Utiliser les liens pour adapter les moteurs de recherche aux spécificités du WEB

Saïd Radhouani, Jean-Pierre Chevallet

Utiliser les liens pour adapter les moteurs de recherche aux spécificités du WEB.

Saïd Radhouani, Jean-Pierre Chevallet
Laboratoire CLIPS-IMAG
B.P. 53 38041 Grenoble cedex 9, France
Said.Radhouani@imag.fr
Jean-Pierre.Chevallet@imag.fr

Résumé : Le WEB, hypertexte mondial, nécessite l'utilisation de moteurs de recherche pour retrouver l'information. Actuellement la notion de page WEB est communément utilisée comme atome d'information retourné à l'utilisateur. Les liens entre pages sont très peu utilisés pour améliorer la qualité des réponses. Nous proposons brièvement une utilisation simple des liens dans les pages WEB, et un niveau d'indexation indépendant de la page WEB, considérée comme granularité physique.
Mots clés : Information, Représentation, Accès et Traitement Intelligent, WEB, Liens Hypertextes.

1. INTRODUCTION

Depuis son apparition, le WEB n'a cessé de progresser pour devenir une source formidable d'informations immédiatement disponibles. Toutefois, l'accès à l'information est devenu de plus en plus difficile car l'internaute est perdu dans la masse disponible. Devant la richesse du WEB en terme de pages et de liens, la navigation est évidement très limitée pour rechercher des informations et l'accès à l'information basé sur une requête semble plus efficace. En effet, les moteurs de recherche sur le WEB sont une aide inestimable pour rechercher une information. Actuellement, ils fonctionnent sur le principe qu'une page HTML ou tout autre fichier document, est l'atome d'information que l'internaute recherche. Ils tiennent peu compte du fait que le WEB est avant tout un système hypermédia issu des hypertextes. L'indexation se réalise donc au niveau de granularité de la page HTML et le réseau de liens entre les pages est peu exploité pour améliorer la qualité de la recherche d'information. Pourtant, il est clair que le découpage physique en page n'est pas forcément lié au contenu de la page. En effet, on trouve sur le WEB, des pages de tailles très différentes, certaines même contiennent l'équivalent d'un ouvrage complet.

Les moteurs de recherche du web sont basés sur des modèles booléens ou vectoriels de RI qui ont été développés pour des documents atomiques. Nous mettons en doute l'adaptation de ces modèles aux spécificités du WEB, en particulier pour des recherches plus précises au niveau des réponses. La question que nous voulons poser à travers ce document concerne le passage des moteurs de recherche actuels orientés plutôt vers le rappel, à des moteurs de recherche plus orientés vers la précision en tenant compte de la structure hypertextuelle de cet espace d'information.

Nous évoquons d'abord quelques méthodes d'utilisations de liens pour la recherche d'information. Ensuite, nous présentons quelques travaux d'évaluation des méthodes précitées, avant de discuter l'état actuel des travaux sur les liens. Nous concluons par une proposition brève pour l'utilisation des liens dans les pages WEB.

2. L'APPARITION DU WEB

En 1989, deux ingénieurs du CERN[1] de Genève, Timothy Berners-Lee et Robert Cailliau, eurent l'idée d'utiliser les liens hypertextes pour permettre la consultation simultanée de tous les sites serveurs d'Internet : par convention, cette date marque la naissance du World Wide Web. En 2000 existaient déjà 2 milliards de pages WEB, ayant chacune un nombre considérable de liens hypertextes. Le lien hypertexte est ainsi devenu le moyen principal pour naviguer sur Internet. Les documents entre lesquels on peut naviguer grâce aux liens sont écrits en langage HTML.

Nous pouvons donc considérer que le WEB est un gigantesque hypermédia issu des systèmes hypertextes. La richesse d'un document hypertexte réside non seulement dans son contenu informationnel mais également dans la richesse de ses liens et c'est ce qui complexifie l'élaboration d'un modèle bien précis pour l'hypertexte [Devlin, 1991]. Dans [Park, 1998], une étude a été menée afin de dégager les propriétés structurelles de l'hypertexte et de la définir dans un langage formel : l'hypertexte étant non structuré vu son aspect non linéaire présente un sérieux problème lors de la navigation [Aleese, 1990].

