Evaluation des performances d'un système de recherche d'information utilisant un algorithme de segmentation thématique de pages Web

Idir Chibane, Bich-Liên Doan

SUPELEC, plateau de Moulon, 3 rue Joliot Curie
91192 Gif sur Yvette
France
Idir.Chibane@supelec.fr, Bich-Lien.Doan@supelec.fr

RÉSUMÉ. *Dans cet article, nous proposons une méthode de segmentation thématique de pages Web qui utilise à la fois des critères visuels et de format (balises <HR>, <H1>, couleur, …) afin d'extraire des segments thématiques. Nous utilisons la segmentation pour améliorer les performances d'un système de recherche d'information. Nous proposons de modéliser une fonction de correspondance qui tienne compte à la fois du contenu d'une page Web et du voisinage de cette page définis par les segments thématiques appelés blocs thématiques qui la référencent. Ce voisinage est calculé dynamiquement en pondérant les liens hypertextes reliant les blocs thématiques aux pages Web en fonction des termes de la requête contenus dans ces blocs thématiques. Notre approche montre de bons résultats sur la collecion TREC.*

ABSTRACT: In this paper, we explore the use of new page segmentation algorithm using both visual and structural mark-up (<H1>,<HR>) to partition web pages into blocks and investigate how to take advantage of block-level evidence to improve retrieval performance in the web. We propose a new ranking function that combines content and link rank based on propagation of scores over links on block-to-page graph. This function propagates scores from blocks of source pages to destination pages in relation with query terms. Our approach shows good results over TREC collections.

MOTS-CLÉS : *recherche d'information, systèmes hypertextes, analyse de liens, web, propagation de pertinence, méthodes de segmentation, analyse thématique.*

KEYWORDS: *information retrieval, hypertext systems, link analysis, web, relevance propagation, segmentation methods, topic analysis.*

1. Introduction

Dans cet article, nous proposons une méthode de segmentation de pages afin d'extraire des blocs à partir de pages Web en utilisant les balises HTML (critères visuels et de présentation). Nous suggérons d'utiliser les critères visuels afin de déterminer les frontières entre les différents segments d'une page Web. Une fois les segments extraits, nous proposons de modéliser une fonction de correspondance d'un système de recherche d'information prenant en compte à la fois du contenu des blocs thématiques constituant une page et du voisinage de cette page. Ce voisinage est calculé dynamiquement en pondérant les liens hypertextes reliant les blocs thématiques aux pages Web en fonction des termes de la requête contenus dans ces blocs thématiques. Avant de proposer notre solution, nous proposons un bref état de l'art sur les approches existantes de segmentation de pages Web à partir de la structure HTML des pages Web, et des méthodes d'analyse de liens. Ensuite nous proposons une méthode de segmentation de pages Web utilisant des critères visuels et de représentation de contenu textuel de pages Web. Nous appliquons notre méthode d'analyse thématique pour déterminer les blocs thématiques de chaque page que nous allons utiliser dans un système de recherche d'information. Enfin, nous décrivons les tests effectués et nous concluons par l'analyse des résultats.

2. Etat de l'art

2.1. Approches de segmentation

Beaucoup de travaux ont porté sur la segmentation de pages Web. Certains chercheurs ont utilisé des techniques issues des bases de données objets pour structurer les données du web (Hammer et al, 1997) (Adelberg et al, 1998) (Ashish et al, 1997). D'autres travaux (Embley et al, 1999) (Chen et al, 2001) (Chakrabarti, 2001) (Chakrabarti et al, 2002) ont porté sur l'extraction d'informations structurelles à partir de l'arbre DOM (Document Object Model) d'une page HTML afin de la découper en plusieurs blocs homogènes. Cependant, en raison de la flexibilité de la syntaxe de HTML, beaucoup de pages web n'obéissent pas aux spécifications HTML de W3C (World Wide Web Consortium), ce qui peut causer des erreurs dans la structure arborescente DOM d'une page Web. Par ailleurs, l'arbre DOM est initialement introduit pour la présentation des pages dans un navigateur plutôt que la description de sa structure sémantique. Un exemple frappant de la non pertinence de la structure DOM est la multifonctionnalité des balises HTML. Par exemple, la balise <TABLE> peut être utilisé comme un tableau de données ou un moyen de représentation de la structure sémantique d'une page Web.

