ProfilDoc, un Système de Recherche d'Information scientifique

PERENON Pascal
Attaché Temporaire à l'Enseignant et la Recherche (université Grenoble 2)
membre associé du laboratoire URSIDOC (université Lyon 1).

1. Introduction

Aujourd'hui, dans un contexte de numérisation de l'information, le nombre de documents numériques est en constante progression. En une décennie, le nombre de pages web sur Internet est passé de quelques centaines en 1992 à plusieurs milliards en 2005[1]. Et il ne s'agit là que des documents dits de "surface" ou "statiques" en référence à la partie immergée d'un iceberg : le WWW possède en réalité une quantité bien plus grande d'information qui n'est pas directement accessible par les moteurs de recherche. Une étude [1] estime que la taille du Web " profond " ou " invisible " est cinq cents fois supérieure à la taille du Web " visible ". Appréhender cet ensemble colossal dans sa totalité est désormais utopique. L'individu en quête d'information doit faire des choix pour sélectionner l'information dont il a besoin. Dans ce monde numérique inadapté aux facultés humaines, des outils de recherche informatique ont été élaborés pour l'assister. L'individu devient un utilisateur lorsqu'il utilise ces outils. Ce sont des Systèmes de Recherche d'Information (SRI). Leur objectif est d'aider l'utilisateur à trouver l'information adaptée à son besoin d'information. Les moteurs de recherche sur Internet sont sans aucun doute les SRI les plus connus et les plus utilisés actuellement. Néanmoins s'ils fonctionnent correctement pour un besoin précis, leur efficacité est déclinante lorsqu'il s'agit de besoins plus complexes. Pour expliquer ce phénomène, il faut remarquer que l'individu exprime son besoin d'information à l'aide d'une requête de mots clés (des mots du langage). Ce faisant, il traduit un état cognitif en une représentation verbale. Le passage de l'un à l'autre entraîne inévitablement une imprécision. Cette imprécision est transmise au SRI qui la reporte à son tour au niveau des documents à sélectionner. Au regard de milliards de documents, cette imprécision peut engendrer la sélection de plusieurs centaines ou milliers de documents plus ou moins éloignés du besoin réel de l'utilisateur : document trop simple ou trop complexe, document déjà lu, document hors sujet. L'individu subit alors une surabondance de documents inégalement pertinents et est insatisfait.
Pour comprendre ce problème, il faut s'intéresser à la conception des SRI d'aujourd'hui. Le SRI est un intermédiaire machine entre une demande d'information et une offre d'information, entre le besoin de l'utilisateur et une source d'information numérique. Pour construire la correspondance, le SRI effectue une simple comparaison entre la représentation du besoin et la représentation de l'information (des documents numériques). Néanmoins, pour comparer, il faut un modèle de représentation commun. Deux approches sont alors envisageables : soit la représentation du besoin de l'utilisateur s'adapte à la représentation de l'information. C'est l'approche orientée " document ". Soit la représentation de l'information s'adapte (ou est adaptée) à la représentation du besoin de l'utilisateur. C'est l'approche orientée " utilisateur ". Il est clair que la première approche est plus facile et rapide à mettre en oeuvre dans un système applicatif. En effet, l'information est concrétisée par un document qui la supporte (article, page Web, mémoire, livre), tandis que le besoin d'information est abstrait, subjectif et individuel. Il est la conséquence d'une activité humaine et est inhérent à

[1] En 2005, le moteur de recherche " Google " recense plus de 8 milliards de pages web. www.google.com

des processus cognitifs encore mal compris par la communauté scientifique. Ajoutons à cela des enjeux économiques attractifs[2] , nous aboutissons alors à une position dominante des SRI orientés " document " et à la problématique citée précédemment. Pourtant, la surabondance de l'information condamne à terme cette orientation. L'approche " utilisateur " est une réponse possible à cette problématique. L'adaptation de l'information à l'utilisateur implique une réponse documentaire personnalisée. Cela nécessite de comprendre son besoin d'information : comment se construit-il ? Dans quelles mesures l'environnement, le contexte, les activités de l'utilisateur influencent-ils son besoin d'information et sa recherche d'information ? Comment expliquer la pertinence d'un document ? Qu'est ce qu'une réponse à un besoin d'information ? Ces questions sont centrales dans une approche utilisateur. Elles sont essentielles dans la conception d'un SRI utilisateur. Quelques projets de recherche pionniers (IOTA [2] , I3R[3] , Metacat[4]) se sont intéressés à la conception de SRI " utilisateur ". Notre prototype " ProfilDoc" [5] [6] [7] [8, 9] est un SRI orienté utilisateur. Il propose de prendre en compte l'utilisateur et le contexte de son besoin d'une part, et le contexte de production et de diffusion d'un document d'autre part. Il s'agit d'ajouter à la prise en compte classique de la requête et des documents, la prise en compte des individus (utilisateurs de SRI et auteurs des documents) qui produisent ces représentations dans leur contexte d'activité. L'objectif de ce SRI est d'augmenter les performances d'une recherche documentaire en diminuant le bruit. Dans cet article, nous décrivons dans la première partie le modèle de représentation de notre prototype. Dans une seconde partie, nous décrivons plus précisément la fonction d'évaluation de la fonction de correspondance du prototype. Enfin, dans une dernière partie, nous proposons d'évaluer cette fonction de correspondance. Nous discutons des résultats, des limites et des perspectives de notre prototype.

