Recherche documentaire sur le Web: Les hyperliens sont-ils vraiment utiles?

Jacques Savoy, Justin Picard

Institut interfacultaire d' informatique
Université de Neuchâtel - Pierre-à-Mazel 7 - 2000 Neuchâtel (Suisse)

Abstract

There is an increasing interest in hypertext systems, digital libraries and the Web. Due to the huge number of pages and links, browsing cannot be viewed as an adequate searching process, even with the introduction of subject directories or other classified lists (e.g., Yahoo!). Therefore, an effective query-based mechanism for accessing information is needed. Nowadays, search engines available on the Web are far from covering all available information, and present many drawbacks. Moreover, most of them ignore hypertext links to enhance their retrieval effectiveness. Recent work in IR on the Web seems to recognize that the hyperlink structure can be very valuable for locating information. This paper exposes some search strategies using hyperlinks. Only a few rigorous experiments deal with such large-sized networked information, and we present some preliminary experiments using a snapshot of around 2.3 Gb extracted from the Web. Our study suggests that the usefulness of interdocument relationships for searching purpose is questionable, at least as implemented actually.

Résumé

Les systèmes hypertexte, les bibliothèques numériques ou le Web connaissent un intérêt grandissant. Pour trouver de l' information pertinente, la navigation à elle seule ne peut pas être vue comme un moyen efficace, surtout si l' on considère le nombre considérable de pages et de liens. Le recours à des moteurs de recherche s' avère essentiel et leur présence à permis au Web de grandir dans les proportions que nous connaissons actuellement. Cependant, ces moteurs de recherche possèdent quelques lacunes et actuellement, la grande majorité d' entre eux ignore les liens hypertexte afin d' améliorer la qualité de leurs réponses. De récents travaux touchant la recherche documentaire sur le Web semblent indiquer que les liens peuvent être utiles pour mieux dépister les documents pertinents. Malheureusement, nous ne disposons pas d' expériences rigoureuses afin de connaître l' efficacité de différentes approches possibles. Cet article présente quelques systèmes de dépistage de l' information utilisant les liens hypertexte afin d' améliorer la qualité de leurs réponses. Basée sur une collection de 2.3 Gb de pages Web et de 100 requêtes ainsi que sur une méthodologie de comparaison rigoureuse, cet article présente quelques évaluations de l' efficacité des liens dans la recherche documentaire sur le Web.

Mots-clés : recherche d' informations, hypertexte, liens, Web, bibliothèques numériques.

1. Introduction

Pour rechercher de l' information dans les hypertextes comme le Web (Bernes-Lee et al. 1994) ou dans des bibliothèques numériques (Lesk 1997), la navigation peut être vue comme une approche possible, mais dès que le nombre de pages (documents) et / ou de liens devient important, cette technique n' est plus satisfaisante. Comme le mentionne Halasz (1988), un outil plus efficace doit être proposé, spécialement pour des collections de taille raisonnable (par exemple, comportant plus de 200 pages). Le recours à différents index, répertoires de vedette-matière ou tables des matières hiérarchisées ("classified lists" comme Yahoo!) s' avère également peu satisfaisant (Alschuler 1989). Un moteur de recherche basé sur des requêtes doit devenir le moyen principal d' accès à l' information, approche complétée par d' autres outils de navigation.

Dans cette optique et comme première solution, le Web dispose de moteurs centralisés d' indexation et de recherche (Schwartz 1998; Leighton et Srivastava 1999; Gordon et Pathak 1999) (par exemple, Alta Vista, Excite, voir http://www.SearchEngineWatch.com/) qui visent

à indexer tous les documents disponibles. Ces outils s' avèrent fort utiles et l' on sait qu' environ 85% des usagers du Web les utilisent en premier lieu pour dépister de l' information. Cependant, ces moteurs présentent quelques lacunes importantes. Ainsi, même des systèmes disposant d' une capacité disque très importante n' indexent qu' une fraction de toute l' information disponible, et la couverture de ces moteurs n' augmente pas aussi rapidement que la taille du Web. Par exemple, Northern Light, moteur disposant selon les dernières estimations de la plus grande couverture du Web, n' indexe que 16% des pages tandis que Lycos couvre environ 2.5% (Lawrence et Lee Giles 1999). Cette difficulté s' accroît si l' on tient compte du fait que les index doivent être dupliqués pour des raisons diverses (temps d' accès, temps de réponse, largeur de bande disponible, pannes). Troisièmement, les pages indexées sont souvent obsolètes ou ont disparu (ce qui représente environ 5% des réponses). Quatrièmement, la page utilisée lors de l' indexation ne correspond pas à la page actuelle car l' auteur a modifié, parfois radicalement, son contenu sémantique. Cinquièmement, une distinction claire entre les sources dignes de foi et les autres n' est actuellement pas possible. Sixièmement, plusieurs sources ou collections de documents ne sont pas librement accessibles (copyrights, collections privées). Finalement, le temps d' accès et de recherche dans un index aussi gigantesque est lent. La conclusion qui s' impose est donc de considérer d' autres approches rapides d' accès à l' information stockée sur un réseau. Dans ce cadre, il conviendrait de tirer parti de la spécificité du Web, c' est-à-dire des liens entres les documents.

