Un modèle à base de chemin de lecture pour la Recherche d'Informations précises sur le Web

Saïd Radhouani*,** — Jean-Pierre Chevallet** — Mathias Géry***

** Centre Universitaire d'Informatique, Université de Genève. 24, rue Général-Dufour - CH 1211 Genève 4-Suisse. Said.Radhouani@cui.unige.ch.*

*** Laboratoire CLIPS-IMAG, B.P. 53, 38041 Grenoble cedex 9, France.*

**** Laboratoire EURISE, Université Jean Monnet de Saint-Etienne, 23, rue du Docteur Paul Michelon, 42023 Saint-Etienne Cedex 2, France. Mathias.Géry@univ-st-etienne.fr.*

RÉSUMÉ. *Actuellement, le nœud hypertexte (document) est utilisé comme la plus petite granularité d'information que l'utilisateur cherche. Nous supposons que le fait de considérer le nœud hypertexte comme unité informationnelle n'as pas toujours un sens, car il s'agit uniquement d'une contrainte physique. Dans la réalité, l'utilisateur peut avoir envie de rechercher un seul paragraphe, ou au contraire un ensemble de pages. Or, les SRI se basent sur la granularité d'un nœud comme unité de base. Cette contrainte physique peut être la cause de résultats non satisfaisants, typiquement des documents "bruités" contenant, en plus de l'information recherchée, d'autres informations non pertinentes. En plus, si nous manipulons les nœuds indépendamment les uns des autres sans prendre en compte les informations dispersées dans plusieurs nœuds, nous aurons probablement un "silence" dans les réponses, c'est-à-dire qu'il y a encore des informations pertinentes mais le SRI n'a pas pu les retrouver car les documents ont perdu leur contexte (les nœuds voisins) et les relations sémantiques entre eux. Dans cet article, nous redéfinissons la notion de document dans un contexte hypertexte et requête précise. Nous proposons, comme réponse précise à une requête, un document virtuel que nous appelons chemin de lecture. Ce dernier reflète la description de l'information sur les hypertextes. Il est formé de zones de texte dispersées dans un ou plusieurs documents connectés. Nous proposons un modèle permettant d'extraire les chemins de lecture. Nous utilisons les liens typés pour regrouper les zones de texte constituant chaque chemin. Pour celà, nous nous basons sur une technique heuristique et sur les valeurs de similarité entre les zones.*

ABSTRACT. *Currently, the hypertext node (document) is used as the smallest information granularity seeked by the user. We suppose that the fact of considering the hypertext node as an informational unit has not always meaning, but is only a physical constraint. In fact, the user may need to search for one simple paragraph, or for a set of pages. However, IRS are based usually on the node granularity as basic unit. This physical constraint can be the cause of non-satisfactory results, typically, documents include irrelevant information for example*

due to their size. Moreover, if we handle nodes independently without taking into account the information dispersed in several nodes, we will probably have a « silence » in the answers. It means that it still has relevant information there but the IRS can not find them because documents lost their context and semantic relations between them. In this article, we redefine the concept of document in a hypertext context and precise query. We propose, as precise answers to a query, virtual documents that we call « reading paths ». They reflect the description of information on hypertexts.

MOTS-CLÉS : *RI, hypertexte, Web, liens, typage de liens, chemins de lecture.*

KEYWORDS: *IR, hypertext, Web, links, link typing, reading paths.*

---

1. Introduction

Avec l'augmentation continuelle de la taille du Web, nous arrivons à une situation parfaitement contradictoire : jamais il n'y a eu autant d'informations disponibles de manière électronique et quasi instantanée, mais retrouver précisément ce que l'on recherche dans cette accumulation devient de plus en plus difficile car pour une requête particulière soumise à un moteur de recherche du Web, le nombre de réponses potentielles augmente en proportion de la taille du corpus. Pour augmenter les chances de succès, les moteurs de recherche doivent se tourner vers la qualité des réponses et donc vers plus de précision.

Dans cet article, nous proposons d'explorer l'utilisation des liens hypertextes typés, en particuliers les « chemins de lecture », afin d'obtenir des réponses plus fines car allant au delà de la page Web et mettant à profit les liens hypertextes. Le résultat est un moteur orienté vers la précision des réponses.

Dans ce contexte, nous appellerons «document» toute unité textuelle qui peut constituer une réponse à une requête utilisateur. Nous proposons également le terme doxel[1] pour exprimer le plus petit élément de texte indexable. Finalement, nous parlerons de noeud, pour désigner un élément du réseau hypertexte connecté par des liens.

Notre objectif est de proposer à l'utilisateur une réponse précise. Nous allons donc redéfinir la notion de document, dans un contexte d'hypertexte et de requête précise, en choisissant une granularité des réponses et en utilisant les liens. Nous proposons comme réponse précise un document virtuel que nous appelons chemin de lecture [Géry, 2002]. Ce document doit contenir le maximum d'informations pertinentes et le minimum d'informations non pertinentes. Dans le but de pouvoir interroger ces chemins de lecture, nous proposons un modèle d'indexation approprié.

---

[1] Le terme "doxel" est construit par analogie au terme "pixel". C'est une suggestion du Pr. Patrick Gallinari du LIP6.

Nous détaillons d'abord notre problématique dans la section 2. Ensuite nous évoquons plusieurs méthodes d'utilisation de liens pour la RI dans la section 3. Dans la section 4, nous proposons un modèle permettant de prendre en compte les liens lors du processus de RI. Dans la section 5, nous présentons la mise en œuvre de notre modèle. Avant de conclure nous présentons les premiers résultats de nos expérimentations.

2. Recherche d'information sur les hypertextes

2.1. Naissance de l'hypertexte

L'idée de cette technologie trouve son origine dans les travaux de Vannevar Bush, qui conçoit en 1945 un système de gestion et d'accès aux connaissances dénommé Mémex [Bush, 1945]. Cette voie fut poursuivie par Ted Nelson, qui conçoit en 1965 un projet de bibliothèque universelle, Xanadu [Nelson, 1965], à l'intérieur de laquelle il serait possible de circuler en utilisant des liens hypertextes.

La navigation est le mode typique de consultation d'un réseau hypertexte. C'est une manière simple qui permet de consulter le contenu des documents hypertextes d'une manière directe. Toutefois, l'utilisateur ne peut que suivre les liens proposés par l'auteur. Pour peu que le réseau soit de taille importante, il doit fournir un grand effort pour une consultation efficace, et il peut éventuellement se perdre.

2.2. Problématique

La richesse d'un document hypertexte réside non seulement dans son contenu informationnel mais également dans son réseau de liens. Devant cette richesse, la navigation se révèle limitée pour rechercher efficacement des informations, et l'accès direct à l'information basé sur une requête est souvent plus efficace. Les SRI représentent donc une alternative efficace à la navigation.

Le processus de RI consiste à retrouver, à partir d'une collection de documents, les documents qui correspondent le mieux au besoin de l'utilisateur. La plupart des systèmes permettent d'exprimer ce besoin sous la forme d'une requête, généralement constituée d'un ensemble de mots-clés, et le SRI doit retrouver les documents qui traitent des sujets décrits par ces mots-clés.

