Règles de propagation pour la création d'ontologies d'annotation de ressources

Lylia Abrouk, Pierre Pompidor
Danièle Hérin, Michel Sala

LIRMM
161, rue ada
34392, Montpellier
{abrouk,pompidor,dh,sala}@lirmm.fr,
http ://www.lirmm.fr/~{abrouk,pompidor,dh,sala}

Résumé. L'annotation se distingue de l'indexation automatique par l'utilisation d'une ou plusieurs ontologies qui définissent un domaine global de référence permettant de cadrer et de normaliser les annotations effectuées, par ailleurs une ressource annotée doit l'être non pas par une liste de mots-clefs, mais bien par une ou plusieurs ontologies.
Malheureusement, il est peu réaliste de penser que les centaines de millions de ressources mises à disposition sur le Web puissent être annotées par leurs auteurs. Pour résoudre ce problème, notre démarche consiste à indexer les documents en se basant sur l'ontologie globale et ensuite propager les annotations en utilisant des documents déjà annotés pour annoter d'autres documents référencés par ceux-ci.
La propagation des annotations suit des règles que nous proposons dans cet article. L'illustration est effectuée sur un corpus de livres dont le thème relève de l'informatique.

1 Introduction

Le Web est devenu dans tous les domaines la plus grande source d'informations, engendrant lors de la recherche d'informations de grandes difficultés pour retrouver les ressources les plus pertinentes, celles-ci provenant d'ailleurs de sources hétérogènes et parfois sécurisées (et dans ce cas accessibles seulement par un fournisseur donné). Une solution est d'annoter les ressources pour décrire non seulement leur contexte de création (noms des auteurs, date de parution, etc.), ce qui ne nous intéresse pas ici, mais également la sémantique de leur contenu. Dans cet article nous nous intéressons uniquement à ce deuxième point.

L'annotation est caractérisée par l'utilisation d'une ou plusieurs ontologies qui définissent un domaine global de référence permettant de cadrer et de normaliser les annotations effectuées. Nous pouvons par exemple citer DMOZ[1] , une ontologie universelle utilisée par certains moteurs de recherche comme Google[2] . Naturellement, l'annotation est d'autant meilleure qu'elle privilégie une intervention humaine, une annotation

[1] www.dmoz.com
[2] www.google.com

excessivement automatisée pouvant engendrer beaucoup de bruit par la sélection de mots-clefs non pertinents. Par ailleurs une ressource annotée doit l'être non pas par une liste de mots-clefs, mais bien par une ou plusieurs ontologies, ces dernières sont une liste de concepts reliés par des relations. L'ontologie globale définit un domaine donné et les ontologies locales sont spécifiques à chaque ressource et sont un sous ensemble de l'ontologie globale. En effet cela permettra non seulement de donner une vue plus ou moins générale sur le contenu de la ressource (les ontologies offrant différents niveaux d'abstraction), mais également de définir le champ d'utilisation d'un concept qui pourrait être ambigüe autrement.

Il est peu réaliste de penser que les centaines de millions de ressources mises à disposition sur le Web puissent être annotées par leurs auteurs, ou par des lecteurs autorisés, même si cette annotation était guidée par des plates-formes de gestions d'ontologies conviviales (Protégé[3]ou Kaon[4]...) et reposant sur des ontologies de domaines plus ou moins exhaustives. Le travail d'annotation étant déjà un travail fastidieux et complexe pour des professionnels tels que les documentalistes, il devient illusoire pour n'importe quel internaute voulant mettre à disposition de manière efficace ses ressources. Notre contribution est donc de proposer des règles de propagations permettant d'annoter une ressource à partir d'annotations d'autres ressources que celle-ci référence.

