Analyse statistique pour la classification des pages WEB

Liang DONG*, Wahiba BAHSOUN*
dong@irit.fr , wahiba.Bahsoun@irit.fr

* Institut de Recherche en Informatique de Toulouse, Equipe SIG,
Université Paul Sabatier 118, route de Narbonne 31062 Toulouse cedex 4

Mots clefs :

Système de Recherche d'Information, système hypertexte, classification

Keywords :

Information Retrieval System (IRS), hypertext system, classification

Palabras claves :

Sistema de la recuperacion de datos, sistema des hypertext, clasificacion

Résumé :

Depuis le début des années 90, la possibilité d'accès au WEB pour un large public, entraîne la prolifération des informations et souvent une certaine redondance. La disponibilité de masses volumineuses d'informations doit être soutenue par des outils d'accès efficaces et ergonomiques. Plusieurs travaux ont été développés dans ce cadre[1], [2]. Dans le cas précis du WEB, les moteurs de recherche retournent habituellement une liste linéaire contenant des milliers de pages susceptibles de répondre aux besoins en information exprimés par l'utilisateur. Cette liste comporte souvent des documents quasi-similaires en terme de localisation géographique, contenu documentaire, versions, etc. Dans ce cadre, nous proposons une technique permettant de présenter à l'utilisateur une vision synthétique et globale des résultats de sa recherche. A cet effet, on s'oriente vers l'utilisation de l'analyse statistique dans les bases de données textuelles. L'analyse consistera essentiellement à mettre en œuvre des outils de classification des pages WEB selon de nombreux critères : localisation géographique, liens, contenu de la page. L'approche proposée nous a conduit à réaliser une expérimentation en l'implémentant sur le système Mercure.

Abstract :

Since the beginning of the Nineties, the web has had a very rapid growth in number of pages and a certain redundancy. The availability of these masses of information must be supported by effective and ergonomic tools of access to information. Several works were developed within this framework [1], [2]. In the precise case of the WEB, the search engines return usually a linear list containing thousands of Web pages likely to answer the requirement of information expressed by the user. This list often includes quasi-similar documents in term of geographical localisation, documentary contents, etc. Within this framework, we propose a technique allowing to present to the user a synthetic and total vision of the results of its research. To that effect, we turn towards utilisation of statistical analysis on the textual databases. The analysis will essentially consist in using tools for classification of Web pages according to many criteria: geographical localisation, links, page contents… The approach suggested led us to carry out an experimentation, and we implant it on the Mercure system.

1 Introduction

Avec le développement exponentiel d'Internet, le nombre des utilisateurs était estimé a plus de 500 millions en 2001[3], et plus de 605 millions en 2002 [4]. Le domaine de la Recherche d'Information (RI) se trouve face à de nouveaux défis pour l'accès à l'information. Les tâches principales de la recherche d'information sont donc la sélection d'informations, l'extraction et la synthèse des données pertinentes.

Les moteurs de recherche créent et maintiennent un index des mots dans des documents qu'ils trouvent sur le Web. Ils renvoient à l'utilisateur une liste ordonnée de documents appropriés comme résultat de la recherche. Peu de ces résultats peuvent satisfaire un utilisateur [5]. On propose plusieurs méthodes pour améliorer le taux de précision des réponses [6].

Pendant que le Web continue à se développer, la plupart des moteurs de recherche sont confrontés à certains problèmes, notamment le fait qu'ils ne peuvent pas indexer tous les documents sur le Web, en raison d'une forte croissance des volumes de données et du nombre de documents disponibles sur le Web.

Cependant, la navigation dans l'espace complexe d'informations comporte un risque pour l'utilisateur de se « perdre ». L'utilisateur hésitera à choisir les pages WEB qui l'intéressent, et il peut être de plus en plus éloigné de son besoin initial en information. Alors l'utilisateur n'est pas toujours satisfait du résultat restitué.