2. Les liens dans la recherche d'informations

Différentes études ont suggéré de tenir compte des liens inter- ou intra- documents afin d' augmenter la qualité du dépistage automatique de l' information, et Frisse (1988) est l' un des premiers à proposer un tel système de dépistage de l' information. Généralement, ces modèles fonctionnent en deux temps. Dans une première étape, un moteur de recherche classique retrouve une liste ordonnée de pages répondant au mieux (aux yeux de la machine) à la requête posée, en fonction des mots de la requête et des termes d' indexation des documents. Cette liste est habituellement triée en fonction du degré de similarité entre la requête Q et le document D_i (notée $SIM(Q,D_i)$). Dans une seconde étape, ces systèmes tiennent compte des liens pour les m meilleurs documents classés (m variant typiquement entre 5 et 200). Dans cette optique, on ne devra pas tenir compte de tous les liens existants dans le Web, mais seulement de ceux qui concernent les m pages les mieux classées. La réduction ainsi obtenue permet de travailler sur un graphe réduit et d' obtenir des temps de traitement très faible.

Dans nos travaux précédents, nous avons analysé essentiellement les liens de référence bibliographique présents dans les articles scientifiques (Garfield 1983). L' hypothèse sous-jacente considère que le but principal d' une référence bibliographique est de citer d' autres travaux antérieurs (modèles, méthodologies, résultats, etc.) dont le sujet est relié directement au document courant. Cette hypothèse semble être valide pour la majorité des liens mais elle ne s' avère pas vérifiée dans tous les cas (Liu 1993).

Dans nos modèles de recherche basés sur l' activation propagée (Cohen et Kjeldsen 1987), nous suggérons d' analyser les liens de référence bibliographiques pour les m meilleurs documents retrouvés. Pour ceux-ci, si nous rencontrons un lien partant du document D_i vers le document D_k, une fraction du score $SIM(Q,D_i)$ (typiquement entre 0.1 et 0.2) est ajoutée au score $SIM(Q,D_k)$. Après inspection des m premiers documents, on peut itérer cette procédure c fois, en prenant garde d' éliminer les cycles et des liens multiples entre deux pages. Finalement, la liste finale est triée en fonction des similarités modifiées et présentée à l' usager.

Nos expériences antérieures ont démontré que la qualité de la réponse est significativement améliorée pour des moteurs booléens de recherche (Savoy 1997). Avec des modèles de recherche plus performants comme l' approche vectorielle ou le modèle probabiliste, l' augmentation de la qualité n' est pas aussi marquée (Savoy 1996).

Le cadre général que nous venons de décrire possède évidemment quelques variantes; on peut accorder un poids identique à tous les liens ou, au contraire, pondérer chaque lien individuellement ou selon son type. L' algorithme peut tenir compte uniquement des liens sortants des m premiers documents, uniquement ceux qui entrent vers cet ensemble ou ignorer l' orientation des liens (on considérera alors tous les liens entrants et sortants).

Les liens de référence bibliographique peuvent être complétés par la prise en compte des liens de couplage bibliographique (Kessler 1963). Dans ce cas, l' hypothèse sous-jacente stipule que si deux articles possèdent une bibliographie similaire, ils devraient posséder un contenu apparenté et traiter du même sujet (comme mesure, on compte le nombre d' articles apparaissant simultanément dans les deux listes de références bibliographiques). Comme deuxième solution, on peut également considérer les liens de co-citation entre D_i et D_k (Small 1973) en comptant le nombre d' articles citant simultanément D_i et D_k. Afin de tenir compte des requêtes passées, nous pouvons également créer des liens de pertinence entre les documents trouvés pertinents à la même requête (Savoy 1994) ou établir des liens du plus proche voisin en liant les deux documents possédant, sur la base de leurs termes d' indexation, la plus forte similarité (Savoy 1997).