Nous nous intéressons à cette problématique d'annotation dans le contexte du projet européen SEMIDE (Système euro méditerranéen d'information sur les savoir-faire dans le domaine de l'eau) Ce système d'informations distribué est basé sur le Web. La démarche décentralisée du SEMIDE a été plébiscité au niveau international (Banque Mondiale, Forum Mondial de l'Eau) comme un bon modèle de coopération et de développement. Aujourd'hui, le SEMIDE axe ces efforts sur l'aide aux pays partenaires méditerranéens pour la mise en oeuvre de Systèmes Nationaux d'Information sur l'eau (SNIE) interopérables. L'objectif est l'intégration transparente dans le système régional SEMIDE de services nationaux. A plus long terme, cette architecture pourrait être appliquée à une échelle géographique plus large.

La recherche d'informations dans ce contexte nécessite une bonne représentation ou annotation des documents. Notre démarche se compose de trois parties, on détaillera la deuxième partie :

- L'indexation : qui sera basée sur une ontologie globale
- la propagation des annotations qui utilise des documents déjà annotés pour annoter d'autres documents qui référencent ces documents déjà annotés,
- La révision des annotations

Disposer d'une masse critique de ressources bien annotées "ontologiquement" (et manuellement), qui recèlent un inter référencement suffisant, ne nous donne pas la garantie que l'annotation réalisée soit suffisante et adéquate, mais plutôt qu'elle épargne une première phase d'annotation laborieuse. Un point fort de notre approche est que les ontologies d'annotations sont révisées de manière cyclique avant de trouver leur point de stabilité.

Cet article se compose de deux parties : dans la première partie, nous définissons les

[3]protege.stanford.edu
[4]http ://kaon.semanticweb.org/

termes d'ontologie, de métadonnées et d'annotations tels que nous les utilisons, dans la deuxième partie, nous décrivons l'approche de propagation. Ensuite nous concluons par quelques perspectives.

2 Ontologies, Métadonnées et Annotations

2.1 Définition des termes utilisés

2.1.1 Ontologie

Les ontologies (Grefenstette 1995) sont au cœur des travaux menés en ingénierie des connaissances, elles visent à établir des représentations à travers lesquelles les machines puissent manipuler la sémantique des informations. La construction d'une ontologie demande à la fois une étude des connaissances du domaine et le choix d'un langage de représentation.

Ontologie (Charlet et al. 2003) : Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts- e.g. entités, attributs, processus-, leurs définitions et leurs interrelations.

Une ontologie est une spécification rendant partiellement compte d'une conceptualisation et incluant un vocabulaire de termes et leurs relations.

2.1.2 Métadonnées annotations

Le terme de métadonnées (Lupovic 1999) désigne les données incluses dans des fichiers informatiques pour fournir les informations sur des ressources électroniques. En terme de documentation, ce sont des informations secondaires apposées à des ressources primaires.

Selon la définition du W3C , le terme annotation est tout commentaire, note, explication ou remarque pouvant être lié à un document web.

2.2 Représentation de l'ontologie

2.2.1 Définition

Un arbre d'ontologie est un arbre orienté depuis la racine dont les noeuds sont les concepts et les arêtes les relations de spécialisation entre ces concepts. De plus à chaque noeud C est associé un ensemble I_C éventuellement vide d'instances vérifiant la propriété suivante : si C_1 est un concept ancêtre de C_2 dans l'arbre d'ontologie alors $I_{C_2} \subseteq I_{C_1}$.

Une annotation d'un document est un arbre d'ontologie ou bien un ensemble d'instances.