Avec l'apparition des hypertextes, les documents possèdent de nouvelles caractéristiques : les données sont hétérogènes et les documents sont des noeuds connectés par des liens hypertextes. Ces caractéristiques exigent de nouveaux critères de pertinence d'un document par rapport à une requête. En particulier, l'information disponible ne se résume plus au contenu textuel, et les traitements d'extraction des index doivent tenir compte de ces aspects hypertextuels. La pertinence basée sur la correspondance entre les mots-clés de la requête et le texte du document semble insuffisante, car elle n'utilise pas l'information portée par les liens.

Or, les SRI actuels ne tiennent pas compte des spécificités des hypertextes, car ils sont basés sur des modèles de RI développés pour des documents textuels

classiques depuis les années 70 [Salton, 1971] [vanRij, 1979] [Salton, 1983]. Ces systèmes se basent sur l'hypothèse que les documents sont atomiques et non structurés. Il n'y a donc pas de différence entre un doxel, l'entité atomique textuelle indexée, et le document, l'entité macroscopique retournée en réponse. Dans la réalité, et surtout sur le Web, cette hypothèse est incorrecte. Par exemple, si les informations recherchées se trouvent dans un document de grande taille, il contient souvent, en supplément des informations pertinentes, d'autres informations non pertinentes. C'est finalement à l'utilisateur qu'il revient de rechercher l'information en parcourant le document proposé. Par symétrie, si les informations recherchées se trouvent dans plusieurs documents connectés par une structure inutilisée par le SRI, c'est encore à l'utilisateur de parcourir les liens entre les documents pour reconstituer toute l'information qu'il désire.

Nous proposons deux nouvelles mesures de qualité des résultats d'un SRI basées sur la densité de bons éléments à l'intérieur d'un document fourni en réponse à une requête. Nous définissons la ***précision élémentaire*** comme le rapport entre le nombre de doxels pertinents présents dans un document et le nombre total de doxels que contient ce document. Par symétrie, le ***rappel élémentaire*** est le rapport entre le nombre de doxels pertinents présents dans un document et le nombre de doxels pertinents dans le corpus. Ces mesures sont les mêmes que le rappel et la précision classique de Salton lorsque l'on a un seul document composé de doxels en réponse. Dans le cas d'un ensemble de documents composés de doxels en réponse, nous avons alors une mesure de rappel/précision globale classique et cette mesure élémentaire est locale pour chaque document. Avec ces définitions, un document possède donc une qualité intrinsèque mesurée par ces deux facteurs : un document à forte précision élémentaire aura tendance à ne contenir que des informations pertinentes. Un document à fort rappel élémentaire aura tendance à contenir à lui seul toute l'information pertinente disponible. Nous disposons donc de quatre mesures de qualité qu'il faut maximiser pour mesurer la qualité d'un SRIS (SRI Structuré) : précision, précision élémentaire, rappel, et rappel élémentaire. Le document entier est pertinent s'il contient au moins un doxel pertinent. Nous pouvons ensuite augmenter la précision élémentaire en réduisant la taille des documents, jusqu'a ce qu'ils ne contiennent qu'un doxel pertinent, ou alors en réorganisant les doxels pour former des documents plus intéressants pour l'utilisateur. C'est cette deuxième option que nous mettons en oeuvre avec la notion de chemin de lecture.

Un des objectifs est alors de modéliser un système capable de retrouver en réponse à une requête, tous les documents pertinents, débarrassés des fragments non pertinents. Si le SRI retourne un document contenant l'équivalent d'un livre pour un utilisateur qui recherche juste une définition, le résultat est de faible précision élémentaire, et l'utilisateur qui a un besoin précis et peu de temps, sera insatisfait d'une telle réponse. Actuellement, les liens sont peu exploités par les moteurs du Web, et les documents sont indexés indépendamment ou presque de ces liens (cf. section 3). Dans ce cas, le résultat de la recherche d'un SRI est un ensemble de

documents indépendants. Par conséquent, si les informations sont dispersées dans plusieurs documents, il se peut que l'utilisateur juge les documents proposés par le système inintéressants car incomplets. Or, comme les informations peuvent être morcelées en pages, il se peut que l'information cherchée soit également éparpillée. Un système qui identifie doxels et documents sans tenir compte des liens sera donc incapable de fournir une réponse optimale, c'est-à-dire, un seul document présentant l'ensemble des informations recherchées.

2.2. Vers plus de précision sur les hypertextes

Nous proposons alors de ne plus considérer la page Web comme la réponse potentielle du SRI (un document), mais de descendre à un niveau de granularité inférieur. Une page Web peut alors être composée de plusieurs doxels. Nous proposons comme réponse pertinente un ***chemin de lecture*** [Géry, 2002] qui est défini comme suit :

Définition : ***un chemin de lecture*** *est une suite de doxels connectés par des liens de navigation et qui décrivent un thème particulier. Ces liens de navigation sont des* ***liens de cheminement***.

Le chemin de lecture contient le maximum d'information pertinente pour remplir le critère de *rappel élémentaire*, et contient aussi le minimum d'information non pertinente, pour remplir le critère de *précision élémentaire*. Notre problématique consiste donc à répondre aux questions suivantes : Comment se détacher de la notion physique de la page Web ? Quelle granularité d'information utiliser concrètement pour un doxel ? Comment mieux définir et exploiter ces liens de cheminement ? Enfin, comment intégrer ces nouveaux concepts dans un modèle de RI orienté précision ?

3. Utilisation des liens pour la Recherche d'Information

Nous proposons donc un *chemin de lecture* comme *réponse précise*. La structure de ces chemins est basée sur les liens hypertextes existants. Un SRI basé sur la notion de chemin de lecture doit donc permettre d'indexer et d'interroger de tels types de documents. Pour ces raisons, nous sommes amenés à étudier les travaux qui utilisent les liens pour la RI.

Principalement, il existe quatre approches dans la littérature. Nous distinguons deux classes : l'utilisation des liens lors de la phase d'indexation d'une part (PageRank [Brin, 1998]), et lors de la phase d'interrogation d'autre part (PAS [Picard, 1998], HITS [Kleinberg, 1998], propagation de pertinence [Crestani, 2000, Savoy, 2000]). La différence 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 notre travail, nous proposons d'utiliser les liens avant la phase d'indexation, lors de la construction des chemins de lecture.

3.1. 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(p) 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 référencent p est important. PR(p) est calculé en fonction des PR(q) des pages qui référencent p.

3.2. 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 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.

3.3. 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. A la deuxième étape, cette valeur de probabilité est modifiée en fonction des valeurs de probabilités des documents voisins au document D_i.

3.4. Algorithme HITS [Kleinberg, 1998]

Cette approche consiste à identifier des *Hubs* (documents « rayonnants ») et des *Authorities* (documents qui font « autorité »). Ces deux concepts sont dépendants (« mutually reinforcing relationships »). L'hypothèse est : « *un document qui pointe vers beaucoup de bonnes Authorities est un bon Hub, et un document pointé par beaucoup de bon Hubs est une bonne Authority* » [Kleinberg99].