.1 Recherche d'information sur le WEB

[1] Centre européen pour la recherche nucléaire

Traditionnellement, le processus de RI consiste à retrouver, à partir d'une collection de documents, les documents qui correspondent le mieux à la requête utilisateur. Dans les modèles les plus simples, une requête est formée par un ensemble de mots clés, et le Système de Recherche d'Information (SRI) doit trouver les documents qui contiennent ces mots.

Avec l'apparition du WEB, les documents disponibles sont devenus très importants. Cette collection a de nouvelles caractéristiques : les données sont hétérogènes et distribuées sur plusieurs sites, de nouveaux médias autre que le texte sont accessibles (image, vidéo, son.), et les documents sont liés par des liens hypertextes. Ces nouvelles caractéristiques exigent de nouveaux critères de pertinence d'un document par rapport à une requête. En particulier l'information disponible dans la collection ne se résume plus au contenu textuel, et les traitements d'extraction des index doivent tenir compte de ces aspects multimédias. La pertinence, basée sur la correspondance entre les mots clés de la requête et le texte du document, est insuffisante. Il faut noter aussi qu'il est en général difficile pour l'utilisateur, de traduire un besoin d'information par une requête : il doit s'adapter à un langage artificiel d'interrogation, même s'il est très simple (quelques opérateurs). Il doit également se faire une image mentale (parfois fausse) sur comment le système produit ses résultats avec une série de mots composés par quelques opérateurs.

Les besoins en RI sur le WEB sont certainement très variés, comme le sont les types d'informations disponibles (cf. [Prime, 2002]). Ces besoins ne sont pas pris en compte, tout comme ne sont pas pris en compte les types de pages différents. De fait, la plupart des moteurs considèrent que le WEB est un ensemble de documents atomiques et indépendants, dont la granularité est celle d'une page HTML. Ils ne tiennent pas compte du fait que le WEB est avant tout un ensemble de documents liés par des liens hypertextes et que les pages HTML indexées indépendamment les unes des autres perdent leur contexte. Pour ces raisons, différentes méthodes ont été développées afin d'utiliser les liens dans le processus de RI sur le WEB.

3. UTILISATION DES LIENS POUR LA RI

Dans la suite, nous évoquons quatre approches d'utilisation des liens pour la RI. Nous citons, également, quelques travaux qui ont concrétisé ces approches avec éventuellement certaines modifications.

.1 Le PageRank [Brin, 1998]

Cette approche est basée sur la notion de propagation de popularité. Le principe est d'évaluer l'importance d'une page en fonction de chaque page pointant vers elle. La propagation met en avant les pages qui jouent un rôle particulier dans le réseau des liens, avec l'hypothèse : "*une page référencée par un grand nombre de pages est une bonne page*". Cette mesure est une distribution de probabilité sur les pages. Elle mesure en fait la probabilité PR pour un internaute navigant au hasard, d'atteindre une page donnée P. Cette probabilité est d'autant plus forte que le nombre de pages qui réfèrent P est important. PR est donc fonction de la somme des probabilités des pages qui référencent P. Il faut aussi tenir compte du fait que les pages qui référencent P ont d'autres liens sortant vers d'autres pages que P. Il faut alors diviser cette probabilité par le nombre C de liens sortant des pages qui référencent P. Finalement, il faut tenir compte du fait qu'un internaute peut arriver sur une page quelconque sans suivre de liens (à partir de ses signets par exemple).

La formule proposée [Brin, 1998] tient ainsi compte de la probabilité d de suivre effectivement les liens. La probabilité d'arriver à la page P sans suivre de liens est donc de (1-d). La formule de calcul du PageRank est alors la suivante :

$$PR(P) = (1-d) + d[(PR(P_1)/C(P_1) + \ldots. + (PR(P_m)/C(P_m))] \quad (1)$$

Avec d est une probabilité constante,
$C(P_j)$ = nombre de liens sortant de la page P_j.