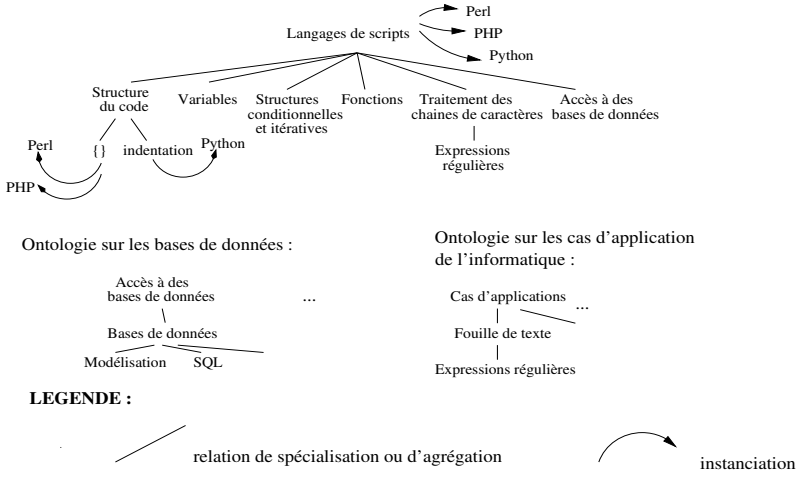

FIG. 1 – *Exemple d'ontologie globale*

2.2.2 Exemple

Fig. 1 représente l'ontologie globale du domaine. L'ontologie globale (OG) est composée de plusieurs ontologies qui représentent des sous domaines. A la racine de l'arbre"langages de script" et sur deux feuilles , , trois instances de langages de scripts ont été associées : "Perl", "PHP" et "Python".(Nous illustrons nos propositions sur un corpus de sept livres électronique de références sur les langages de script).

2.3 Sources d'annotations

On annote les documents de deux manières ; manuellement ou semi automatiquement en se basant sur une ontologie globale.

Notre travail se divise en trois partie : l'indexation des documents, la propagation des annotations et enfin la révision, notre papier traite essentiellement de la deuxième problématique, le résultat du travail de propagation est lui même une indexation des documents. Dans ce qui suit nous développons un état de l'art des travaux existants avant de présenter notre approche.

3 La propagation des annotations

La propagation consiste à reporter tout ou une partie de l'ontologie locale d'annotation de la ressource référence (source) sur la ressource référencée (ressource cible). Selon que la ressource cible ait déjà été annotée ou non, les règles de propagation de l'ontologie de la source se divisent en deux groupes :

– les règles qui régissent un report simple de l'ontologie sur la cible non annotée (hormis par des instanciations éventuelles)
– les règles qui régissent une jointure de l'ontologie sur la cible déjà annotée

3.1 Etat de l'art

Nous présentons brièvement dans ce qui suit quelques travaux dans le domaine de la propagation

3.1.1 Analyse des liens (Marchiori 1998)

Si une ressource R du Web a des métadonnées (mots clés) associées A.v, indiquant que le mot clé A a un poids v et s'il existe une Ressource R' dans le Web avec un hyperlien vers R, alors les métadonnées de R sont propagées à R'. L'idée est que l'information contenue dans R sont accessibles par R', étant donné qu'il existe un lien.

La pertinence de R' en ce qui concerne le mot clé A n'est pas identique à R(v), puisque l'information de R est seulement potentiellement accessible de R', mais pas directement contenu dans R'. La solution à ce problème est de baisser la valeur v de l'attribut en multipliant par "un facteur d'affaiblissement" f . Ainsi, dans l'exemple ci-dessus R' peut avoir sa liste de mots clés avec A :v·f. Le même raisonnement est alors appliqué périodiquement. Ainsi, si nous avons une autre ressource R" avec un lien vers R' nous pouvons propager les métadonnées A :v·f exactement de la même manière ; l'ensemble de métadonnées sera A :v·f·f.

La propagation des métadonnées dans notre solution s'appuie partiellement sur cette idée, mais nous considérons en plus des hyperliens, les liens représentés par des références bibliographiques.

3.1.2 La propagation de popularité : PageRank

Cette technique (Arasu et al. 2001) s'appuie sur l'hypothèse qu'une page référencée par un grand nombre de pages est une bonne page. Elle est utilisée par le moteur de recherche Google qui s'appuie sur l'algorithme *PageRank*. Le *PageRank* (PR)d'un document B dépend de 3 facteurs :
– le nombre de pages faisant un lien vers B,
– le PageRank (PR) de chaque page,
– le nombre de liens sortants de chaque page

le PageRank d'une page A est définit comme suit

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + PR(T_n)/C(T_n))$$

– une page A reçoit des liens émis par les pages $T_1 ... T_n$.
– Le paramètre d est un facteur d'amortissement pouvant être ajusté entre 0 et 1.
– C(T) est le nombre de liens émis par la page A (liens sortants)

3.1.3 La méthode de co-citation

La méthode de co-citation (Garfield 1993), utilisée en bibliométrie depuis 1973 a pour but de créer, à partir d'articles scientifiques d'un même domaine de recherche et en utilisant leurs références bibliographiques, des relations entre ces articles. Cette méthode repose sur l'hypothèse que deux références bibliographiques de date quelconque, fréquemment citées ensemble ont une parité thématique. Un lien hypertexte dans une publication peut constituer une citation et être une relation intéressante entre la page d'origine et la page destination. La limite de cette méthode est que le lien peut être complètement hors contexte comme une publicité ou un lien vers un site.

