De l’exploitation des liens hypertexte en recherche d’information

Malika ABCHICHE, Chantal SOULE-DUPUY
IRIT-SIG
Campus Univ. Toulouse III
118, route de Narbonne – F-31062 Toulouse Cedex 4
E-mail*:* {abchiche, soule}@irit.fr

1. Introduction

Un système de recherche d'informations (SRI) a pour objectif d'organiser et de manipuler des collections de documents. L’objectif, à la base, est l'identification des documents pertinents répondant aux besoins en information des utilisateurs. Les contenus textuels des documents, et par la suite les contenus des requêtes, sont analysés et représentés par des ensembles de mots clés extraits en se basant généralement sur des méthodes statistiques (fréquences d'apparition des mots dans les documents et dans la collection, …). Le paradigme de recherche de l’information pertinente est alors basé sur la comparaison des représentations des documents et des représentations des requêtes. Les documents suffisamment ressemblants à la requête sont restitués à l'utilisateur dans l'ordre décroissant de leurs valeurs de ressemblance avec la requête.

Le web est une collection différente des collections habituellement manipulées par les SRI traditionnels. Des estimations récentes montrent que la quantité d'informations disponible sur le web croît de façon considérable.

Cette explosion est problématique pour les systèmes utilisant l'approche de recherche standard qui ne prend en considération que les mots clés des documents [21]. En effet, différentes études ont montré que les requêtes utilisateurs sont généralement courtes (deux à trois mots) et imprécises [11], elles sont de ce fait générales. En conséquence, étant donné le volume du web, les résultats obtenus en réponse à ces requêtes sont trop volumineux et de ce fait, les listes des documents restituées deviennent peu maniables. Bien évidemment, si le système de recherche

est très performant et la sortie bien organisée, alors les premiers documents de la liste peuvent être une réponse adéquate pour la requête de l'utilisateur. Cependant, ils ne sont pas tous aussi efficaces et les documents retrouvés ne sont pas toujours satisfaisants. Les utilisateurs se retrouvent donc noyés dans des listes de quelques centaines à quelques milliers de documents retournés en réponse à leurs requêtes. Pour trouver les informations qui les intéressent, ils doivent parcourir toute la liste. Or, souvent, ils ne vont pas au delà de la deuxième page écran risquant ainsi de perdre des documents intéressants.
Pour ces différentes raisons, les approches standards de recherche d'informations ne sont donc pas adaptées à des collections volumineuses telles que le web, d'où la nécessité de nouvelles alternatives de recherche. Récemment, dans le but d'améliorer les performances de la recherche d'informations dans un environnement hypertexte, différents travaux se sont intéressés à l'exploitation de la structure hypertexte des documents [6] [5] [14] [7] [2] [19].
L'objectif de ce papier est de discuter des effets de la prise en compte des liens hypertexte dans l'évaluation des performances du processus de recherche d'informations. Nous avons étudié différentes stratégies de recherche et de restitution d'informations intégrant l'analyse du contenu textuel des documents avec l'analyse des liens entrants et sortants de ces documents. Plus particulièrement, notre objectif est d'essayer d'améliorer la précision sur l'ensemble des premiers documents restitués en réponse à une requête afin de réordonner cet ensemble (sachant que seuls les premiers documents restitués d'une liste sont effectivement consultés).
Après avoir proposé un état de l'art sur l'exploitation de la structure hypertexte de documents en recherche d'informations, ce papier est consacré à la définition de notre modèle de recherche et de restitution d'informations basé sur l'exploitation combinée des liens hypertexte et des contenus textuels des documents.
Nous terminerons sur une présentation et une discussion des résultats de nos expérimentations.

