Annotation sémantique de documents basée sur la relation de citation

Lylia Abrouk, Abdelkader Gouaich, Danièle Hérin

Annotation sémantique de documents basée sur la relation de citation

Lylia Abrouk, Abdelkader Gouaïch
Danièle Hérin

LIRMM, Laboratoire d'Informatique, de Robotique
et de Microlectronique de Montpellier
161 rue Ada, 34392 Montpellier Cedex 5
{abrouk, gouaich,dh}@lirmm.fr
http://www.lirmm.fr

Résumé. L'annotation sémantique est le cœur du principe du Web sémantique. Effectivement, l'annotation des ressources par des métadonnées rend possible l'exploitation des ressources par les utilisateurs et les agents logiciels, même en étant dans un contexte distribué et de taille importante comme le Web. Cependant, l'annotation ne devrait pas être considérée comme garantie, particulièrement avec le nombre croissant des ressources où il est quasiment impossible d'effectuer une annotation manuelle. Nous proposons dans cet article, une approche semi-automatique d'annotation de ressources basée sur le contexte de citation et la propagation des annotations entre les ressources. L'approche présentée dans cet article a été expérimentée et évaluée sur la base de documents scientifiques Citeseer. L'évaluation montre l'avantage d'utiliser, non seulement le contenu des ressources pour l'annotation, mais également les liens de citation comme contexte pour décrire les thèmes des ressources.

1 Introduction

Le Web Sémantique, comme vision, entend fournir un médium d'échange d'information et de savoir en partageant des ressources. Afin de faciliter le partage de ces ressources, il est impératif de pouvoir les décrire avec des données additionnelles, appelées des métadonnées. Ces métadonnées seront par la suite exploitées pour localiser et extraire les ressources pertinentes, soit par des utilisateurs humains ou bien par des agents logiciels (Charlet et al., 2003). Cette vision du Web Sémantique repose sur un certain nombre d'hypothèses. Par exemple, on doit disposer d'un langage de description à la fois expressif, pour pouvoir décrire les différents domaines d'application, et disposant d'une sémantique et d'une structure exploitable automatiquement par les agents logiciels. Une hypothèse est que les ressources soient effectivement annotées avec les metadonnées. Cette hypothèse est loin d'être garantie avec le nombre important de ressources mises à disposition dans le médium.

Par conséquent, un système automatique ou semi-automatique d'annotation de ressources composé non seulement avec des métadonnées liées à la structure du document, comme l'au-

teur ou la date de création, mais également les métadonnées liées au contexte thématique des ressources, est essentiel pour la réalisation de la vision du Web Sémantique.

Cet article présente une approche pour faciliter le partage de ressources parmi des partenaires distribués sur plusieurs pays et membres du Système Euro-Méditerranéen d'Information sur les savoir-faire dans le Domaine de l'Eau (SEMIDE[1]). L'idée principale est de faciliter le partage des ressources en facilitant leur annotation avec des métadonnées par un procédé semi-automatique.

Annoter des ressources sans avoir accès au contenu est une tâche difficile. Effectivement, dans un système tel que le SEMIDE où l'accès aux documents est restreint, généralement pour des raisons de confidentialité, les fournisseurs de ressources ne mettent à disposition qu'un ensemble de métadonnées.

Partis du constat qu'une ressource référence généralement une autre ressource, nous avons défini une approche d'annotation complètement indépendante du contenu. En utilisant le contexte de citation, nous propageons les annotations des références afin d'annoter le document citant.

Etant dans un système distribué et collaboratif, l'environnement linguistique est nécessairement hétérogène. La langue du pays ne doit pas être une barrière au partage de l'information. L'utilisation de simples mots clés marque vite la limite d'une telle solution. L'annotation dans notre approche est basée sur une ontologie qui définit le domaine. L'annotation de la ressource peut être faite en utilisant les annotations des références, l'idée étant de faire hériter une ressource des concepts d'autres ressources (références) en restant dans l'hypothèse que tous les fournisseurs partagent une ontologie.

Le reste de l'article est organisé de la manière suivante. Dans la section 2 nous proposons un survol les travaux liés à l'annotation des documents et nos motivations pour une nouvelle approche. La section 3 présente notre approche, et la section 4 les résultats de nos expérimentations. Enfin, une conclusion est proposée dans la section 5.

