Expérimentations sur un modèle de recherche d'information utilisant les liens hypertextes des pages Web

Bich-Liên Doan*, Idir Chibane**

* Supélec, Plateau de Moulon, 3 rue Joliot Curie, 91 192 Gif/Yvette, France
Bich-Lien.Doan@supelec.fr
** Supélec, Plateau de Moulon, 3 rue Joliot Curie, 91 192 Gif/Yvette, France
Idir.Chibane@supelec.fr

Résumé. La fonction de correspondance, qui permet de sélectionner et de classer les documents par rapport à une requête est un composant essentiel dans tout système de recherche d'information. Nous proposons de modéliser une fonction de correspondance prenant en compte à la fois le contenu et les liens hypertextes des pages Web. Nous avons expérimenté notre système sur la collection de test TREC-9, et nous concluons que pour certains types de requêtes, inclure le texte ancre associé aux liens hypertextes des pages dans la fonction de similarité s'avère plus efficace.

1 Introduction

Les systèmes de recherche d'information (SRI) sont composés essentiellement de deux modules. Un module d'indexation qui représente les documents, et un module d'interrogation qui représente la requête. La fonction de correspondance permet de calculer le degré d'appariement entre les termes de la requête et les termes d'indexation des documents afin d'évaluer la pertinence des documents par rapport à la requête. Avec le succès grandissant du Web (Google recense plus de 4 milliards de pages Web) le classement des réponses devient critique. Aussi des fonctions de correspondance prenant en compte les liens hypertextes ont vu le jour. En réalité, la plupart des fonctions de correspondance utilisées par les systèmes de recherche hypertextes combinent une mesure de pertinence calculée en fonction du contenu de la page et de la requête utilisateur avec une mesure de popularité de la page qui elle, est indépendante de la requête. Cette dernière mesure repose sur la structure du Web, considéré comme un graphe orienté de pages et de liens. L'hypothèse (Savoy et Rasolof 2000) stipule qu'une page est supposée être de bonne qualité si elle a beaucoup de liens entrants, en particulier, si les pages qui pointent vers elle sont aussi de bonne qualité. Un certain nombre de systèmes qui tiennent compte de la structure du web dans la fonction de correspondance ont été développés. Les systèmes les plus connus sont InDegree, PageRank (Brin et Page 1998) utilisé dans Google, HITS (Kleinberg 1998) et SALSA (Lempel et Moran 2000). Dans ces systèmes, la fonction de correspondance, qui relie la requête aux documents est remplacée par une fonction de classement des résultats qui elle est indépendante de la requête. L'étude des systèmes existants nous a permis de conclure que toutes les fonctions de correspondance basées sur les liens hypertextes ne dépendent pas des termes de la requête. Cela a diminué considérablement la précision des résultats retrouvés. En effet, l'analyse du comportement des utilisateurs dans leur recherche montre qu'ils ne s'intéressent pas aux pages populaires, si ces dernières ne contiennent aucun terme de la requête.

Dans cet article, nous proposons une fonction de correspondance qui dépend en plus du contenu textuel de la page, de sa popularité par rapport aux termes de la requête. Nous partons de cette hypothèse : une page est connue pour un terme t de la requête si celle-ci contient des liens entrants et/ou sortants de pages qui elles aussi contiennent le terme t de la requête. Nous limitons l'analyse du contenu des pages pointant ou pointés au seul terme lié à l'ancre des liens entrants et/ou sortants. Nous avons évalué notre système par rapport à un système plat, dans notre cas le modèle vectoriel (Salton et al 1975), et par rapport à d'autres systèmes dont la fonction de correspondance repose sur les liens entrants et/ou sortants indépendamment des termes de la requête. Cette étude se termine par l'analyse des résultats et une conclusion.