2. Etat de l'art

Dans le domaine bibliométrique, qui concerne l'étude de documents écrits reliés entre eux par des citations, différents travaux se sont focalisés sur l'utilisation des citations dans le but d'estimer l'importance ou la popularité de papiers scientifiques. L'idée de base est que les citations traduisent des jugements délibérés des auteurs : si l'auteur d'un document cite un autre document alors celui-ci pense que ce document contient des informations utiles par rapport au thème développé dans son document. Pour estimer l'importance de papiers scientifiques, Garfield [9] utilise une mesure appelée « facteur d'impact ». Ce facteur représente, pour une année donnée, le rapport entre le nombre de citations sur le nombre d'articles publiés par un journal, sur une période de référence de deux ans. Il mesure donc la fréquence moyenne avec laquelle l'ensemble des articles de ce journal est cité pendant une durée définie. L'analyse des citations telle qu'elle a été introduite par Garfield a été critiquée par de nombreux auteurs [12] [8]. Principalement, à cause du fait que des groupes tendent à se citer les uns les autres par déférence plutôt que par pertinence. Cependant, ces critiques liées au facteur d'impact de Garfield restent propres au domaine bibliométrique. En effet, comme il a été argumenté par Bharat et Henzinger [2], ces problèmes se posent moins dans le contexte du web car la communauté est diverse et distribuée, et le droit de publication ne peut pas être restreint à des groupes.

Les travaux sur les citations ont été largement appliqués sur le web [15]. En particulier, l'analyse des liens hypertexte a été essentiellement utilisée comme une alternative au processus standard de recherche d'informations (basé sur les mots clés) [6] [5] [14] [7] [2] [19]. Le modèle proposé par Carrière et Kazman [6] dans le but de réordonner les pages web est basé principalement sur le dénombrement des liens entrants et Fifth sortants. Le rang d'une page peut être interprété comme une valeur de popularité ou de qualité basée sur les liens entourant cette page. A l'opposé, dans [5] les liens entourant une page ne sont pas comptés de façon égale. Le rang d'une page, appelé PageRank, est calculé en utilisant un simple algorithme itératif qui correspond à un vecteur principal de valeurs propres de la matrice normalisée des liens du web. Les PageRank sont alors utilisés pour réordonner la liste des résultats du SRI. L'algorithme proposé par Kleinberg [14], quant à lui, est restreint à l'ensemble

constitué par l'ensemble des pages en réponse à une requête augmenté des pages pointées (ou qui pointent) par (ou vers) ces pages. Kleinberg définit deux notions : « page autorité » et « page centrale ». Une page centrale est une page qui contient des liens vers des pages pertinentes, une page autorité est une page dont le contenu est pertinent. Un processus itératif est utilisé pour calculer, pour chaque page, son poids en tant que page autorité (source de contenu pertinent) et un poids en tant que page centrale (source de liens pertinents). Une des différences entre l'algorithme de Brin et Page et le modèle de Kleinberg est que, la qualité (ou autorité) d'une page est passée directement de pages autorités à d'autres pages autorités, sans interposer une notion de pages centrales.

D'autres études ont introduit d'autres approches combinant les informations extraites des contenus textuels des documents et les informations dérivées de la structure hypertexte [7] [2] [19].

L'algorithme proposé par Chakrabarti et al. est une modification de l'algorithme de Kleinberg incluant des informations issues de l'analyse des textes entourant les liens *href* (ou « anchor text » en anglais). Ils considèrent que le texte autour d'un lien *href* vers une page p est descriptif du contenu de cette page. A chaque lien est alors assigné un poids fonction de la ressemblance du texte qui l'entoure avec la requête. Ces poids sont ensuite combinés avec les informations issues de l'analyse des liens. Les algorithmes proposés dans [2] et [19] sont également des modifications de l'algorithme de Kleinberg. A la différence de la version de Chakrabarti et al., leurs modifications utilisent les textes intégraux des documents.

L'implémentation de Bharat et Henzinger considère une nouvelle requête qui est une extension de la requête originale avec les mots clés des premiers documents restitués par le système. Les poids des liens sont alors calculés par rapport à cette nouvelle requête.

En revanche, le modèle proposé par Silva et al. [19], basé sur les réseaux inférentiels bayésiens, ne nécessite aucune expansion de requête et est de ce fait moins coûteux en temps de traitement car la requête est plus ourte.

L'implémentation de Bharat & Henzinger, utilise un serveur de connexions [1] contenant 1 billion de liens correspondant à 100 millions de pages générées à partir du moteur de recherche Altavista.