Tout SRI est composé de trois éléments:

1. une représentation du document
2. une représentation du besoin d'information de l'utilisateur
3. une fonction de correspondance entre les deux représentations dont l'objectif est d'évaluer le jugement de pertinence de l'utilisateur.

La typologie d'un SRI repose sur le choix de ces trois éléments.
Dans une approche SRI orientée " document ", la représentation de l'utilisateur est minimale. Il s'agit dans la plupart des cas d'une requête de mots clés représentant le sujet de la recherche de l'utilisateur. La représentation du document est par contre plus élaborée. Non seulement le contenu du document est pris en compte pour la représentation (notamment par l'opération de l'indexation) mais d'autres caractéristiques sont aussi prises en compte. Par exemple, les hyperliens d'une page web, la structure logique (titre, auteur, date, figures, paragraphes) et physique (nombre de pages, taille des polices,...) d'un document. Le troisième élément du SRI orienté document est la fonction qui a pour objectif de faire correspondre les deux représentations : une représentation restreinte du besoin de l'utilisateur avec une représentation plus complexe[3] du document, c'est-à-dire de trouver les documents qui apportent une réponse à l'utilisateur. Cependant, malgré des techniques toujours plus évoluées et complexes, la performance de la fonction de correspondance est limitée par la représentation restreinte du besoin de l'utilisateur. En effet, comment distinguer un utilisateur qui, en tapant la requête " les particules élémentaires ", recherche les références d'un roman, de celui qui recherche un texte de vulgarisation scientifique sur la physique des particules ? C'est sans doute le point faible des SRI d'aujourd'hui massivement orientés " document "

[2] Le moteur de recherche " Google " pèse 36 milliards de dollars de CA (source : journal du net http://www.journaldunet.com/)
[3] Pondération probabiliste, vectorielle. Reformulation de requête. Réseau connexionniste.

comme les moteurs de recherche[4] sur Internet entre autres. Cette prise de conscience est effective. Ainsi, Wilkinson [10] affirme que le besoin d'information est représenter trop simplement (quelques mots clés) par les moteurs de recherche mais c'est pour cette même raison qu'elle est utilisée. Il argumente qu'une représentation étendue associée à une implémentation simple représente le meilleur choix pour les moteurs de recherche. Autre cas, la conférence TREC [11] , qui réunit chaque année depuis 1990 des équipes scientifiques concepteurs de SRI, propose depuis l'année 2003 une nouvelle évaluation " HARD track " ajoutant à la recherche sur le contenu des métadonnées précisant le besoin d'information de l'utilisateur [12] . Que ce soit au niveau du document et maintenant au niveau du besoin d'information, une représentation plus précise de l'offre et de la demande d'information offre de nouvelles pistes pour améliorer la précision d'une recherche d'information. Ainsi, l'idée d'un SRI orienté " utilisateur " comme le projet ProfilDoc constitue l'une de ces nouvelles voies.

2. Le SRI ProfilDoc

Pour développer la représentation du besoin de l'utilisateur (en priorité) et celle du document, le SRI ProfilDoc propose de prendre en compte le contexte. Le contexte de la requête de l'utilisateur (l'activité pour laquelle l'utilisateur recherche de l'information, l'usage de l'information, l'utilisateur lui-même) et le contexte de production et de diffusion du document (l'auteur du document, le type de document produit).
Pour notre projet ProfilDoc, nous avons choisi de décomposer le contexte d'un besoin d'information en catégories : utilisateur, activité liée à la recherche d'information, usage cognitif et fonctionnel de l'information, et thème de l'information recherchée. De même, nous avons décomposé le contexte de production et de diffusion d'un document en catégories : l'auteur du document, le document lui-même, le contexte éditorial et le sujet du document. Pour représenter ces catégories, nous avons défini un ensemble de variables (objectif de recherche, profession de l'utilisateur, taille du document, profession de l'auteur du document,...) (tableau 1).

variables sur l'utilisateur		variables sur le document	
catégories	**variables**	**catégories**	**variables**
utilisateur	sexe	**auteur du document**	type d'auteur
	age		sexe
	niveau d'études		age
	formation		niveau d'études
	Profession		formation
	communauté professionnelle		profession
	domaines expertises		communauté professionnelle
	langue(s) lue(s)		domaines expertises
activité du besoin	type d'activité liée à la recherche	**document**	langue du doc
	organisme liée à l'activité		date de création
			taille du document
besoin	type de besoin		structure logique
			format numérique
usage cognitif	type d'apprentissage		granularité du document
			forme du document
			type discours dominant
temporel	contrainte de temps	**source**	
		contexte édition initiale	environnement éditorial
thème du besoin d'information	Domaine disciplinaire recherché		organisme éditeur
	thème de recherche (requête)		
		thème du document	domaine disciplinaire du document
			descripteurs

Tableau 1. Variables décrivant l'utilisateur et le document dans leur contexte

[4] Google, Yahoo recherche, Altavista, Hotbot.