3.5. Évaluation des approches existantes

De nombreuses expériences, notamment dans le cadre de la conférence TREC[2], ont montré que l'utilisation des liens pour la RI n'apporte pas de gain significatif par rapport aux méthodes basées sur le contenu textuel seul [Savoy, 2000], [Savoy, 2001], [Gao, 2002]. Le système PAS (cf. section 3.3) 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 et les meilleurs sortant ont été utilisés. Dans ces

[2] http://trec.nist.gov/

deux travaux, les résultats ont été inférieurs aux résultats des méthodes basées sur le contenu textuel.

Pour l'algorithme HITS et la méthode de PageRank, les résultats sont médiocres [Savoy, 2000]. Les auteurs ont conclu que les méthodes basées sur les liens n'apportent pas d'amélioration [Savoy, 2000]. La même conclusion a été tirée du travail de sept autres chercheurs ayant combiné les méthodes basées sur le contenu textuel 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 ceux 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 en général la *page principale* d'un site. Les auteurs ont comparé l'efficacité de deux méthodes de classement dans la recherche de site :

– La méthode 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'ancres 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. Les résultats de la méthode de classement basée sur les ancres sont deux fois meilleurs que ceux de la méthode classique. L'information des ancres semble donc plus utile que celle du contenu dans un processus de recherche de site.

3.6. Discussion des insuffisances des méthodes actuelles

Dans les approches présentées ci-dessus, nous remarquons que les liens sont utilisés comme une simple surcouche lors de l'indexation (respectivement, lors de l'interrogation), c'est-à-dire après avoir indexé (respectivement, interrogé) les documents à l'aide de méthodes classiques basées sur le contenu textuel. Les index des documents ne sont plus modifiés une fois qu'ils sont construits par les méthodes classiques. Ces raisons ont poussé Aguiar à s'interroger sur l'utilisation de l'information externe à un nœud qui peut être utile lors de son indexation [Aguiar, 2002]. Aguiar pense que le fait d'indexer un nœud à partir de son seul contenu produit un index qui ne révèle pas précisément l'information véhiculée par le nœud. Il propose de retrouver de l'information provenant du contexte d'un nœud, et de la prendre en compte lors de l'indexation et de l'interrogation.

Nous pensons que l'échec de ces approches peut être imputé à la manipulation de 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 Web comme étant la plus petite granularité d'information. 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. Il existe

aussi différents types de liens : au niveau syntaxique, nous trouvons des liens inter-pages, des liens externes au site Web, etc. Au niveau sémantique, nous trouvons les liens de publicités, les liens « voir aussi », etc.

Suite aux résultats médiocres de méthodes utilisant les liens pour la RI, nous pouvons déduire que soit les liens sont effectivement inutiles à la RI, soit ils sont utiles mais mal exploités. Nous allons dans cette dernière direction en proposant de discerner les différents types de liens avant de s'en servir en RI. Notre approche privilégie les *liens de cheminement*.

4. Modèle de Recherche d'Information sur les chemins de lecture

Nous présentons dans cette section le modèle de RI pour l'indexation des chemins de lecture sur le Web.

4.1. Définition d'un chemin de lecture sur les doxels

Le doxel est une zone de texte matérialisée, par exemple, par un paragraphe. A l'aide de cette nouvelle granularité, nous proposons pour une requête, une réponse précise contenant le minimum d'informations non pertinentes. Cette réponse remplit le critère de *précision élémentaire* et évite le *bruit*.

Nous proposons comme résultat d'une recherche, un document avec une structure nous permettant d'optimiser le compromis entre *rappel* et *précision élémentaire*. Nous appelons ce document un *chemin de lecture* : c'est une navigation prévue par l'auteur qui fait sens pour le lecteur. Un chemin de lecture est matérialisé par une suite de zones de textes contiguës connectées par des *liens de cheminement*. Nous décidons alors que ces zones de textes sont atomiques du point de vue de l'indexation, ces zones de textes sont donc nos doxels.

Structurellement, le chemin de lecture est construit en regroupant des passages de texte situés dans un ou plusieurs documents. On le modélise par :

$ch = <Z_1, L_{1,2}, Z_2, \ldots, Z_i, L_{i,j}, Z_j, \ldots, Z_{n-1}, L_{n-1,n}, Z_n>$.

Avec : Z_i les zones de texte formant le chemin *ch*, et $L_{i,j}$ le *lien de cheminement* entre Z_i et Z_j.

4.2. Indexation des chemins de lecture

En vue d'interroger les *chemins de lecture*, nous les indexons à l'aide du modèle vectoriel [Salton, 1983]. Comme nous verrons dans la suite, ce modèle est inadapté à l'indexation des chemins de lecture en tant que tels. Par contre, il a été développé pour l'indexation de fragments de textes, comme ceux qui composent un chemin de lecture.

L'indexation d'un chemin ch_j en un vecteur $\overrightarrow{ch_j}$ consiste à extraire les termes t_i représentatifs du chemin et à leur affecter une pondération $w_{i,chj} \in [0..1]$ calculée à l'aide d'une fonction de type *tf*idf.*

$\forall\ ch_j \in CH,\ \overrightarrow{ch_j} = (w_{1,chj}, w_{2,chj}, \ldots, w_{i,chj}, \ldots, w_{m,chj})$.

Avec : CH l'ensemble de tous les chemins de lecture, *m* est le nombre de termes dans le chemin ch_j. .La fonction de pondération se base sur les informations suivantes :

Fréquence locale : $tf_i ch_j$ est le nombre d'occurrences du terme t_i dans le chemin de lecture ch_j, ce qui correspond à la somme des occurrences du terme t_i dans toutes les zones de texte constituant ce chemin : $tf_i ch_j = \sum_{z \in zones(chj)} tf_{i,z}$.

Avec : zones(ch_j) est l'ensemble de toutes les zones appartenant au chemin ch_j, et z une zone appartenant à ch_j.

Fréquence documentaire : $df_i ch_j$ est le nombre de chemin dans lesquels le terme t_i apparaît.

Taille du corpus : N_{ch} est le nombre de chemins de lecture dans le corpus.

La particularité du chemin de lecture par rapport aux documents classiques réside dans le calcul de $df_i ch_j$. En effet, le calcule du *df* fait intervenir la fréquence documentaire d'un terme, ce qui nécessite de savoir dans quels documents il apparaît. Dans le cas des chemins de lecture, on ne peut pas considérer le corpus comme un ensemble de documents indépendants, car les chemins ne sont pas disjoints : deux chemins peuvent partager une ou plusieurs zones de texte (cf. *Fig. 1*). Si on procède de la manière classique de calcul du *df*, à chaque apparition d'un terme dans un chemin de lecture, son *df* sera incrémenté de 1. Dans l'exemple de la *Fig. 1*, supposons qu'un terme t_i appartient au chemin ch_1 car il apparaît dans la zone z_3 ou z_4. Dans ce cas, le terme t_i appartient aussi au chemin ch_2 à travers les zones z_3 et z_4. Pour calculer le *df* du terme t_i par rapport au chemin ch_1, on doit compter le nombre de chemins auxquels appartient le terme t_i, dans ce cas le *df* est égal à 2. La particularité c'est que l'appartenance du terme t_i au chemin ch_2 est faite via les zones z_3 et z_4 qui appartiennent en même temps au chemin ch_1. Le *df* est incrémenté de 1 bien que le terme n'appartienne pas à deux chemins de lecture indépendants. Afin de prendre en compte cette particularité, nous proposons de tenir compte des zones de texte partagées par les chemins de lecture ch_1 et ch_2. Au lieu d'utiliser la valeur 1 pour incrémenter le *df* à chaque fois où le terme apparaît dans un chemin de lecture, on utilise une valeur égale au nombre de zones de ch_1 contenant le terme t_i divisé par le nombre total des zones de ch_1, plus le nombre de zones de ch_2 contenant le terme t_i et n'appartenant pas à ch_1 divisé par le nombre total des zones de ch_2. Dans l'exemple de la *Fig.1*, si un terme appartient à toutes les zones des deux chemins, alors son *df* est augmenté de (6/6 + 3/5) pour ch_1et (5/5 + 4/6) pour ch_2.

