Propagation de métadonnées par l'analyse des liens

C. Prime-Claverie, M. Beigbeder

Laboratoire RIM - SIMMO
Ecole Nationale Supérieure des Mines de Saint-Etienne
158, cours Fauriel 42023 Saint-Etienne Cedex, FRANCE
Email : {prime,mbeig}@emse.fr
Tél : +33 4 77 42 66 12 Fax :

T. Lafouge

Laboratoire RECODOC
Université Claude Bernard Lyon 1
43, bd du 11 novembre 1918 69622 Villeurbanne Cedex, FRANCE
Email : lafouge@enssib.fr
Tél : +33 4 72 44 58 34 Fax :

Résumé

La Toile apparaît comme une véritable mine d'information et une des difficultés pour ses utilisateurs est de retrouver les documents répondant à leur besoins. Outre le problème de pertinence thématique, les documents rendus par les moteurs ne sont pas toujours en adéquation avec les attentes de l'utilisateur : document trop généraliste, ou contraire d'un niveau élevé, d'un genre différent de celui attendu par l'utilisateur, etc. Nous pensons que l'ajout de métadonnées aux pages pourrait considérablement améliorer la recherche d'information sur la Toile. Dans cet article nous proposons une méthode permettant d'ajouter ces métadonnées de manière semi-automatique. Elle se base sur la propagation des métadonnées dans le graphe de co-citation formé à partir du graphe web.

Abstract

The World Wide Web currently has a huge amount of pages and it is extremely difficult to retrieve documents corresponding to one's informational needs. In addition to the problem of thematic relevance, documents returned by the search engines do not correspond to the user expectations, documents are either too difficult or on the contrary too easy. We realize that the way to clearly improve information retrieval on the Web is to add metadata to web pages (thematic or non-thematic). In this paper, we present a method able to add metadata in a semi-automatic way. It is based on the propagation in the co-citation graph coming from the Web graph.

1 Introduction

Le Web apparaît comme une véritable mine d'information regroupant des ressources très différentes les unes des autres, aussi bien au niveau de leur contenu thématique, que de leur genre, leur langue, leur niveau, etc. Une des difficultés pour ses utilisateurs est de retrouver les ressources pertinentes à leurs besoins. Contrairement aux bases de données documentaires traditionnelles qui sont gérées et organisées par une même autorité, le Web est un espace d'expression libre qui ne connaît aucune organisation. Une des manières d'améliorer l'accès à son contenu serait d'ajouter de manière systématique des méta-informations aux pages web. Bien que prévu par les langages HTML et maintenant XML et malgré tous les efforts de normalisation (Dublin Core [Dublin, 2003]) l'utilisation de métadonnées est encore peu répandue. Ces métadonnées d'auteurs sont d'ailleurs assez mal utilisées, soit par un manque de pratique ou d'objectivité de la part des auteurs honnêtes, soit détournées de leur objectif initial pour permettre une meilleure visibilité par ceux qui les maîtrisent.

Pour que les pages web soient décrites de manière uniforme et systématique, nous pensons que ce sont les systèmes de recherche d'information eux-mêmes qui doivent affecter les méta-informations, de la même manière que ce sont les professionnels de la documentation qui effectuent le catalogage et l'indexation. Précisons que ces opérations documentaires sont effectuées manuellement et sont donc très coûteuses, et étant donné le nombre de pages disponibles sur le Web, leur volatilité, il n'est pas envisageable que les métadonnées soient affectées manuellement. Il faut donc s'orienter vers des méthodes automatiques ou semi-automatiques.

Cet article présente nos recherches en cours concernant la possibilité de propager des métadonnées aux documents web par l'analyse du graphe formé par les liens hypertextuels.

2 La représentation des documents Web

Les principaux outils de recherche d'information (moteurs) disponibles sur la Toile s'appuient sur les techniques des SRI (Système de Recherche d'Information) traditionnels notamment pour la représentation des documents et des requêtes, et pour le calcul des fonctions de correspondance. Rappelons toutefois que les SRI traditionnels travaillent sur des corpus de documents, que l'on appelle aussi collections. Une collection est un ensemble de documents sélectionnés, rassemblés par une même autorité et parfois classés. Les collections constituent donc des ensembles cohérents et homogènes où les documents partagent des propriétés communes (collections d'articles scientifiques, de brevets, etc.). Dans de telles collections la priorité est donc d'appréhender l'apport informationnel de chaque document et c'est pourquoi ceux-ci sont représentés sémantiquement au cœur du SRI par des mots-clés. Or sur le Web, espace hétérogène, il serait dommage de se limiter à une simple représentation thématique des documents comme le font les moteurs et les annuaires. C'est pour cela que nous envisageons d'ajouter aux pages en plus de leur représentation sémantique des métadonnées non thématiques.