3. La fonction de correspondance du SRI ProfilDoc

Pour construire notre fonction de correspondance, nous avons ensuite émis des hypothèses de relations de dépendance entre ces variables. Nous avons alors testé la validité de nos relations au cours d'une expérience [13] [14] de recherche d'information réelle effectuée par 50 participants. La méthodologie a consisté à récolter des couples (requête, document jugé pertinent pour cette requête), décrire ces couples par des variables et enfin calculer statistiquement les relations de dépendances entre ces variables. Le tableau 2 indique les relations trouvées entre les catégories de variables.

document / besoin	contexte				
	auteur	document	édition	thème	total
utilisateur	2	0	0	0	2
activité	7	1	2	0	**10**
usage	4	1	1	0	**6**
thème	1	0	0	0	1
total	**14**	2	3	0	**19**

Tableau 2. Dépendances entre catégories de variables

Les résultats indiquent que les relations concernant l'usage, l'activité motivant la recherche d'information de l'utilisateur et l'auteur du document sont particulièrement dépendantes.
Nous souhaitons à présent construire notre fonction de correspondance à partir des relations de dépendances obtenues et les évaluer.
Les relations entre variables sont dues aux liaisons fortes entre les modalités des deux variables. Suite aux résultats de l'expérimentation, quatre variables de la représentation du besoin sont prises en compte : " activité liée à la recherche ", " organisation liée à l'activité ", " profession de l'utilisateur ", " type d'apprentissage ".
En examinant les relations au niveau de leurs modalités, nous pouvons observer des liaisons fortes d'attirance et/ou de répulsion entre deux modalités (distance du chi deux partiel ou distance euclidienne dans une Analyse Factorielle des Correspondances). Le cumul de ces liaisons détermine l'indépendance ou non de la relation entre deux variables.
Ainsi, à partir de l'observation des relations dépendantes, nous avons observé au total 65 liaisons d'attraction ou de répulsion. A titre d'exemple, la variable " profession de l'utilisateur " totalise 10 liaisons fortes. Parmi celles-ci, la modalité " chercheur " est impliquée dans 4 liaisons : attirance avec la modalité " chercheur " de la variable " profession de l'auteur " (profession auteur, chercheur), répulsion avec (profession auteur, indéterminé), attraction avec (communauté professionnelle, universitaire), répulsion avec (communauté professionnelle, associatif).
Au final, toutes les liaisons sont intégrées dans la fonction de correspondance de notre prototype. Celle-ci, en fonction de la représentation de du besoin de l'utilisateur et la représentation du document, effectue un filtrage en fonction des relations contextuelles entre les deux représentations. Associé à un filtrage classique sur le sujet du document, le filtrage contextuel permet d'affiner la sélection des documents réponses. Concrètement, le filtrage est basé sur une formule mathématique qui calcule un indice d'attraction entre deux variables. Nous voyons plusieurs possibilités pour calculer cette fonction " score ".
Stratégie 1 : La stratégie la plus simple pour construire notre fonction est de compter le nombre de liaisons entre les modalités d'un besoin d'information et les modalités des métadonnées du document réponse à évaluer. Ainsi, plus une représentation de document possède de liaison avec la représentation du besoin, plus un score élevé lui sera attribué. Par exemple, une recherche d'information avec 10 liaisons attractives fortes obtient un score de

10. Une recherche d'information avec 3 liaisons attractives fortes obtient un score de 3 et ainsi de suite. Les documents réponses sont ensuite classés dans l'ordre décroissant de leur score.

$$Score = \sum A(q,d)$$

Stratégie 2 : La première stratégie a le mérite de la simplicité mais ne prend pas en compte les liaisons de répulsion. Cette stratégie calcule le score d'un document en additionnant les liaisons attractives et en soustrayant les liaisons répulsives. Formellement, nous avons la formule suivante :

$$Score = \sum A(q,d)$$

Score d = somme (Attraction) – somme (Répulsion)

D'autres stratégies sont bien entendu possibles en augmentant les paramètres et la complexité de la formule. Mais parce que ce travail aurait pu être un sujet d'étude en soi, nous nous limitons pour cette évaluation à l'utilisation de la stratégie numéro 2.

4. Evaluation de la fonction de correspondance

1. Méthodologie

A partir des liaisons entre modalités, nous avons construit notre fonction de correspondance. A présent, nous souhaitons tester et évaluer cette fonction et donc, par conséquent, l'exploitation des relations par le SRI ProfilDoc. Pour cela, nous proposons de comparer le classement des documents réponses d'un SRI témoin (le moteur de recherche web " Google ") avec le classement du SRI ProfilDoc. Le classement du SRI ProfilDoc repose sur un score construit à partir des liaisons entre modalités de variables dépendantes.
Notre test d'évaluation se décompose en plusieurs étapes :

1. Sélection des documents réponses : Nous sélectionnons à partir d'un moteur de recherche existant, un ensemble de documents répondant à un besoin d'information. Par exemple, les 50 premiers documents réponses pour la requête " réseaux locaux " correspondant au besoin " construction d'un cours sur les réseau locaux ". Ce moteur de recherche a classé les documents du plus pertinent au moins pertinent selon sa fonction de correspondance.
2. Sélection des documents pertinents. Dans cette phase, il s'agit de repérer les documents réellement pertinents. Nous demandons à l'utilisateur participant à l'évaluation et producteur de la requête de noter les documents pertinents pour son besoin. A l'aide de cette information, nous calculons deux indicateurs : la précision de l'ensemble des documents réponses et la position des documents pertinents dans le classement effectué par le SRI témoin.
3. Construction de la représentation étendue des documents réponses. Cette phase est une préparation au filtrage ProfilDoc. Il s'agit de construire manuellement la représentation étendue (sur le contexte) de chaque document réponse. Il s'agit d'associer, sous formes de métadonnées, les couples (variable, modalité) à chaque document. Néanmoins, seules les variables jugées significatives par l'expérimentation seront prises en compte, c'est-à-dire en majorité les variables relatives à l'auteur du document et la variable " organisme éditeur ".
4. Evaluation. Elle repose sur deux outils :
   (a) La comparaison visuelle des deux classements obtenus.
   (b) La comparaison des courbes rappel/précision