Finalement le *df* d'un terme t_i appartenant à un chemin ch_j est égal à la somme de la valeur, que nous venons d'introduire, calculée pour chaque chemin :

$$df_{i,j} = \frac{|\{z \in ch_j | t_i \in z\}|}{|\{z \in ch_j\}|} + \sum_{ch \in CH} \frac{|\{z \in ch | t_i \in z \text{ et } z \notin ch_j\}|}{|\{z \in ch\}|}.$$

Avec, ch_j : un chemin de lecture, t_i : un terme appartenant à ch_j, CH : l'ensemble de tous les chemins de lecture autre que ch_j, et z : une zone de texte.

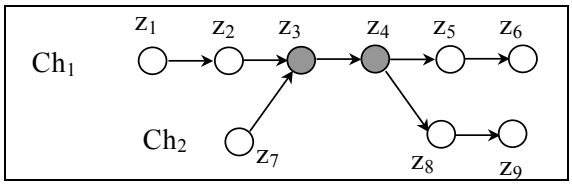

Figure 1. *Exemple de deux chemins de lecture qui partagent deux zones de texte. Ch_1 est formé des zones z_1, z_2, z_3, z_4, z_5 et z_6. Ch_2 est formé des zones z_7, z_3, z_4, z_8 et z_9.*

Bien que l'indexation d'un chemin lecture soit faite d'une manière structurée, elle se réduit en fin de compte à un vecteur classique. Elle est utilisée à l'interrogation de manière classique en calculant le produit scalaire entre le vecteur de la requête et celui du chemin.

5. Application du modèle sur le Web

Le modèle proposé s'applique dans un contexte de documents hypertextes. Nous présentons dans cette section une application possible sur le Web, qui représente lui-même un gigantesque hypertexte. Nous n'abordons pas le problème de son application à une grande échelle, qui est pourtant un problème important pour une application au Web. Notre objectif est plutôt de réduire l'utilisation de notre modèle à un sous ensemble précis du Web comme les pages d'un site ou domaine particulier ou encore les pages d'une communauté d'utilisateurs.

La construction des chemins de lecture est basée sur les hypothèses suivantes :

Hypothèse 1 : "Un chemin de lecture contient des zones de texte à propos d'un même thème".

Hypothèse 1a : "Le thème d'une zone de texte est défini par les termes qu'elle emploie".

Hypothèse 1b : "La similarité entre deux zones de texte est basée sur le partage des mêmes termes, ou le partage de termes en rapport entre eux".

Hypothèse 2 : "Un chemin de lecture est une proposition délinée d'un même auteur dans le même site Web".

Hypothèse 3 : "Dans une page Web, l'auteur suit un ordre d'écriture séquentiel du haut de la page vers le bas". Cet ordre correspond à l'organisation séquentielle de la structure logique d'un document. Elle implique l'existence d'un sens de lecture linéaire imposé par l'auteur : le premier paragraphe est à lire avant le deuxième, le deuxième avant le troisième, etc.

Hypothèse 4 : "Pour décrire ses informations à travers des pages Web, l'auteur suit un ordre descendant, non strict, suivant la hiérarchie des répertoires du site". C'est-à-dire qu'il crée d'abord une page dans un répertoire et, à partir de cette page,

il crée des liens vers d'autres pages dans le même répertoire ou dans des sous-répertoires, qui sont donc situés plus bas dans la hiérarchie du site.

5.1. Typage de liens sur le Web

Hormis la succession des paragraphes qui constitue un lien de cheminement implicite (cf hypothèse 3), les liens du Web ne comportent malheureusement pas de typage de cheminement. Un typage de lien à été introduit avec la norme XLink. Il permet de typer les liens dans des documents XML. La navigation peut s'effectuer en utilisant le langage XPath. Ce langage permet de sélectionner ou d'adresser une ou plusieurs parties des documents XML en définissant un chemin de lecture. Actuellement XPath n'est pas encore réellement utilisé sur le Web. Dès qu'il le sera, son utilisation facilitera grandement la mise en œuvre de notre modèle.

Pour pouvoir mettre en place de manière pratique notre modèle, nous proposons une heuristique de recherche des liens probables de cheminement. Dans [Géry, 2002], nous avons identifié trois types de liens sur le Web utiles à la RI : 1) les liens de composition entre les pages (le découpage d'un document, par exemple livre, en un ensemble de page par exemple chapitre), ou à l'intérieur d'une seule page (le découpage d'un document, par exemple chapitre, en un ensemble de sections). 2) les liens de référence : ils référencent d'autres documents à l'extérieur d'un site Web. 3) les liens de cheminement : ils induisent un sens de lecture particulier.

5.2. Typage de liens : premier filtrage

Notre heuristique de typage commence par l'extraction des liens de composition, puis nous appliquons un filtrage basé sur la similarité des vecteurs de termes des zones de texte.

5.2.1. *Typologie physique des liens*

Nous avons identifié cinq types physiques de liens, groupés en deux classes : liens externes et liens internes. Leur extraction se fait en fonction de leur rôle dans la structuration des informations.

- Les liens externes pointent vers une page située en dehors du site Web source du lien.
- Les liens internes sont groupés en deux sous-classes : les liens entre deux pages et les liens à l'intérieur d'une même page. Commençons par les liens entre deux pages :
    - Les liens hiérarchiques sont des liens de composition entre pages. Ils suivent un ordre descendant, non strict, suivant la hiérarchie des répertoires du site. Nous avons constaté expérimentalement [Géry 2001] qu'il y a plus de liens qui descendent dans la hiérarchie des répertoires du site que de liens qui remontent. Les liens du même niveau sont supposés descendre la hiérarchie d'un document structuré. Par structure, nous entendons la structure logique d'un document : un

chapitre est constitué de sections, une section est constituée de paragraphes, etc.

– Les liens retour-page suivent un ordre ascendant, suivant le sens inverse de la hiérarchie des répertoires du site.

La deuxième sous-classe des liens internes contient les liens à l'intérieur d'une même page. On trouve deux types de liens dans cette sous-classe :

– Les liens bas-de-page relient deux passages de texte à l'intérieur d'une même page. Le sens de ce lien est du haut de la page vers le bas.

– Les liens haut-de-page sont l'inverse des précédents. Ils pointent du bas de la page vers le haut.

Dans la figure.2, nous présentons un ensemble de pages connectées par les différents types de liens que nous venons de décrire. Les pages présentées dans cette figure appartiennent au même site.