Cette méthode consiste à calculer, pour chaque couple de pages leur fréquence de co-citation C_ij (le nombre de fois où les pages ont été citées ensemble par d'autres pages). L'indice de similarité est l'indice d'équivalence

$$E_{ij} = \frac{C_{ij}^2}{C_i * C_j}$$

où C_{ij} est la fréquence de cocitation des pages i et j
et C_i la fréquence de citation de la page i

3.1.4 Propagation des métadonnées par l'analyse des liens

Les travaux (Prime-Claverie et al. 2003) ont été effectués dans la propagation de métadonnées en se basant sur les liens. Elle repose sur l'hypothèse que deux pages proches par l'indice de co-citation partagent des métadonnées communes. Cette méthode s'appuie sur le graphe des co-citations, elle comporte deux étapes :

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

4 Notre approche

Nous proposons de traiter la propagation des annotations en se basant sur les références ou les ressources citées dans un document.

Trois problèmes importants se posent, i. La partie du document couvert par les annotations issues de la ressource référencée, ii. Synthétiser les différentes annotations des différentes parties pour en avoir une vision générale, iii. maîtriser l'impact de la propagation qui peut très rapidement s'avérer exponentiel et incohérente du fait que la propagation peut remettre en cause plusieurs fois et successivement les métadonnées d'un document.

Validation des noeuds des ontologies locales

Lors de l'étape de propagation, des concepts peuvent apparaître dans l'ontologie locale du document à annoter mais ne qui ne sont pas validés, la validation entraîne

l'ajout du concept dans l'ontologie. Un concept doit être validé pour être ajouté dans l'ontologie du document à annoter.

Nous avons défini cinq règles de propagation que nous illustrons sur un corpus de sept livres

4.1 Les règles

Etant donnés deux documents D_1 et D_2, il y a quatre cas de référencement :
- Le document D_2 référence le document D_1 dans son ensemble.
- Le document D_2 référence une partie (un chapitre par exemple) du document D_1.
- Une partie du document D_2 référence le document D_1.
- Une partie du document D_2 référence une partie du document D_1.

Nous avons retenu dans notre exemple le chapitre comme unité de découpage (partie=chapitre)

Voici la liste des règles de propagation

4.2 Définition des règles de propagation

Les règles 1 à 3 décrivent la propagation d'annotations lorsqu'un document source S référence un document cible C. La règle 4 traite le cas où une partie d'un document source référence un document cible, la règle 5 traite le cas où une partie d'un document source référence une partie d'un document cible. On annote A_s l'annotation du document source et A_c l'annotation du document cible.

4.2.1 Règle 1

prérequis :
A_s est un arbre d'ontologie et A_c est un ensemble d'instances.
résultat :
Soit T_s l'arbre d'ontologie de A_s, r la racine de T_s et I_r l'ensemble d'instances associé à r. Soit I l'ensemble d'instances de A_c.
A_c un arbre d'ontologies réduit à un unique noeud r et dont l'ensemble d'instances associé est $I \cap I_r$

4.2.2 Règle 2

prérequis :
A_s est un arbre d'ontologie et A_c est un ensemble d'instances.
résultat :
Soit T_s l'arbre d'ontologie de A_s, et I l'ensemble d'instances de A_c.
A_c est remplacé par le sous arbre T_c de T_s formé des noeuds p dont l'ensemble d'instances associé I_p contient au moins un élément de I. Pour chaque noeud p de T_c l'ensemble d'instances qui lui est associé dans A_c est $I_p \cap I$.