1. Calcul d'une courbe d'évaluation

Pour évaluer un SRI, la courbe rappel/précision est couramment utilisée. Cependant, cet outil nécessite un grand nombre de requêtes pour évaluer un système. Il n'est pas adapté à notre contexte où nous n'avons que quatre requêtes pour l'évaluation. Nous proposons donc de construire une évaluation pour représenter une seule requête dont le paramètre variable est le nombre de documents pertinents pris en compte pour calculer le rappel/précision.

Ainsi, notre courbe d'évaluation représente une requête. Chaque point représente un calcul (précision/rappel) pour un ensemble de documents réponses. Cet ensemble est variable en fonction du nombre de documents pertinents inclus dans l'ensemble. Chaque ensemble est un sous ensemble de l'ensemble total des documents réponses (50 au total). Le premier ensemble intègre le document classé au rang 1 et les documents des rangs suivant jusqu'au rang du premier document jugé pertinent. Le deuxième ensemble débute au document classé au rang 1 et termine au rang du deuxième document jugé pertinent. Etc... Au total, il existe autant d'ensembles que de documents jugés pertinents par l'utilisateur et leur taille est croissante. L'ensemble des couples de valeurs (précision/rappel) forme une courbe dans un tableau précision / rappel.

D'une manière plus formelle, nous définissons :

N = ensemble des 50 documents réponses, $N = Card(N) = 50$.

p = nombre de document pertinent.

Pour $i \in [1..p]$ nous définissons E_i le plus petit ensemble des premiers documents du classement contenant i documents pertinents. Alors, la courbe est constituée de p points $(x_i, y_i), (x_2, y_{2i}), ..., (x_p, y_p)$

où chaque point (x_i, y_i) $(i \in [1..p])$ de la courbe est calculé de la façon suivante:

$$x_i = \frac{i}{Card(E_i)} \quad et \quad y_i = \frac{i}{p}$$

Quatre tests ont été effectués au total. Ils correspondent à quatre besoins d'information d'utilisateurs.

2. Résultats

1. Test 1 : requête " réseaux locaux "

Notre premier test concerne le besoin d'information lié à la construction d'un cours général sur les réseaux locaux. L'utilisateur est un doctorant qui exerce une fonction d'enseignement conjointement à son activité de recherche. Nous créons le profil de l'utilisateur en affectant les modalités correspondant à son besoin. Le thème (ou sujet) du besoin n'est pas pris en compte dans le profil.

Profil de l'utilisateur (tableau 3):

Activité	Cours
Organisation liée à l'activité	Universitaire
Profession de l'utilisateur	Chercheur
Type d'apprentissage	Approfondissement

Tableau 3. Représentation du besoin de l'utilisateur

En fonction de ce profil, nous construisons un tableau des modalités attractives et répulsives des variables sur le document. Le tableau 4 récapitulatif des modalités à prendre en compte par rapport au profil donné.

Variables	**Modalités attractives**	**Modalités répulsives**
Communauté professionnelle	Universitaire	Association
Forme discursive du document	Argumentatif	Descriptif
Organisme éditeur	Individuel, presse, scientifique	Individuel
Profession de l'auteur	Autres, chercheurs	Non réponse
Typé d'auteur	Individuel	Organisation
Niveau d'études	3ème cycle	indéterminé

Tableau 4. Variables et modalités du profil utilisateur pour la requête " réseaux locaux "

Etape 1 : 50 documents ont été sélectionnés à partir du moteur " Google "
Requête : " réseaux sans fils " Nous récupérons les 50 premiers URL retournés pas Google.
Etape 2 : Parmi les 50 documents, l'utilisateur évalue les documents réellement pertinents. Après lecture, 8 documents sont retenus. Le tableau 4 montre les 50 documents retournés par Google ainsi que les documents pertinents surlignés en gris. Le classement des documents dans le tableau est issu du classement Google.

Figure 5. Classement " Google " pour le test 1
Nous pouvons observer que les documents pertinents ne sont pas classés en tête de classement mais sur l'ensemble du classement en général et notons particulièrement deux documents pertinents en fin de classement (document 48 et 49).
Etape 3 : les 50 documents ont été enrichis en métadonnées.
Etape 4 : évaluation Pour chaque document réponse, nous avons compté les liaisons d'attraction entre les modalités et les liaisons de répulsion. Le score total est calculé selon la formule suivante :
Score = total attraction - total répulsion.
La figure 6 indique le score ProfilDoc de chaque document réponse et classe les documents dans l'ordre décroissant du score. Les documents réponses sont surlignés en gris.