2.4 Utilisation des liens dans la recherche d'information

Au début de la recherche d'information reposant sur l'analyse des liens, le contenu des documents et la structure des liens ont été utilisés séparément. C'est le cas des approches comme (Hawking, 2000) (Craswell et al., 2003) (Craswell et al., 2004) qui utilisent d'une part le modèle vectoriel (Salton et al., 1975) pour calculer un degré de pertinence reposant sur le contenu du document et l'autre part, les liens hypertextes pour calculer un indice de popularité indépendant de la requête (exemple PageRank (Brin et al., 1998)). Ensuite, ces deux scores sont combinés pour classer les documents retrouvés. Ces dernières années, plusieurs méthodes prenant en compte une certaine dépendance entre le contenu des documents et la structure des liens ont été développées. (Qin et al., 2005) distinguent deux catégories de techniques d'analyse des liens. La première catégorie est celle qui utilise l'information du contenu pour améliorer les performances des algorithmes d'analyse des liens (Kleinberg, 1998) (Lempel et al., 2000) (Haveliwala, 2002). Dans cette catégorie, les documents sont retournés selon l'existence ou non des termes de la requête utilisateur dans ces documents, puis des algorithmes d'analyse des liens qui s'inspirent des études des graphes sont appliqués pour classer ces documents. L'autre catégorie est la propagation de pertinence qui propage l'information du contenu à travers la structure du web (Mcbryan, 1994) (Song et al., 2004) (Shakery et al., 2003) .

3. Notre système

Nous proposons notre algorithme de segmentation des pages Web qui nous permet d'extraire des blocs thématiques qui seront utilisés dans notre fonction de correspondance. Tout d'abord, nous commençons par la méthode de segmentation thématique des pages Web.

3.1. Algorithme de segmentation thématique à critères visuels

Dans ce qui suit, nous proposons une solution pour la segmentation d'une page Web. Cette solution repose sur une évaluation de plusieurs segmentations en utilisant une méthode d'analyse thématique. Le but est de trouver une segmentation à base de critères visuels (ligne, la couleur) et de représentation du contenu (paragraphe, sous-titres) qui permet d'avoir des segments thématiques. Ces critères de délimitation de segments permettent dans la plupart des cas de passer d'une section à une autre ou de changer une idée exposée dans une section précédente. Nous voulons combiner ces deux critères afin de segmenter les pages Web de sorte que les sections soient distantes entre eux et homogènes à l'intérieur de leur contenu. Afin de pouvoir calculer ces distances, nous disposons de deux mesures : l'une est appliquée à l'intérieur d'un segment qui repose sur la cooccurrence entre les termes appartenant au même segment. Et l'autre repose sur la mesure du cosinus entre deux vecteurs segments. Ces deux mesures sont définies comme suit :

3.1.1 Mesure de cohérence d'un bloc

Nous supposons que les termes les plus fréquents dans un bloc constituent le thème du bloc. Nous avons fixé le nombre de termes constituant un thème à dix termes. La cohérence à l'intérieur d'un bloc est calculée de la manière suivante :

$$Coh(b) = \frac{1}{n^2} \sum_{t_i \in b} \sum_{t_j \in b} Coocur(t_i, t_j)$$

$$avec \quad Coocur(t_i, t_j) = \frac{Nbdoc(t_i \cap t_j)}{Nbdoc(t_i) + Nbdoc(t_j) - Nbdoc(t_i \cap t_j)}$$

Nous remarquons que plus la cooccurrence entre les termes d'un bloc est grande, plus la cohérence à l'intérieur du bloc est élevée.

