Hyperliens et recherche d'information sur le Web

Jacques Savoy, Yves Rasolofo

Institut interfacultaire d'informatique
Université de Neuchâtel – Pierre-à-Mazel 7 – 2000 Neuchâtel (Suisse)
Jacques.Savoy@unine.ch

Abstract

Today search engines are very useful tools for retrieving information on the Web. When submitting a query however users may not always be interested in retrieving long lists of sites; rather they may prefer only a single Web site (e.g., when looking for a given homepage, an on-line service or a named page). Or, they may be only seeking useful starting points from which to browse a given topic; thus the system should respond with a list of key resources on a given subject (topic distillation). This paper describes our search system that would respond to these two search types and how its reliance on hyperlinks enhances retrieval effectiveness. Using a relatively large Web test collection (18.1 GB), an evaluation of our strategy indicates that the presence of hyperlinks improves search precision and that our system provides performance levels superior to those of other models.

Résumé

Les moteurs de recherche s'avèrent des outils indispensables afin de retrouver de l'information sur Internet. Cependant, les usagers ne désirent pas toujours une liste de sites répondant à leurs requêtes mais souhaitent obtenir un site comme unique réponse (recherche d'une page d'accueil ou d'un service en-ligne). D'autre part, nous voulons parfois extraire une liste des ressources clés ayant trait à un domaine précis. Dans ce cas, nous ne cherchons pas toutes les pages sur un thème mais souvent des bons points de départ pour la navigation. Cet article décrit notre système de recherche capable de répondre à ces deux types d'interrogation en recourant aux hyperliens afin d'accroître la qualité des réponses. Une évaluation basée sur une collection de pages Web (18,1 GB) indiquent que la présence des hyperliens permet d'accroître la qualité du dépistage.

Mots-clés : recherche d'information, Web, recherche de service en-ligne, recherche de ressources clés.

1. Introduction

La présence de moteurs de recherche de bonne qualité a certainement été l'un des facteurs qui a permis à Internet de grandir tout en permettant un accès à un volume très important de documents. Des études récentes (Spink *et al.*, 2001) démontrent que les requêtes adressées à ces moteurs varient avec les années passant d'un Internet ludique à un Web plus commercial. Cependant, ces interrogations s'expriment toujours avec un nombre très restreint de mots. De plus, les internautes restent peu soucieux du rappel c'est-à-dire de dépister un grand nombre de pages pertinentes à leur requête. Ils attachent une grande importance à la simplicité d'accès (Nielsen, 2000) et à un temps de réponse minimal.

Actuellement, les usagers d'Internet attendent plus des moteurs de recherche. Par exemple, en réponse à la demande « cinéma Paris », le moteur de recherche devrait proposer une liste des films à l'affiche correspondant au lieu indiqué explicitement ou implicitement, tout en tenant compte de la date et de l'heure de l'émission de la requête (Watters et Amoudi, 2003). Parfois, les internautes ne désirent pas des pages relatives au thème indiqué, mais une réponse précise comme, par exemple, pour la demande « Quelle est la hauteur de la tour Eiffel ? » ou « Qu'est-ce qu'un neutron ? » (Radev *et al.*, 2002). De plus, on sait souvent que la réponse

attendue correspond à un site précis, comme lorsque l'on recherche une page personnelle, le service de réservation d'une compagnie de chemin de fer ou la page d'accueil d'une entreprise (*homepage and named page searching*). Enfin, on souhaite parfois que le moteur de recherche puisse extraire d'Internet une liste de sites proposant un contenu très pertinent ou un bon point de départ pour la navigation sur un thème donné (*topic distillation*).

Afin de répondre à ces deux derniers types d'interrogation, nous avons développé un moteur de recherche spécialement adapté à la recherche de pages précises ou capable d'extraire les ressources clés du Web sur un thème donné. La deuxième section explique en grandes lignes les principes de notre système de recherche. La troisième section décrit avec plus de détails notre stratégie de dépistage de sites spécifiques et évalue nos propositions avec l'aide d'une collection de pages du Web (1'247'753 pages pour un volume d'environ 18,1 GB). La quatrième section expose notre système d'extraction des bonnes pages de départ pour la navigation, évalue notre proposition et la compare à diverses autres stratégies.