4.2.3 Règle 3

prérequis :
A_s et A_c sont deux arbres d'ontologies T_s et T_c de racines respectives r_s et r_c et il existe un unique chemin entre r_s et r_c dans l'ontologie globale.
résultat :
Soit P_{sc} le chemin entre r_s et r_c dans l'ontologie globale. T_c est complétée par l'ajout de la trace du chemin P_{sc} sur l'ensemble des noeuds de T_s et T_c. C'est à dire que les noeuds de P_{sc} ne se trouvant ni dans T_s ni dans T_c ne sont pas ajoutés à T_c.

4.2.4 Règle 4

prérequis :
A_s et A_c sont deux arbres d'ontologies T_s et T_c de racines respectives r_s et r_c et il existe un unique chemin entre r_s et r_c dans l'ontologie globale.
résultat :
Soit P_{sc} le chemin entre r_s et r_c dans l'ontologie globale. T_c est complétée par le chemin P_{sc}. Les noeuds du chemin P_{sc} n'appartenant ni à T_c ni à T_s sont aussi ajoutés à T_c.

4.2.5 Règle 5

prérequis :
A_s et A_c sont deux arbres d'ontologies T_s et T_c de racines respectives r_s et r_c. Il existe un unique noeud q de T_c tel qu'il existe un chemin P_{cq} de r_c vers q. De plus ce chemin est unique.
résultat :
T_c est complétée par le chemin P_{cq}. Les noeuds du chemin P_{cq} n'appartenant ni à T_c ni à T_s sont aussi ajoutés à T_c.
remarque :
Lorsque $q \neq r_c$ cette règle conduit à un multi héritage.

Chaque règle est appliquée sur les annotations d'un document source (document de référence) et instancie des métadonnées d'un document cible (document référencé).

4.3 Application des règles de propagation

Notre application se base sur un corpus de sept livres qui s'inter-réferencent.

1. Langages de scripts sous unix (Blaess 2004)
2. Programmation en Perl (Wall et al. 2001)
3. Programmation en Python (Lutz 2002)
4. Maîtrise des expressions réguliéres (Jeffrey et Friedl. 2003)
5. Pratique de php et MySql (Rigaux 2003)
6. Perl en action (Christiansen et Torkington 1999)
7. Python cookbook (Ascher et Martelli 2002)

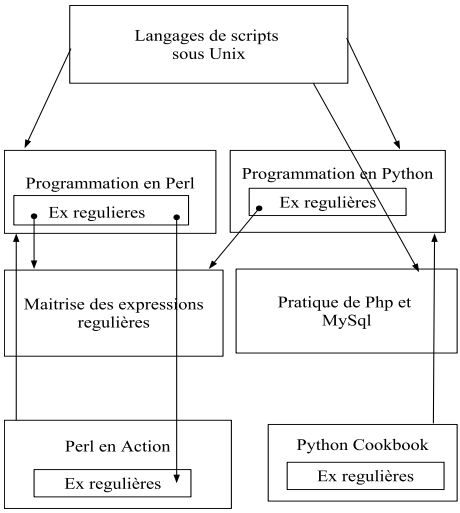

FIG. 2 – *L'inter-référencement des ressources*

Pour la commodité de l'illustration, nous avons en partie aménagé les références présentes dans ces livres suivant le schéma ci-dessous :

L'existant (avant propagation)

Les figures suivantes décrivent les ontologies locales associées aux sept ressources, chaque document en référençant d'autres

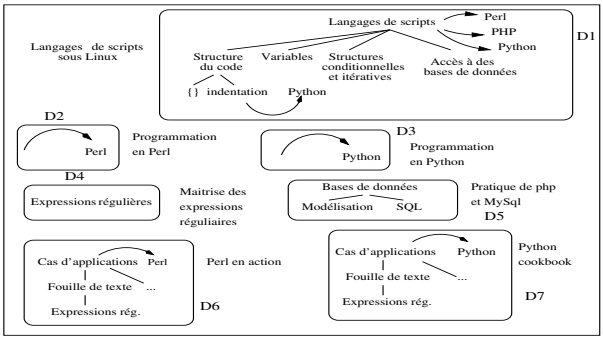

FIG. 3 – *Les ontologies locales associées aux sept ressources*

4.4 Synthèse du premier cycle de propagation

La figure suivante illustre toutes les règles définies sur notre exemple.