3.1.2 Mesure de distance entre deux blocs adjacents

Il n'existe pas de vraie distance entre deux vecteurs blocs. Dans notre système, nous avons calculé une valeur qui peut être interprétée comme une distance entre deux blocs. Cette mesure est basée sur la mesure de similarité entre blocs. L'inverse de la similarité peut être considérée comme une distance entre deux blocs. Cette mesure est définie comme suit :

$$Dist(b_i, b_j) = \frac{1}{Sim(V_{b_i}, V_{b_j})} = \frac{1}{\cos(V_{b_i}, V_{b_j})} = \frac{\sqrt{\sum_{i=1}^{n} w_{k,b_i}^2} \times \sqrt{\sum_{i=1}^{n} w_{k,b_j}^2}}{\sum_{k=1}^{n} w_{k,b_i} \times w_{k,b_j}}$$

Où V_{bi} et V_{bj} sont deux vecteurs de termes des deux segments b_i et b_j respectivement. Le poids de chaque terme est calculé en utilisant TFIDF.

3.1.3 Fonction d'évaluation d'une segmentation

La fonction d'évaluation d'une segmentation est calculée à partir des deux mesures : cohérence du contenu des blocs d'une page et la distance entre ces blocs. Cette mesure est décrite de la manière suivante :

$$EvalSegm(S_i, P) = \left[\frac{1}{n} \sum_{1 \le i \le n} Coh(b_i) \right] * \left[\frac{1}{n-1} \sum_{1 \le i \le n-1} Dist(b_i, b_{i+1}) \right]$$

Où S_i est une solution de segmentation de la page *P* reposant sur un critère visuel candidat à la segmentation. Elle est composée de *n* blocs thématiques. La meilleure solution de segmentation est celle qui a une grande valeur de la fonction *EvalSegm(S_i,P).* C'est cette solution qui sera retenue afin de segmenter la page Web.

3.2 Processus de segmentation thématique à critères visuels

Le processus de segmentation fonctionne comme suit : pour chaque page de la collection, un index est créé en suivant les étapes d'indexation standard (extraction

des mots, lemmatisation et suppression des mots vides). Puis, une matrice de cooccurrence entre termes est déduite à partir de l'index. Cette matrice est utilisée pour évaluer les différentes solutions de segmentation générées à partir de la même page. Ensuite, pour chaque document de la collection, on extrait les différents segments qui le composent en utilisant des différents délimiteurs de segments contenus dans la liste des critères. Une solution par critère est générée. Le résultat est un ensemble de solutions de segmentation. L'évaluation est faite de manière à ce que les segments soient homogènes. La meilleure solution de segmentation thématique est retenue et l'index bloc est créé. Un graphe de blocs est construit à partir de deux matrices : matrice des liens entre blocs et les pages pointées par ces blocs et la matrice d'importance entre la page est ces blocs. Les relations page-bloc sont déterminées par l'analyse de la topographie de la page, et les relations bloc-page sont déterminées par l'analyse des liens qui relient les blocs à des pages Web. Le but est de construire un graphe sémantique de telle sorte que chaque noeud représente exactement un seul thème sémantique. Ce graphe peut mieux décrire la structure sémantique du Web. Une fois que l'importance donnée à un bloc est calculée, l'information est utilisée dans une fonction de voisinage d'une page Web.

3.3 Fonction de voisinage

La nouveauté dans notre modèle est l'utilisation d'une fonction de correspondance qui dépend en plus du contenu textuel des pages, de son voisinage. Cette dépendance permet une meilleure adéquation des résultats retrouvés par un modèle classique de RI avec un besoin utilisateur. Notre fonction de correspondance repose sur deux mesures : l'une repose sur le contenu seul de la page qui donne de meilleurs résultats et largement utilisées dans les systèmes actuels. C'est la mesure OKAPI 25.

