143-147.

[Lain 96] Lainé-Cruzel Sylvie, Lafouge Thierry, Lardy Jean-Pierre, Ben Abdallah Nabil. Improving information retrieval by combining user profile and document segmentation. Information Processing and management -1996- vol 32 n 3 - pp. 305-315.

[Poli 94] Polity Y. Evaluation des modes de recherche en langage naturel. Documentaliste. Sciences de l'information - 1994 - 31 (3) - pp. 136-142.

[Rada 88] Radasoa H. Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles. Thèse. Université Paris Sud. Centre d'Orsay - 28 Novembre 1988 - pp. 156.

[Tagu95] Tague-Sutcliffe J. Mesuring information. An information services perspectives. Academic Press. - 1995 - pp. 206.