2 Etat de l'art

Une bonne recherche d'informations nécessite une bonne description. Dans notre contexte, ceci se traduit par des documents annotés par un ensemble de termes représentatifs suivant ou pas un vocabulaire contrôlé. Deux approches d'annotations ont été retenues tout au long de notre travail : l'*annotation par le contenu*, et ceci en utilisant uniquement le contenu du document pour créer son annotation et l'*annotation par le contexte*.

2.1 Annotation par le contenu

Le premier type d'annotation, qu'on appelle aussi *annotation statique* ou indexation, utilise uniquement le contenu du document pour le décrire.

L'indexation (Salton, 1971), (Rijsbergen, 1979) d'un texte consiste à repérer dans son contenu certains mots ou expressions particulièrement significatifs (appelés termes d'indexation) dans un contexte donné, et à créer un lien entre ces termes et le texte d'origine. Il existe trois types d'indexation :

[1] www.semide.org

1. manuelle : lorsque le document est analysé par un spécialiste du domaine ou un documentaliste ;
2. automatique : lorsque le processus d'indexation est complètement informatisé ;
3. semi-automatique : lorsqu'une première sélection de termes est réalisée automatiquement mais le choix final reste au spécialiste. Les systèmes les plus simples et les plus répandus sont basés sur la sélection de mots-clés dans les textes (Enguehard et al., 1992).

L'indexation sémantique est en fait est une spécialisation de l'indexation classique qui prend en compte la sémantique des mots au travers des relations entre les termes indexés.

Nous définirons l'indexation sémantique comme l'utilisation d'ontologies de domaine ou de thésaurus afin d'effectuer le processus d'indexation, une ontologie étant un ensemble de concepts reliés par une relation de "spécialisation/généralisation" définissant un domaine donné. Chaque concept est dénoté par un terme.

2.2 Annotation par le contexte

C'est à ce deuxième type d'annotation que nous nous intéressons particulièrement dans cet article, et nous allons présenter les travaux associés aux annotations utilisant les liens de citation ou de référencement considérés alors comme le contexte du document.

2.2.1 contexte de citation

La figure 1 illustre les différentes relations de citations entre les documents :

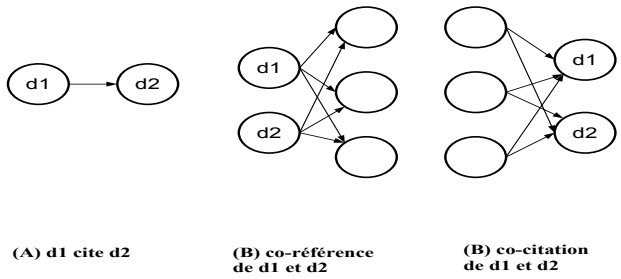

FIG. 1 – *Les relations entre les documents*

– la relation de citation : lorsqu'un document d_1 référence un document d_2. Généralement l'analyse de citation détermine l'impact d'un auteur dans un domaine particulier, en déterminant le nombre de fois où cet auteur a été cité ;
– le relation de co-référence : représente les documents qui partagent une ou plusieurs références bibliographiques ;
– la relation de co-citations : représente des documents qui sont cités par les mêmes documents.

L'analyse des citations dans le domaine de la bibliométrie examine les relations entre les auteurs et les publications. C'est en premier Kessler (Kessler, 1965) qui a utilisé les citations comme relation entre les documents techniques. Le principe de son approche est basé sur l'hypothèse que deux articles qui citent un ou plusieurs documents communs, ont une relation significative avec une force d'association traduite par le nombre d'articles en commun. Cette méthode n'a pas été utilisée par la suite à cause du volume important des données (Rostaing, 1996). Garfield et Small ont repris l'idée mais en utilisant la méthode de co-citations (Garfield, 1993). Elle permet de créer des relations entre des articles scientifiques d'un même domaine de recherche et en utilisant leurs références bibliographiques. Cette méthode repose sur l'hypothèse que deux références bibliographiques de dates quelconques, fréquemment citées ensemble ont une parité thématique.

De nombreux commentaires ont été faits sur la méthode de co-citations et plus généralement sur les citations (Rostaing, 1996) :

- les citations erronées, quand un auteur cite un travail mais en référençant une autre source que l'auteur principal ;
- la nature des citations qui peut être critique. Un auteur peut citer un document afin de critiquer le travail développé dans ce document ;
- l'auto-citation ;
- la période entre la citation et la publication ; en effet un document qui vient de paraître ne peut être cité immédiatement, il faut attendre une longue période avant qu'il ne soit cité ;
- le nombre de citations dépend de la nature de la publication ;
- le choix de l'auteur des citations (par exemple, il cite plus facilement des personnes de son pays ou des auteurs de référence, qui sont beaucoup cités ailleurs).