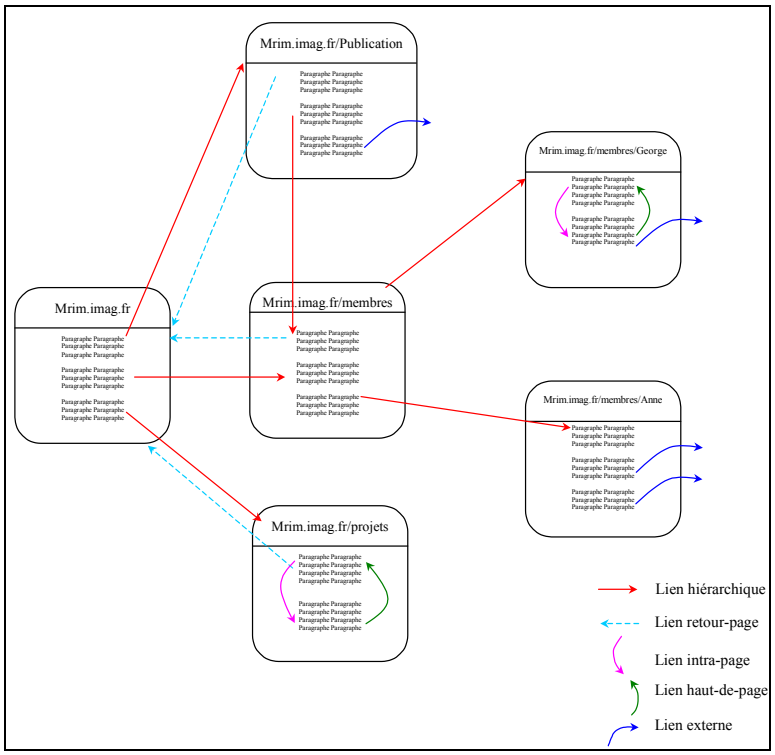

Figure 2. *Exemple de typologie des liens hypertextes.*

5.2.2. *Extraction des zones de teste*

Pour extraire les zones de texte (les doxels), nous découpons chaque page en paragraphes (par exemple, le texte entre les balises et du code HTML[3]). Ainsi nous aurons, pour chaque page, un ensemble de paragraphes indépendants débarrassés du code HTML : ce sont ces "zones de texte" qui constituent les doxels (cf. *Fig.3*).

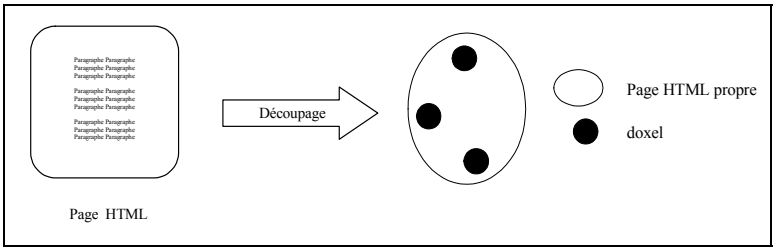

Figure 3. *Représentation graphique de la nouvelle granularité d'information*

5.2.3. *Extraction des liens de composition*

Après avoir découpé chaque page en doxels, le réseau des pages Web initial devient un graphe de pages divisées en doxels et connectées par des liens hypertextes. Les liens connectent directement les doxels et non les pages. Un lien pointe soit vers le début de la page cible, soit vers un paragraphe précis de la page cible. Dans le premier cas, le lien sera entre la zone courante et la première zone de la page cible. Dans le deuxième cas, le lien sera entre la zone courante et une zone précise à l'intérieur de la page cible. À ce niveau, seuls les liens hypertextes sont utilisés. Dans la *Fig.4*, nous présentons un graphe constitué d'un ensemble de pages (présentées par des ellipses). Chaque page est constituée d'un ensemble de zones de texte liées par des liens hypertextes (présentés par des flèches).

[3] Nous identifions également d'autres balises HTML permettant de décrire les paragraphes.

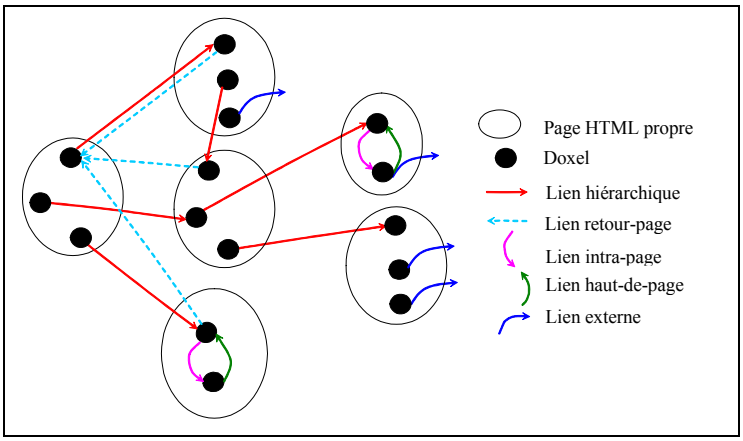

Figure 4. *Un graphe de documents divisés en doxels connectées par des liens.*

Le découpage des pages est fait indépendamment des types de liens. À ce niveau, le graphe contient tous les types que nous avons identifiés. La prochaine étape consiste donc à sélectionner les liens de composition (hiérarchiques, bas-de-page), qui sont des liens potentiels pour la construction des chemins de lecture. Ainsi, nous procédons aux étapes suivantes :

– Enlever les liens retour-pages : en se basant sur l'Hypothèse 4, nous considérons que le rôle de ce lien est d'aider le lecteur à naviguer et non à découvrir les informations (par exemple, "retour au home page" ou "retour à la page précédente", etc.).

– Enlever les liens haut-de-pages : en se basant sur *l'Hypothèse 3*, nous considérons que ces liens sont créés juste pour aider le lecteur dans sa navigation et que l'auteur suit un ordre séquentiel descendant pour décrire ses informations dans une page.

Ces étapes produisent un graphe acyclique.

– Enlever les liens externes : en se basant sur *l'Hypothèse 2*, nous supposons qu'un chemin de lecture est proposé volontairement par l'auteur à l'intérieur d'un même site et que ces liens (externes) sont des liens de référence vers d'autres pages d'un autre site.

– Considérer les relations implicites : en se basant sur *l'Hypothèse 3*, nous matérialisons les *relations de séquence* entre les zones de la même page par des liens que nous appelons *liens de séquence*. La création de ces liens se fait séquentiellement, en liant la première zone à la deuxième, la deuxième à la troisième, etc. Ce type de lien permet de garder la connectivité du graphe.