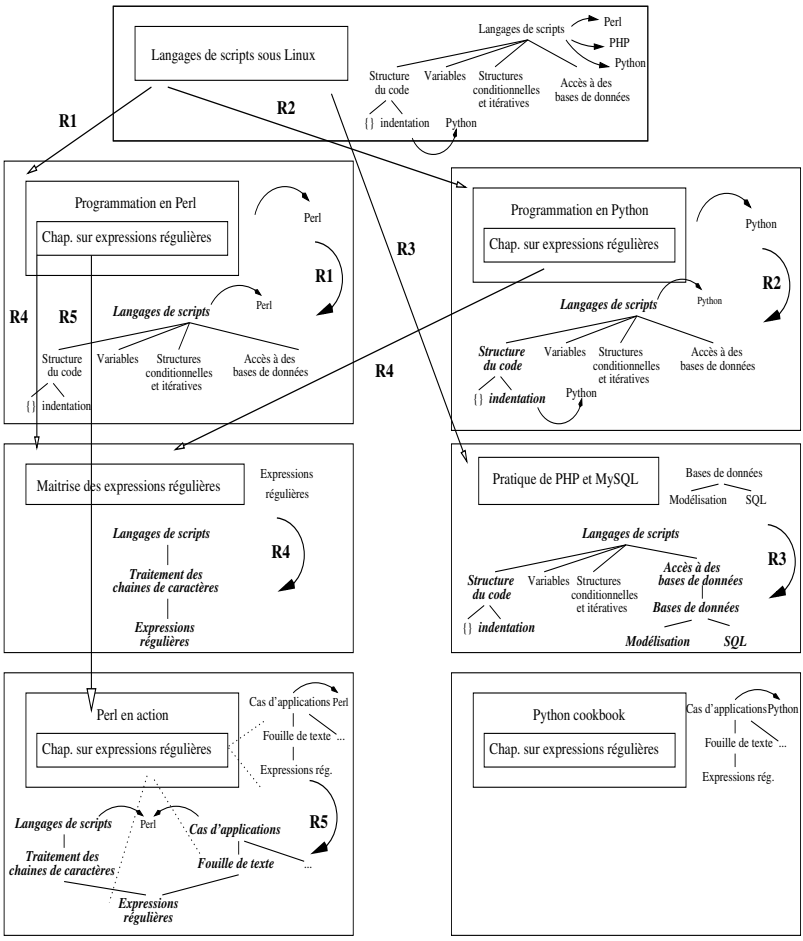

FIG. 4 – *Synthèse du premier cycle de propagation*

5 Conclusion et perspectives

La mise en pratique des règles de propagation sur un corpus significatif de ressources est en cours d'expérimentation sur des documentations informatiques accessibles en ligne et prochainement sur le fond documentaire du Semide (Système euro méditerranéen d'information sur les savoir-faire dans le domaine de l'eau). Bien que nos premières expérimentations démontrent que la propagation des annotations est satisfaisante, les prochaines étapes de notre travail consisteront à mettre en place la pondération des ontologies locales et la révision de l'ontologie globales (ou des ontologies globales) dans le but de l'améliorer.

En effet, la propagation des annotations comporte une part d'incertitude liée à la portée des concepts d'annotations dans les documents annotés. En effet un terme/concept peut avoir une portée dépassant celui du chapitre où il est référencé, tandis qu'au contraire il ne peut concerner qu'un paragraphe de celui-ci. La situation est encore plus critique pour les concepts issus d'ontologies d'annotations concernant des documents référencés en introduction ou en conclusion. Une solution consiste à associer une pondération à chaque concept, en fonction des retours de satisfaction des utilisateurs au fur et à mesure des propagations des ressources (cette pondération pourrait par exemple être calculée en fonction d'éléments tels que la non soumission de nouvelles requêtes de recherche de document). Ces pondérations seraient affaiblies au fur et à mesure des propagations.