Si l' on considère le Web, les documents (ou pages Web) possèdent également des liens entre eux dont le but premier est de faciliter la navigation à l' intérieur d' un site (liens d' organisation définissant la structure d' un site, structure arborescente pour l' essentiel). Cependant, ces liens établissent également des relations entre des pages appartenant à des sites différents et pouvant être vus comme des liens de proximité sémantique entre pages Web.

Ainsi, pour Chakrabarti et al. (1999):

> "Citations signify deliberate judgment by the page author. Although some fraction of citations are noisy, most citations are to semantically related material. Thus the relevance of a page is a reasonable indicator of the relevance of its neighbors, although the reliability of this rule falls off rapidly with increasing radius on average. Secondly, multiple citations from a single document are likely to cite semantically related documents as well." (Chakrabarti 1999, p. 550-551)

et une telle hypothèse est reprise, dans les grandes lignes, par d' autres auteurs (Kleinberg 1998; Bharat et Henzinger 1998).

La démarche proposée par Marchiori (1997) est similaire à notre approche, mais présentée dans la communauté du Web. En plus du contenu textuel, les liens entrants ou sortants peuvent fournir des indications précieuses afin de revoir le classement des pages à retourner à l' usager. Par exemple, la visibilité d' une page se mesure par le nombre de liens pointant vers cette page et correspond à une indication de sa valeur. Cependant, cette indication de facto ne fournit pas directement une valeur sur le contenu informatif de la page et il ne serait pas judicieux d' établir un lien entre la popularité d' une page et sa qualité. Les liens ne doivent fournir qu' une indication secondaire (surtout les liens entrants) et le contenu textuel restera la clé première d' accès. De plus, l' influence d' une page sur une autre diminue exponentiellement avec le nombre de liens qui les relie. Selon Marchiori (1997), le post-traitement devrait (1) s' effectuer une seule fois (c' est-à-dire que l' on considére uniquement les voisins immédiats);

(2) modifier le score pour les 100 premiers documents retrouvés; (3) ignorer les liens entrants; (4) pondérer les liens sortant d' un facteur 0.75.

Pour Kleinberg (1998), il s' avère utile de distinguer les pages centrales ("hub") des pages qui font autorité ("authoritative"). Les premières correspondent aux pages possédant un nombre important de liens sortants (comme, par exemple, une page répertoire) tandis que les secondes correspondent aux pages souvent citées par d' autres sites (donc possédant une importance reconnue par beaucoup d' auteurs). Par transitivité, on peut ajouter qu' une page centrale est meilleure qu' une autre si elle pointe vers de nombreuses pages qui font autorité. Si une page pointe vers de nombreux sites qui ne font pas autorité, elle ne sera pas considérée comme une bonne page centrale. En bibliothéconomie, les documents qui font autorité (donc cités par de nombreux auteurs) correspondent souvent à une description de l' état de l' art, à un article fondamental ou à une méthodologie largement utilisée. Sur cette idée, on en déduit que les pages faisant autorités possèdent une plus grande chance d' être pertinente, tandis que les "hub pages" contiennent des liens vers des pages pertinentes.

Sur la base des travaux de Kleinberg (1998), Bharat et Henzinger (1998) reconnaissent qu' il s' avère difficile d' un point de vue technique de vouloir indexer toutes les pages du Web et ce défi s' accroît quand on constate que le contenu des pages est loin d' être très statique. Le recours aux hyperliens n' est donc pas exempt de problèmes comme : (1) la faible densité de liens d' un site ou vers d' autres sites; (2) une page cite parfois plusieurs pages différentes d' un même site (et la valeur d' autorité de ce site serait donc trop forte); (3) les outils de génération automatique de liens ne tiennent pas vraiment compte de la proximité sémantique; (4) les voisins de certaines pages ne sont que marginalement pertinents à la page de départ (par exemple, "My favorite Links" d' une page personnelle ou les liens "MailTo").