Cette mesure a tendance à mettre en valeur les pages qui sont référencées par des pages à forte probabilité. On remarque que la formule est récursive : la valeur du PageRank (score) d'une page P, est calculée par un algorithme itératif. Initialement, toutes les pages sont équiprobables, leur valeur de PageRank est alors égale à 1/N avec N est le nombre de documents de la collection. Un cycle d'itération propage les probabilités sur les liens. L'algorithme s'arrête théoriquement lorsqu'une nouvelle itération ne produit plus de modifications dans le graphe des valeurs de PR.

Ce calcul se fait lors de la phase d'indexation (indépendamment des requêtes). Cette valeur peut alors être utilisée à l'interrogation pour modifier l'ordre de pertinence des pages évaluées par le seul critère de comparaison du contenu. L'approche suivante tient compte de la requête dans la propagation de la pertinence sur les liens.

.2 L'approche de propagation de pertinence [Crestani, 2000, Savoy, 2000]

Le principe de cette approche consiste à propager des valeurs de similarité de documents par rapport à une requête avec l'hypothèse suivante : "*un document référencé par un grand nombre de documents pertinents est un bon document*". La typologie du graphe des liens entre les pages est ici prise en compte au moment du calcul de la valeur de pertinence (Relevance Status Value RSV) des documents. Elle est alors fonction de la valeur de pertinence du document D_i par rapport à la requête Q, mais elle va aussi dépendre des valeurs de pertinence des documents liés au document D_i par rapport à la requête. Ainsi la valeur de pertinence finale notée RSV_{fin} se calcule comme suit :

$$RSV_{fin}(Di,Q) = RSV_0(Di,Q) + \lambda \cdot \sum_{j=1}^{k} RSV_{fin-1}(Dj,Q) \cdot \quad (2)$$

Avec :
D_i : un document.
D_j : les documents liés à D_i.

Q : une requête.
λ : une constante destinée à atténuer la propagation.
K : le nombre de document reliés au document D_j.
RSV_{fin} : la pertinence finale du document Di par rapport à la requête Q.
RSV_0 : la pertinence initiale du document Di basée sur le contenu.
RSV_{fin-1} : est la pertinence du document Dj par rapport à la requête Q lors de l'avant dernière itération.

La première étape consiste à calculer la similarité entre une requête donnée et les documents de la collection. Cette valeur est notée $RSV_0(D_i,Q)$. La deuxième étape consiste à propager cette valeur dans le réseau de documents à travers un certain nombre de cycles en utilisant des facteurs de propagations. A chaque cycle, la pertinence d'un document change en fonction de la pertinence des documents voisins. Cette propagation peut être aussi mise en œuvre en propageant une pertinence binaire, c'est à dire, une valeur qui est à 1 si le document est jugé pertinent, et à 0 s'il ne l'est pas.

.3 Le système probabiliste d'argumentation (PAS) [Picard, 1998]

Dans cette approche, au lieu de propager la valeur de similarité d'un document par rapport à une requête, on propage la probabilité qu'il soit pertinent. La première étape consiste à calculer la probabilité de pertinence d'un document D_i notée $P(D_i|rank)$. A la deuxième étape, cette valeur de probabilité de pertinence est modifiée en fonction des valeurs de probabilités des documents voisins au document D_i.

La contribution d'un document D_j lié à D_i est égale à $P(D_j|rank).P(Link)$ au lieu de $\lambda.\sum_{j=1}^{k} RSV_i(Dj,Q)$ utilisé dans l'approche de propagation de pertinence (cf. .3).

La probabilité d'un document D_i est multipliée par les probabilités des liens entrants noté $P(Link_{in})$ et des liens sortants $P(Link_{out})$. Le degré de support d'un document noté DSP est alors calculé comme suit :

$$DSP(D_i) = 1 - (1-P(D_i \mid rank)) * [1 - P(D_{in} \mid rank) * P(link_{in})] * [1 - P(D_{out} \mid rank) * P(link_{out})]. \quad (3)$$

Dans [Savoy, 2000], cette approche a été mise en œuvre en utilisant seulement les sources[2] d'évidence importantes en choisissant les meilleurs documents entrant D_{in} et les meilleurs sortant D_{out} au lieu d'utiliser tous les documents voisins à un document D_i.