Par ailleurs, l'évolution de (ou des) ontologies sous-jacente(s) à l'annotation s'avère nécessaire dans deux cas :

- lorsque les producteurs ont du mal à annoter les documents uniquement avec les ontologies importées, proposeraient à cet effet de nouveaux mots clefs (dans la phase d'annotation).
- lors de l'utilisation dans les requêtes de mots clefs fréquents non référencés (et donc là lors de la phase d'interrogation)

L'intégration de nouveaux mots clefs correspondants à de nouveaux concepts (ou à leurs spécialisations) serait semi-automatique, dans la mesure où si nous pouvons avoir, en analysant les requêtes, le contexte général d'application du mot clef, son insertion précise dans l'ontologie requiert l'intervention de l'expert. Là encore, la création ou la réutilisation d'un gestionnaire graphique d'ontologies serait nécessaire, soit pour permettre à l'expert de rattacher librement les nouveaux mots clefs (de préférence hiérarchisés), à une branche de l'ontologie globale, ou pour lui signaler les points de connexion avec celle-ci dans le cas ou plusieurs mots clefs y apparaissent déjà.

En définitive, ces améliorations ne peuvent avoir lieu qu'avec la participation, directe ou indirecte de tous les acteurs du système (producteurs et utilisateurs), motivés par une meilleure efficacité du processus global.

Références

Arasu A., Novak J., Tomkins A. et Tomlin J. (2001), Pagerank computation and the structure of the web : Experiments and algorithms, 2001.

Garfield E. (1993). Co-citation analysis of the scientific literature : Henry small on mapping the collective mind of science. Essays of an Information Scientist : Of Nobel Class, Women in Science, Citation Classics and Other Essays,15(19), 1993.

Grefenstette G. (1995). Comparing two language identification schemes. In Proc. of Analisi Statistica dei Dati Testuali (JADT), pages 263-268, 1995.

Charlet J., Laublet P. et Reynaud C. (2003). Action spécifique 32 web sémantique rapport final. Technical report, CNRS/STIC, Octobre 2003.

Lupovic C. (1999) . Identification des ressources sur internet et métadonnées : diversité des standards. Documentaliste : sciences de l'information, 36(6), 1999.

Marchiori M. (1998) . The limits of web metadata, and beyond. In Proceedings of the Seventh International World Wide Web Conference, pages 1-9, Australia,1998.

Prime-Claverie C., Beigbeder M. et Lafouge T.(2003). Propagation de métadonnées par l'analyse des liens. In Journees Francophones de la Toile - JFT2003, FRANCE, juillet 2003.

Blaess C.(2004),Scripts sous Linux ,Eyrolles ,2004.

Wall L., Christiansen T. et Orwant J.(2001), Programmation en Perl, O'Reilly, 2001.

Lutz M.(2002), Python, O'Reilly, 2002.

Jeffrey E. et Friedl F. (2003), Maîtrise des expressions régulières, O'Reilly, 2003.

Rigaux P.(2003), Pratique de MySQL et PHP, O'Reilly, 2003.

Christiansen T. et Torkington N.(1999), Perl en action, O'Reilly, 1999.

Ascher D. et Martelli A.(2002), Python cookbook, O'Reilly, 2002.

Summary

In this paper, we define the rules that should be followed by the annotations propagation. The annotation is distinguished from automatic indexation by the use of one or several ontologies which define a total field of reference allowing standardizing the annotations carried out. Moreover, an annotated resource is not a list of keywords, but is more likely one or more ontologies.
Unfortunately, it is not very realistic to think that the hundreds of millions of resources available on the Web can be annotated by their authors. To solve this problem, our first step consists in indexing the documents while being based on a global ontology, and then propagating the annotations by using documents already annotated with he aim to annotate other documents to whose are referred, and finally, the last step is to revise -in a cyclic way -ontologies annotations. An illustration is carried out on a data-processing corpus of books.