2 Modélisation de la fonction de correspondance

Notre système est composé de deux modules : le module d'indexation qui représente les documents et le module d'interrogation qui représente les requêtes. La nouveauté dans notre modèle est l'utilisation d'une fonction de correspondance qui dépend en plus du contenu textuel des pages, de popularité de celle-ci par rapport aux termes de la requête. Cette dépendance permet une meilleure adéquation des résultats retrouvés avec les besoins utilisateurs. Notre fonction de correspondance repose sur deux mesures : l'une est classique et utilisée dans les systèmes actuels. C'est la mesure « cosine » qui calcule le cosinus de l'angle entre le vecteur représentant la requête et celui représentant le document. L'autre mesure utilise la structure hypertexte des pages Web, elle est composée des deux fonctions suivantes :

$$Sim_{struct1}\left(\vec{D},\vec{Q}\right)=\log\left(1+\frac{\left|IC(D,t)\right|}{\left|I(D)\right|}\right) \qquad Sim_{struct2}\left(\vec{D},\vec{Q}\right)=\log\left(1+\frac{\left|IC(D,t)\right|}{\left|C(t)\right|}\right)$$

La première fonction correspond à la fraction du nombre de documents qui citent le document D et contiennent au moins un terme t de la requête Q par rapport au nombre de documents qui citent le document D. Cette mesure favorise les documents qui contiennent moins de liens entrants. La seconde fonction correspond au nombre de documents qui citent le document D et qui contiennent au moins un terme t de la requête Q par rapport au nombre de documents qui contiennent les termes de la requête Q. Cette deuxième mesure favorise les documents qui contiennent des termes peu fréquents dans toute la collection. Plusieurs variantes peuvent être envisagées, par exemple en ne tenant compte que des liens sortants, des liens entrants ou les deux ensemble. Ces deux mesures peuvent être combinées en une seule comme suit :

$$Sim_{struct}=Max\left(Sim_{struct1},Sim_{struct2}\right)$$

Dans notre système, nous n'avons considéré que les liens où le texte ancre contient au moins un terme de la requête, i.e IC (P,t) représente l'ensemble des pages qui comprend le terme t dans le texte ancre du lien qui pointe la page P. Notre fonction de correspondance entre document et requête est:

$$Sim\left(\vec{D},\vec{Q}\right)=\alpha\cdot Sim_{\cos ine}+\beta\cdot Sim_{struct1}$$

où α et β sont deux facteurs tel que α + β = 1. Ces deux facteurs nous permettent d'accorder de l'importance à l'une des deux mesures selon la structure du Web. Ils nous permettent

aussi de comparer les résultats et de déterminer l'importance d'une mesure par rapport à l'autre. Dans notre évaluation ces deux facteurs sont fixés à 0.5. Nous pourrions tester notre système avec différentes valeurs de α et β. Apres avoir décrit notre système qui intègre une nouvelle fonction de correspondance, nous allons maintenant passer à l'expérimentation et à l'évaluation de notre système.

3 Expérimentations sur la collection TREC-9 (WT10g)

Dans le cadre de nos expérimentations, nous avons choisi comme collection de tests la collection WT10g en raison de son statut de collection standard dans le domaine de la recherche d'information. D'une façon générale la collection a été conçue dans l'objectif de modéliser une recherche réelle sur le Web et de permettre une évaluation fiable des méthodes de recherche d'information reposant sur l'analyse des liens. Nous expliquons la façon dont nous avons exploité les données composant la collection pour les adapter à nos tests. Puis nous détaillons les tests effectués. Nous avons exécuté 50 requêtes et comparé sept algorithmes. Nous concluons avec une analyse des résultats.