2.2.2 contexte de référencement

Le lien de référencement se réfère au graphe du Web. Il se présente comme un système hypertexte sous la forme d'un graphe orienté où les nœuds correspondent aux pages, et les arcs aux liens hypertextes. Il représente le lien de citation sur le Web sous la forme d'un hypertexte.

Marchiori (Marchiori, 1998), a été le premier à utiliser les liens entre les documents pour propager un ensemble de mots clés d'un document cité vers un document citant.

L'hypothèse de Marchiori est la suivante : si une ressource P 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 P' dans le Web qui possède un hyperlien vers P, alors les métadonnées de P sont propagées à P'. L'idée est que l'information contenue dans P est accessible par P', étant donné qu'il existe un lien de référencement. Marchiori applique un facteur d'affaiblissement f sur le poids des mots clés propagés mais, à moins que ce facteur soit important, la portée de la propagation est rapidement ingérable et exponentielle. De la même façon, appliquer un facteur d'affaiblissement important peut supprimer des mots clés dans les références qui peuvent être importants.

L'affectation des métadonnées aux pages est une tâche difficile, c'est pour cela que Prime (Prime-Claverie, 2004) s'est intéressée à l'attribution de ces métadonnées en les propageant dans le graphe du Web. Ces métadonnées ne représentent pas des mots clés mais le type d'autorité, le type d'information et le type de site. Tout comme Marchiori, Prime se base sur l'hypo-

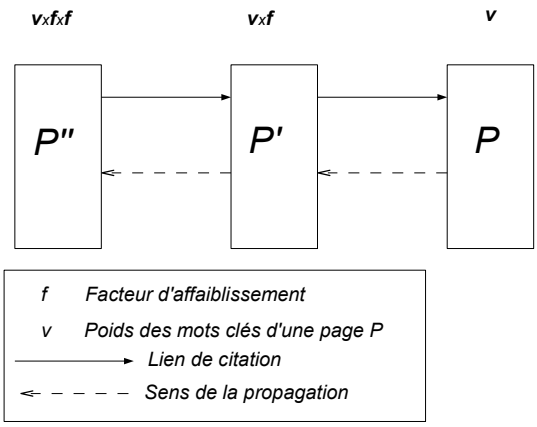

FIG. 2 – *Propagation de métadonnées (Marchiori)*

thèse que, si une page P contient un lien vers une autre page P', alors ces deux pages partagent des métadonnées communes. La propagation se fait en identifiant les pages les plus proches, en utilisant la méthode de co-citations.

2.3 Analyse des travaux de l'état de l'art

L'annotation sémantique utilisant des ontologies est la méthode la plus pertinente et prometteuse pour pallier au problème d'hétérogénéité et de multilinguisme des documents. L'apport d'une telle approche est à deux niveaux :

- l'indexation des documents en utilisant un vocabulaire contrôlé et spécialisé dans un domaine bien défini permet d'éviter d'avoir du bruit (trop de réponses dans le processus de recherche) ;
- la recherche d'informations en exploitant les relations entre les termes pour retrouver les documents les plus pertinents.

Utiliser uniquement le contenu du document génère un manque qui peut être dû aux raisons suivantes :

- un auteur peut vouloir transmettre une information dans un document et ne pas utiliser les termes de l'ontologie ;
- un document peut contenir beaucoup d'informations et ne pas avoir beaucoup de contenu. Par exemple, un document décrivant les présentations d'une conférence va contenir des références vers les présentations ;
- enfin, la raison qui nous pousse essentiellement à nous intéresser au contexte des documents est le cas du SEMIDE où on ne dispose pas toujours du contenu des documents mais d'un ensemble de métadonnées associées.

Quelques chercheurs se sont intéressés au contexte d'un document en exploitant les liens qui existent entre ce document et les autres documents. Les travaux sont basés sur l'idée que, quand un auteur cite un autre document c'est qu'il estime que ce dernier est pertinent pour le contenu

ou pour l'information qu'il veut faire passer. Une bonne annotation prend en compte tout le contexte d'un document, mais contrairement aux travaux décrits nous pensons qu'on ne peut pas simplement propager des mots clés sur le document citant et cela pour plusieurs raisons :

- les citations n'ont pas toutes la même importance dans un document ;
- un document peut traiter de plusieurs thèmes sans qu'ils aient forcément la même importance, cette importance ayant une relation avec les thèmes des documents cités.