Pour Brin et Page (1999), les liens permettent d' établir un classement des pages les plus populaires du Web selon la probabilité d' arriver à cette page en naviguant au hasard sur le Web (système PageRank). Ainsi, plus une page est référée par d' autres sites, et plus ces autres sites sont référés par d' autres sources, plus la chance de tomber sur la page analysée augmente et plus son importance s' accroît. Les premières expériences tendent à démontrer que les sites très populaires (IBM, Microsoft, voir (Bray 1996) pour une liste des sites centraux du Web) sont les mieux classés, mais sont-ils les plus pertinents à une requête?

Chakrabarti et al. (1999) reconnaissent l' importance du défi technologique de vouloir indexer tout le Web à l' aide d' un seul système centralisé. Ils suggèrent plutôt de segmenter son contenu par domaines d' intérêt très précis, division faite sur la base d' une taxonomie (à l' image de Yahoo!) et d' un ensemble de pages pertinentes au thème considéré. Ces auteurs relèvent que la machine s' avère capable de créer une telle classification de manière semi-automatique. Cette dernière permet de traiter beaucoup plus facilement un volume d' information plus spécifique et réduit ainsi que son évolution. Au besoin, une hiérarchie de brokers serait à même de collaborer pour dépister l' information souhaitée sur le Web.

Le recours presque exclusif aux liens pour dépister de l' information a été proposé par Dean et Henzinger (1999) dans l' intention de répondre à des requêtes du type "trouves-moi des pages similaires à celle-ci (ou similaire à tel URL)". Basé sur les idées de Kleinberg (1998), ces auteurs proposent deux algorithmes de navigation dont le résultat semble être meilleur que le dépistage proposé par l' option "What' s related" du navigateur Netscape.

3. Évaluation

Afin d' évaluer ces différentes propositions, nous avons recouru à une méthodologie classique en recherche d' informations (Salton et McGill 1983, Chapter 5) et qui est identique à celle utilisée dans la conférence TREC (Harman 1995). Sur la base d' une collection d' environ 247' 491 pages HTML du Web-Track (environ 2,3 Gb provenant de 969 sites différents) et d' un lot de 50 requêtes, la machine nous a calculé la précision moyenne pour les différentes approches étudiées. La précision mesure la qualité de la réponse fournie par l' ordinateur et s' obtient en divisant le nombre de documents retrouvés et pertinents par le nombre de documents retrouvés. Ceci suppose que pour les 50 requêtes utilisées, nous connaissions les pages HTML pertinentes à la question posée (évidemment, cette information n' est pas connue par le moteur de recherche évalué). Les requêtes ne sont pas limitées à un domaine précis mais couvrent un éventail de thèmes (par exemple "Falkland petroleum exploration", "journalist risks", "El Niño", "piracy"). Finalement, pour distinguer si un système de dépistage est meilleur qu' un autre, on admet comme règle d' usage qu' une différence de 5% dans la précision moyenne peut être considérée comme significative.

Notre première interrogation consiste à savoir si les moteurs de recherche actuellement proposés sur Internet apportent une réponse de qualité comparable aux moteurs développés par différents laboratoires de recherche. Selon l' étude récente de Hawking et al. (99a), une différence de précision moyenne d' environ 37% sépare les deux catégories et ceci en défaveur des moteurs de recherche d' Internet. Pour Gordon et Pathak (1999), la précision actuelle des moteurs commerciaux est relativement faible; sur les dix premiers documents retrouvés, le mode du nombre de pages pertinentes s' élève à un. On peut relativiser cette constatation en tenant compte du fait que ces derniers se préoccupent surtout d' un temps de réponse bref plutôt que de la qualité de la réponse fournie (Lesk 1998).

Dans nos expériences, nous avons utilisé le modèle de recherche probabiliste OKAPI (Roberston et al. 1995) reconnu pour posséder une performance très intéressante (première ligne de notre tableau). Sur la base de ce modèle, nous avons procédé à une expansion automatique des requêtes (Buckley et al. 1996), technique reconnue comme améliorant significativement la qualité de la réponse fournie par la machine. En analysant cette technique, nous désirons comparer sa performance avec une technique basée sur les hyperliens. Les résultats de la deuxième ligne de notre tableau confirment l' efficacité de l' expansion automatique des requêtes, que ces dernières soient courtes (dont le nombre moyen de termes s' élève à 2.4 par requête) ou longues (longueur moyenne de 5.7 mots).

Dans une première série d' expériences (troisième ligne de notre tableau), nous avons suivi les liens sortants pour les cinquante pages les mieux classées. La variation de la précision moyenne qui en résulte n' est pas significative.