D'autres travaux sont orientés dans d'autres directions, en particulier, l'utilisation des liens pour la classification de documents dans un environnement hypertexte [3] [22] [16] [10]. La classification est une technique d'organisation de la collection de documents ayant pour objectif l'identification de groupes (ou classes) de documents similaires. La similarité entre documents est généralement mesurée en tenant compte des termes qu'ils ont en commun. Les premiers travaux dans ce domaine se sont basés sur l'hypothèse de Van Rijsbergen selon laquelle les documents associés entre eux tendent à être pertinents pour les mêmes requêtes [20]. Botafogo [3] a proposé un algorithme de classification basé sur les liens hypertexte pour générer des classes de documents.
Dans leur modèle, la similarité entre deux noeuds hypertexte est proportionnelle au nombre de chemins indépendants entre eux. Weiss et al. [22] ainsi que Pirolli et Pitkow [16] ont proposé de nouvelles méthodes de classification basées sur la combinaison des informations dérivées du contenu textuel et de la structure hypertexte pour regrouper ensemble et catégoriser des pages sur le web. Plus récemment, Gibson et al. ont utilisé l'algorithme de Kleinberg [14] pour explorer la structure de communautés (classes) de pages centrales et de pages autorités sur le web en implantant une forme de décomposition en valeurs propres sur une matrice de connexions générée à partir du voisinage immédiat de l'ensemble des résultats en réponse à une requête.

Notre modèle présenté dans ce papier est orienté vers l’utilisation de la structure des liens hypertexte dans le but d’améliorer les performances de la recherche. Plus précisément, notre objectif est d’améliorer le taux de précision. A la différence des modèles présentés dans [14] [5], qui ignorent le contenu textuel des documents, notre schéma combine des informations extraites des contenus des documents et des informations dérivées de la structure hypertexte. Dans notre étude, l’analyse du contenu textuel est effectuée par un systèmes de recherche d’informations basé sur un modèle connexionniste [4] et porte sur la totalité des contenus des documents au lieu d’être restreinte à des portions de textes entourant les liens *href* comme dans [7]. Par ailleurs, nous n’utilisons pas d’expansion de requête, à la différence du modèle de Bharat

et Henzinger, ce qui permet d'avoir des temps de traitement plus courts.

3. Intégration des liens hypertexte dans le processus de recherche

3.1. Analyse du contenu textuel : le système Mercure

Les méthodes traditionnelles de recherche d'informations sont basées sur la comparaison des représentations des documents avec les représentations des requêtes. Un document (ou une requête) est représenté par un ensemble de termes pondérés. Les fonctions de pondération favorisent les termes représentatifs des contenus textuels des documents mais aussi les termes qui les différencient [18].

Dans notre étude, l'analyse du contenu textuel est effectuée par le système de recherche d'informations Mercure [4]. Celui-ci est une implémentation du modèle vectoriel utilisant l'approche neuronale. Le modèle de pondération des termes utilise des poids normalisés avec trois facteurs : la fréquence relative des termes (*tf*), la fréquence inverse (*idf*) et un facteur inversement proportionnel à la taille des documents. La fréquence relative d'un terme dans un document est le nombre d'occurrences de celui-ci dans ce document. La fréquence inverse d'un terme est inversement proportionnelle au nombre de documents de la collection dans lesquels celui-ci apparaît. Le facteur taille du document est utilisé pour compenser les grandes fréquences de termes dans le cas de gros documents conduisant généralement à la restitution des longs documents au détriment des plus courts (les fréquences d'apparition des termes étant plus élevées dans les documents longs). La fonction de pondération dans Mercure, inspirée des travaux de Robertson sur le projet OKAPI [17], est définie comme suit [4] :

$$w_{ij} = \frac{(1+\log(tf_{ij})).\left[h_1+h_2.(\log\left[\frac{M}{n_i}\right]\right]}{h_3+h_4.\frac{dl_j}{\Delta}}$$

Où ;

tf_{ij} : fréquence d'apparition du termes t_i dans le document d_j

M :nombre de documents de la collection,

n_i :nombre de document contenant le terme t_i

dl_j : nombre de termes d'indexation du document d_j

Δ : nombre moyen de termes dans un document,

h_1, h_2, h_3, h_4 : sont des paramétres dépendant de la collection et probablement de la nature des requêtes. Leur valeurs ont été obtenues par expérimentation sur la collection TREC-5 :h_1=0.8, h_2=0.2, h_3=8.8, h_4=0.2.

Le mécanisme de recherche d'informations de Mercure est basé sur un processus d'activation/propagation [4].
Après une succession de traitements, les documents dont les valeurs d'activation sont inférieures à un seuil prédéfini sont considérés non pertinents et ceux dont les valeurs d'activation sont supérieures au même seuil sont considérés pertinents. Ces derniers sont restitués dans l'ordre décroissant de leur valeur d'activation (qualifiée alors de valeur de pertinence).