Nous proposons une approche de propagation d'annotation sans connaissance du contenu, basée sur les liens de citations entre les documents. Cette annotation résout le problème de l'annotation par le contenu, et répond à la faiblesse d'une simple propagation de mots clés. Pour les raisons que nous avons déjà citées, nous utilisons une ontologie du domaine.

3 Approche : annotation basée sur les références

L'utilisation d'annotations pour décrire les documents est obligatoire afin de pouvoir décrire et utiliser au mieux les ressources. Notre proposition consiste à propager les annotations en utilisant les liens de citation entre les documents. Ces annotations consistent en un ensemble de mots clés issus de l'ontologie du domaine.

Afin d'utiliser les documents cités pour l'annotation, nous devons répondre à trois questions :

- quelles citations retenir pour effectuer la propagation ? Toutes les citations ne sont pas significatives pour le document source et pour déterminer son thème ;
- comment annoter le document ?
- et enfin, comment fusionner les annotations issues des références sélectionnées ?

Pour ajouter un nouveau document, noté d, dans la base initiale, nous procédons de la manière suivante :

1. récupérer l'ensemble des documents cités par d. Cet ensemble est noté Ref_d ;
2. regrouper par thème les documents de l'ensemble Ref_d afin de déterminer les groupements thématiques les plus pertinents et éviter ainsi les références non pertinentes mais présentes dans Ref_d ;
3. importer les annotations des documents cités par d ;
4. sélectionner parmi les annotations importées les plus pertinentes pour les proposer comme annotation du document d, en appliquant un ordre d'importance.

Dans la suite, nous détaillons les différentes étapes.

3.1 Regroupement thématique des documents

Un document, notamment lorsqu'il est technique ou scientifique, peut faire référence à plusieurs autres documents. En utilisant ce contexte de citation, cela nous offre la possibilité de situer thématiquement le document (Garfield, 1993). Par la suite, nous utilisons cette caractéristique pour retrouver les thèmes les plus importants d'un document sans avoir accès à son contenu mais en utilisant simplement les références de celui-ci. Un thème est un des sujets principaux du document, il est représenté par un ensemble de concepts issus de l'ontologie du domaine.

La première étape de notre démarche consiste à ne retenir que les références qui serviront pour l'annotation du document cible. Le but de cette première étape consiste à ne garder que les références qui traitent du même thème que le document source. Afin de déterminer les thèmes les plus importants dans l'ensemble des références d'un document, nous utilisons l'hypothèse de la co-citations. Cette mesure, importante dans le domaine de la bibliométrie (Rostaing, 1996) tient compte de l'hypothèse suivante : si deux documents sont souvent cités ensemble alors ils sont thématiquement proches.

Dans un premier temps, la matrice de co-citations représentant le graphe de co-citations des documents références Ref_d est calculée. Ensuite, les valeurs de similarités entre les couples de documents sont déduites, ceci représente la distance thématique entre les documents. Enfin, des classes de documents qui correspondent aux ensembles de documents partageant le même thème sont construites.

3.1.1 Construction du graphe de co-citations

Le graphe de co-citations est crée à partir du graphe de citation où les nœuds sont les documents et les arcs valués par le nombre de fois où les noeuds (documents) sont cités ensemble. Par exemple dans la figure 3, la valuation 2 entre les documents d_1 et d_3 indique que ces deux documents sont cités ensemble par deux documents.

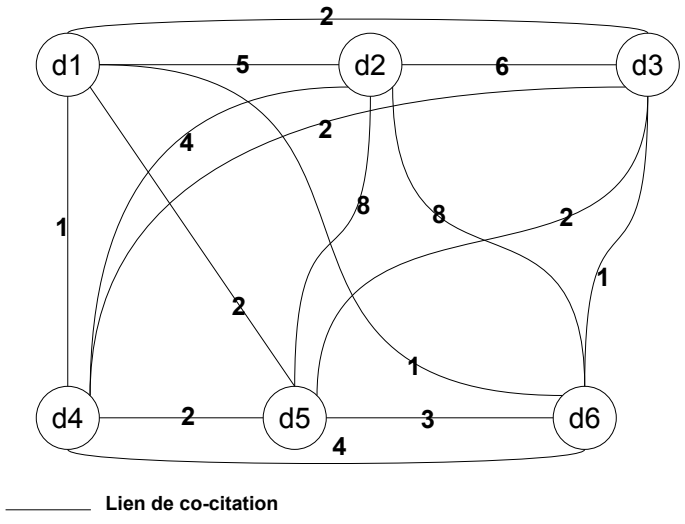

FIG. 3 – *Le graphe de co-citations*

La matrice de co-citations est une représentation du graphe de co-citations, elle correspond à une matrice carrée, cette matrice ne prend pas en compte les documents citants.