2. Idée de base de notre stratégie de dépistage

Afin de concevoir un moteur de recherche d'information efficace pour le Web, des études précédentes ont démontré que le modèle probabiliste Okapi (Robertson *et al.*, 2000) possède une bonne qualité de réponse (Savoy et Picard, 2001). Cependant, dans ces expériences, il s'agissait de trouver tous les documents répondant à une requête. Dans le cas présent, nous ne désirons pas retourner toutes ces pages mais seulement un site correspondant au service ou contenant l'information exigée. Ainsi, face à la requête « Air France », notre système doit retourner la page permettant une réservation en-ligne auprès de cette compagnie aérienne et non pas un ensemble de documents concernant de près ou de loin cette compagnie.

Dans ce but, nous avons décidé de modifier ce modèle probabiliste afin de pouvoir tenir compte de plusieurs représentations (mais au maximum trois) des pages Web. Ainsi, nous pourrions, par exemple, générer une première indexation des pages basée uniquement sur les balises <TITLE> et <META>, une deuxième indexation sur les étiquettes <H*i*> et une troisième sur tout le texte. Lors de l'estimation du score d'une page D_i, valeur dénotée $RSV(D_i)$, nous calculons une somme pondérée du score obtenue par chaque représentation, soit :

$$RSV(D_i) = \alpha \cdot \sum_{j=1}^{m} w_{ij}^{(1)} \cdot qw_j + \beta \cdot \sum_{j=1}^{m} w_{ij}^{(2)} \cdot qw_j + \gamma \cdot \sum_{j=1}^{m} w_{ij}^{(3)} \cdot qw_j \quad (1)$$

dans laquelle $w^{(1)}_{ij}$ indique le poids attribué au terme t_j dans le document D_i selon la première représentation ($w^{(2)}_{ij}$ et $w^{(3)}_{ij}$ indique cette même pondération relative à la deuxième, respectivement, troisième représentation), et les paramètres α, β, γ indiquent l'importance accordée à chaque représentation. Enfin, le poids du terme t_j dans la requête est indiqué par qw_j et m indique le nombre de mots communs entre le document et la requête.

Habituellement en recherche d'information, il se révèle utile de permettre des appariements entre des formes variant en genre ou en nombre. Ainsi, si un article possède le terme « chevaux », il est raisonnable de penser qu'une requête contenant le mot « cheval » devrait pouvoir dépister cette page. Comme notre moteur doit posséder une haute précision (c'est-à-dire être capable de retourner uniquement des sites correspondant aux critères de la requête), nous avons renoncé à un enracineur éliminant la plupart des dérivations morphologiques (e.g., « reliability »→« reliable »). Nous avons inclus un traitement morphologique simple supprimant la marque du pluriel soit le « s » final pour l'anglais (Harman, 1991).

Pour encore améliorer la qualité des réponses, nous avons également inclus une fonction de proximité entre termes (pour les détails, voir (Rasolofo et Savoy, 2003)). L'idée de base est la suivante. Si tous les termes de la requête apparaissent dans un document, le score attribué à cette page doit être augmenté. De plus, si ces termes se présentent proches les uns des autres ou s'ils apparaissent fréquemment, le score sera accru. À l'inverse, si seulement une partie de ces termes apparaissent, le score n'est pas modifié. Sachant que les requêtes soumises par les internautes sont très courtes (en moyenne 2,4 mots), ce calcul de proximité se limite souvent à contrôler si deux termes apparaissent bien dans un document donné.

Afin de savoir si l'emploi de cette composante s'avère bénéfique ou non, nous avons inclus le paramètre δ dans notre modèle. Si $\delta = 0$, cette fonction de proximité est ignorée ; par contre, si $\delta > 0$, la valeur de la fonction de proximité entre termes, multipliée par δ, sera ajoutée au score de la page Web calculé selon l'équation 1.

3. Dépistage des pages d'accueil et services en-ligne

Le dépistage d'une page unique correspond à une demande fréquente sur Internet. Ainsi, la réponse appropriée aux requêtes « sncf », « postes ouverts à la Cour suprême », « barbara mikulski », ou « observatoire du Mont-Blanc » ne correspond pas un ensemble de pages Web mais à accès direct au service (réservation en-ligne, consultation d'horaires), à la liste (films à l'affiche, postes ouverts) ou à la page d'accueil ou personnelle spécifiée.