.4 Algorithme de Kleinberg (HITS) [Kleinberg, 1998]

Cette approche consiste à calculer la popularité (*Hub*) et l'autorité (*Authority*) d'un document et ce pour classer les documents résultats par rapport à une requête. L'hypothèse est : "*Un document qui pointe vers beaucoup de bonnes Authorities est un bon Hub, et un document pointé par beaucoup de bons Hubs est une bonne Authority*". Le principe est de calculer le nombre de documents qui pointent vers un document D_i et celui des documents que D_i pointe. Les *Hubs* (H) et les *Authorities* (A) sont calculés de la façon suivante :

$$A^{c+1}(D_i) = \sum_{D_j \in parent(D_i)} H^c(D_j) \cdot \quad (4)$$

$$H^{c+1}(D_i) = \sum_{D_i \in fils(D_j)} A^c(D_j) \cdot \quad (5)$$

Avec :
c+1 : le nombre d'itérations.
Parent(D_i) : l'ensemble des documents qui pointent vers le document D_i.
Fils(D_j) : l'ensemble des documents que D_i pointe.

Les *Hubs* et les *Authorities* sont initialisés à 1. Durant *c*+1 itérations, les documents sont répartis dans deux listes, l'une classée selon les *Hub*, l'autre selon les A*uthorities*. La liste finale des documents classés peut être l'intersection de ces deux listes.

.5 Autres utilisations des liens pour la RI

Les approches présentées ci-dessus ont été concrétisées dans des travaux avec éventuellement des modifications. Par exemple, le moteur de recherche google [Brin, 1998] utilise le principe de PageRank : tout lien pointant de la page A à la page B est considéré comme un "vote" de la page A en faveur de la page B. Il procède également à une analyse de la page qui contient le lien. Les liens présents dans des pages jugées importantes ont plus de "poids", et contribuent ainsi à " élire" d'autres pages [Google].

Au lieu d'utiliser tous les liens entrants et sortants d'une page, [Gurrin, 2000] a identifié deux types de liens, les liens structurels et les liens fonctionnels et utilise seulement ce dernier type pour la propagation de pertinence. Il a séparé les liens du web en deux types larges basés sur leur fonctions voulues quand ils sont créés : les liens structurels qui aident l'utilisateur à naviguer à l'intérieur d'un site et les liens fonctionnels qui peuvent être vus essentiellement comme des liens entre un document source et un document cible qui possèdent une similarité dans le contenu.

.6 Synthèse

L'utilisation des liens pour la RI sur le WEB a été étudiée dans plusieurs travaux. Nous pouvons classer ces travaux en deux classes :

- l'utilisation des liens lors de la phase d'indexation (le PageRank),
- l'utilisation des liens lors de la phase d'interrogation (le système PAS, l'algorithme de Kleinberg, et l'approche de propagation de pertinence).

La différence entre ces deux classes est que lors de la phase d'indexation, les calculs se font indépendamment de la requête utilisateur, tandis que lors de l'interrogation, les calculs dépendent de la requête.

Dans les approches présentées ci-dessus, nous remarquons que l'utilisation des liens est utilisée comme une simple surcouche lors de l'indexation (respectivement lors de l'interrogation). C'est à dire, l'utilisation des liens est faite après avoir indexé (respectivement interrogé) les documents en se basant

[2] Les documents qui ont les meilleurs valeurs de probabilité de pertinence.

sur les méthodes basées sur le contenu. L'utilisation des liens sert, dans ces cas, à classer les documents résultats par rapport à une requête, et les index des documents ne sont plus modifiés une fois qu'ils sont construits par les méthodes classiques[3]. Ces raisons ont poussé Aguiar à s'interroger sur l'utilisation de l'information externe à un nœud (page) qui peut être utile lors de son indexation. Il pense que le fait d'indexer un nœud à partir de son seul contenu, risque d'avoir un index qui ne révèle pas précisément l'information véhiculée par ce nœud. Ainsi il propose de retrouver de l'information qui donne un contexte au contenu d'un nœud et la prendre en compte lors de l'indexation et lors de l'interrogation [Aguiar, 2002].