Finalement, nous obtenons les liens de composition (les liens hiérarchiques et bas-de-page). Nous avons ainsi un graphe acyclique orienté (Directed Acyclic

Graph : DAG), dont les nœuds représentent les doxels et les arcs représentent soit des liens hypertextes soit des liens de séquence :

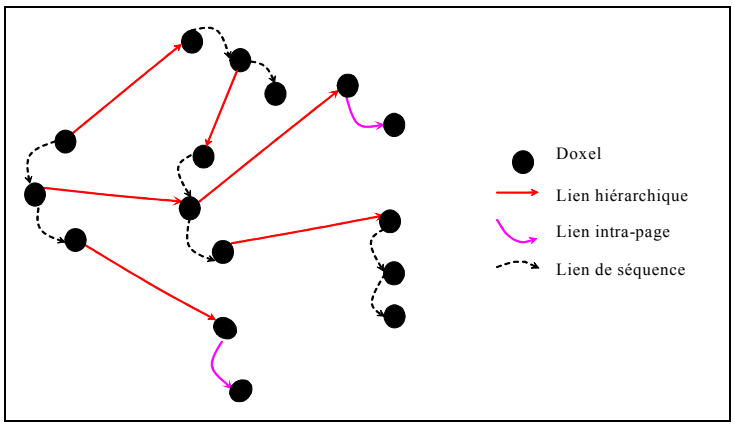

Figure 5. *Graphe orienté acyclique constitué de doxels reliés par des liens hypertextes et des liens de séquence***.**

Dans la suite, nous ne ferons plus de différence entre les types de liens. Nous parlerons de "liens inter-doxels". Ainsi, le résultat est un DAG constitué de doxels connectés par des liens inter-doxels :

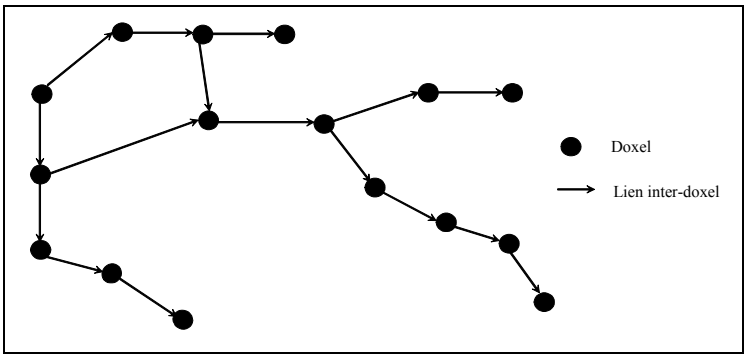

Figure 6. *Graphe de doxels Orienté Acyclique (DAG).*

5.3. Typage de liens : deuxième filtrage

Après l'identification des liens de composition (hiérarchiques, bas-de-page), nous faisons un deuxième filtrage basé sur la similarité des vecteurs de termes des zones de texte. Cela permet d'identifier les *liens de cheminement*, ce qui implicitement revient à construire les chemins de lecture.

5.3.1. *Algorithme d'extraction de chemins de lecture*

Notre algorithme effectue un parcours total du DAG en calculant, à chaque fois, la similarité entre les zones successives deux à deux. Si la valeur de similarité est supérieure à un seuil, les deux zones correspondantes sont ajoutées au chemin en cours. Le résultat final est un ensemble de chemins de lecture. Chaque chemin est constitué de doxels liés par des *liens de cheminement* :

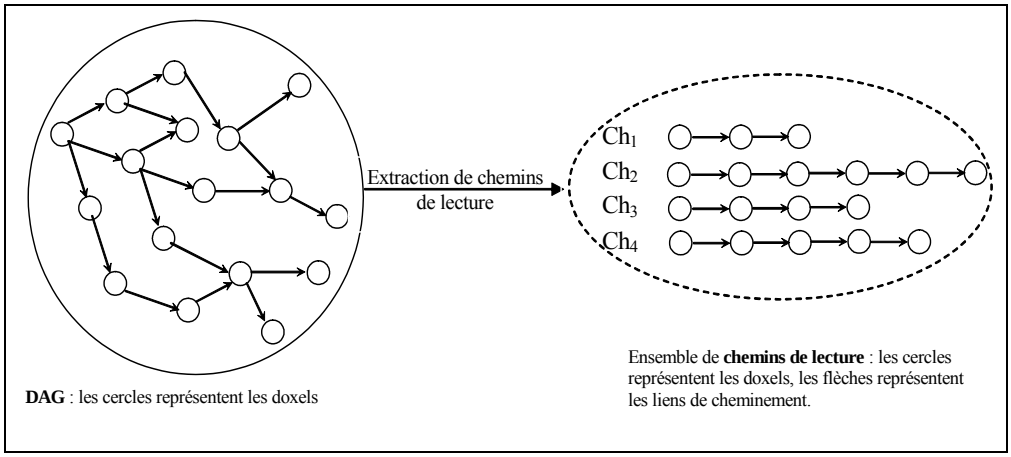

Figure 7. *Exemple de résultat de l'algorithme d'extraction de chemins de lecture à partir d'un DAG.*

La similarité entre deux nœuds d'un chemin de lecture, basée sur les hypothèses 1,1a et 1b est une combinaison de deux formes de similarité :

- *La similarité statistique* : entre deux zones qui partagent les mêmes termes.
- *La similarité sémantique* : entre deux zones qui ne contiennent pas exactement les mêmes termes mais des termes sémantiquement similaires. Par exemple, le mot "personne" et le mot "individu" ne sont pas similaires statistiquement, mais sémantiquement ils sont quasi-synonymes.

5.3.2. *Calcul de similarité*

La similarité entre deux zones de texte $Zone_i$ et $Zone_j$ est calculée par la formule suivante :

$$SIM(Zone_i, Zone_j) = \alpha \cdot SIM\_stat(Zone_i, Zone_j) + \beta \cdot SIM\_sem(Zone_i, Zone_j).$$

Avec :

- $SIM(Zone_i, Zone_j)$: la similarité globale entre les deux zones $Zone_i$ et $Zone_j$.
- $SIM\_stat(Zone_i, Zone_j)$: la similarité statistique du modèle vectoriel [Salton, 1983]. $SIM\_stat \in [0,1]$: la valeur de similarité statistique entre deux zones qui

partagent exactement les même termes est égale à 1, et celle entre deux zones qui ne partagent aucun terme est égale à 0.

- SIM_sem($Zone_i$, $Zone_j$) : la similarité sémantique calculée, en plus des analyses statistiques, en analysant les relations sémantiques entre zones de texte. Ces analyses sont basées sur un thésaurus ou une ontologie[4]. SIM_sem $\in$ [0,1] : la valeur de similarité sémantique entre deux zones est égale à 1 si chaque terme de la première zone est en rapport avec un terme de la deuxième zone, et à 0 si aucun terme de la première zone n'est en rapport avec un terme de la deuxième zone.
- Si deux zones sont similaires sémantiquement et statistiquement, alors la valeur de similarité globale (SIM) est égale à 1. Nous utilisons deux constantes α et β tel que $\alpha+\beta=1$. Avec SIM_stat et SIM_sem $\in$ [0,1], nous aurons ainsi ($\alpha \cdot$SIM_stat + $\beta \cdot$SIM_sem) $\in$ [0,1].

5.3.2.1. Calcul de la similarité statistique

La similarité statistique est calculée en utilisant le *tf.idf*[5] du modèle vectoriel [Salton,1983]. Dans notre modèle, chaque zone de texte est traitée comme une unité. La fréquence d'un terme à l'intérieur d'une zone est comparée à sa fréquence dans toute la collection des zones. Si un terme est jugé fréquent mais seulement dans quelques zones de texte, il aura un poids plus élevé que s'il était non fréquent, ou bien fréquent dans toute la collection. Donc si deux zones adjacentes partagent les mêmes termes, et si ces derniers ont des poids élevés, on peut conclure que ces deux zones sont similaires.