Dans la pratique, les documents retournés en réponse à la requête de l'utilisateur comportent des informations redondantes. Cette redondance entraîne une liste de documents trop longue [7]. En effet, certaines données sont copiées entièrement ou partiellement dans différents sites pour diverses raisons techniques, commerciales, culturelles, et d'autres sont révisées légèrement. Selon les résultats statistiques du groupe Naryanan de l'Université de Stanford (http://www-db.stanford.edu/), le nombre de pages similaires représente environ 22% du nombre des pages Web d'une part, et d'autre part, les informations sont parfois disponibles sur plusieurs types de médias et sous différents formats. Par exemple, le même article avec le même contenu peut être présenté sous des formes différentes comme html, pdf, post-script.

A cause de ces problèmes, l'utilisateur trouve difficilement des informations en relation avec ses besoins dans les listes trop longues de documents restitués. Donc, la recherche d'information pertinente devient un problème crucial, mais aussi les performances d'un SRI sont largement dépendantes de la qualité des méthodes de recherche d'information.

Notre travail consiste à synthétiser les résultats fournis par les moteurs de recherche et ensuite à les classifier pour les présenter à un utilisateur donné. Dans notre étude, nous avons bénéficié des apports du système PageRank et de l'algorithme HITS, qui sont les plus connus et utilisés dans le domaine de la propagation de pertinence sur le Web.

2 Modèle proposé

Notre modèle d'exploitation de la structure hypertexte en recherche d'information utilise conjointement les liens entrants et/ou sortants et le contenu textuel des documents reliés. La figure 1 présente le contexte et l'enchaînement du processus, l'utilisateur est au centre du processus de recherche d'information, dans l'ordre suivant :