L' algorithme proposé par Kleinberg (1998) ne donnant aucun résultat probant (baisse significative de la précision moyenne), nous avons décidé d' utiliser les hyperliens afin de favoriser le rappel, c' est-à-dire de dépister tous les pages pertinentes à une requête. Dans ce but, nous avons appliqué notre algorithme pour les pages retrouvées depuis le rang 51 jusqu' au rang 1000. Les résultats de notre système sont présentés dans la dernière ligne ci-dessus et semblent indiquer que les hyperliens n' ont qu' une importance marginale dans la qualité de la réponse.

	Précision moyenne (% changement)		
modèle \ type de requêtescourtes	courtes	longues	
probabiliste OKAPI	29.84	29.84	36.41
OKAPI avec expansion automatique des requêtes (Buckley et al. 1996)		35.62 (+19.37%)	40.50 (+11.23%)
OKAPI avec prise en compte des liens hypertexte (1 à 50)	30.11 (+0.90%)	35.94 (+0.90%)	40.98 (+1.19%)
OKAPI avec prise en compte des liens hypertexte (51 - 1000)	30.21 (+1.24%)	36.01 (+1.09%)	40.98 (+1.18%)

Tableau 1 : Précision moyenne avec et sans liens hypertextes

Dans notre collection extraite du Web, nous avons trouvé 1'171'795 liens au total (lien reliant deux pages appartenant à notre collection) dont 36'002 (3.07%) qui relient deux pages appartenant à des sites différents. Le nombre moyen de liens par page s'élève à 4.73 liens et cette valeur passe à 0.15 lien par page si l'on considère uniquement les liens sortant vers un autre site. Bray (1996) a obtenu des chiffres comparables en indiquant qu'environ 80% des sites Internet ne possèdent pas de lien vers d'autres sites.

Finalement, en utilisant l'algorithme PageRank de Brin et Page (1999) sur notre collection, la valeur moyenne de popularité d'une page pertinente s'élève à 0.33 tandis que cette valeur moyenne est de 0.36 pour les pages jugées non pertinentes. La comparaison de ces deux moyennes indique clairement que le recours à l'algorithme PageRank n'apporte pas de modification notable à la précision moyenne.

4. Conclusion

Confronté à une masse importante et sans cesse croissante d'information disponible sur le Web, les moteurs de recherche actuels ont adopté une stratégie basée sur une indexation centralisée. Cette option présente plusieurs lacunes et, à moyen terme, ces inconvénients devraient nous pousser à considérer d'autres approches. Dans cette perspective, les metamoteurs (Dreilinger et Howe 1997) (comme MetaCrawler, SavySearch) ainsi que les modèles de recherche conçus pour les environnements distribués (Hawking et Thislewaite 1999b; Fuhr 1999) ont recours à plusieurs moteurs de recherche afin de dépister plus de pages pertinentes en réponse à une requête.

Dans cet article, nous avons considéré une autre démarche automatique utilisant l'information contenue dans les liens hypertextes afin d'augmenter la qualité de la réponse fournie par la machine. Cette approche ne requiert que peu de traitement informatique supplémentaire. Cependant, basé sur une méthodologie d'évaluation rigoureuse (collection de pages d'environ 2,3 Gb tirées du Web avec 50 requêtes), l'augmentation de la qualité de la réponse n'est pas aussi importante que prévue. Cette constatation, partiellement négative, ne doit pas fermer la porte à toute considération des liens hypertexte; ces informations peuvent fournir des éléments fort utiles pour organiser le contenu du Web par thèmes et nous permettre de passer du paradigme "rechercher et extraire" à une métaphore "organiser et naviguer". De telles applications existent en bibliothéconomie (Garfield 1983; Egghe et Rousseau 1990) et des expériences préliminaires commencent à apparaître (Chakrabarti et al. 1999).

Remerciements

Cette recherche a été subventionnée en partie par le Fonds national suisse (subventions 20-50' 578.97 et 20' 55' 771.98).

Références

Alschuler L. (1989). Hand-crafted Hypertext - Lessons from the ACM experiment. In E. Barrett editor, *The Society of Text, Hypertext, Hypermedia, and the Social Construction of Information*. The MIT Press.

Bernes-Lee T., Cailliau R., Luotonen A., Nielsen H. F., and Secret A. (1994). The World-Wide Web. *Communications of the ACM*, 37(8):76-82.