3.1. *Notre modèle de recherche*

Afin d'atteindre une grande précision, notre moteur de recherche s'appuie sur le modèle Okapi avec trois représentations pour chaque document (voir section 2). La première indexation s'appuie sur l'ensemble du texte d'une page Web, ensemble comprenant également les phrases comprises dans les balises <TITLE> et les étiquettes <META>.

Cette première indexation repose uniquement sur le contenu d'une page et elle ignore les hyperliens indiquant des relations sémantiques entre pages. Par exemple, différentes pages à travers le monde pointent vers le département d'informatique de l'Université de Montréal (www.iro.umontreal.ca) et ces hyperliens sont indiqués dans ces pages sous la forme « la <href = "www.iro.umontreal.ca"> linguistique computationnelle possède un degré … ». Nous avons fait l'hypothèse que le texte encadré par la balise d'ancrage de l'hyperlien (soit « linguistique computationnelle » dans l'exemple précédent) décrit de manière concise et précise le contenu sémantique de la page pointée. Notons que ce texte possède aussi l'avantage d'être écrit par d'autres personnes que l'auteur de la page référencié, autorisant ainsi un enrichissement du vocabulaire d'indexation. En s'appuyant sur cette hypothèse, nous avons construit une deuxième indexation de chaque document sur la base de tous les textes d'ancrages d'une part et, d'autre part, du champ <TITLE> des pages pointant vers la page courante. Ainsi si la page P_k pointe vers P_i, le titre de P_k ainsi que le texte d'ancrage annonçant la présence d'un lien vers P_i sera inclus dans la deuxième représentation de P_i. Comme la page P_k n'est pas unique, nous concaténons les éléments extraits des pages P_k pour construire le représentant de P_i. Notre hypothèse s'appuie des études précédentes démontrant que les textes d'ancrage des hyperliens fournissent habituellement une bonne description de la page pointée (Craswell *et al.*, 2001 ; Westerveld *et al.*, 2002 ; Kraaij *et al.*, 2002).

Comme troisième indexation, nous avons concaténé le texte encadré par les balises <TITLE>, <H1> et <BIG> de toutes les pages P_k pointant vers la page courante P_i. Cette indexation renforce le rôle tenu par les pages faisant référence au document indexé.

3.2. *Évaluation*

Afin d'évaluer notre moteur de dépistage des pages d'accueil et de services en-ligne, nous avons utilisé la collection test de TREC 2003 comprenant des sites extraits du domaine « .gov » (1'247'753 pages pour un volume de 18,1 GB). En plus de ces pages, 300 requêtes avec leur réponse exacte sont incluses dans cette collection. Pour être précis, nous avons 352 bonnes réponses pour les 300 requêtes, soit en moyenne 1,173 bonne réponse par requête (médiane : 1, maximum : 6).

Comme mesure de performance, nous avons retenu l'inverse du rang moyen de la première réponse exacte retournée par la machine (mesure notée IRM). Ainsi, si le système place la bonne réponse en deuxième position, il obtient 1/2 point. Par contre, si cette réponse exacte apparaissait en 5^{e} rang, 1/5 de point sera accordé.

La table 1 résume l'ensemble des expériences que nous avons menées. Dans la première colonne, nous avons regroupé les paramètres indiquant l'importance relative des trois représentations (α, β, γ) ainsi que le paramètre pondérant le calcul de proximité entre termes (δ). La deuxième colonne indique la valeur de l'IRM pour les 300 requêtes de notre évaluation (maximum 1 ou 100 %, minimum 0). La dernière indique le nombre de bonnes réponses dans les dix premiers rangs (sur un maximum 300) ainsi que le pourcentage correspondant.

Paramètres	IRM	dans les 10
$\alpha = 1;\ \beta = 0;\ \gamma = 0;\ \delta = 0$	0,335	163 (54,3%)
$\alpha = 0{,}6;\ \beta = 0{,}4;\ \gamma = 0;\ \delta = 0$	0,666	252 (84%)
$\alpha = 0{,}6;\ \beta = 0{,}4;\ \gamma = 0;\ \delta = 0{,}1$	0,691	251 (83,7%)
$\alpha = 0{,}6;\ \beta = 0{,}4;\ \gamma = 0{,}05;\ \delta = 0$	0,707	258 (86%)
$\alpha = 0{,}6;\ \beta = 0{,}4;\ \gamma = 0{,}05;\ \delta = 0{,}1$	**0,720**	**259** (86,3%)
$\alpha = 0{,}7;\ \beta = 0{,}3;\ \gamma = 0{,}05;\ \delta = 0{,}1$	0,700	258 (86%)
$\alpha = 0{,}8;\ \beta = 0{,}2;\ \gamma = 0{,}05;\ \delta = 0{,}1$	0,676	252 (84%)