Chaque zone est représentée par un vecteur dans un espace à n dimensions. Une composante $w_{t,zi} \in [0,1]$ du vecteur de la zone Z_i représente le poids *tf.idf* d'un terme t dans cette zone. La valeur de similarité entre deux zones de texte est calculée en utilisant le cosinus :

$$\text{SIM\_stat}(Z_i, Z_j) = \text{COS}(Z_i, Z_j) = \frac{\sum_{t=1}^{n} W_{t,zi} W_{t,zj}}{\sqrt{\sum_{t=1}^{n} W_{t,zi}^2 \sum_{t=1}^{n} W_{t,zj}^2}}$$

5.3.2.2. Calcul de la similarité sémantique

Pour calculer cette valeur de similarité, nous prenons en compte les relations sémantiques qui mettent en évidence le rapport entre deux termes, par exemple la relation d'équivalence, la relation générique/spécifique, etc. Ces relations sont stockées dans une ontologie comme WordNet, qui consiste en la définition d'un certain nombre de termes d'un domaine, généralement appelés *concepts* et la représentation de *relations sémantiques*. Nous pouvons alors calculer la similarité entre deux zones en fonction des relations entre les termes qu'elles contiennent.

---

[4] Par exemple WordNet. L'utilisation d'une ontologie pour la RI dans un contexte Web est problématique notamment au niveau de l'ambiguïté des concepts et de la diversité de leur sens en fonction du point de vue.

[5] tf : term frequency, idf : invert document frequency.

Initialement, chaque zone est représentée par le vecteur correspondant aux termes qu'elle contient. Ensuite, nous créons pour chaque zone un nouveau vecteur basé sur son vecteur initial, « augmenté » des termes qui sont en rapport avec les termes initiaux. Enfin, nous calculons le cosinus entre les nouveaux vecteurs.

6. Expérimentations

A travers ces expérimentations, nous voulons valider les hypothèses de base de notre modèle de chemin de lecture. Pour extraire ces chemins, notre algorithme se base sur les valeurs de similarité calculées entre les zones de texte. Nous voulons donc montrer à quel point les valeurs de similarité représentent une source d'information utile pour l'extraction des chemins.

6.1. Collection de chemins de lecture

Nous utilisons une collection de chemins de lecture construite à partir d'articles scientifiques en faisant l'hypothèse qu'un article est un chemin de lecture : l'auteur suit "le fil" de son article pour décrire l'introduction, l'état de l'art, ses travaux, ses expérimentations, etc. Nous découpons chaque article en zones de texte (fragmentation) et notre algorithme a pour objectif de retrouver toutes les zones de texte qui constituent un article (reconstruction), c'est-à-dire de retrouver, dans l'ordre, toutes les zones de texte appartenant à l'article en question.

Nous utilisons 22 articles téléchargés du Web, écrits en langue française. Nous les analysons pour les découper en paragraphes. Afin d'évaluer l'impact de la taille de zone de texte sur les résultats de l'algorithme, nous utilisons deux types de collection contenant le même nombre de zones : i) une collection de zones de texte dont chacune est constituée d'un seul paragraphe, ii) une collection de zones de texte dont chacune est constituée de quatre paragraphes.

Les articles sont découpés et stockés dans l'ordre : d'abord les zones du premier article, ensuite celles du deuxième, et ainsi de suite. De telle sorte que la dernière zone de l'article N° n et la première zone de l'article N° n+1 sont adjacentes.

6.2. Calcul de similarité statistique

Pour chaque zone de la collection, l'algorithme calcule les valeurs de similarité entre elle et le reste des zones. À partir des valeurs de similarité calculées, l'algorithme cherche, pour chaque zone, la zone avec laquelle la similarité est maximale par rapport aux autres zones. Pour reconstruire un chemin de lecture, l'algorithme doit trouver que les zones similaires (les plus semblables) sont adjacentes et appartiennent au même chemin (article).

La *Fig. 8* présente les valeurs de similarité calculées entre les zones de texte adjacentes de trois articles de notre collection[6]. L'axe des abscisses représente les numéros séquentiels des zones de texte dans la collection. L'axe des ordonnées

[6] Nous nous sommes inspirés du travail de [Hearst, 1993] sur le Passage Retrieval.

représente les valeurs de similarité calculées entre les zones. Par exemple, la valeur de similarité entre la zone N° 37 et la zone N° 38 est égale à 0,4.

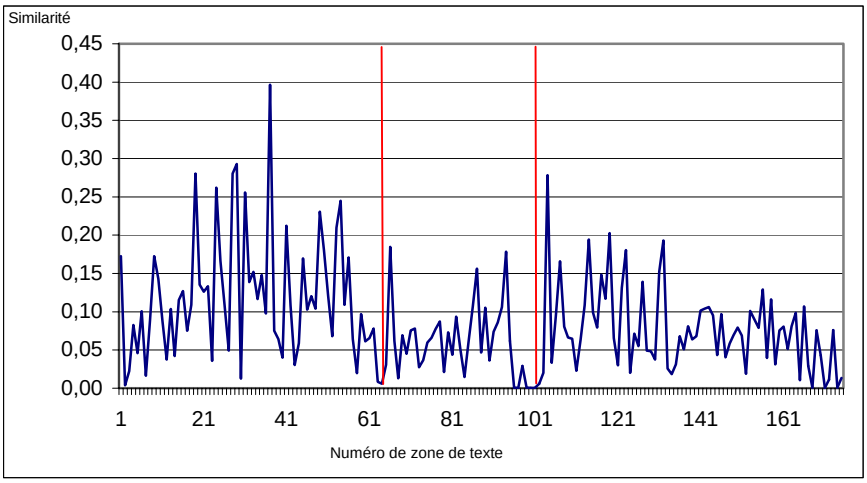

Figure 8. *Valeurs de similarité calculées pour les zones de trois articles, contenant respectivement, 65, 37, et 73 zones. Les lignes verticales représentent les bornes de chaque article.*

6.3. Résultats et discussion

Voici les résultats fournis par l'algorithme appliqué sur les deux collections, dont les zones contiennent, respectivement, 1 paragraphe et 4 paragraphes : pour la première collection, l'algorithme a trouvé 19,70 % de zones similaires et adjacentes, donc ce pourcentage représente le taux de réussite de la reconstruction de chemins de lecture. Pour la deuxième collection, le taux de réussite de reconstruction des chemins de lecture est de 51,61 %.

Les valeurs de similarité calculées entre les zones à l'intérieur d'un article (entre les lignes représentant les bornes de l'article) sont presque toujours supérieures à zéro. Les valeurs de similarité minimales sont celles calculées entre la dernière zone d'un article et la première zone de l'article suivant dans la collection. Ces résultats montrent que les zones de texte du même article sont homogènes, et que les valeurs de similarité peuvent représenter une source d'information utile pour reconstruire les chemins de lecture : les valeurs de similarité minimales indiquent la rupture d'un chemin, et les valeurs de similarité plus importantes indiquent la continuité à l'intérieur des chemins.