Bharat K. and Henzinger M. (1998). Improved algorithms for topic distillation in hyperlinked environments. In Croft W. B., Moffat A., van Rijsbergen C. J., Wilkinson R. and Zobel J., editors, *Proc. of ACM-SIGIR'98*, pages 104-111.

Bray T. (1996). Measuring the Web. In *Proc. of WWW5*.

Brin S. and Page L. (1999): The anatomy of a large-scale hypertextual Web search engine. In Mendelzon A., editor, *Proc. of WWW8*, pages 107-117.

Buckley C., Singhal A., Mitra M. and Salton G. (1996). New retrieval approaches using SMART. In Harman D., editor, *Proc. of the TREC' 4*pages 25-48.

Chakrabarti S., Van den Berg M. and Dom B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. In Mendelzon A., editor, *Proc. of WWW8*, pages 545-562.

Cohen P. R. and Kjeldsen R. (1987). Information retrieval by constrained spreading activation in semantic networks. *Information Processing & Management*, 23(4):255-268.

Dean J. and Henzinger M. R. (1999). Finding related pages in the World Wide Web. In Mendelzon A., editor, *Proc. of WWW8*, pages 389-401.

Dreilinger, D. and Howe A. E. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, 15(3):195-222.

Egghe L. and Rousseau R. (1990). Introduction to Informetrics. Quantative Methods in Library, Documentation and Information Science. Elsevier.

Frisse M. E. (1988). Searching for information in a Hypertext medical handbook. *Communications of the ACM*, 31(7):880-886.

Fuhr N. (1999). A decision-theoretic approach to database selection in networked IR. *ACM Transactions on Information Systems*, to appear.

Garfield E. (1983). *Citation Indexing: Its Theory and Application in Science, Technology and Humanities*. The ISI Press, 2nd edition.

Gordon M. and Pathak P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, 35(2):141-180.

Halasz F. G. (1988). Reflections on NoteCards: seven issues for the next generation of Hypermedia systems. *Communications of the ACM*, 31(7):836-852.

Harman D. (1995). Overview of the second TExt Retrieval Conference (TREC-2). *Information Processing & Management*, 31(3):271-289.

Hawking D., Craswell N., Thistlewaite P. and Harman D. (1999). Results and challenges in Web search evaluation. In Mendelzon A., editor, *Proc. of WWW8*, pages 243-252.

Hawking D. and Thistlewaite P. (1999). Methods for information server selection. *ACM Transactions on Information Systems*, 17(1):40-76.

Jansen B. J., Spink A., Bateman J. and Saracevic T. (1998). Real life information retrieval: a study of user queries on the Web. *ACM-SIGIR Forum*, 32(1):5-17.

Kessler M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10-25.

Kleinberg J. (1998). Authoritative sources in a hyperlinked environment. In *Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pages 668-677.

Lawrence S. and Lee Giles C. (1999). Accessibility of information on the Web. *Nature*, 400 (6740):107-110.

Leighton H. V. and Srivastava J. (1999): First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*, 50(10):870-881.

Lesk M. (1997). *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann.

Lesk M. (1998). "Real world" searching, panel at SIGIR 97. *ACM-SIGIR Forum*, 32(1):1-4.

Liu M. (1993). The complexities of citation practice: a review of citation studies J*ournal of Documentation*, 49(4):370-408.

Marchiori M. (1997). The quest for correct information on the Web: hyper search engines. In *Proc. of WWW6*.

Picard J. (1998). Modeling and combining evidence provided by document relationships using probability argumentation systems. In Croft W. B., Moffat A., van Rijsbergen C. J., Wilkinson R. and Zobel J., editors, *Proc. of ACM-SIGIR'98*, pages 182-189.

Robertson S. E., Walker S. and Hancock-Beaulieu M. M. (1995). Large test collection experiments on an operational, interactive system: OKAPI at TREC. *Information Processing & Management*, 31(3):345-360.

Salton G. and McGill M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Savoy J. (1994). A learning scheme for information retrieval in Hypertext. *Information Processing & Management*, 30(4):515-533.

Savoy J. (1996). Citation schemes in Hypertext information retrieval. In Agosti M. and Smeaton A. editors, *Information Retrieval and Hypertext*. Kluwer.

Savoy J. (1997). Ranking schemes in hybrid Boolean systems: a new approach. *Journal of the American Society for Information Science*, 48(3):235-253.

Schwartz C. (1998). Web search engines. *Journal of the American Society for Information Science*, 49(11):973-982.

Small H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265-269.