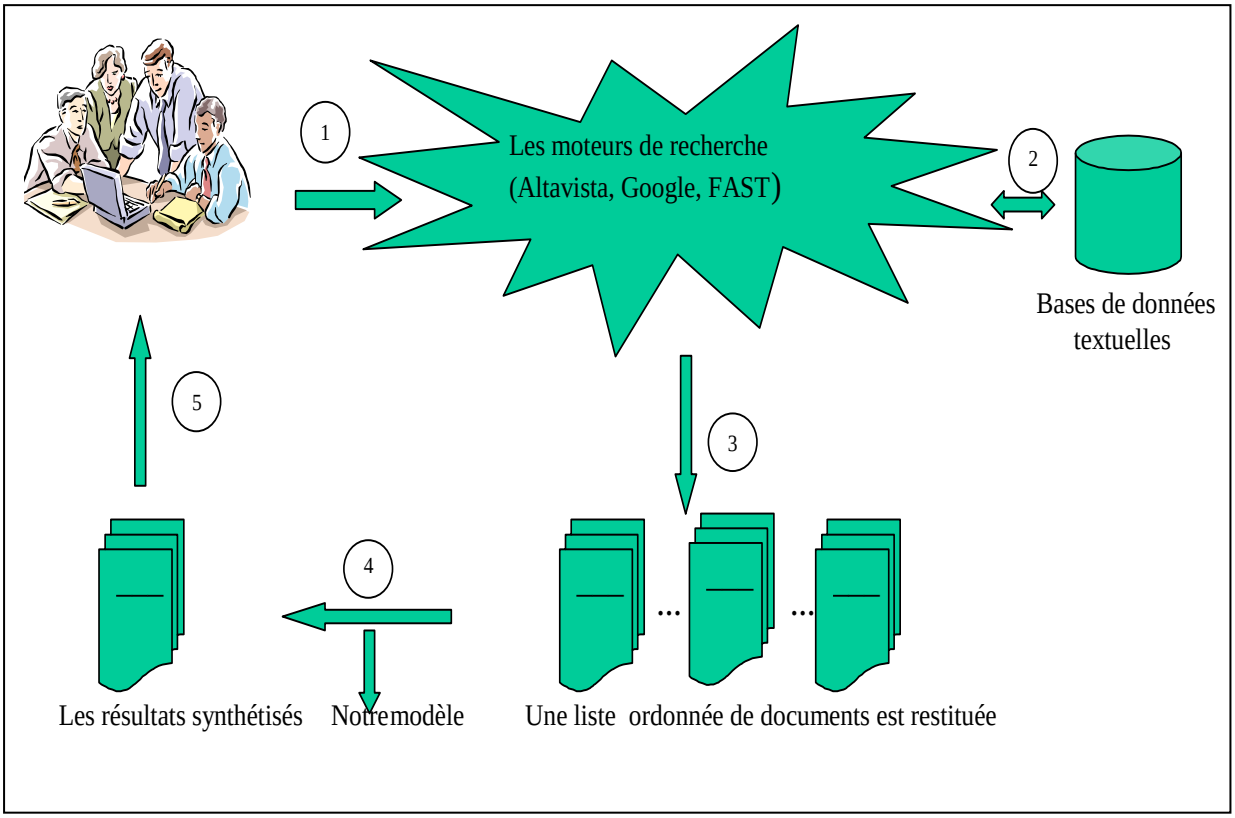

Figure 1. *Le contexte de modèle de la recherche d'information*

1) Il se connecte à Internet à partir de n'importe quel terminal. Il exprime son besoin de recherche. Il choisit un moteur de recherche et formule sa requête.

2) Le moteur de recherche reçoit une requête, recherche l'information dans les bases de données textuelles.

3) Le moteur de recherche choisi, restitue une liste ordonnée de documents, qui répond au besoin en information exprimé par l'utilisateur.

4) Dans notre modèle, on utilise l'analyse statistique afin de proposer une technique permettant de présenter à l'utilisateur une vision synthétique et globale des résultats de sa recherche tout en contrôlant le nombre de documents restitués.

5) On retourne les résultats synthétisés à l'utilisateur.

Dans notre étude, l'analyse du contenu textuel est effectuée par un Système de Recherche d'Informations (SRI) basée sur une classification réalisée sur la métrique du cosinus, ensuite on intègre des hyperliens dans le processus de Recherche d'Informations afin de fournir de meilleurs résultats de recherche à l'utilisateur.

Notre étude est constituée de 2 étapes, la figure 2 présente le contexte et l'enchaînement du processus.:

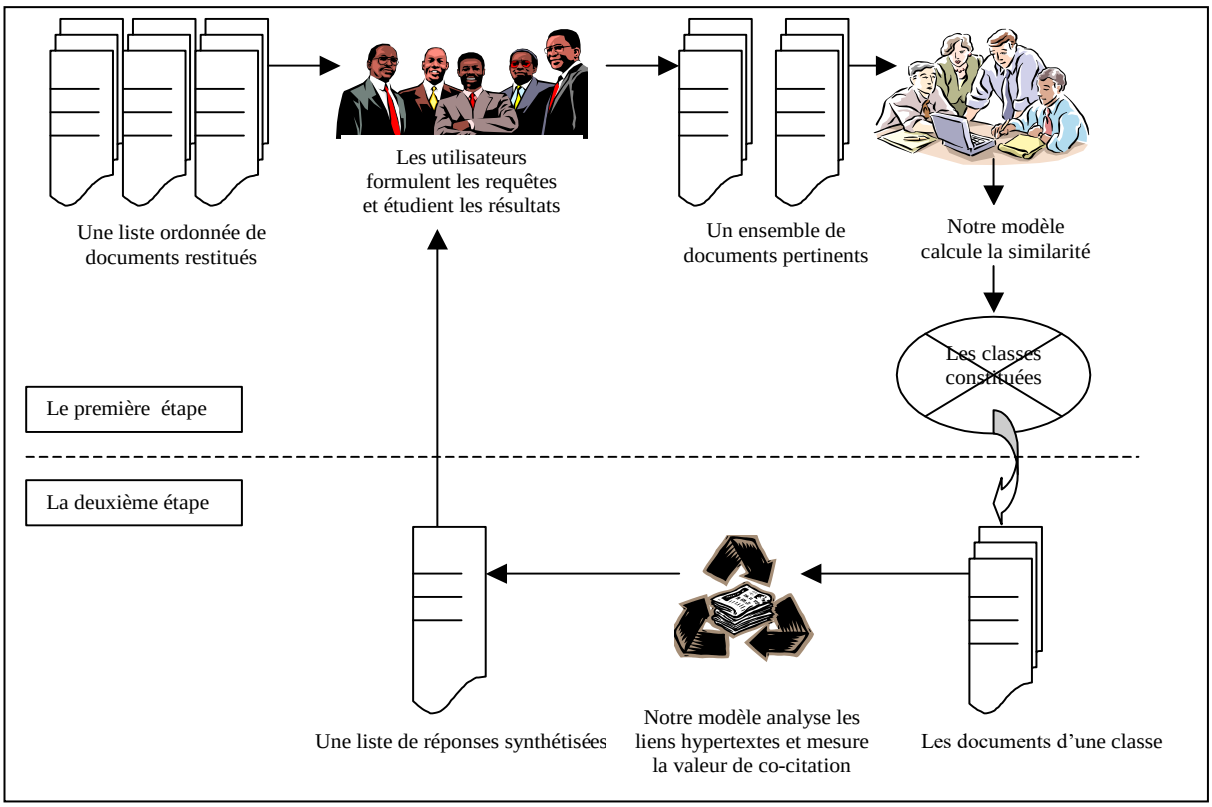

Figure 2. *Le contexte de notre modèle d'exploitation*

2.1 La première étape de l'étude

Cette première étape concerne l'identification des classes. On identifie les groupes ou classes de la collection de documents similaires afin de synthétiser les documents restitués. Pour cette étape, on s'est inspiré de la méthode d'analyse du contenu textuel qui offre un fondement pour regrouper les documents similaires.
L'approche adoptée est constituée des points suivants :

- Selon le Système de Recherche d'Informations, le moteur restitue une liste ordonnée de documents ou de pages dont le degré de pertinence par rapport à la requête est calculé en fonction des mots-clés (de façon classique), on utilise l'approche de la classification.

- Sélectionner les *n* premiers documents ou pages restitués dans le classement de la recherche initiale. Ce sous-ensemble correspond à un ensemble local, il est relativement petit et riche en documents pertinents. On peut donc contrôler le volume de données.

- Calculer la similarité entre tous les documents pris deux à deux, en utilisant la fonction du cosinus. Cette fonction se base sur un modèle vectoriel [8] de représentation de l'espace des données, sa dimension est égale au nombre de termes d'indexation de la collection. Chaque document d est représenté au moyen d'un vecteur d^{VS} [1] = (d_1^{VS}, d_2^{VS},… $d_{|T|}^{VS}$), appelé profil lexical, dans lequel la $j^{ème}$ composante d_j^{VS} représente le poids, dans le document d, du terme d'indexation t_j associé à la $j^{ème}$ dimension de l'espace vectoriel.
D'une façon générale, la mesure du poids utilisée est le plus souvent une fonction de la fréquence du terme dans le document et intègre de plus une pondération locale (dans le document), une pondération globale (dans l'ensemble des documents à classer) et un facteur de normalisation par rapport à la longueur du document. Cette pondération reflète l'idée qu'un terme est important s'il est fréquent dans un document et peut fréquent dans les autres.
La fonction de cosinus est une mesure largement utilisée et développée à partir du cosinus de l'angle entre les deux vecteurs représentatifs des deux documents. Elle fournit une valeur réelle entre 0 et 1.

[1] vectoriel standard. L'espace vectoriel est l'ensemble de termes des documents Collectés.

$$cos(d_i,d_j)=\frac{d_i \bullet d_j}{|d_i| \times |d_j|} \qquad [a]$$

Si l'angle entre deux vecteurs est petit, la valeur du cosinus est grande. Ces deux documents sont supposés proches l'un de l'autre. Les documents sont regroupés dans une classe lorsqu'ils présentent une similitude assez importante.