Tableau 1. Évaluation de différentes variantes de notre stratégie de recherche (300 requêtes)

Si l'on ignore les hyperliens, nous devons indexer chaque page en fonction uniquement de son contenu. La performance d'une telle stratégie s'avère décevante (voir la deuxième ligne de la table 1 soit « $\alpha = 1; \beta = 0; \gamma = 0; \delta = 0$ »). La bonne réponse est retrouvée pour seulement 163 requêtes sur 300 (soit 54,3 %) si l'usager consulte les dix premières réponses extraites. La meilleure stratégie de recherche consiste à accorder une importance relativement marquée au contenu de la page elle-même (paramètre $\alpha = 0{,}6$) ainsi qu'à l'indexation basée sur les textes d'ancrage et champ <TITLE> des pages pointant vers la page considérée (paramètre $\beta = 0{,}4$). La présence de notre troisième représentation n'augmente que marginalement l'efficacité et la valeur de son paramètre ($\gamma = 0{,}05$) reste faible par rapport aux deux autres paramètres. L'usage de notre fonction de proximité ($\delta = 0{,}1$) ne permet pas d'augmenter le nombre de bonnes réponses trouvées (dernière ligne), mais la valeur IRM (deuxième colonne) augmente lorsque $\delta = 0{,}1$, signalant que l'ordinateur classe mieux les bonnes réponses qu'il a trouvées.

4. Extraction de ressources clés

En réponse à des requêtes telles que « virus informatique », « avantages et inconvénients de la légalisation de la marijuana ? » ou « Corée du Nord », les usagers désirent parfois obtenir une liste de ressources clés sur le thème spécifié. Il s'avère assez difficile de donner une définition opérationnelle d'une telle ressource. Certes, une page dont le contenu correspond au thème spécifié représente certainement une ressource à retenir et que l'on peut extraire à l'aide d'un

moteur de recherche classique. De bonnes pages centrales (c'est-à-dire des pages pointant vers des pages dont le contenu correspond au thème recherché) constituent aussi des ressources clés. Ainsi, au lieu de retourner plusieurs pages possédant la même page parente, il s'avère opportun d'extraire uniquement cette page parente. De manière similaire, si un site propose plusieurs pages possédant un contenu approprié, il est plus judicieux de retourner la page d'accueil du site ou celle correspondante à l'un de ses sous-sites.

4.1. *Notre modèle d'extraction de ressources clés*

Le modèle de recherche que nous proposons repose aussi sur trois indexations indépendantes des pages Web. Le premier représentant est identique au modèle précédent et regroupe tout le texte de la page à indexer (avec les balises <TITLE> et <META>). Cette stratégie ignore complètement les hyperliens. Une bonne ressource peut aussi être identifiée par les pages qui pointent vers elles d'une part, et d'autre part, par les textes d'ancrage des hyperliens quittant la page courante. Par exemple, si la page P_k pointe vers P_i qui pointe vers P_j, lors de l'indexation de P_i nous incluons les textes d'ancrage de P_k vers P_i de même que les textes d'ancrage dans P_i allant vers P_j. Enfin, le troisième représentant correspond à la concaténation de tous les champs <TITLE> et <H1> extraits des pages pointées par la page courante (c'est-à-dire de toutes les pages P_j de notre exemple précédent).

Afin d'évaluer la performance d'un système capable d'extraire les ressources clés sur un thème donné, nous avons retenu la précision obtenue après l'extraction de cinq éléments (notée « Précision à 5 ») et celle atteinte après la présentation de dix pages (indiquée sous la colonne « Précision à 10 »). Ces mesures indiquent le pourcentage de ressources clés extraites après cinq ou dix éléments retournés. Afin de mesurer ces précisions à l'aide de notre corpus « .gov », nous avons 50 requêtes avec leurs 516 ressources clés (soit, en moyenne, 10,32 ressources par requête, médiane : 8, maximum : 86, minimum : 1).