Nous avançons que l'échec de ces approches repose sur le fait de manipuler tous les liens (respectivement toutes les pages) de la même manière sans aucune distinction. Nous pensons qu'il ne faut pas utiliser les liens sans avoir analysé leur nature. Il faut comprendre leur rôle, en terme de description de l'information, avant de les utiliser. Aussi, il ne faut pas considérer systématiquement la page HTML comme étant la plus petite granularité d'information. Il faut comprendre la nature de chaque page ainsi que son rôle avant de l'utiliser. De manière évidente, il existe différents types de page : une page index est différente d'une page de contenu ou encore, une page de publicité est différente d'une page professionnelle, etc.

4. ÉVALUATION

Plusieurs travaux ont été menés sur l'utilisation des liens pour la RI sur le WEB mais, jusqu'à maintenant de nombreuses expériences ont montré qu'il n'y a pas de gain significatif par rapport aux méthodes de recherche basées seulement sur le contenu [Savoy, 2000], [Savoy, 2001], [Gao, 2002].

Le système PAS (cf. .4) a été testé dans [Savoy, 2001] en utilisant tous les documents voisins à un document D_i. Dans [Savoy, 2000], seuls les meilleurs documents entrant D_{in} et les meilleurs sortant D_{out} ont été utilisés. Dans ces deux travaux, les résultas ont été inférieurs aux résultats des méthodes de recherche classiques basées sur le contenu.

Pour l'approche de propagation, [Savoy, 2001] ont utilisé tous les liens entrant et sortant des documents. Dans [Savoy, 2000], les liens ont été sélectionnés avant de faire la propagation. Seuls les liens qui sortent des documents les mieux classés par rapport à la requête sont utilisés. Dans ces deux travaux, les résultats sont nettement inférieurs à ceux des méthodes classiques.

Pour l'algorithme de Kleinberg et la méthode de PageRank, les résultats sont médiocres [Savoy, 2000]. A travers leurs travaux d'évaluation, les auteurs ont pu conclure que les méthodes de recherche basées sur les liens n'apportent pas des améliorations [Savoy, 2000]. Ils se posent donc les questions suivantes :

- Existe-t-il d'autres travaux qui peuvent confirmer ces résultats ?
- Est-ce qu'on a bien choisi les valeurs des paramètres ?

La même conclusion a été tirée du travail de sept autres chercheurs bien qu'ils aient combiné les méthodes basées sur le contenu avec les méthodes basées sur les liens [Gao, 2002].

Toutefois, dans [Craswell, 2001], les résultats issus des méthodes basées sur les liens ont été bien meilleurs que celles des méthodes basées sur le contenu. Le cas traité dans ce travail est la recherche de site et non d'information. Le résultat d'une requête est une URL, en général, la *home page* d'un site. Les auteurs ont comparé l'efficacité de deux méthodes de classement dans la recherche de site :

- La méthode classique de classement basée sur le contenu.
- La méthode de classement basée sur les liens (texte de l'ancre).

Dans le cas de l'utilisation du texte de l'ancre seulement, pour chaque page *p*, ils construisent un *document d'ancre* qui contient tous les textes des ancres des liens qui pointent vers *p*. Ce document représente la description de la page *p*. Dans le cas basé sur le contenu, une indexation classique a été utilisée.

Dans leurs expérimentations, les auteurs ont utilisé 3 collections différentes de document web. Pour chaque collection, 100 requêtes ont été fixées. Les deux méthodes utilisent le même système de recherche et le même algorithme de classement. La différence réside dans le document index pour chaque page. Dans le cas de la méthode classique, un index classique a été utilisé. Tandis que dans le cas de la méthode basée sur les ancres, l'index c'est le *document ancre*. Les résultats de la méthode de classement basée sur les ancres sont deux fois meilleurs que ceux de la méthode classique. Donc l'information des ancres est plus utile que celle du contenu, et ce dans le processus de recherche de site.