- A partir de ces valeurs de similarité obtenues, on peut rassembler les documents dans des classes, en adoptant un certain seuil. La valeur choisie est précisée dans le chapitre qui traite de l'expérimentation. Les classes sont disjointes.

2.2 La deuxième étape de l'étude

Dans cette seconde étape, on extrait les documents pertinents à partir des classes déjà construites afin de contrôler le volume et la qualité de l'ensemble des documents restitués. Pour atteindre cet objectif, on analyse la structure des liens hypertextes entre les documents, en s'intéressant plus particulièrement aux co-citations car on tient compte des liens entrants et sortants à partir des classes déjà construites pour faire la sélection.
Pour chaque classe :

- on identifie les liens hypertextes entre les documents,

- s'il n'y a pas beaucoup de liens dans un document, ce document est considéré comme un document informatif, on choisit ce document dans la liste des réponses,

sinon, on mesure une valeur de co-citation entre ces documents,

- on choisit le(s) document(s) le(s) plus co-cité(s),

Dans notre modèle d'exploitation, pour chaque classe déjà construite, on supprime les documents qui possèdent les mêmes parents, afin de contrôler le volume de l'ensemble des documents restitués. De plus, on extrait les meilleurs pages en tant qu'un bon point de départ pour une navigation dans une zone de pertinence.

3 Exemple simple

Avant de présenter une expérimentation sur une collection assez significative de documents, nous proposons l'exemple suivant : Soient **cinq** documents, le nombre maximum des termes d'un document est de **trois.** Le poids de chaque terme est présenté dans ce tableau :

documents / poids / termes	Document1	Document2	Document3	Document4	Document5
T1		0.3	0.3		0.3
T2	0.2				0.3
T3		0.52		0.3	

A partir de cette matrice de termes pondérés, on peut donc désigner l'espace vectoriel multidimensionnel.

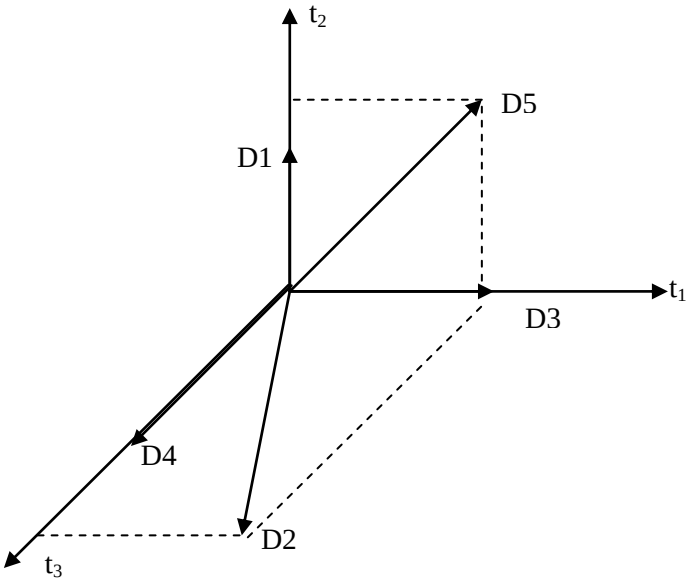

Figure 3. Espace vectoriel multidimensionnel

Si on applique le principe de calcul exposé dans cette figure et la formule [a], on obtient le tableau suivant:

	D1	D2	D3	D4	D5
D1	1	0	0	0	0,707
D2	0	1	0,5	0,866	0,353
D3	0	0,5	1	0	0,707
D4	0	0,866	0	1	0
D5	0,707	0,353	0,707	0	1

Lorsque la valeur du cosinus est égale à zéro (cosinus = 0), cela signifie que ces deux documents sont distincts. Cependant lorsque la valeur du cosinus est égale à un (cosinus = 1), cela signifie que ces deux documents ont les mêmes termes, et sont similaires.
Les documents sont regroupés dans une classe lorsqu'ils présentent une similitude assez importante ou la valeur du cosinus entre les documents est assez forte. A partir de cette matrice, on construit le graphe de classement.
Dans le graphe, les nœuds sont des documents restitués. Il existe un lien entre deux nœuds d_i et d_j si $\cos(d_i,d_j) \geq seuil$. Ici, on prend seuil = 0,6 et on associe ces 5 documents.