3.1 Les tests effectués

Dans cette section nous décrivons les tests que nous avons effectués. Nous avons testé notre système sur un ensemble significatif de sites présents dans la collection WT10g qui sont pertinents à au moins un topique de l'ensemble des topiques que nous avons utilisés. Avant de détailler notre méthodologie expérimentale nous rappelons quelques chiffres sur la collection utilisée. L'idée de départ consistait à tester les 871 sites de la collection TREC. Cependant, pour une question d'espace mémoire et de calcul nous avons sélectionné pour les tests les sites contenant au plus 700 pages. Avec cette limite sur le nombre de pages nous sommes passés de 871 sites à 490 sites. Notre collection de tests contient 91460 documents et 29793 liens hypertextes, en moyenne 0.33 liens par pages, pas assez pour un système hypertexte. C'est l'un des inconvénients de notre collection de tests TREC. En ce qui concerne notre index, Nous avons un index de 489219 termes différents après l'élimination des mots vides, ce qui mène à une moyenne de 146,16 termes par document. Nous avons comparé les fonctions de correspondance suivantes :

(i) celles qui tiennent compte du contenu textuel de la page seulement : calcul des fréquences des termes (simple modèle vectoriel). La fonction de correspondance calcule le cosinus de l'angle entre le vecteur représentant le document et celui représentant la requête.

(ii) celles qui tiennent compte de la popularité de la page indépendamment de la requête : on distingue trois fonctions différentes : La première est celle qui ne prend compte que du nombre de liens entrants, la deuxième ne tient compte que du nombre de liens sortants et la dernière du nombre de liens entrants et sortants.

$$F_1 = Sim_{\cos ine}(\vec{D},\vec{Q}) + \left|I(D)\right|$$

$$F_2 = Sim_{\cos ine}(\vec{D},\vec{Q}) + \left|O(D)\right|$$

$$F_3 = Sim_{\cos ine}(\vec{D},\vec{Q}) + \left|I(D) + O(D)\right|$$

(iii) celles qui tiennent compte de la popularité de la page par rapport aux termes de la requête. Dans ce cas de figure, on distingue trois types de fonctions comme auparavant. Elles

sont calculées à partir des textes ancres des liens entrants et/ou sortants qui dépendent des termes de la requête utilisateur. Par exemple, la fonction décrite ci-dessous ne tient compte que des liens entrants.

$$F_4 = \frac{1}{2.}\left(\frac{\sum_{t_i \in D \cap Q} w_{t_i,D} \cdot w_{t_i,Q}}{\sqrt{\sum_{t_i \in D} w_{t_i,D}^2 \cdot \sum_{t_i} w_{t_i,Q}^2}} + \log\left(1 + \frac{|IC(D,t)|}{|I(D)|}\right)\right)$$

Même chose pour les liens sortants et la combinaison des deux types de liens.
Le processus d'évaluation est le suivant : nous avons exécuté 50 requêtes sur chaque système et nous avons calculé, pour chaque algorithme, la précision aux 11 niveaux standard du rappel : 0%, 10%, 20%,…,100%. Les résultats de nos expériences sont montrés dans Tab 1 et Tab 2. La figure ci-dessous présente des résultats de quelques requêtes avec des courbes de comparaison entre trois fonction de correspondance : le contenu seulement, les liens, et notre fonction de correspondance reposant sur le texte ancre.

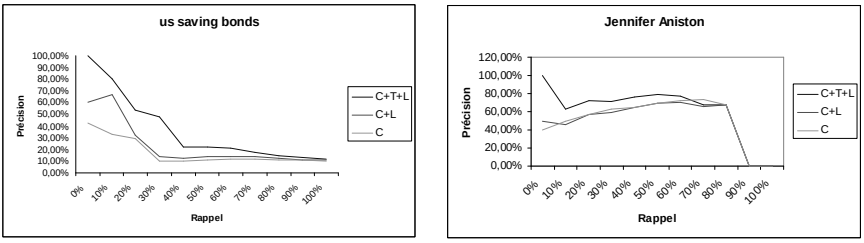

FIG. 1 - *Comparaison des requêtes : contenu (C) , contenu + lien (C+L), contenu + lien + ancre (C+T+L)*

Le tableau suivant présente un extrait des résultats obtenus pour les 50 requêtes exécutées sur notre système.