La deuxième mesure est celle du voisinage qui tient compte de la structure du Web composée des liens hypertextes. Afin de comprendre notre démarche, nous partons de l'hypothèse suivante : *on considère qu'une page est bien connue pour un terme T de la requête Q si celle-ci contient beaucoup de liens entrants émis à partir des pages qui elles aussi contiennent le terme T de la requête* (Doan et *al.*, 2005). Cette mesure tient compte du nombre de termes de la requête contenus dans les pages Web. L'idée principale de notre mesure de voisinage est de pondérer les liens entrants selon le nombre des termes contenus dans la page source d'un lien entrant. L'hypothèse qu'on a fixée au départ stipule que le poids d'un lien émis par une page contenant *n* termes de la requête *Q* est deux fois plus important que le poids d'un lien contenant *n-1* termes de la requête *Q*. La mesure de voisinage que nous avons proposée est décrite comme suit : Supposons une requête *Q* contenant nbt_q termes et une page *P* retournée par un système traditionnel de recherche d'information. Soit $nbt_q(P)$ le nombre de termes de la requête *Q* contenus dans la page *P*. Notons E_P le nombre de liens entrants de la page *P*.

$$RankLR(P,Q) = \sum_{Pi \to P} \frac{Poids(P_i, P) * RankDR(P_i, Q)}{E_P}$$

Avec *Poids(Pi,P)* est la pondération du lien entre la page *Pi* et la page *P*. Plus la page *Pi* contient de termes de la requête *Q*, plus le poids du lien entre *Pi* et *P* est grand. Ce poids est défini comme suit :

$$Poids(P_i, P) = \frac{2^{nbt_q(P_i)}}{2^{nbt_q}} * \beta = \frac{\beta}{2^{nbt_q - nbt_q(P_i)}}$$

β un paramètre compris entre 0 et 1 qui vérifie la condition suivante :

$$\sum_{k=1}^{nbt_q} \frac{\beta}{2^{nbt_q - k}} = 1 \quad \Rightarrow \quad \beta * \sum_{k=1}^{nbt_q} \frac{1}{2^{nbt_q - k}} = 1 \quad \Rightarrow \quad \beta * \left(\frac{1}{2^{nbt_q - 1}} + \frac{1}{2^{nbt_q - 2}} + ... + \frac{1}{2} + 1 \right) = 1$$

Nous avons une suite géométrique de rayon ½ d'où la somme d'une telle suite est donnée par la formule suivante:

$$\left(\frac{1}{2^{nbt_q - 1}} + \frac{1}{2^{nbt_q - 2}} + ... + \frac{1}{2} + 1 \right) = \frac{1 - \left(\frac{1}{2}\right)^{nbt_q}}{1 - \frac{1}{2}} = 2 * \left(1 - \left(\frac{1}{2}\right)^{nbt_q} \right) \quad \Rightarrow \quad \beta = \frac{1}{2 * \left(1 - \left(\frac{1}{2}\right)^{nbt_q} \right)}$$

En remplaçant β par sa valeur dans l'équation du poids d'un lien, nous obtiendrons :

$$Poids(P_i, P) = \frac{2^{nbt_q(P_i)}}{2^{nbt_q + 1} * \left(1 - \left(\frac{1}{2}\right)^{nbt_q} \right)}$$

Après la transformation de la fonction de calcul du voisinage d'une page en remplaçant β par sa valeur, nous obtiendrons la fonction suivante qu'on a utilisée dans nos expérimentations :

$$RankLR(P,Q) = \sum_{P_i \to P} \frac{2^{nbt_q(P_i)} * RankDR(P_i, Q)}{2^{nbt_q + 1} * \left(1 - \left(\frac{1}{2}\right)^{nbt_q} \right) * E_P}$$

Le même calcul s'effectue aux niveaux des blocs thématiques. La fonction de voisinage des pages reposant sur les blocs thématiques est définie comme suit.