3 L'analyse des liens

Actuellement, deux communautés scientifiques s'intéressent de près à l'analyse des liens du Web : les bibliométres et les informaticiens. Les premiers, dont l'un des objectifs est de structurer l'univers du savoir à partir de grands volumes d'information, étudient les équivalences entre les concepts établis en bibliométrie et le graphe du Web [Ingwersen, 1998], [Björneborn and Ingwersen, 2001], [Aguillo, 1999], [Egghe, 2000]. En effet, comme dans le réseau des publications scientifiques [Garfield, 1972] un lien hypertexte peut matérialiser une citation et indiquer une relation intéressante entre la page d'origine et la page vers laquelle il pointe. Les seconds utilisent les méthodes mathématiques de la théorie des graphes dans l'objectif d'améliorer la recherche d'information sur le Web. Parmi les applications les plus connues nous pouvons citer les algorithmes de classement de Google [Brin and Page, 1998], la découverte de communautés d'intérêts [Kumar et al., 1999].

Marchiori [Marchiori, 1998] propose dès 1998, une méthode permettant de propager des métadonnées de classification (thématique) le long des liens. Dans cette méthode, chaque page est décrite par des métadonnées thématiques (mots-clés) pondérées par un coefficient variant entre 0 et 1 (1 lorsque la métadonnée décrit parfaitement la page, 0 lorsqu'elle est inappropriée). Son hypothèse est la suivante : si une page P (décrite par une métadonnée A pondérée par le coefficient ν) est citée par une page P', alors on peut supposer que P sert à expliciter (à appuyer) des idées évoquées dans la page P'. Les métadonnées de P peuvent donc être propagées à P' avec un facteur d'affaiblissement f ($0 < f < 1$). La métadonnée A décrit à présent le document P' avec le coefficient $\nu \times f$.

Nous pensons comme lui que si un document P contient un lien hypertexte vers un document P', il existe (au moins pour le créateur de la page P) une association entre ces deux documents. Celle-ci se traduit par des valeurs identiques pour une ou plusieurs métadonnées (c'est-à-dire que deux pages reliées dans le Web partagent au moins un point commun, même origine géographique, même thème, même niveau...). Cependant nous pensons qu'une analyse plus poussée du graphe, utilisant des relations plus complexes que la simple relation "citant-cité" peut permettre d'extraire des ensembles de pages très homogènes partageant une majorité de métadonnées identiques.

4 Notre méthode

Nous envisageons comme Marchiori de propager des métadonnées en utilisant les relations entre les pages du Web. Notre méthode ne s'appuie pas directement sur le graphe Web, mais sur un graphe obtenu de manière indirecte, le graphe des co-citations (fig. 1). La méthode proposée comporte deux étapes :

- la structuration du corpus par la méthode des co-citations en vue d'obtenir une hiérarchie de sous-corpus que nous supposons homogènes,
- la propagation de métadonnées dans les sous-corpus.