Topic	Les 11 niveaux du rappel standard										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
451	50,00%	50,00%	15,38%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
452	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
453	100,00%	60,00%	58,33%	50,00%	48,28%	26,98%	27,27%	11,88%	0,00%	0,00%	0,00%
454	46,15%	56,25%	48,28%	51,06%	45,31%	37,23%	32,81%	28,82%	0,00%	0,00%	0,00%

TAB 1 - *Extrait de la table récapitulative des résultats obtenus en exécutant les 50 requêtes*

La figure suivante présente un tableau comparatif des résultats obtenus selon les différentes fonctions de correspondance.

Fonctions de correspondance	Les 11 niveaux du rappel standard										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Contenu+texte ancre+liens in	45,96%	16,45%	12,92%	11,48%	9,07%	7,03%	6,65%	5,88%	4,12%	2,55%	1,75%
Contenu+texte ancre+liens out	44,51%	15,44%	12,67%	11,17%	9,05%	7,00%	6,64%	5,87%	4,12%	2,55%	1,75%
Contenu+nombre de liens in	40,04%	14,96%	10,57%	9,68%	8,19%	6,57%	6,32%	6,04%	4,07%	2,52%	1,76%
Contenu+nombre de liens out	38,65%	12,64%	10,94%	8,88%	8,06%	6,45%	0,06%	5,82%	4,23%	2,54%	1,77%
Contenu+nombre de liens in et out	39,34%	12,56%	10,69%	8,87%	7,86%	6,33%	6,05%	5,92%	4,15%	2,51%	1,75%
Contenu seulement	46,41%	16,27%	13,04%	11,24%	9,06%	7,03%	6,66%	5,88%	4,12%	2,55%	1,75%

TAB 2 - *La moyenne de la précision en fonction du rappel pour les 50 requêtes exécutées*

3.2 Analyse des résultats

L'analyse des résultats de nos expérimentations montre que les algorithmes reposant sur le texte ancre dans le classement des résultats sont meilleurs par rapport à ceux qui ne le considèrent pas. La plupart des résultats obtenus avec notre formule pour le top du rappel sont au dessous des autres algorithmes. Avec une précision moyenne au niveau de 10% du rappel dépassant 45% (45.96% pour les liens entrants, 44,51 pour les liens sortants et 45,10 en tenant compte des deux liens), notre système est légèrement au-dessous du système basé sur le contenu seulement avec 46.41% de moyenne. Cependant, les mauvaises performances ont été observées dans les systèmes qui tiennent compte de la popularité de la page (liens entrants et sortants) indépendamment de la requête avec moins de 40% de précision moyenne (liens entrants 40.04 %, liens sortants 38,65 et deux types de liens 39,34). En ce qui concerne les requêtes exécutées dans notre système, nous avons constaté que 13 sur 49 ont une précision de 100% au rappel de 10%, c'est à dire que tous les documents retrouvés correspondant au 10% du rappel, pour les 13 requêtes, sont pertinents. Ce chiffre est très important vu que les utilisateurs du Web sont plus intéressés par les premières pages retournées par les moteurs de recherche que par les autres pages classées au milieu ou à la fin. En plus, les résultats obtenus par notre fonction de correspondance pour quelques requêtes sont meilleurs par rapport aux résultats obtenus par le modèle vectoriel. Un extrait des résultats des requêtes pour lesquelles notre système produit de bons résultats est illustré dans la figure FIG 1. Nous pensons que selon certain type de requête, notre système prenant compte des ancres des liens dans la fonction de correspondance peut s'avérer plus efficace. C'est le cas pour les requêtes « Jennifer Anniston » ou « US saving bonds ». Ce résultat peut s'expliquer pour des requêtes générales ou très courantes (beaucoup de pages référencent le sujet). Les requêtes très spécifiques et rares (termes peu présents dans toute la collection) donnent de moins bons résultats avec notre système, il est préférable d'utiliser alors le modèle vectoriel. Si on compare les différentes fonctions de correspondance étudiées en fonction de la rapidité de trouver les documents pertinents, on constate que notre proposition n'est pas loin du modèle vectoriel. Avec une précision de 31.19 % pour notre système contre 31.27 pour le modèle vectoriel pour les cinq documents pertinents retrouvés. Les systèmes dont la fonction de correspondance ne dépend pas de la requête sont loin par rapport au système que nous avons proposé avec une précision moyenne pour les cinq documents pertinents trouvés qui avoisine 24%. Cette nouvelle technique permettrait de mieux répondre aux besoins exprimés par des utilisateurs et limiterait les nouvelles pratiques pour fausser les calculs comme le spamming (achat des liens ou ajout des mots clé blancs dans l'entête de la page). Enfin, nous n'avons pas eu l'occasion d'évaluer notre système par rapport aux autres résultats obtenus par les différents algorithmes exécutés sur la collection de tests pour deux raison : la première est la restriction du nombre de sites testés dans notre système par rapport à d'autres qui ont testé leur système sur toute la collection. Deuxièmement, le choix de la façon d'évaluer notre système. Nous avons choisi l'utilisation de la précision en fonction des 11 niveaux standard du rappel. D'autres systèmes se basent sur l'évaluation en fonction des 10 premiers documents obtenus pour chaque algorithme.