Paramètres	Précision à 5	Précision à 10
$\alpha = 1;\ \beta = 0;\ \gamma=0;\ \delta = 0$	8,00	6,20
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0;\ \delta = 0$	16,40	10,80
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0;\ \delta = 0,1$	16,00	11,00
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0,03;\ \delta = 0$	16,40	10,80
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0,03;\ \delta = 0,1$	16,00	11,00
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0,1;\ \delta = 0$	15,60	11,40
$\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0,1;\ \delta = 0,1$	16,00	**11,60**
$\alpha = 0,7;\ \beta = 0,3;\ \gamma = 0;\ \delta = 0$	15,20	10,20
$\alpha = 0,7;\ \beta = 0,3;\ \gamma = 0;\ \delta = 0,1$	16,00	10,20
$\alpha = 0,7;\ \beta = 0,3;\ \gamma = 0,1;\ \delta = 0$	14,00	11,40
$\alpha = 0,7;\ \beta = 0,3;\ \gamma = 0,1;\ \delta = 0,1$	14,00	11,40

Tableau 2. Évaluation de différentes stratégies d'extraction (50 requêtes, TREC 2003)

Si nous disposons uniquement du contenu d'une page Web afin de l'indexer, la précision à 5 s'élève à 8 % (ou à 6,2 % pour la précision à 10) (voir deuxième ligne de la table 2, « $\alpha = 1;\ \beta = 0;\ \gamma=0;\ \delta = 0$ »). Comme dans notre modèle précédent, il convient d'utiliser deux ($\alpha > 0$ et $\beta > 0$) ou trois ($\alpha > 0$, $\beta > 0$ et $\gamma > 0$) représentations afin d'améliorer la performance comme l'indique la table 2. Dans les grandes lignes, le meilleur choix se situe aux environs des valeurs « $\alpha = 0,5;\ \beta = 0,5;\ \gamma = 0,1;\ \delta = 0,1$ ». Le recours à la fonction de proximité (δ) n'apporte qu'une modification marginale de la performance. En effet, dans le cas présent, il ne s'agit plus de dépister une réponse unique mais plusieurs et cette fonction ne possède pas à elle seule la capacité de distinguer les bonnes réponses. Le troisième représentant (champ

<TITLE> et <H1> des pages pointées) n'apporte souvent qu'une légère amélioration de la précision à 10.

4.2. *Activation propagée, HITS et PageRank*

Pouvons-nous améliorer cette précision qui demeure relativement faible ? Comme une ressource clé est également un bon point de départ à la navigation, nous avons utilisé les listes de résultats obtenues précédemment afin de tenir compte des hyperliens d'une deuxième manière. Comme première stratégie, nous pouvons recourir à l'activation propagée (Savoy et Picard, 2001) qui transmet une fraction du score (notée λ) de chacune des pages extraites du Web vers ses voisines. Après avoir donné la requête à un moteur de recherche (dans notre cas, selon le modèle décrit en section 4.1), nous obtenons une liste triée de pages Web. Sur la base de cette liste, nous propageons les scores des documents selon les hyperliens sans tenir compte de leur orientation. Cette étape achevée, nous pouvons calculer le nouveau score pour chaque document en fonction de l'équation suivante :

$$RSV'(D_i) = RSV(D_i) + \lambda \cdot \sum_{j=1}^{r} RSV(D_j) \quad (2)$$

dans laquelle $RSV'(D_i)$ indique le nouveau score de la page D_i dépendant de son score initial (noté $RSV(D_i)$, calculé par l'équation 1) et du score de ses r voisins. Dans une version plus sophistiquée de cette stratégie, le facteur de propagation λ pourrait dépendre du type de lien reliant les deux documents ou du contenu respectif de D_i et de D_j. De plus, au lieu de propager une fraction du score vers tous les voisins, on peut limiter cette propagation au r « meilleurs » voisins, c'est-à-dire aux voisins les mieux classés par le moteur de recherche.

Comme deuxième stratégie d'extraction s'appuyant sur les hyperliens, nous pouvons recourir à l'algorithme HITS proposé par Kleinberg (1999). Dans ce modèle, on estime qu'une page pointant vers d'autres sources d'information est perçue comme une bonne « page centrale » (*hub*) tandis qu'un document possédant beaucoup d'hyperliens entrants sera analysé comme une bonne « page de référence » (*authority*). De manière récursive, si une page pointe vers plusieurs bonnes « pages de référence », elle obtiendra une meilleure évaluation en tant que « page centrale ». De manière duale, si une page est pointée par de nombreuses bonnes « pages centrales », sa valeur, comme « page de référence », augmentera.