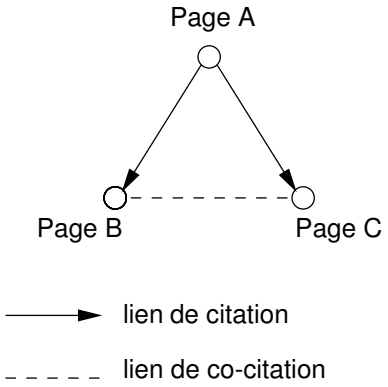

FIG. 1 – Liens de citation et de co-citation sur le Web

4.1 Structuration du corpus par la méthode des co-citations

4.1.1 La méthode des co-citations

La méthode des co-citations, utilisée en bibliométrie depuis 1973 [Marshakova, 1973] [Small, 1973], a pour objectif de créer à partir d'articles scientifiques d'un même domaine de recherche, et plus précisément de leurs références bibliographiques, des cartes relationnelles de documents ou d'auteurs qui reflètent à la fois les liens sociologiques et thématiques de ce domaine. Cette méthode repose sur l'hypothèse que deux références bibliographiques de date quelconque, fréquemment citées ensemble ont une parité thématique. Le lien hypertexte lui aussi peut matérialiser une citation, et plusieurs auteurs [Larson, 1996] [Pitkow and Pirolli, 1997] [Prime et al., 2002a] se sont intéressés à la transposition de la méthode des co-citations de documents pour caractériser les univers du Web. Ils mettent en évidence les limites théoriques et techniques de l'analogie, mais ont montré l'intérêt de la structuration pour rapprocher thématiquement les pages. Une des limites de cette analogie est de considérer tous les liens hypertextes comme des liens de citation ou de référence. En effet, il faut aussi prendre en compte les liens de publicité, mais surtout ceux qui servent à se déplacer dans un même site web : les liens de navigation interne. C'est pourquoi notre méthode ne tient compte que des liens inter-serveurs entre les pages citantes et citées espérant ainsi supprimer la majorité des liens de navigation.

La première phase de la méthode consiste à déterminer la proximité des pages entre elles. Pour cela on définit un indice de similarité qui doit traduire mathématiquement l'idée suivante : deux pages P_i et P_j sont proches, si par rapport à leurs fréquences de citation respectives (C_i et C_j), leur fréquence de co-citation (C_{ij}) est importante. Il existe plusieurs indices possibles qui par convention varient de 0 à 1 : 1 lorsque les pages sont toujours citées ensemble, et 0 lorsque celles-ci ne le sont jamais. L'indice que nous utilisons est l'équivalence :

$$E_{ij} = \frac{{C_{ij}}^2}{C_i \times C_j}. \quad (1)$$

Dans la suite de l'article nous utiliserons la distance d_{ij} entre les pages, plutôt que leur proximité avec $d_{ij} = 1 - E_{ij}$. Toutes les valeurs d_{ij} sont inscrites dans une matrice de co-citations à partir de laquelle on peut construire le graphe de co-citations, graphe valué où les nœuds sont les pages et les arcs les liens de co-citations entre les pages valués.

La seconde phase, le regroupement des pages les plus proches, utilise des méthodes de classification automatique issues de l'analyse de données. Nous utilisons une classification hiérarchique ascendante. Plusieurs choix sont possibles : le simple lien (voisin le plus proche), le lien complet (voisin le plus éloigné), le chaînage moyen. Cette classification permet de créer une hiérarchie de classes (agrégats de pages). Les documents les plus similaires sont regroupés dans des classes au plus bas niveau, tandis qu'au plus haut niveau les documents sont tous regroupés ensemble. La hiérarchie obtenue peut être visualisé graphiquement par un dendrogramme (fig. 2).

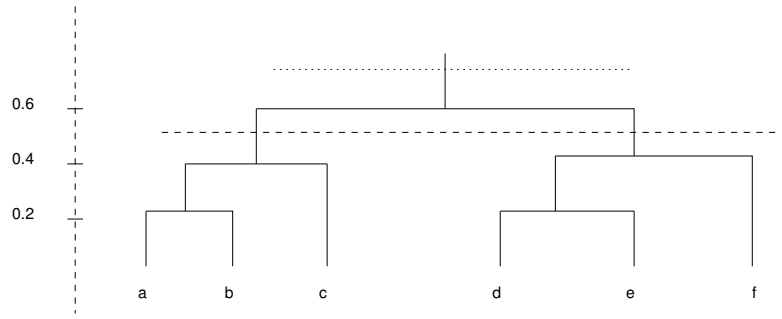

FIG. 2 – Exemple d'un dendrogramme

Une des difficultés de la méthode consiste à déterminer le niveau de coupure du dendrogramme qui donnera à la fois des classes de taille importante et les plus homogènes possible.

4.1.2 Expérience et résultats

Nous avons mené en 2001 une expérience sur un corpus contenant des pages relatives au thème de l'astronomie [Prime et al., 2002b]. Nous avons classé 198 pages par la méthode des co-citations que nous avons indexées à la main pour des métadonnées liées au genre (type) de document. Nous nous sommes intéressés à l'homogénéité des classes obtenues par la méthode du lien complet. Les résultats observés ont été très encourageants.