4 Conclusion et perspectives

Plusieurs travaux ont été menés sur l'utilisation des liens dans la recherche d'information 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. Ce que nous avons proposé dans ce papier est un moteur de recherche utilisant à la fois le modèle vectoriel et les liens hypertextes. La nouveauté dans notre système est l'utilisation d'une fonction de correspondance qui tient compte de la popularité d'une page par rapport aux termes de la requête. Les résultats obtenus montrent qu'ils sont proches des résultats obtenus par le modèle vectoriel basé sur le contenu textuel des pages. De plus, la précision pour certaines requêtes est grande au top du classement par rapport aux autres systèmes testés. Donc, le texte ancre peut être un facteur déterminant pour comprendre le contenu d'une page. Cependant, notre système prévoit quelques limitations dues à notre implémentation. La limitation la plus évidente de notre système concerne le modèle de requête que nous avons proposé. D'une part pour certains besoins d'information, la traduction de la requête utilisateur peut s'avérer très difficile. D'autre part, nous avons tendance à penser (vu les caractéristiques des requêtes observées dans les moteurs de recherche très utilisées du Web) que l'utilisateur ne serait pas prêt à fournir l'effort pour bien formuler sa requête. Malgré cela, les expérimentations que nous avons menées avec des requêtes bien formulées montrent que notre modèle pourrait s'avérer utile. Nous poursuivons actuellement un travail pour tester notre système sur l'ensemble de la collection afin de le comparer par rapport aux résultats obtenus par les différents systèmes testés sur la collection TREC. En plus, nous allons faire varier les coefficients présentés dans cet article afin de trouver un compromis entre la mesure basée sur le contenu et celle basée sur la structure.

5 Bibliographie

Brin S. et Page L. (1998), The anatomy of a large-scale hypertextual Web search engine, In Proceeding of WWW7, 1998.

Kleinberg L. (1998), Authoritative sources in a hyperlinked environment, In Proceeding of 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

Lempel R. et Moran S. (2000), The stochastic approach for link-structure analysis (SALSA) and the TKC effect, In Proceeding of 9th International World Wide Web Conference, 2000.

Savoy J. et Rasolof Y. (2000), Link-Based Retrieval and Distriuted Collections, Report of the TREC-9 experiment: Proceedings TREC-9, 2000.

Salton G., Yang C.S. et Yu C.T. (1975), A theory of term importance in automatic text analysis, Journal of the American Society for Information Science and Technology, 1975

Summary

The matching function is essential for any information retrieval system because it enables the requested documents to be both selected and sorted. We propose to model a matching function taking into account both the content and the hypertext links between the Web pages. We tested our system over the benchmark TREC-9 and we conclude that for certain types of queries, including the anchor text of hypertext links in the matching function contributes to improving the overall effectiveness of the system.