numéro doc	URL document
1	http://prix.materiel.be/liste/130/
2	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html
3	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html
4	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html
5	http://www.commentcamarche.net/ccmbugs/?Bug_url=%2Fwireless%2Fwlintro.php3
6	http://www.commentcamarche.net/wireless/wlintro.php3
7	http://www.supinfo-projects.com/2002/reseaux_sans_fil_securite/
8	http://www.rd.francetelecom.fr/fr/technologies/ddm200207/dossier.php
9	http://www.rd.francetelecom.fr/fr/technologies/ddm200207/techfiche3.php
10	http://www.planet.ch/services/wireless2.asp
11	http://www.planet.ch/services/wireless2.asp
12	http://www.rue-hardware.com/prix/details/5592/
13	http://www.rue-hardware.com/prix/details/5592/
14	http://www.domobox.fr/
15	http://www.xasa.com/grupos/fr/thread.php?group=fr.comp.reseaux.sans-fils
16	http://www.xasa.com/grupos/fr/1.htm
17	http://www.linux-gull.ch/projets/reseau/mail/msg00035.html
18	http://www.rue-montgallet.com/prix/75/details/5605/
19	http://www.rue-montgallet.com/prix/75/details/5605/
20	http://www.digital-home-concept.com/devenir_revend_gb.asp
21	http://www-rp.lip6.fr/infradio/inf-prog/Mairie_de_Paris/1
22	http://www-rp.lip6.fr/site_rp/reunions_aff.php3?num=4&langue=fr
23	citi.insa-lyon.fr/powerpoint/ Katia_La%20Planif%20WLAN%20_%20conf_thesards.pp
24	citi.insa-lyon.fr/~iguerinl/sujet_DEA_freq.html
25	http://www.bag.admin.ch/strahlen/nonionisant/emf/h_frequence/f/rlan.php
26	http://www.sans-fils.org/index.php
27	http://www.sans-fils.org/dossiers/article.php3?id_article=40
28	http://www.atollplus.fr/dossiers/wifi.php
29	http://rev.inrialpes.fr/intech/Registration?op=511&meeting=15
30	http://rev.inrialpes.fr:8080/intech/2003-06-26/Dynetcom.ppt
31	www.cccure.net/modules.php?name=Topics
32	http://www.abcs-international.fr/Wifi/Reseaux_radio.htm
33	http://www.macwifi.com/index.php?cat=12
34	http://www.rennes-wireless.org/
35	http://blog.netpartoo.com/index.php/2003/10/02
36	http://blog.netpartoo.com/index.php/2003/10/22
37	http://www.ulb.ac.be/polytech/elecgen/
38	http://www.silicon.fr/click.asp?id=904
39	http://www.swissup.com/art_content.cfm?upid=FR3133
40	http://www.swissup.com/art_content.cfm?upid=FR3112
41	http://www.indexel.net/listedoc.jsp?categorie=104&scategorie=260
42	http://240plan.ovh.net/~angerswi/v3/
43	http://www.ges.fr/wireless/wireless-carte-PCI.asp
44	http://www.ges.fr/wireless/wireless-carte-PCI.asp
45	http://www.dupuis-informatique.ch/p582.html
46	www.sos-pc.ch/prix/connexion-reseaux.asp
47	www.ldlc.fr/recherche/ id_71161_s_souris_logitech_sans_fils_noire.html
48	http://solutions.journaldunet.com/0111/011115_faq_sansfil.shtml
49	http://solutions.journaldunet.com/itws/031125_it_smartmedia.shtml
49	http://www.insa-lyon.fr/Departements/TC/Ects/5/5TC-MOB.html
50	http://www.ina.fr/formation/fiche.php?ID=1332&Filiere=TE