3.2. Exploitation combinée des liens hypertexte et des contenus textuels des documents

3.2.1. Analyse des liens : notations

L'analyse de la structure des liens repose sur l'hypothèse que la structure hypertexte a un rôle sémantique dans l'interprétation des contenus des documents. En d'autres termes, lorsqu'un document référence, via un lien hypertexte, un autre document cela peut impliquer que les deux documents traitent de sujets similaires ou présentent des concepts reliés.
Le point de départ de notre analyse est l'ensemble des n premiers documents retrouvés par le SRI « Mercure » dans la phase d'analyse du contenu textuel. En suivant les liens hypertexte des documents de cet ensemble de départ, ce dernier peut être étendu en un nouvel ensemble de documents que l'on notera R_e.
En se basant sur l'hypothèse que les documents reliés ont des contenus similaires, les documents reliés à des documents non pertinents sont également non pertinents.

Pour cette raison, nous restreignons l'ensemble de départ aux *n* premiers documents. Dans nos expérimentations, nous avons fixé la valeur de *n* à 200.
L'expansion de l'ensemble de départ peut être faite en utilisant un ou plusieurs niveaux de profondeur tout en prenant garde d'éliminer les cycles. La profondeur entre deux documents *di* et *dj* reliés par des liens hypertexte est mesurée par la distance entre ces deux documents, c'est à dire le nombre de liens parcourus à partir de *di* pour atteindre *dj*.
Une collection de documents dans un environnement hypertexte peut être représentée sous forme d'un graphe orienté. Les noeuds correspondent aux documents et les liens correspondent aux liens hypertexte entre ces documents. De façon formelle, le graphe associé à l'ensemble R_e est défini comme suit :

G= (R_e,L)

Avec:

R_e est l'ensemble étendu de documents

L est l'ensemble des liens hypertexte entre les documents de R_e.

Les liens sont orientés : un lien $(d_i , d_j) \in L$ indique que le document d_i, contient un lien hypertexte vers le document d_j Ce lien est dit «sortant » pour d_i, et « entrant » pour d_j Les documents induits par les liens entrants (resp. sortants) et pour lesquels il existe un chemin vers un document *d* constituent l'ensemble des parents (resp. descendants) de *d* noté Pd (resp. Dd d).

3.2.2. Mesure de pertinence des documents restitués

La mesure que nous proposons pour le calcul de la pertinence d'un document hypertexte, en réponse à une requête, intègre des paramètres permettant de quantifier l'influence (positive ou négative) de la prise en compte de ses liens hypertexte entrants et sortants : le nombre de parents (liens entrants), le nombre de descendants (liens sortants) et la distance minimale entre le document considéré et ses parents et descendants.
Nous adoptons l'hypothèse que la pertinence d'un document est inversement proportionnelle à la distance minimale qui

le sépare de ses parents et descendants, c'est à dire, plus la distance minimale est grande, plus la pertinence du document décroît. D'autre part, la pertinence d'un document est d'autant plus forte que ses parents et descendants sont pertinents pour la même requête.

Nous définissons : $distm_{ij}$ ≡distance minimale du document di au document dj Comme les liens sont bidirectionnels, on a généralement $distm_{ij} \neq distm_{ij}$

De façon formelle, la mesure de pertinence que nous proposons est exprimée par l'équation suivante :

$$\text{Pertinence}(di) = S(di,q) + \beta \cdot \sum_{dj \in Pd_{di}} c(distm_{ij}) \cdot s(dj,q) + \gamma \cdot \sum_{dj \in Dd_{ii}} c(distm_{ij}) \cdot S(dj,q) \quad (1)$$

avec:

S('di ,q):similarité entre le document di et la requête q

Pdj : ensemble des parents de dj,

Ddj : ensemble des descendants de dj

Le coefficient $c(distm_{ij})$ est une fonction décroissante de la distance minimale entre le document *di* et le document *dj* telle que : $0 \leq c(distm_{ij}) \leq 1$. Nous utilisons la fonction $c(distm_{ij}) = e^{-distm_{ij}+1}$. Notre choix pour la fonction exponentielle est motivé par le fait qu'elle permet d'incurver la courbe. Elle est de ce fait plus discriminante pour les différentes distances minimales entre les documents. De plus, cette fonction assure que si le lien entre un document *di* et un document *dj* est direct, c'est à dire que la distance minimale entre eux est égale à 1, alors la contribution du document *dj* par rapport au poids de *di* est maximale. L'équation (1) inclut différents paramètres afin de permettre différentes expérimentations. Par exemple, nous pouvons considérer différentes stratégies selon que l'on prend en compte uniquement les liens entrants ($\gamma=0$), uniquement les liens sortants ($\beta=0$) ou les deux types de liens à la fois ($\gamma \neq 0 \text{ and } \beta \neq 0$). Cela nous permettra en particulier de mesurer l'impact des liens entrants et des liens sortants dans la mesure de pertinence d'un document donné. De même, nous pouvons utiliser plusieurs niveaux de profondeur (nombre maximum de liens à parcourir entre les documents).

4. Evaluation

L'objectif de nos expérimentations est d'évaluer les effets de l'intégration des liens hypertexte sur la précision des résultats de recherche d'informations.
Pour effectuer nos expérimentations, nous avons utilisé la collection de test WT2g de Trec-8 (plate-forme internationale de tests : http://www.nist.gov/) et un ensemble de 50 requêtes couvrant des domaines variés. Les caractéristiques de la collection WT2g sont données dans le tableau ci-dessous (Tableau 1). Des exemples de document et de requête sont donnés en annexe.

Taille de la collection	2,3 GB
Nombre de documents	247,491
Nombre de sites de provenance des documents	969
Nombre total de liens	1,166,702
Nombre maximum de liens entrants d'un document	3206
Nombre maximum de liens sortants d'un document	800
Nombre moyen de liens entrants par document	4,71
Nombre moyen de liens sortants par document	4,71

Tableau 1. Caractéristiques de la collection WT2g

L'analyse des liens hypertexte est utilisée afin de réordonner les 200 premiers documents restitués par le système Mercure.
Pour évaluer les performances de la recherche, le système calcule la précision moyenne (AvgPr) sur l'ensemble des documents restitués. La précision, qui mesure la qualité des résultats de recherche d'un SRI, est égale au nombre de documents retrouvés et pertinents sur le nombre de documents retrouvés. Pour chacune des 50 requêtes, les documents pertinents sont connus via la plate-forme Trec (cette information n'est utilisée qu'a posteriori pour évaluer les résultats de la recherche).

L'évaluation de la précision est effectuée sur l'ensemble des 50 requêtes. La précision moyenne est calculée sur les 5, 10, 15 et 20 premiers documents restitués. Pour mesurer l'efficacité de l'exploitation des liens, les résultats obtenus par différentes stratégies sont comparés au résultat du système Mercure, basé uniquement sur l'analyse du contenu textuel des documents. Les meilleures performances sont obtenues en utilisant un niveau de profondeur égal à 1 et des valeurs de paramètres β=0,11 et γ=0,1 .

	AvgPr@5		AvgPr@10		AvgPr@15		AvgPr@20		AvgPr@30	
Contenu textuel (Mercure)	0,4800		0,4640		0,4107		0,3840		0,3273	
Contenu textuel + liens entrants	0,4960	3,3%	0,4820	3,9%	0,4240	3,3%	0,3840	0%	0,3340	2,1%
Contenu textuel + liens sortants	0,5080	5,8%	0,4460	-3,9%	0,4027	-1,9%	0,3630	-5,5%	0,3180	-2,8%
Contenu textuel + liens entrants et liens sortants	0,4960	2,5%	0,4460	-3,9%	0,3960	-3,6%	0,3620	-5,7%	0,3080	-5,9%

Table 2. Variations de la précision moyenne (AvgPr) – Premier niveau de profondeur.