Dans la deuxième collection, les valeurs de similarité à l'intérieur de chaque article sont toujours supérieures à zéro, ce qui n'est pas toujours le cas pour la première collection. Ceci fait que la valeur de similarité moyenne entre les zones de chaque article est plus importante dans le cas de la deuxième collection. Les bornes

de chaque article (là où les valeurs de similarité sont minimales) sont alors mieux détectées. Elles correspondent bien aux valeurs de similarité minimales.

La différence des résultats correspondants à la première et à la deuxième collection est peut être dûe à un problème lié au modèle vectoriel : quand les vecteurs sont trop petit, la mesure de leur similarité est difficilement utilisable. Nous ajoutons que le fait de ne pas retrouver toutes les zones similaires et adjacentes pour chaque article ne veut pas dire que les zones adjacentes ne sont pas cohérentes. Il est possible que les termes qu'elles contiennent ne soient pas discriminants et se trouvent dans d'autres zones plus loin dans l'article.

À partir de ces résultats, nous pouvons conclure que les valeurs de similarité calculées entre les zones de texte représentent une source d'information utile pour l'extraction de chemins de lecture. Les valeurs de similarité minimales peuvent aider l'algorithme à détecter le début et la fin de chaque chemin, et les valeurs de similarité supérieures à la valeur de similarité minimale peuvent aider à retrouver la continuité des zones de texte correspondantes à chaque chemin. Nous avons aussi montré que la taille des zones de texte a une influence positive sur la détection des chemins : les résultats sur une collection contenant des zones de texte formées de quatre paragraphes sont meilleurs que ceux de la collection de zones de texte formées d'un seul paragraphe. Ces résultats sont encourageants bien qu'insuffisants du fait de la trop petite taille de la collection testée.

7. Conclusion

Dans ce travail, nous sommes parti de la notion de *chemin de lecture* introduite dans [Géry, 2002]. Nous avons proposé une nouvelle granularité d'information en se détachant de la notion physique de document (nœud) hypertexte. Nous avons proposé alors un nouveau type de réponse pour le SRI : *le chemin de lecture*. Ce dernier contient le minimum d'information non pertinente, et en même temps, le maximum d'information pertinente. Nous avons proposé un algorithme pour construire ces chemins de lecture en se basant sur les valeurs de similarité (statistique et sémantique) entre les zones qui le constituent. En vue d'interroger les chemins de lecture, nous avons proposé un processus d'indexation approprié [Radhouani, 2003].

Nous avons expérimenté la notion de rupture et suivit de chemin basée sur le contenu des zones de texte afin de tester la faisabilité de notre approche. Nous avons montré que les valeurs de similarité entre les zones de texte représentent une source d'information utile pour la construction des chemins de lecture. Nous avons mis en œuvre un filtrage du typage de liens, et utilisé les valeurs de similarité entre les zones de texte pour la construction des chemins de lecture.

Nous envisageons actuellement de poursuivre les expérimentations en utilisant un thésaurus ou une ontologie pour prendre en compte la valeur de similarité sémantique entre les zones de texte. Pour cela, une étude doit être menée pour choisir une ontologie adéquate à nos besoins. Une deuxième perspective concerne

l'étude des facteurs introduits dans ce travail : *la précision* et *le rappel élémentaires*, et la mise en œuvre d'une technique permettant d'évaluer un SRI en tenant compte de ces deux mesures de qualité.

8. Bibliographie

[Aguiar, 2002] Fernando Aguiar, *Modélisation d'un Système de Recherche d'Information pour les systèmes hypertextes. Application à la Recherche d'Information sur le World Wide Web*. Saint-Etienne, Thèse de Phd, Ecole Nationale Supérieure des Mines, Juin 2002.

[Brin, 1998] S. Brin, L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. WWW, 1998, 107-117.

[Bush, 1945] Vannevar Bush. *As we may think*. The atlantic Monthly, vol. 167, pp. 101-108, Juillet 1945.

[Craswell, 2001] N. Craswell, D. Hawking, and S. Robertson. *Effective site finding using link anchor information*. 24ème ACM SIGIR'01, pp. 250-257. Nouvelle-Orléans, Louisiane, Etats-Unis, Septembre 2001.

[Crestani, 2000] F. Crestani, P. L. Lee. *Searching the Web by Constrained Spreading Activation*. Information Processing & Management, 36(4), 2000, 585-605.

[Fürnkranz, 1998] J. Fürkranz. *Using links for classifing Web-pages*. Technical report TR-OEFAI-98-29, Austrian Research Institute for Artificial Intelligence, 1998.

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

[Géry 2001] M. Géry, J.P. Chevallet, *Toward a Structured Information Retrieval System on the Web: Automatic Structure Extraction of Web Pages*, International Workshop on Web Dynamics, 3 janvier, 2001

[Géry, 2002] M. Géry. *Indexation et interrogation de chemins de lecture en contexte pour la Recherche d'Information Structurée sur le WEB.* Thèse de Phd, Grenoble, Université Joseph Fourier, Octobre 2002.

[Gurrin, 2000] C. Gurrin, A.F Smeaton. *Dublin City University Experiment in connectivity Analysis for TEC-9*. 9ème Text Retrieval Conference (TREC'00). Gaithersburg, Maryland, Etats-Unis, Novembre 2000.

[Hawking, 2000] D. Hawking. *Overview of the TREC-9 WEB Track*. 9ème Text Retrieval Conference (TREC'00). Gaithersburg, Maryland, Etats-Unis, Novembre 2000.

[Hearst, 1993] M. A. Hearst, C. Plaunt. *Subtopic structuring for full-length document access*. In Proceedings of the 16th ACM SIGIR conference on Research and Development in Information Retrieval, 1993, pp. 59-68

[Kleinberg, 1998] J. Kleinberg. *Authoritative Sources in a hyperlinked Environment*. Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 668-677. Washington.

[Kleinberg, 1999] J. Kleinberg. *Authoritative Sources in a Hyperlinked Environnement*. Journal of the ACM, vol. 46, Septembre 1999, pp. 604-632.

[Nelson, 1965] Ted H. Nelson. *The hypertext*. World Document Federation (WDF'65). 1965

[Picard, 1998] J. Picard. *Modeling and combining evidence provided by document relationships using PAS systems*. ACM-SIGIR'98, 182-189.

[Radhouani, 2003] S. Radhouani, *Extraction et Indexation de chemins de lecture pour la recherche d'information sur le WEB.* Rapport de DEA, MRIM-CLIPS-IMAG, Juin, 2003.

[Salton, 1971] G. Salton. *The SMART retriever system: experiments in automatic document processing.* Prentice Hall, 1973.

[Salton, 1983] G. Salton, M. J. Mcgill. *Introduction to the modern Information Retrieval*. McGraw-Hill, Janvier 1998.

[Savoy, 2000] J. Savoy, Y. Rasolof. *Report of the TREC-9 experiment: Link-Based Retrieval and Distributed Collections*. TREC-9, NIST, Washington, Novembre 2000.

[Savoy, 2001] J. Savoy, J. Picard. *Retrieval effectiveness on the Web*. Information Processing & Management, vol. 37, 2001, 543-569.

[vanRij, 1979] C. J. van Rijsbergen. *Information Retrieval.* Butherworths, Londres, Janvier 1979.