5. DISCUSSION DES INSUFFISANCES DES MÉTHODES ACTUELLES

Les méthodes, qui ont été proposées pour extraire de l'information à partir des liens de navigation, traitent le WEB comme un graphe orienté dont les nœuds sont les pages HTML, et les arcs sont les liens de navigation entre les pages. Elles ne font pas de distinction entre les pages ou entre les liens. Pourtant il y a une différence entre une page contenant un livre et une page contenant quelques paragraphes, entre une page de contenu textuel et une page d'index ne contenant que des liens, entre une page personnelle et une page professionnelle, etc. De même pour les liens, il y a une différence entre un lien intra-page et un lien inter-page, entre un lien de publicité et un lien vers du texte, etc.

Dans la suite, nous détaillons les insuffisances des méthodes existantes en présentant quelques limites des moteurs de recherche existants.

[3] Pour l'indexation, les méthodes classiques indexent les documents seulement en se basant sur le contenu textuel. Pour l'interrogation, les méthodes classiques mesurent la correspondance entre les termes de la requête et les termes du document atomique.

.7 Limites des moteurs de recherche actuels

Les moteurs de recherche actuels se basent sur l'hypothèse que la plus petite granularité d'information recherchée par l'internaute est une page HTML. Pour cette raison, l'indexation se fait souvent au niveau de la page HTML d'une manière atomique et indépendante. Dans la réalité, cette hypothèse est souvent mise en échec. Par exemple, si les informations recherchées se trouvent dans une page[4] qui contient, en supplément des informations pertinentes, d'autres informations non pertinentes, l'utilisateur risque de devoir rechercher par lui même l'information qu'il recherche en parcourant le document proposé par le système. Pour rendre compte de cet état de fait, nous proposons une nouvelle mesure de qualité des résultats d'un SRI basé sur la densité de bonnes "informations" à l'intérieur d'un document fournit en réponse à une requête. Nous définissons la *précision informationnelle* comme le rapport entre le nombre d'informations pertinentes présentes dans le document et le nombre total d'informations que contient ce document. Par symétrie nous proposons la notion de *rappel informationnel*, comme le rapport entre le nombre d'informations présentes dans un document et le nombre d'informations pertinentes dans le corpus. Avec ces définitions, un document possède une qualité intrinsèque mesurée par ces deux facteurs : un document à forte précision informationnelle aura tendance à ne contenir que des informations pertinentes. Un document à fort rappel informationnel aura lui tendance à contenir à lui seul toute l'information pertinente disponible.

..1 L'atome de la page HTML : source de précision informationnelle faible

Dans un processus de RI, si le moteur de recherche retourne une page contenant tout un livre pour un utilisateur qui recherche juste une définition, le résultat est de faible précision informationnelle (donc bruité). Donc la notion de page HTML comme élément atomique d'une réponse de SRI n'est pas un bon choix. En fait, ce n'est pour nous qu'une simple contrainte physique. Pour cette raison, nous jugeons utile de se détacher de cette contrainte physique pour redéfinir la notion de "document" sur le WEB. Ainsi nous proposons de la manipuler avec plus de finesse en s'intéressant plus à sa structure logique interne.

..2 La page HTML indépendante : source de rappel informationnel faible

Lorsque nous indexons des pages HTML indépendamment les unes des autres, le résultat de la recherche est toujours une page indépendante. Par conséquent, si les informations sont dispersées dans plusieurs pages (ce qui est le cas fréquent), il y aura un silence dans le résultat. Le moteur de recherche ne va pas fournir toutes les informations, demandées par l'utilisateur, dans un seul document. Donc la contrainte d'une page HTML indépendante devient une source d'un rappel informationnel faible. L'hypothèse que la page HTML ou tout autre fichier document, est l'atome d'information que l'internaute recherche, est prise encore une fois à l'échec. Pour éviter le silence et fournir à l'utilisateur toutes les informations qu'il cherche, nous pensons que l'indexation ne doit plus se réaliser au niveau de granularité de la page HTML, et que le réseau de liens doit être exploité pour améliorer la qualité de la RI. Les pages doivent être indexées en fonction de leurs liens pour quelles ne perdent pas leurs contextes. Avant de pouvoir utiliser les liens il nous faudra les analyser car, comme nous allons voir dans la section suivante, les liens hypertextes sont typés implicitement et n'ont pas tous le même rôle.