Figure 5. Classement " Google " pour le test 1

numéro doc	URL document	TOTAL attraction	TOTAL répulsion	SCORE
6	http://www.commentcamarche.net/wireless/wlintro.php3	6	0	**6**
7	http://www.supinfo-projects.com/2002/reseaux_sans_fil_securite/	6	1	**5**
22	http://www-rp.lip6.fr/site_rp/reunions_aff.php3?num=4&langue=fr	5	1	**4**
39	http://www.swissup.com/art_content.cfm?upid=FR3133	5	1	**4**
40	http://www.swissup.com/art_content.cfm?upid=FR3112	5	1	**4**
21	http://www-rp.lip6.fr/infradio/inf-prog/Mairie_de_Paris/1	4	0	**4**
30	http://rev.inrialpes.fr:8080/intech/2003-06-26/Dynetcom.ppt	3	1	**2**
48	http://solutions.journaldunet.com/0111/011115_faq_sansfil.shtml	3	1	**2**
49	http://solutions.journaldunet.com/itws/031125_it_smartmedia.shtml	3	1	**2**
49	http://www.insa-lyon.fr/Departements/TC/Ects/5/5TC-MOB.html	3	2	**1**
38	http://www.silicon.fr/click.asp?id=904	1	0	**1**
25	http://www.bag.admin.ch/strahlen/nonionisant/emf/h_frequence/f/rlan.php	4	4	**0**
29	http://rev.inrialpes.fr/intech/Registration?op=511&meeting=15	4	4	**0**
37	http://www.ulb.ac.be/polytech/elecgen/	4	4	**0**
17	http://www.linux-gull.ch/projets/reseau/mail/msg00035.html	1	2	**-1**
2	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html	0	2	**-2**
3	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html	0	2	**-2**
4	http://www.usenet-fr.net/fur/chartes/comp.reseaux.sans-fils.html	0	2	**-2**
5	http://www.commentcamarche.net/ccmbugs/?Bug_url=%2Fwireless%2Fwlintro.php3	0	2	**-2**
8	http://www.rd.francetelecom.fr/fr/technologies/ddm200207/dossier.php	0	2	**-2**
9	http://www.rd.francetelecom.fr/fr/technologies/ddm200207/techfiche3.php	0	2	**-2**
10	http://www.planet.ch/services/wireless2.asp	0	2	**-2**
11	http://www.planet.ch/services/wireless2.asp	0	2	**-2**
14	http://www.domobox.fr/	0	2	**-2**
15	http://www.xasa.com/grupos/fr/thread.php?group=fr.comp.reseaux.sans-fils	0	2	**-2**
16	http://www.xasa.com/grupos/fr/1.htm	0	2	**-2**
23	citi.insa-lyon.fr/powerpoint/ Katia_La%20Planif%20WLAN%20_%20conf_thesards.ppt	0	2	**-2**
24	citi.insa-lyon.fr/~iguerinl/sujet_DEA_freq.html	0	2	**-2**
28	http://www.atollplus.fr/dossiers/wifi.php	0	2	**-2**
31	www.cccure.net/modules.php?name=Topics	0	2	**-2**
32	http://www.abcs-international.fr/Wifi/Reseaux_radio.htm	0	2	**-2**
33	http://www.macwifi.com/index.php?cat=12	0	2	**-2**
36	http://blog.netpartoo.com/index.php/2003/10/22	0	2	**-2**
41	http://www.indexel.net/listedoc.jsp?categorie=104&scategorie=260	0	2	**-2**
44	http://www.ges.fr/wireless/wireless-carte-PCI.asp	0	2	**-2**
46	www.sos-pc.ch/prix/connexion-reseaux.asp	0	2	**-2**
47	www.ldlc.fr/recherche/ id_71161_s_souris_logitech_sans_fils_noire.html	0	2	**-2**
50	http://www.ina.fr/formation/fiche.php?ID=1332&Filiere=TE	0	2	**-2**
43	http://www.ges.fr/wireless/wireless-carte-PCI.asp	1	4	**-3**
1	http://prix.materiel.be/liste/130/	0	3	**-3**
12	http://www.rue-hardware.com/prix/details/5592/	0	3	**-3**
13	http://www.rue-hardware.com/prix/details/5592/	0	3	**-3**
18	http://www.rue-montgallet.com/prix/75/details/5605/	0	3	**-3**
19	http://www.rue-montgallet.com/prix/75/details/6271/	0	3	**-3**
20	http://www.digital-home-concept.com/devenir_revend_gb.asp	0	3	**-3**
45	http://www.dupuis-informatique.ch/p582.html	0	3	**-3**
26	http://www.sans-fils.org/index.php	0	4	**-4**
27	http://www.sans-fils.org/dossiers/article.php3?id_article=40	0	4	**-4**
34	http://www.rennes-wireless.org/	0	4	**-4**
35	http://blog.netpartoo.com/index.php/2003/10/02	0	4	**-4**
42	http://240plan.ovh.net/~angerswi/v3/	0	5	**-5**

Figure 6. Classement « ProfilDoc » pour le test 1

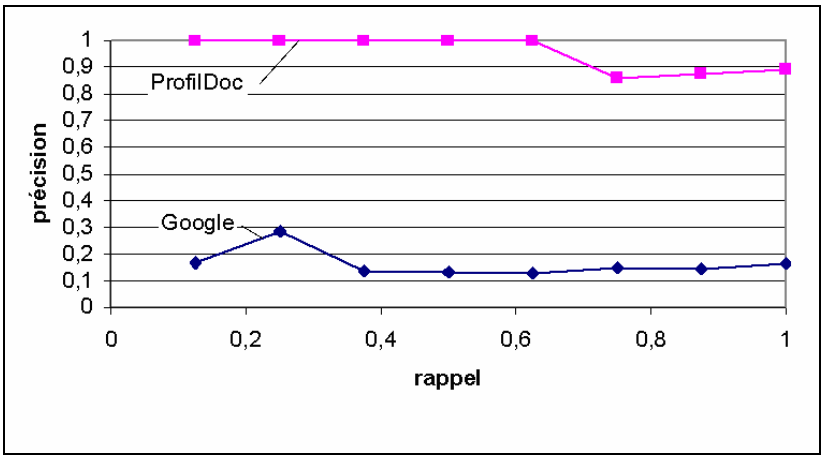

Figure 7. Courbe d'évaluation du test 1

Nous constatons, en tête de classement, les 8 documents réponses parmi les 9 premières places du classement. La courbe (figure 7) synthétise ces résultats et indique une courbe au dessus (évaluation supérieure) pour le SRI "profildoc".

Pour consolider ce premier résultat concluant, nous avons réalisé trois tests supplémentaires auprès de trois utilisateurs différents mais de profil semblable (chercheur) et pour un type d'activité identique : réalisation d'un état de l'art. Voici les résultats de ces tests.