4.2 Propagation par l'analyse des liens

La méthode de propagation que nous proposons repose sur l'hypothèse que deux pages proches par l'indice de co-citation partagent des métadonnées communes. Elle permet de propager la (ou les) valeur(s) d'une (ou de plusieurs) métadonnée(s). Contrairement à Marchiori, il n'est pas nécessaire d'utiliser des métadonnées pondérées car notre méthode n'influe pas sur les pondérations. Elle s'appuie à la fois sur le dendrogramme obtenu par la classification et le graphe des co-citations. En effet, l'expérience menée montre que les classes ne se forment pas toutes "à la même vitesse". Certaines sont déjà de taille importante et bien homogènes à un seuil relativement bas dans le dendrogramme, alors qu'à ce même seuil persistent encore beaucoup de singletons. A un seuil plus élevé, certains singletons ont pu se regrouper ou rejoindre d'autres classes pour former des ensembles homogènes, tandis que d'autres classes qui étaient homogènes se sont "bruitées". C'est pourquoi nous trouvons dommage de ne travailler qu'à un seul niveau de coupure du dendrogramme et de ne pas utiliser toute la richesse de la hiérarchie.

Nous définissons un seuil S à partir duquel nous supposons que les classes sont déjà de taille importante et que celles-ci ne sont pas encore trop bruitées. Ce seuil dépend de la méthode d'agrégation choisie, plus la distance inter-classe est exigeante, plus le seuil S pourra être élevé. Au niveau de coupure S, nous obtenons une partition du corpus de départ. Chaque classe induit un sous-graphe du graphe de co-citations (figure 3). Pour chacune d'elles nous identifions le couple d'éléments les plus éloignés par la distance dans un graphe valué[1].

La figure 3 montre le graphe qui a permis de générer le dendrogramme de la figure 2 avec la méthode du simple lien. Pour un seuil supérieur à 0,6, nous obtenons une classe à 6 éléments. Les deux éléments les plus éloignés sont b et f (distant de $1,6$). Au seuil 0,6 cette classe sera divisée en deux créant ainsi deux sous-graphes à trois éléments. Les éléments b et f sont indexés pour les trois métadonnées "type de site", "type d'autorité" et "thème" avec les valeurs respectives *site de ressources*, *entreprise*, *astronomie* et *site de ressources*, *association* et *astronomie*.

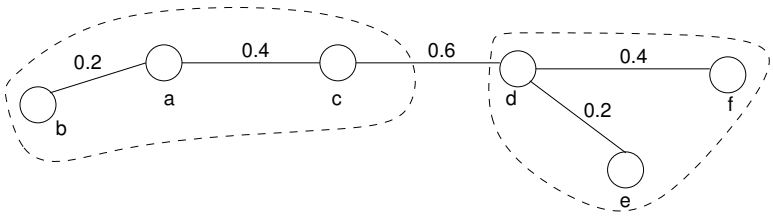

FIG. 3 – Visualisation du graphe de la classe de la figure 2 pour un seuil supérieur à 0,6

Pour chaque classe, nous examinons les valeurs de métadonnées du couple d'éléments les plus distants. Lorsqu'un de ces éléments n'est pas indexé, ce qui est le cas au départ, nous le faisons manuellement. Ces valeurs sont comparées. Le principe de notre méthode est de propager les valeurs de métadonnées du couple lorsqu'elles sont identiques, aux autres éléments de la classe. En effet ceux-ci ont une forte probabilité de partager les mêmes valeurs, puisqu'ils sont plus proches les uns des autres. Lorsque l'on travaille avec plusieurs métadonnées le couple peut partager les mêmes valeurs pour certaines métadonnées et avoir des valeurs différentes pour les autres. Dans l'exemple ci-dessus, les éléments b et f partagent les deux valeurs *site de ressources* et *astronomie*, alors que les valeurs de la métadonnée "type d'autorité" diffèrent. Deux choix sont possibles :

- propager les valeurs de métadonnée de manière individuelle. Dès que le couple partage une valeur commune de métadonnée, celle-ci est propagée aux autres éléments. Pour ce choix et dans l'exemple ci-dessus, les valeurs *astronomie* et *site de ressources* sont propagées aux éléments a, c, d, et e.
- propager les valeurs de métadonnées en groupe, c'est-à-dire, exiger qu'une partie ou la totalité des valeurs de métadonnées du couple soient identiques pour les propager aux autres éléments de la classe. Dans l'exemple ci-dessus, les valeurs de la métadonnée "type d'autorité" sont différentes. Si l'on exige que toutes les valeurs de métadonnées de couple soient identiques 2 à 2, alors aucune valeur n'est propagée, et la classe est divisée en deux au seuil 0,6.