6. LES LIENS SUR LE WEB SONT TYPÉS IMPLICITEMENT

Implicitement, il y a différents types de liens. Sémantiquement, les auteurs incluent parfois implicitement le type pour exprimer leurs intentions et ce par la description, dans le texte, de ce qui est pointé par le lien [Allan, 1996]. Nous trouvons, par exemple, "pour voir la définition cliquer ici" ou "pour voir les résultats cliquer ici", etc. Syntaxiquement, il y a les liens inter-pages (liens entre les pages dans le même site WEB), les liens intra-pages (liens entre les paragraphes dans la même page), les liens externes au site WEB, les liens de publicités, etc. Nous trouvons également des liens dont le but est de guider le lecteur dans sa navigation, par exemple, "haut de page" ou "page précédente", etc.

Nous pouvons donc conclure qu'il existe différents types de liens ou encore différentes raisons de créer les liens. Il faut les analyser et en fonction de leurs rôles, les utiliser dans le processus de RI.

Actuellement, les liens contiennent des informations qui sont encore cachées et qu'il faudrait expliciter pour les utiliser lors du processus de RI. Le texte de l'ancre de la page source contient parfois une description sur la page cible. Le contexte dans lequel l'ancre apparaît et les étiquettes qui, structurellement, précèdent la section là où le lien apparaît, contiennent des informations sur la pages cible [Fürnkranz, 1998]. Syntaxiquement, l'analyse comparative des URLs source et destination contient aussi des informations.

7. SOLUTION PROPOSÉE

Comme solution nous proposons d'utiliser une nouvelle granularité d'information au lieu de la page atomique, par exemple, le paragraphe. Nous indexons les pages en fonction des liens entre elles pour qu'elles ne perdrent pas leur contexte. Comme réponse pertinente, nous proposons *un chemin de lecture* qui est un document virtuel formé de paragraphes dispersés dans une ou plusieurs pages. Ce document peut contenir le maximum d'information pertinente pour remplir le critère de *rappel informationnel* et éviter le *silence*, il peut aussi contenir le minimum d'information non pertinente, pour remplir le critère de *précision informationnelle* et éviter le *bruit*.

[4] Bien sûr une page de grande taille

Ainsi nous allons manipuler les sites WEB comme étant un ensemble de page inter-reliées contenant des information complémentaires. Nous allons essayer de regrouper ces informations dans un document en utilisant les liens typés et une nouvelle granularité. Ce document peut être une réponse contenant le maximum d'information susceptibles d'être pertinentes par rapport à une requête.

Les questions que nous posons donc sont : comment extraire les types de liens ? Quel(s) type(s) utiliser lors du processus de RI ? Comment se détacher de la notion physique de la page HTML ? Quelle granularité d'information utiliser ? Comment l'extraire ? Comment prendre en compte les informations dispersées ? et enfin, comment intégrer ces nouveaux concepts dans un modèle de RI, et plus précisément lors de la phase d'indexation ?

8. CONCLUSION

A travers ce papier, nous avons présenté quatre approches d'utilisation des liens pour la RI sur le WEB. Des expérimentations sur ces méthodes ont montré qu'il n'y a pas un gain significatif par rapport aux méthodes basées sur le contenu textuel seulement. Toutefois, de bons résultats ont été atteints dans un processus de recherche de site. Les approches actuelles utilisent le contenu textuel pour la recherche de document et les liens ne sont utilisés que pour la classification des pages. Ces approches sont utilisées comme des surcouches aux méthodes de RI classiques. Nous pensons aller beaucoup plus loin, en retenant les solutions existantes basées sur le texte et en les combinant avec de nouvelles solutions basées sur les liens qui représentent l'une des spécificités du WEB. De nouvelles pistes ont été ouvertes comme celle de l'indexation des chemins de lecture [Gery, 2002]. Il reste donc à prouver les potentialités des liens cachées pour adapter les moteurs de recherche actuels aux spécificités du WEB.