2. Test 2 : recherche d'information sur la requête :" zipf's law entropy " (loi d'entropie de Zipf)

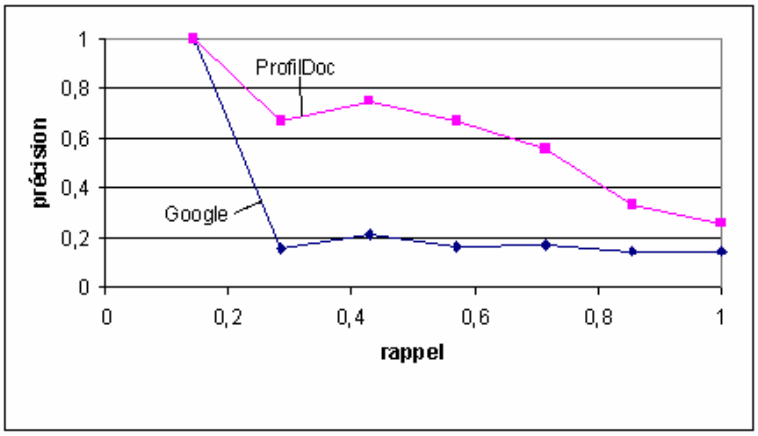

Figure 8. Courbe d'évaluation du test 2

Ce graphique (figure 8) indique les courbes rappel/précision d'une recherche d'information d'un utilisateur. Nous constatons que la courbe en rose (point carré) est toujours supérieure à la courbe bleue (représentant " Google "). Cela signifie que le classement du SRI " ProfilDoc " est constamment supérieur à celui de l'autre SRI

3. Test 3 : recherche sur la requête : " segmentation (thématique OR thème) texte"

Remarque : cette requête inclut un opérateur booléen (OR).

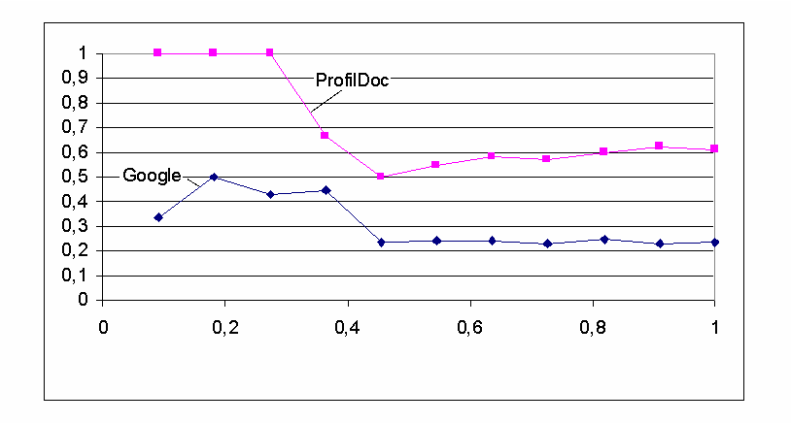

Figure 9. Courbe d'évaluation du test 3

Cette courbe (figure 9) indique un résultat identique à la courbe précédente : une performance supérieur du classement " ProfilDoc " sur celui du SRI témoin.

4. Test 4: recherche sur la requête : " livre électronique dans les universités américaines "

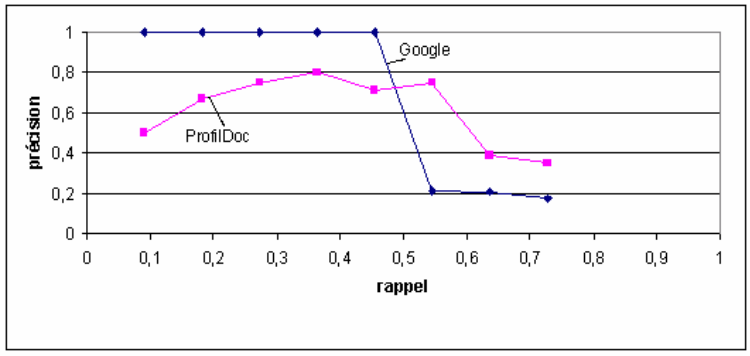

Figure 10. Courbe d'évaluation du test 4

La figure 10, indique un résultat partagé. En effet, on constate un renversement de la position des deux courbes. Dans la première partie du tableau, le " SRI " Google est plus performant mais décroît fortement lorsque le rappel dépasse les 0.5. A partir de ce point, la courbe de " ProfilDoc " plus constante, passe et se maintient au-dessus de la courbe concurrente. Après une analyse du classement " ProfilDoc " par l'utilisateur, celui-ci révèle que le premier document n'est pas pertinent car est hors sujet, le thème ne correspond pas. En fait, c'est la conséquence d'un nombre faible de documents réponses sur le sujet. Les SRI commerciaux comme " Google ", dans ce cas de figure, élargissent leurs critères de filtrage pour augmenter le nombre de réponses comme par exemple la tolérance sur la proximité des termes. C'est ce qui ce passe dans ce cas précis et l'on retrouve en fin de classement des documents hors sujet, qui contiennent bien le mot " livre " et le mot " électronique " mais pas dans la même phrase : le document ne parle aucunement du livre électronique mais du " livre papier " et de " l'électronique des transistors ". Ainsi, sans tenir compte de ce document hors sujet, nous obtenons une courbe globalement supérieure pour le SRI " Profildoc ".
Nous retrouvons une courbe " rappel/précision " du SRI " ProfilDoc " globalement supérieure à la courbe concurrente.

Conclusion

En conclusion, la prise en compte d'une représentation étendue par notre fonction de correspondance a entraîné une augmentation visible des performances du SRI " ProfilDoc " par rapport au SRI témoin. Nous expliquons ce résultat par l'augmentation de la précision des représentations : celle du besoin d'information et celle du document. Une meilleure précision des représentations entraîne une correspondance plus précise. Le score ainsi calculé permet de distinguer plus précisément les documents réponses entre eux.