Le tableau 2 montre clairement que nous pouvons améliorer la précision de la recherche d'informations sur les 5, 10, 15, 30 premiers documents restitués. Les meilleurs résultats sont obtenus en utilisant uniquement les liens entrants des documents en combinaison avec les contenus textuels (2 ème ligne du tableau 2). L'utilisation des seuls liens sortants combinés aux contenus textuels (3 ème ligne du tableau 2) améliore la précision uniquement sur les 5 premiers documents retrouvés (1 ère colonne du tableau 2) mais entraîne une nette réduction des performances dans tous les autres cas de figures (autres colonnes du tableau 2). Nous obtenons les mêmes types de

résultats en prenant en compte les liens entrants et les liens sortants avec les contenus textuels (4 ème ligne du tableau 2).
En revanche, nous pouvons observer qu'aucune des stratégies utilisant les liens n'améliore la précision moyenne à 20 documents (4 ème colonne du tableau 2).

5. Discussion et perspectives

A l'issue de ces tests, nous pouvons en déduire que l'on peut améliorer la précision moyenne de la recherche d'informations grâce à l'intégration des liens hypertexte dans le processus de recherche d'informations. En particulier, cette amélioration est toujours effective grâce à l'utilisation des liens entrants. D'autre part, toutes les combinaisons améliorent la précision moyenne sur les 5 premiers documents restitués.
Cependant, l'exploitation des hyperliens n'est pas exempte de problèmes. En particulier, l'hypothèse selon laquelle deux documents reliés par un lien hypertexte ont des contenus ressemblants n'est pas toujours vérifiée, il faudrait définir une typologie des liens.
En effet, on peut remarquer qu'il existe au moins deux types de liens qui dépendent du contexte dans lequel ils ont été créés : les liens référentiels (fonctionnels) et les liens structurels. Les liens référentiels établissent des relations de sémantiques très diverses entre documents. En revanche, les liens structurels, appelés aussi organisationnels, sont généralement créés pour permettre la navigation à l'intérieur d'un document.
Intuitivement, il nous semblerait plus intéressant d'étudier l'influence des liens référentiels que celle des liens structurels. Or, dans la collection que nous avons utilisée, le nombre moyen de liens référentiels est très faible (en moyenne 0,19 lien par document). Ce qui ne nous permet pas ici de tirer des conclusions généralisables.
Actuellement, nous continuons nos expérimentations sur une autre collection de 10 GB de Trec-9 extraite du web. Etant plus volumineuse, nous espérons que cette collection est plus représentative de la structure des hyperliens du web.

6. Annexes

Exemple de document HTML de la collection Trec WT2g :

```
<DOC>
<DOCNO>WTX104-B01-1</DOCNO>
<DOCHDR>
http://msfcinfo.msfc.nasa.gov              :80/nmo/nmonasa.html
192.112.225.4 19970215104446 text/html 1014
HTTP/1.0 200 Document follows
Date : Sat, 15 Feb 1997 10 :37 :04 GMT
Server : NCSA/1.5
Content-type :text/html
</DOCHDR>
<HTML>
<HEAD>
<TITLE> Instructions to NASA Sponsors </TITLE> </HEAD>
<BODY>     <H1><STRONG>Instructions     to     NASA     Sponsors
</STRONG></H1><P>   <H3>JPL   is   under   the   institutional
management   of   the   Office   of   Space   Science   at   NASA
Headquarters. NASA.
Centers or activities contemplating the placement of research
and development work at the Jet
Propulsion Laboratory may contact the NASA Contracting Office
(<A href="mailto :
vstickley@nmo.jpl.nasa.gov">  vstickley@nmo.jpl.nasa.gov"  </a>
at the NMO for more details or the Research and Administration
Division  of  the  Office  of  Space  Science,  Code  SP  at  NASA
Headquarters.
</H3><HR>[<A     HREF="nmohome.html">NMO     Procurement     Home
Page</A>]<P>  Please  send  comments  and  questions  to  <A
href="mailto :kwolf@nmo.jpl.nasa.gov">
kwolf@nmo.jpl.nasa.gov</A><BR> Curator and Owner : Katherine
M. Wolk<BR>Last update to this page : september 15, 1995 @ 3
:23 p.m. PDT
</BODY>
</HTML>
</DOC>
```

Exemple de requête :

```
<top>
<num> Number : 428
<title> declining birth rights
<desc> Description :
Do any countries other than the U.S. and China have a
declining birth rate ?
<narr> Narrative :
To be relevant, a document will name a country other than the
U.S. or China in which the birth rate fell
```

from the rate of the previous year. The decline need not have occurred in more than the one preceding
year.
</top>

7. Références bibliographiques

[1] K. Bharat, A. Broder, M. Henzinger, P. Kumar, S. Venkatasubramanian, *"The connectivity server :* fast access to linkage information on the web ". Proceeding of the 7 th International World Wide Web Conference (WWW7), pages 469-477, Brisbane, Australia, 1998.