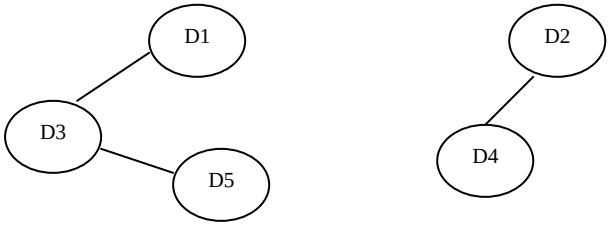

Figure 4. Le graphe de classement

A partir de ce graphe de classement de la figure 4, on peut rassembler les nœuds (documents) dans des classes selon une certaine stratégie. Ici, on adopte la stratégie suivante :
- une classe est un ensemble de documents,
- les classes sont disjointes,

Si on suit le principe de ce graphe de classement, l'application de cette stratégie donne la classe suivante :
Les classes : (D1, D3, D5) et (D2, D4)

La deuxième étape consiste à analyser la structure des liens hypertextes entre des documents d'une même classe. Si deux documents ont le même premier niveau de l'adresse URL, cela signifie que ces documents sont dans le même site Web. Le schéma suivant montre un exemple de structure du site de la bibliothèque d'Université Paul Sabatier.

http://www.biu-toulouse.fr/ups/

- http://www.biu-toulouse.fr/ups/biblio_sciences.htm
- http://www.biu-toulouse.fr/ups/biblio_sante.htm
- http://www.biu-toulouse.fr/ups/biblio_associees.htm

Figure 5. La structure du site

Le premier niveau de l'adresse URL de ces liens hiérarchiques (les liens de navigation interne) est le même : http://www.biu-toulouse.fr/ups/.

Pour éviter la répétition, on élimine ainsi la majorité des liens hiérarchiques. Et puis, on utilise la méthode des co-citations. On identifie les liens hypertextes entre les documents. Supposons qu'on ait des liens comme le tableau suivant :

	U	BF1	BF2	BF3	BF4	BF5	BF6	BF7	BF8
B1	√		√	√			√		√
B2	√		√		√	√	√	√	
B3	√	√				√	√	√	

On remonte à un ensemble de « B » parents du document « u ». Après, on récupère les descendants de chacun des éléments parents, soit « BF », autres que « u ».

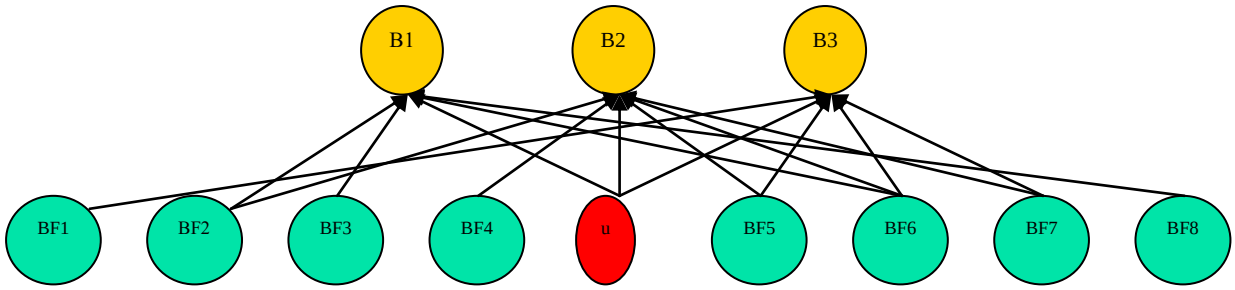

Figure 6. Le graphe des liens hypertextes entre des documents

Pour chaque élément de « BF », on détermine le degré de co-citation avec « u » . Le degré de co-citation entre deux nœuds est égal au nombre de parents qu'ils ont en commun.
A partir des observations précédentes, on peut en déduire le tableau suivant:

BF	1	2	3	4	5	6	7	8
degré	1	2	1	1	2	3	2	1

A partir des valeurs de co-citation de ce tableau, on extrait le(s) document(s) plus co-cité(s), en l'(les) intégrant à la liste des réponses. Dans cet exemple, les documents que l'on a intégré à la liste des réponses sont les documents « u » et « BF6 ».

4 Expérimentation et Evaluation

Pour réaliser notre expérimentation, on a utilisé la collection de documents du système Mercure (plate-forme de tests : /usr/local/fc4500-0/RI/Mercure/). Le système Mercure est disponible dans l'équipe de recherche. Ce système sert uniquement à récupérer les documents et utiliser les requêtes existantes. En effet, il existe 50 requêtes. Ces requêtes couvrent des domaines variés. La taille de la collection est 2,73 Gigaoctets ce qui correspond à 5164 documents Web.

La démarche qu'on adopte dans le processus de recherche et que nous utilisons pour les expérimentations est réalisée à partir des deux étapes précédemment citées.

Etape 1 : Analyse des documents pour identifier des classes.

Les contenus textuels des documents de la collection sont trouvés par le système Mercure. Pour une requête utilisateur, le système calcule le poids de chaque document dans la collection, et il restitue une liste ordonnée de documents à l'utilisateur. Cette liste est définie par le degré de pertinence par rapport à la requête.

Cette étape est décomposée en quatre phases :

- Dans cette première expérimentation préliminaire, on a sélectionné uniquement les dix premiers documents restitués par le système Mercure, c'est-à-dire, **n=10**.
- On utilise la fonction de Mercure---« terme_doc1 » afin de trouver le nombre de termes dans le document et la fréquence d'un terme dans ce document.
- En utilisant la formule du cosinus, on calcule la similarité entre tous les documents pris deux à deux,
- Et enfin, on constitue des classes en fonction des valeurs de similitude calculées dans la phase précédente.

Etape 2 : Analyse des liens hypertextes pour extraire des documents pertinents

A partir de ces classes, on analyse des liens hypertextes de chaque document dans une classe.

Cette étape est décomposée de cinq phases :

- Identifier les liens hypertextes entre les documents par les liens HREF dans la page HTML.
- Selon les liens identifiés, on construit la structure des liens entrants et des liens sortants d'un document, en utilisant la fonction de Mercure --- « docid_to_url ».
- Pour éviter la répétition, on élimine la majorité des liens hiérarchiques. Ensuite, les documents informatifs, sont ajoutés dans la liste des réponses.
- A partir de cette structure, on mesure la valeur de co-citation entre les documents dans une classe donnée.
- On choisit le(s) document(s) le(s) plus co-cité(s), en ordonnant les documents en fonction de leur degré de pertinence, afin de restituer cette liste de documents ordonnée à l'utilisateur.

Dans l'expérimentation, la méthode de classification calcule la similarité entre tous les documents restitués, et les résultats sont comparés avec un certain seuil.

Dans notre cas, on utilise dix documents. Si on choisit un seuillage égale à 0,95, on peut obtenir 3 classes différentes. C'est-à-dire, que la valeur du cosinus entre deux documents est supérieure à ce seuil, on peut donc sélectionner ce document dans cette classe. Ensuite, on choisit une autre valeur de seuillage égale à 0,99. Avec cette valeur, on obtient 5 classes différentes. En résumé, on constate que si la valeur du seuillage est grande, le lien entre deux documents est plus faible, alors, on peut obtenir plusieurs classes.