Néanmoins, certaines réserves sont à observer :

- Les relations entre les variables sont issues d'une population restreinte dans un contexte spécifique (celui de l'expérience) et dans un espace-temps fixé. Le calcul du score est directement lié aux relations obtenues pendant l'expérience. Celles-ci ne peuvent être généralisables et intégrées définitivement dans la fonction de correspondance.
- Certaines relations de dépendances sont absentes (date, objet à construire,...). Elles n'ont pu être observées par manque de données. La fonction de correspondance ProfilDoc est donc incomplète.
- Les relations sont observées à partir d'un référentiel collectif. Les particularités individuelles ne sont pas prises en compte. C'est sans doute un des points les plus important de nos résultats. La réduction du bruit documentaire passe par la personnalisation d'une recherche d'information. Par extension nous pensons que le SRI doit prendre en compte les particularités individuelles de l'utilisateur. Or les relations obtenues sont issues d'une moyenne statistique sur un groupe d'utilisateurs. Ce sont des relations collectives et non individuelles. Il est tout à fait envisageable d'obtenir un ensemble de relations spécifiques à un utilisateur au prix d'une observation individuelle.
- La présentation étendue du document (contexte de production et de diffusion) a été recherchée et implémentée manuellement pour chaque document. C'est une solution coûteuse en temps et inappropriée pour un corpus documentaire de grande échelle.

Suite à ces observations, le passage à l'échelle du prototype ProfilDoc passe par la résolution des limites précédentes. En particuliers, il nous semble que la construction automatique de métadonnées est une étape décisive pour l'exploitation à l'échelle de ce projet. La présence progressive de métadonnées dans les documents numériques est une solution possible. Plusieurs projets de recherche proposent une nomenclature pour définir ce que les métadonnées décrivent et comment les intégrer dans un document. Le projet Dublin Core [15] propose un standard de métadonnées pour décrire un document web. Le projet Du W3C appelé " web sémantique "[16] propose des outils (RDF : Resource Description Framework pour la création de relation entre entités, URI : Uniform Resource Identifier pour identifier une identité, RDF Schema pour construire un ensemble de relations entre entités (document, auteur, organisation, thème,...). Cependant, l'intégration des métadonnées dans un document par son auteur est aléatoire. Une autre solution, plus solide, est la construction de métadonnées au moment de l'indexation. Les récents travaux de Prime-Claverie, sur la propagation des métadonnées [17] d'un document référence à d'autres documents aux caractéristiques proches de celui-ci, sont un exemple parmi les voies à explorer pour associer des métadonnées automatiquement à des documents numériques.

References

[1] Bergman, M. K.-The Deep Web: Surfacing Hidden Value.-in The Journal of Electronic Publishing Vol.7, n°1, 2001.

[2] Chiaramella, Y., Defude, B., Bruandet, M.-F. and Kerkouba, D. - IOTA : a full text information retrieval system - in: Conference on Research and development in Information Retrieval, Pisa - pp. 207-213.

[3] Croft, W. B. and Thompson, R. H.-I3R: A new approach to the design of document retrieval systems.-in Journal of the American Society for Information Science and Technologie Vol.38, 1987.-pp.389-404.

[4] Chen, H.-Knowledge-based document retrieval: framework and design.-in Journal of Information Science, Principles & Practice Vol.18, n°4, 1992.-pp.293-314.

[5] Lainé-Cruzel, S.-Conception de systèmes de recherche d'informations : accès aux documents numériques scientifiques.- 2001. Habilitation à diriger des Recherches - Université Claude Bernard Lyon 1

[6] Ben Abdallah, N.-Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information: vers un système d'information évolué.- 1997. Thèse de doctorat - Université Lyon 1

[7] Michel, C.-Evaluation de Système de Recherche d'Information, comportant un fonctionnalité de filtrage, par des mesures endogènes. Réalisation et évaluation d'un prototype de Système de Recherche d'Information avec filtre selon les profils des utilisateurs.- 1999. Thèse de doctorat - Université Lumière Lyon 2

[8] Lainé-Cruzel, S., Lafouge, T., Lardy, J. P. and Ben Abdallah, N.-Improving information retrieval by combining user profile and document segmentation.-in Information Processing and Management Vol.32, n°3, 1996.-pp.305-315.

[9] Lainé-Cruzel, S.-Documents, ressources, données: les avatars de l'information numérique.-in revue I3 Vol.4, n°1, 2004.-pp.105-120.

[10] Wilkinson, R. - User Modeling for Information Retrieval on the Web - in: 2nd workshop on adaptative systems and user modeling on the WWW, - pp. 117-119.

[11] TREC.-Text REtrieval Conference- - http://trec.nist.gov/

[12] Allan, J. - HARD Track Overview in TREC 2003 High Accuracy Retrieval from Documents - in: proceedings of the Twelfth Text REtrieval Conference (TREC 2003), 2003, Gaithersburg, Maryland - pp.

[13] Perenon, P. - Relation entre le contexte de la requète et le contexte de production du document - in: INFORSID, biarritz, 2004 - pp. 423-438.

[14] Perenon, P.-Profil d'utilisateur et métadonnées associés dans un système de recherche d'information scientifique.- 2004. thèse de doctorat - Université Lyon 1

[15] OCLC.-Dublin Core- - http://dublincore.org/

[16] W3C.-Semantic Web- 2001. - http://www.w3.org/2001/sw/

[17] Prime-Claverie, C., Beigbeder, M. and Lafouge, T. - Propagation de métadonnées par l'analyse des liens - in: Journées Francophones de la Toile, 30 juin, 1 et 2 juillet 2003, Tours - France - pp. 257-264.