Dans cette approche, chaque page D_i possède deux scores, soit celui comme « page de référence », noté $A^{c+1}(D_i)$, et celui comme « page centrale », noté $H^{c+1}(D_i)$, valeurs obtenues après c+1 itérations. Ces scores sont calculés selon les formules suivantes :

$$A^{c+1}(D_i) = \sum_{D_k \in \text{parent}(D_i)} H^c(D_k) \quad \text{et} \quad H^{c+1}(D_i) = \sum_{D_j \in \text{fils}(D_i)} A^c(D_j) \quad (3)$$

dans lesquelles $\text{parent}(D_i)$ indique l'ensemble des pages pointant vers D_i et $\text{fils}(D_i)$ l'ensemble des pages pointées par D_i.

Comme pour l'activation propagée, cet algorithme débute par l'établissement d'un ensemble de pages jugées pertinentes par un moteur de recherche. De cette liste, on extrait les r meilleures pages (ensemble noté σ) ainsi que toutes les pages pointées (tous les fils de σ) et toutes les pages pointant vers l'un des éléments de σ (tous les ancêtres directs). Sur cet ensemble élargi, on calcule les scores selon l'équation 3 et on les normalise (division par la somme des carrées de toutes les valeurs). Lors de la phase initiale, on attribue une valeur unitaire aux deux scores

de chaque page. Finalement, on itère durant cinq cycles car après ces cinq étapes, le classement des documents ne se modifie généralement plus.

Brin et Page (1998) ont proposé un modèle nommé PageRank débutant également par une liste triée de pages. Cependant, au lieu de classer ces pages en fonction de leur score (selon l'équation 1), PageRank va réordonner cet ensemble de pages dépistées en fonction de leur connectivité. En présence de la page D_i ayant les documents D_1, D_2, … D_m pointant vers elle, nous calculons son score PageRank, noté $PR^{c+1}(D_i)$ au cycle c+1, par l'équation suivante :

$$PR^{c+1}(D_i) = (1\text{-}d) \cdot [1 / n] + d \cdot [(PR^c(D_1) / C(D_1)) + \ \ldots \ + (PR^c(D_m) / C(D_m))] \quad (4)$$

dans laquelle d est un paramètre (fixé à 0,85), n indique le nombre de documents dans la collection et $C(D_k)$ indique le nombre d'hyperliens sortant du document D_k.

L'idée sous-jacente est la suivante. La valeur PageRank d'une page correspond à une estimation de la probabilité qu'un internaute la rencontre en naviguant de manière aléatoire. Ce dernier peut, avec une probabilité (1-d), choisir n'importe quel document sur le Web (et chaque page possède la même probabilité 1/n d'être choisie). D'un autre côté, l'internaute peut, avec une probabilité d, choisir de suivre un des hyperliens qu'il voit sur la page courante et chaque hyperlien possède la même chance d'être sélectionné. Dans ce cas, la probabilité d'aboutir sur la page D_i dépendra donc de la probabilité d'être sur une des pages D_j (notée $PR^c(D_j)$) et de l'inverse du nombre d'hyperliens inclus dans cette page D_j (soit $C(D_j)$).

Comme dans l'algorithme de Kleinberg, le score PageRank des pages se calcule de manière itérative et, habituellement, on effectue cinq itérations. Après chaque itération, on normalise les valeurs $PR^c(D_i)$ en les divisant par la somme des valeurs $PR^c(D_j)$ et, comme valeurs initiales, on fixe $PR^0(D_i)$ à 1/n.