BIBLIOGRAPHIE

[Aleese, 1990] Aleese, R., Green, C.: "Hypertext: State of the Art". Oxford, Angleterre, Intellect, (1990).

[Allan, 1996] Allan, J.: "Automatic Hypertext Link Typing". Proceedings of ACM Hypertext '96, Washington DC, P. 42-52, (1996).

[Aguiar, 2002] Aguiar, F.: "Modélisation d'un Système de Recherche d'Information pour les Systèmes Hypertextes : Application à la Recherche d'Information sur le World Wide Web", Thèse de Phd, Ecole Nationale Supérieure des Mines, Saint-Étienne (2002).

[Brin, 1998] Brin, S., Page, L.: "The Anatomy of a Large-Scale Hypertextual Web Search Engine". WWW8, P. 107-117, (1998).

[Chakrabarti, 2001] Chakrabarti, S.: "Integrating the Document Object Model with Hyperlinks for enhanced Topic Distillation and Information Extraction", 10ème World Wide Web Conference (WWW'01). Hong-Kong, Chine, (2001).

[Craswell, 2001] Craswell, N., Hawking, D., Robertson., S.: "Effective site finding using link anchor information". 24 eme ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01), New Orleans, Louisiana, Etats Unis, P. 250-257, (2001).

[Crestani, 2000] Crestani F., Lee, P.: "Searching the Web by Constrained Spreading Activation. Information". Processing & Management, P. 585-605, (2000).

[Devlin, 1991] Devlin, J., Berk, E.: "Hypertext/Hypermedia Handbook". Intertext Publications/McGraw-Hill Publishing Company., New York, États-Unis, (1991).

[Fürnkranz, 1998] Furnkranz, J.: "Using links for classifying Web-pages". Technical Report TR-OEFAI-98-29, Austrian Research Insitutue for Artificial Intelligence, (1998).

[Gao, 2002] Gao, J., Cao, G., He, H., Zhang, M., Nie,J-Y., Walker, S., Robertson, S.; TREC-10 Web track experiments at MSRA, TREC-10, NIST Special Publication 500-250, (2002).

[Gery, 2002] Gery, M.,:"Indexation et interrogation de chemins de lecture en contexte pour la Recherche d'Information Structurée sur le Web", Thèse de Phd, Université Joseph Fourier, Grenoble, (2002).

[Google] www.google.fr/intl/fr/why_use.html.

[Gurrin, 2000] Gurrin, C., Smeaton, A.,: "Dublin City University Experiments in Connectivity Analysis for TREC-9". 9eme Text Retrieval Conference (TREC'00). Gaithersburg, Maryland, United States, (2000).

[Henzinger, 2000] Henzinger, M.,: "Link Analysis in Web Information Retrieval". Bulletin of the IEEE computer Society Technical Committee on Data Engineering, (2000).

[Kleinberg, 1998] Kleinberg, J.: "Authoritative sources in a hyperlinked environment". Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, P. 668-677, Washington, (1998).

[Park, 1998] Park, P.: "Structural Properties of Hypertext". In UK Conference on Hypertext, P. 180 187, (1998).

[Picard, 1998] Picard, J.: "Modeling and combining evidence provided by document relationships using PAS systems". ACM-SIGIR'98, P. 182-189, (1989).

[Prime, 2002] Prime, C., Beigbeder, M., Lafouge, T. : "Clusterisation du Web en vue d'extraction de corpus homogènes". in actes de INFORSID, 20e congrès informatique des organisations et des systèmes d'information et de décision, Nantes, France, (2002).

[Salton, 1971] Salton G.: "The SMART retrieval system: experiments in automatic document processing", Prentice Hall, (1971).

[Salton, 1983] Salton, G., Mcgill, M-J.: " Introduction to modern Information Retrieval". McGraw-Hill, (1983).

[Savoy, 2000] Savoy, J. Rasolof, Y.: "Link-Based Retrieval and Distriuted Collections". Report of the TREC-9 experiment :, Proceedings TREC-9, NIST, Washigton D.C., (2000).

[Savoy, 2001] Savoy, J., Picard, J.: "Retrieval effectiveness on the web. Information". Processing & Management, vol. 37, P. 543-569, (2001).