$$RankBLR(P,Q) = \sum_{B_{i,j} \to P} \frac{2^{nbt_q(B_i)} * RankBR(B_{i,j}, Q) * IMP(B_{i,j}, P_i)}{2^{nbt_q + 1} * \left(1 - \left(\frac{1}{2}\right)^{nbt_q} \right) * E_P}$$

Avec $B_{i,j}$ est le bloc thématique J de la page P_i et $IMP(B_{i,j}, P_i)$ est une fonction d'importance qui assigne une valeur d'importance d'un bloc thématique par rapport à la page qui le contient. Cette fonction est calculée de la manière suivante :

$$IMP\left(B_{i,j}, P_i\right) = \frac{dl\left(B_{i,j}\right)}{dl\left(P_j\right)}$$

Où $dl(B_{i,j})$ et $dl(P_i)$ représente la taille en nombre de terme du bloc $B_{i,j}$ et de la page P_i respectivement.

3.4. Expérimentations sur la collection TREC

Dans le cadre de nos expérimentations, nous avons choisi comme collection de tests la collection WT10g. Afin d'adapter la collection TREC à nos expérimentations sur les blocs thématiques, nous avons considéré tous les blocs thématiques des documents pertinents comme des blocs pertinents. La mesure principale d'évaluation de nos expérimentations est la précision moyenne aux 11 niveaux standard du rappel qui sont 0%, 10%, 20%,…,100% du rappel.

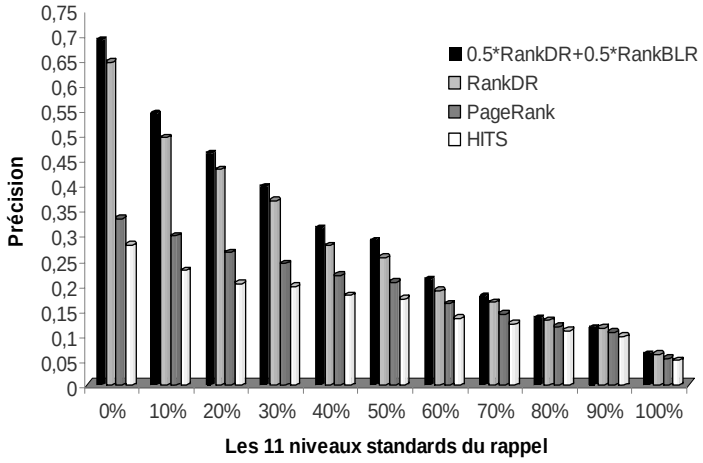

Figure 1. *La précision moyenne aux 11 niveaux standard du rappel pour les fonctions de correspondance (contenu seul des pages Web, PageRank, HITS et combinaison du contenu des pages et du voisinage basé sur les blocs)*

La figure 1 montre les résultants expérimentaux obtenus sur la collection TREC en utilisant quatre fonctions de correspondance qui reposent sur la page considérée comme unité d'information. La première fonction repose sur le contenu textuel seul de la page (RankDR). Elle représente l'algorithme de base de nos évaluations. La

deuxième fonction repose sur la combinaison du contenu textuel de la page et de son voisinage par rapport aux blocs thématiques qui pointent cette page (0.5*RankDR+0.5*RankBLR). La troisième et la quatrième fonction reposent sur la popularité d'une page Web. Pour cela, nous avons utilisé deux algorithmes les plus connus des algorithmes d'analyse des liens qui sont PageRank et HITS. Le voisinage d'une page apporte plus de précision dans les résultats retournés par un moteur de recherche classique.