La matrice de co-citations de l'exemple présenté dans la figure 3 est la suivante :

$$\begin{pmatrix} 0 & 5 & 2 & 1 & 2 & 1 \\ 5 & 0 & 6 & 4 & 8 & 8 \\ 2 & 6 & 0 & 2 & 2 & 1 \\ 1 & 4 & 2 & 0 & 2 & 4 \\ 2 & 8 & 2 & 2 & 0 & 3 \\ 1 & 8 & 1 & 4 & 3 & 0 \end{pmatrix}$$

- C_{34} (3ème ligne, 4ème colonne) correspond à la fréquence de co-citations de d_3 et d_4, elle est égale à 2 car ils sont cités ensemble par deux documents ;
- C_{14} est égale à 1 car d_1 et d_4 ne sont cités ensemble que par le document d.

3.1.2 Calcul de la similarité thématique

La deuxième étape pour le choix des références à utiliser pour la propagation des annotations est le calcul de la similarité entre les documents références.

Un exemple est donné par Prime (Prime-Claverie, 2004) qui propose une fonction de distance comme suit :

$$S_{i,j} = 1 - \frac{{C_{(i,j)}}^2}{C_i \times C_j} \tag{1}$$

Dans l'équation 1 :

- $C_{(i,j)}$ représente l'indice de co-citations qui est défini comme le nombre de fois où les documents i et j sont cités ensemble ;
- C_i représente le nombre de fois où le document i est cité ;
- C_j représente le nombre de fois où le document j est cité ;

Cependant, dans cette fonction de distance, le dénominateur était conçu à l'origine pour normaliser la fraction, le résultat étant dans l'intervalle $[0, 1]$.

En effet, les documents i et j sont indépendants et peuvent par exemple apparaître à des périodes différentes (si nous supposons que le document i est plus ancien que j). Dans ce cas, le document i peut être cité plusieurs fois et nous aurons par conséquent un grand C_i. Cependant, si j est récent, alors dans ce cas C_j sera petit et comme $C_{i,j} \leq C_j$, nous aurons également $C_{i,j}$ petit.

Dans ce cas de figure, si nous supposons qu'à chaque fois que le document j est cité, le document i est cité dans le même document, on s'attend à une proximité thématique. Cependant, on ne retrouve pas ce résultat si on applique la fonction de distance de Prime. En effet, comme $C_{i,j} \leq C_j << C_i$ alors $S_{i,j} \approx 1$. Ceci laisse entendre une disparité entre le document i et j. Or, même si le document j est cité à chaque fois avec le document i, cette disparité ne peut se résorber car le paramètre C_i est complètement indépendant du document j. Ceci est illustré par l'exemple suivant :

- un document d_1 a été publié récemment et il est cité 10 fois, ce qui donne $C_1 = 10$;
- un document d_2 a été publié depuis longtemps et est cité 500 fois, ce qui donne $C_2 = 500$;
- les documents d_1 et d_2 sont cités 10 fois, ce qui inclut qu'à chaque fois que d_1 a été cité alors d_2 a été cité aussi.

A partir de ces données, on s'attend à avoir un résultat qui démontre que les deux documents sont proches, or en appliquant le fonction 1 le résultat est $S_{i,j} = 0.98 \approx 1$.

Afin de calculer la distance thématique $S_{(i,j)}$, nous définissons la mesure suivante :

$$S_{(i,j)} = \frac{1}{{C_{(i,j)}}^2} \tag{2}$$

L'équation 2 prend en compte simplement l'indice de co-citations entre deux documents afin de déterminer leur proximité thématique. Ainsi, plus deux documents sont cités ensemble, plus la distance $S_{(i,j)}$ sera proche du zéro. Quand les références d'un document d sont récupérées, nous construisons le graphe de distances GC_d.

$$GC_d =< Ref_d, Ref_d \times Ref_d \times [0,1] > \tag{3}$$

Tel que décrit dans l'équation 3, le graphe de distances est un graphe complet où les nœuds représentent les documents cités dans d. Un lien entre deux documents i et j, est un lien valué avec la fonction de distance $S_{(i,j)}$ présentée dans l'équation 2. La représentation de ce graphe peut également être vue comme une matrice, appelée matrice de distance, MC, définie comme suit :

$$\begin{array}{l} MC_d : |Ref_d| \times |Ref_d| \\ \forall i,j \in Ref_d,\ MC_d(i,j) = \left\{ \begin{array}{ll} S_{(i,j)} & \text{si} \quad i \neq j \\ 0 & \text{sinon} \end{array} \right. \end{array} \tag{4}$$