La figure 7 montre la situation des liens hypertextes sur **dix** documents.

Le nombre maximum de liens sortants d'un document dans une classe	5498
Le nombre minimum de liens sortants d'un document dans une classe	1366

Figure 7. La situation des liens hypertexte

La figure 8 montre le nombre de classes obtenues pour chaque seuillage.

Résultat d'échantillonnage	Seuil = 0.95	Seuil = 0.99
Le nombre de classes	3	5
Le degré maximum de co-citation entre deux documents dans chaque classe	4792	5227
La probabilité d'équivalence avec le SRI Mercure	95%	90%

Figure 8. Résultats de l'échantillonnage

Enfin, les résultats obtenus sont comparés avec les documents pertinents du système Mercure.
A l'issue de ces tests, les résultats montrent qu'on peut présenter à l'utilisateur une vision synthétique et globale des résultats de sa recherche grâce à l'utilisation combinée des liens hypertextes et des contenus textuels des documents.

5 Conclusion et perspective

Confronté à une masse importante et sans cesse croissante d'informations disponibles sur le Web, la plupart des systèmes d'accès à l'information sont basés sur des modèles classiques (mots-clés). Pour améliorer les performances des Systèmes de Recherche d'Informations classiques, il faut prendre en compte des hyperliens dans le processus de recherche d'informations. L'étude que nous présentons utilise les résultats restitués par un Système de Recherche d'Informations. Cet ensemble est étendu suivant les liens de l'hypertexte.
La principale caractéristique de notre étude concerne la mise en œuvre d'un processus pour calculer la similitude afin de classer précisément des documents distincts et éviter la duplication d'informations. On a également considéré l'information contenue dans les liens hypertextes afin d'augmenter la qualité de la réponse fournie par le moteur de recherche.
L'analyse du contenu textuel des documents retrouve l'ensemble de documents répondant à ces requêtes.
L'approche des liens hypertextes permet l'exploitation de la structure hypertexte. A la différence des moteurs de recherche traditionnels, l'utilisation des liens hypertextes permet une consultation non-linéaire d'une collection de documents par le moyen des liens ainsi qu'une détection des documents des co-citations. Enfin, on extrait des documents pertinents qu'on retourne à l'utilisateur, par rapport aux critères choisis comme le nombre des liens sortants et l'importance des documents pères.
En somme, cette approche permet à l'utilisateur de naviguer aisément dans la liste restituée, améliore le taux de précision des réponses, évite la redondance d'informations et réduit des listes trop longues de documents restitués. La perspective d'évolution de notre travail est d'implémenter une solution pour résoudre les problèmes d'optimisation liées aux temps de réponse, et d'ajuster l'emploi de la stratégie de recherche afin d'obtenir une meilleure recherche.

6 Bibliographie

[1] Salton G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer,* Addison-Wesley, Reading, 1989

[2] Robertson S. E., Walker S., Hnacock-beaulieu M. M., *Large test collection experiments on an operational, interactive system: Okapi at TREC,* Information Processing & Management, 31(3), 1995

[3] NUA *Internet Surveys, http://www.nua.ie/surveys,* Rapport technique, Dublin, Irlande, NUA, août 2001

[4] NUA *Internet Surveys, http://www.nua.ie/surveys,* Rapport technique, Dublin, Irlande, NUA, août 2002

[5] Mizzaro S., *Relevance: The whole history*, Journal of the American Society for Information Science, 48(9): 810-832, 1997

[6] Lawrence S., *Context in web Search,* IEEE Data Engineering Bulletin, Volume 23, Number 3, pp. 25-32, 2000

[7] Abchiche M., *Intégration des liens hypertextes dans la recherche d'information*, 19ème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID'01), pp 253-266, Martigny, Suisse, Mai 2001

[8] Salton et al., *Introduction to Modern Information Retrieval*, McGraw Hill. 1983.