[1]La distance $d(x, y)$ entre 2 sommets x et y est la longueur du plus court chemin entre x et y. Dans un graphe valué, c'est la somme des valuations des arêtes de ce chemin.

Tant que les valeurs de métadonnées sont différentes, nous "descendons" dans le dendrogramme au seuil qui partionnera cette classe et recommençons l'opération espérant ainsi intervenir le moins possible manuellement et propager le plus de valeur de métadonnées automatiquement.

5 Limites et perspectives

La méthode présentée ci-dessus est en cours de test sur le corpus utilisé dans l'expérience antérieure pour les 3 métadonnées "type de site", "type d'autorité", "type d'information". Nos premiers résultats avec une méthode qui propage les valeurs de métadonnées en groupe, nous donnent une bonne qualité de propagation (peu d'erreurs) mais une très faible rentabilité : l'indexation manuelle est trop importante par rapport à l'indexation automatique par propagation.
D'autre part, nous savons qu'une des limites de cette méthode est le faible taux de pages indexées. En effet, sur le Web de nombreuses pages ne sont pas citées par des pages hébergées sur d'autres sites, si bien qu'elle ne peuvent être *a fortiori* co-citées et classées. Il faudra donc envisager une méthode d'indexation et de propagation de métadonnées au sein des sites Web. Notons aussi que notre indice de similarité (l'équivalence) ne dépend pas du nombre de liens émis par les pages citantes. Or sur le Web le nombre de liens émis par chaque page est extrêmement variable, et il serait judicieux d'en tenir compte pour calculer la proximité entre les pages. Actuellement, nous commençons une expérience de plus grande envergure sur un corpus contenant 5 millions de pages qui correspond au Web francophone de décembre 2000 (collecté par M. Géry et D. Vaufreydaz du laboratoire CLIPS `http://www-clips.imag.fr`), pour identifier clairement le pourcentage de pages co-citées et le nombre de pages pouvant être ainsi classées.

Références

[Aguillo, 1999] Aguillo, I. (1999). Statistical indicators on the internet : The european science technology industry system in the world-wide web. *at http://diotima.math.upatras.gr/weborg/aguillo2.*

[Björneborn and Ingwersen, 2001] Björneborn, L. and Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1) :65–82.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International WWW Conference*. IW3C2.

[Dublin, 2003] Dublin (2003). Dublin core metadata initiative (dcmi). *at http://dublincore.org consulté en février 2003.*

[Egghe, 2000] Egghe, L. (2000). New informetric aspects of the internet : some reflections, many problems. *Journal of information science*, 26(5) :329–335.

[Garfield, 1972] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, (178) :471–479.

[Ingwersen, 1998] Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2) :236–243.

[Kumar et al., 1999] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. In *Proceedings of the Eighth World Wide Web Conference.*

[Larson, 1996] Larson, R. (1996). Bibliometrics of the world wide web : An exploratory analysis of the intellectual structure of the cyberspace. In *Proceedings of the Annual Meeting of the American Society of Information Science*, Baltimore.

[Marchiori, 1998] Marchiori, M. (1998). The limits of web metadata and beyond. In *Proceedings of the Seventh International WWW Conference*. IW3C2.

[Marshakova, 1973] Marshakova, I. V. (1973). Document coupling system based on references taken from science citation index. *Russia, Nauchno - Teknicheskaya Informatsiya*, 2(6,3).

[Pitkow and Pirolli, 1997] Pitkow, J. and Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97*, pages 118–125.

[Prime et al., 2002a] Prime, C., Bassecoulard, E., and Zitt, M. (2002a). Co-citations and co-sitations : a cautionary view on an analogy. *Scientometrics*, 54(2) :291–308.

[Prime et al., 2002b] Prime, C., Beigbeder, M., and Lafouge, T. (2002b). Clusterisation du web en vue d'extraction de corpus homogènes. In *Actes du 20ème congrès INFORSID*, pages 229–242, Nantes.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science*, 24 :265–269.