5. Conclusion

Dans cet article, nous avons proposé une méthode de segmentation thématique de pages Web. Cette méthode de segmentation nous permet d'extraire des blocs thématiques à partir des pages Web. Nous avons proposé un modèle de propagation de pertinence. La nouveauté dans notre modèle est l'utilisation d'une fonction de correspondance qui tient compte à la fois le contenu de la page et de son voisinage reposant sur les blocs thématiques qui la référencent. Ce voisinage est calculé dynamiquement en pondérant les liens hypertextes reliant les blocs thématiques à des pages Web en fonction du nombre de termes de la requête contenus dans ces blocs. Nous avons expérimenté notre système sur la collection de test WT10g. Les résultats obtenus avec la combinaison du contenu de la page et son voisinage par rapport aux blocs thématiques qui la pointent dans la fonction de correspondance montrent de meilleurs résultats par rapport à ceux qui reposent sur le contenu seul ou sur les algorithmes d'analyse des liens (PageRank et HITS).

6. Bibliographie

Adelberg, B., NoDoSE: A tool for semiautomatically extracting structured and semistructured data from text documents, In Proceedings of ACM SIGMOD Conference on Management of Data, 1998, pp. 283-294.

Ashish, N. and Knoblock, C. A., Wrapper Generation for Semi-structured Internet Sources, SIGMOD Record, Vol. 26, No. 4, 1997, pp. 8-15.

Brin S., Page L., «The anatomy of a large-scale hyper textual Web search engine», *In Proceeding of WWW7*, 1998.

Chakrabarti, S., Punera, K., and Subramanyam, M., Accelerated focused crawling through online relevance feedback, In Proceedings of the eleventh international conference on World Wide Web (WWW2002), 2002, pp. 148-159.

Chen, J., Zhou, B., Shi, J., Zhang, H.-J., and Wu, Q., Function-Based Object Model Towards Website Adaptation, In Proceedings of the 10th International World Wide Web Conference, 2001.

Craswell N., Hawking D., « Overview of the TREC 2003 Web Track », *in the 12th TREC*, 2003.

Craswell, N., Hawking, D. « Overview of the TREC 2004 Web Track », *in the 13th TREC*, 2004.

Doan B-L., Chibane I., « Expérimentations sur un modèle de recherche d'information utilisant les liens hypertextes des pages Web », *Revue des Nouvelles Technologies de l'Information (RNTI-E-3), numéro spécial Extraction et Gestion des Connaissances (EGC'2005)*, Vol. 1:245-256, Cépaduès-Editions, January 2005, pp.257-262 .

Embley, D. W., Jiang, Y., and Ng, Y.-K., Record-boundary discovery in Web documents, In Proceedings of the 1999 ACM SIGMOD international conference on Management of data, Philadelphia PA, 1999, pp. 467-478.

Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., and Crespo, A., Extracting Semistructured Information from the Web, In Proceedings of the Workshop on Management for Semistructured Data, 1997, pp. 18-25.

Haveliwala T.H., « Topic-Sensitive Pagerank: A Context-Sensitive Ranking Algorithm for Web Search ». *In Proceedings of the eleventh international conference on World Wide Web*, pages 517-526, ACM Press, 2002.

Hawking D., « Overview of the TREC-9 Web Track », *in the 9th TREC*, 2000.

Kleinberg J., « Authoritative Sources in a Hyperlinked Environment », *Journal of the ACM*, Vol. 46, No. 5, pp. 604-622, 1999.White H.D., McCain K.W., « Bibliometrics », *Annual Review of Information Science and Technology*, 119-186, 1999

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

Mcbryan O., « GENVL and WWW: Tools for Taming the Web », *In Proceedings of the 1st WWW*, 1994.

Qin T., Liu T.Y., Zhang X.D., Chen Z., Ma W.Y., A Study of Relevance Propagation for Web Search**,** *The 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

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

Shakery A., Zhai C.X., « Relevance Propagation for Topic Distillation UIUC TREC 2003 Web Track Experiments », *in the 12th TREC,* 2003.

Song R., Wen J.R., Shi S.M., Xin, G.M., Liu T.Y., Qin T., Zheng X., Zhang J. Y., Xue G. R., Ma W.Y., « Microsoft Research Asia at Web Track and Terabyte Track of TREC 2004 », *in the 13th TREC,* 2004