Paramètres	Précision à 5	Précision à 10
$\alpha = 0,5$; $\beta = 0,5$; $\gamma = 0,1$; $\delta = 0,1$	16,00	**11,60**
$\lambda = 0,02$; r = 50	17,60	12,00
$\lambda = 0,05$; r = 50	**17,60**	12,20
$\lambda = 0,1$; r = 50	13,60	12,00
$\lambda = 0,05$; r = 200	19,20	12,40
$\lambda = 0,05$; r = 300	16,40	**14,00**
$\lambda = 0,05$; r = 400	16,00	13,80
HITS, *hub*, σ = 50 premiers	3,60	2,60
HITS, *authority*, σ = 50 premiers	0.80	0.60
PageRank, d = 0,85	2,00	1,60

Tableau 3. Évaluation de stratégies utilisant les hyperliens (50 requêtes, TREC 2003)

Afin de mesurer la performance de ces trois solutions, nous avons repris, dans la première ligne de la table 3, la meilleure solution obtenue par notre modèle d'extraction (voir tableau 2). Sur cet ensemble ordonné de pages, nous avons appliqué notre algorithme d'activation propagée en faisant varier le facteur de propagation λ ainsi que le nombre r de pages sur lesquelles cette propagation est appliquée. De même, nous avons extrait les 50 premières pages pour former l'ensemble σ, puis nous avons élargi cet ensemble et avons calculé le score des pages centrales (*hub*) et celui des pages de référence (*authority*). Enfin, nous avons utilisé l'algorithme PageRank. Les résultats de la table 3 sont clairs. Les algorithmes HITS ou PageRank n'apportent pas la réponse souhaitée.

5. Conclusion

Différentes études ont démontré que les hyperliens ne permettaient pas, ou de façon marginale, d'améliorer la qualité des moteurs de recherche sur le Web (Savoy et Picard, 2001). Or, si le type d'interrogation correspond à la recherche de service en-ligne (ou de page d'accueil) d'une part, ou, d'autre part, à l'extraction de ressources clés, l'inclusion des textes d'ancrage dans l'indexation des pages Web permet d'augmenter significativement les performances obtenues, du moins dans le cadre de notre proposition d'indexation combinée. Dans le cadre du dépistage de ressources clés, l'algorithme HITS ou PageRank n'apporte pas une solution intéressante tandis que notre approche basée sur le principe de l'activation propagée permet une augmentation de la précision de l'ordre de 10 à 20 %.

Références

Brin S. et Page L. (1999). The anatomy of a large-scale hypertextual Web search engine. In Mendelzon A.(Ed.), *Proceedings of the WWW8* : 107-117.

Craswell N., Hawking D. et Robertson S. (2001). Effective site finding using link anchor information. In Kraft D.H., Croft W.B., Harper D.J. et Zobel J. (Eds), *Proceedings of ACM-SIGIR'2001* : 250-257.

Harman D. (1991). How effective is suffixing? *Journal of the American Society for Information Science,* vol. (42/1) : 7-15.

Kleinberg J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, vol. (46/5) : 604-632.

Kraaij W., Wes/terveld T. et Hiemstra D. (2002). The importance of prior probabilities for entry page search. In Baeza-Yates R., Beaulieu M., Järvelin K. et Myaeng S.H. (Eds), *Proceedings of ACM-SIGIR'2002* : 27-34.

Nielsen J. (2000). *Designing Web usability: the practice of simplicity*. New Riders.

Radev D.R., Libner K. et Fan W. (2002). Getting answers to natural language questions on the Web. *Journal of the American Society for Information Science and Technology*, vol. (53/5) : 359-364.

Rasolofo Y. and Savoy J. (2003). Term proximity scoring for keyword-based retrieval systems. In Sebastiani F. (Ed.), *Proceedings of ECIR 2003* : 207-218.

Robertson S.E., Walker S. et Beaulieu M. (2000). Experimentation as a way of life: OKAPI at TREC. *Information Processing & Management*, vol. (36/1) : 95-108.

Savoy J. et Picard J. (2001). Retrieval effectiveness on the Web. *Information Processing & Management*, vol. (37/4) : 543-569.

Savoy J. et Rasolofo Y. (2003). Report on the TREC-11 experiment: Arabic, named page and topic distillation searches. In Voorhees E.M. et Buckland L.P. (Eds), *Proceedings of TREC 2002* : 765-774.

Spink A., Wolfram D., Jansen M.B.J. et Saracevic T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*, vol. (52/3) : 226-234.

Watters C. et Amoudi G. (2003). GeoSearcher: location-based ranking of search results. *Journal of the American Society for Information Science and Technology*, vol. (54/2) : 140-151.

Westerveld T., Kraaij W. et Hiemstra D. (2002). Retrieving Web pages using content, links, URLs and anchors. In Voorhees E.M. et Harman D.K. (Eds), *Proceedings of TREC 2001* : 663-672.