[2] K. Bharat and M. R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment" Proceedings of the 21 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Distributed Retrieval, pp. 104-111, 1998. Fifth International Symposium on Programming and Systems

[3] R. A. Botafogo, "Cluster analysis for hypertext systems". In ACM 16 th Annual International SIGIR'93, Pitsburgh, 1993.

[4] M. Boughanem, C. Chrisment, J. Mothe, C. Soulé-Dupuy and L. Tamine, "Connectionist and genetic approaches to achieve IR". In *Soft Computing in Information Retrieval Techniques and Applications* Editorial, F. Crestani and G. Pasi in Physica Verlag, (Springer Verlag), ISBN 3-7908-1299-4, pp. 173-198, 2000.

[5] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine". In Proceedings of the 7 th International World Wide Web Conference (WWW7), pages 107-117, Brisbane, Australia, 1998.

[6] J. Carriere and R. Kazman, "WebQuery : Searching and visualizing the web the web through connectivity". In Proceedings of the 6 th International World Wide Web Conference, 1997.

[7] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg, "Automatic ressource

compilation by analysing hyperlink structure and associated text". In Proceedings of the 7 th International World Wide Web Conference (WWW7), pages 65-74, Brisbane, Australia, 1998.

[8] B. Cronin and H. W. Snyder, "Citation indexing's achilles heel ? Evaluative bibliometrics and non coverage of the monographic literature". http://www.slis.indiana.edu/Research/cronin-achilles.html

[9] E. Garfield, "Citation analysis as a tool in journal evaluation". Science, pages 471-479, 1972. http://www.garfield.library.upenn.edu/essays/V1p527y1962-73

[10] D. Gibson, J. Kleinberg and P. Raghavan, "Inferring Web communities from link topology". In Proceedings of the 9 th Conference on Hypertext and Hypermedia, 1998.

[11] A. Grefenstette "SQLET : Short Query Linguistic Expansion Techniques : Palliating one or two-word queries by providing intermediate structure to www pages". In Proceedings of RIAO, 1997.

[12] H. Hauffe, "Is citation analysis a tool for evaluation of scientific contributions? ". In 13 th Winterworkshop on Biochemical and Clinical Aspects of Pteridines, St Christoph/Arlberg, Feb 25, 1994.

[13] M. Jansen, A. Spink, J. Bateman and T. Saracevic, "Real life information retrieval : A study of user queries on the web". In Proceedings of the 21 h Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 5-17, 1998.

[14] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment". In Proceedings of the 19 th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 668-677, San Francisco, California, 1998.

[15] R. Larson, "Bibliometrics of the World Wide Web : An exploratory analysis of the intellectual structure

of cyberspace". In Annual Neeting of the American Society in Information Science, 1996.

[16] J. Pirolli, J. Pitkow and R. Rao, "Silk from a sow's ear : Extracting usable structures from the web". In Proceedings of ACM SIGCHI Conference on Human Factors in Computing, 1996.

[17] S. E. Robertson, S. Walker, M. Beaulieu, "Okapi at TREC-7 : automatic ad hoc, filtering, VLC and interaction". In 7 th International Conference on Text Retrieval TREC-7, pages 253-264, 1998.

[18] G. Salton and M. J. McGill, "Introduction to modern Information Retrieval". Mc Graw Hill International Book Company, 1983.

[19] I. Silva, B. Ribeiro-Neto, P. Calado, E. Moura and N. Ziviani, "Link-based and content-based evidential information in a belief network model". In Proceedings of the 23 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 96-103, Athens, Greece, 2000.

[20] C.J. Van Rijsbergen, "Information Retrieval". 2 o Edition, Butterworths, Londres (UK), 1979.

[21] G. Venditto, "Search engine showndown". Internet World, Vol. 7, No 5, May 1996.

[22] R. Weiss, B. Velez, M. A. Sheldon, C. Manprempre, P. Szilagyi, A. Duda and D. K. Gifford, "HyPursuit : A hierarchical Network Search Engine that exploits Content-Link Hypertext Clustring". Proceedings of the 7 th ACM Conference on Hypertext, New York, 1996.