La matrice de distance de l'exemple présenté dans la figure 3 est :

$$\begin{pmatrix} 0 & 0.04 & 0.25 & 1 & 0.25 & 1 \\ 0.04 & 0 & 0.027 & 0.0625 & 0.015 & 0.015 \\ 0.25 & 0.027 & 0 & 0.25 & 0.25 & 1 \\ 1 & 0.0625 & 0.25 & 0 & 0.25 & 0.0625 \\ 0.25 & 0.015 & 0.25 & 0.25 & 0 & 0.11 \\ 1 & 0.015 & 1 & 0.0625 & 0.11 & 0 \end{pmatrix}$$

Après avoir calculé la similarité thématique entre les documents en nous basant sur la méthode de co-citations, il s'agit de regrouper les documents partageant le même thème et de séparer les documents traitant de thèmes différents. Pour cela, nous utilisons les méthodes de classification. Les documents regroupés dans la même classe ont une proximité thématique et par cela des annotations proches.

3.1.3 Construction des classes

Cette étape consiste en la construction de classes de thèmes des documents référencés. Pour cela on attribue une classe à chaque groupe de documents ayant le même thème, à partir de la matrice de distances, en utilisant un algorithme de groupement *fuzzy c-means* (Dunn, 1974) basé sur la théorie des ensembles flous. Cet algorithme est utilisé afin d'autoriser le chevauchement des classes.

Afin d'utiliser la matrice de distances comme entrée à ces algorithmes, il faut veiller à ce que les valuations des liens définissent bien une distance au sens mathématique. Rappelons que les propriétés d'une distance sont les suivantes :

1. $d(x, y) \geq 0$ *(positivité)*
2. $d(x, y) = 0$ si et seulement si $x = y$ *(identité)*
3. $d(x, y) = d(y, x)$ *(symétrie)*
4. $d(x, z) \leq d(x, y) + d(y, z)$ *(inégalité triangulaire)*

Dans notre cas, comme les documents sont indépendants et que le calcul de $S_{(i,j)}$ ne prend en compte que l'indice de co-citations de ces documents, la spécification d'une distance au sens mathématique peut ne pas être satisfaite. En effet, le matrice de distances peut mener à des cas où :

$$MC_d(i, j) > MC_d(i, k) + MC_d(k, j), \qquad k \notin \{i, j\}$$

Plus généralement, on peut avoir une distance cumulée sur un chemin reliant deux documents qui est inférieure à la distance directe entre deux documents. La figure 4 représente le graphe de distances de notre exemple.

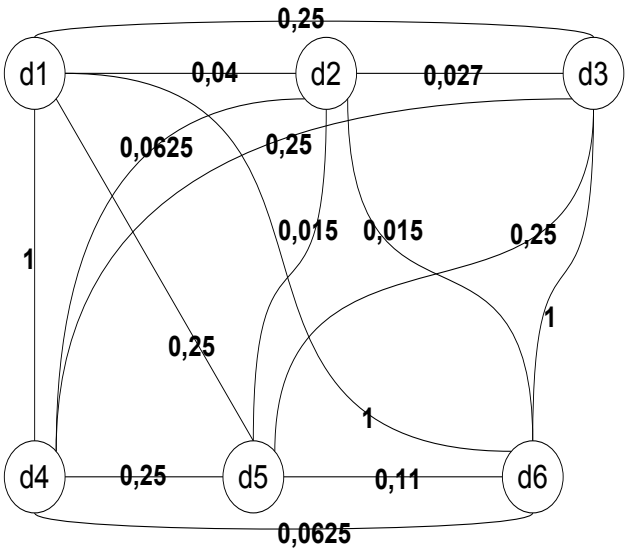

FIG. 4 – *graphe de distance*

Dans cet exemple, nous avons :

$$MC_d(1, 3) > MC_d(1, 2) + MC_d(2, 3),$$

effectivement

$$0.25 > 0.04 + 0.027,$$

Dans ce cas, le graphe de distance ne respecte pas les propriétés d'une distance et l'utilisation de l'algorithme de classification ne sera pas appropriée.

Pour résoudre ce problème nous transformons la matrice de distance pour que la distance entre deux documents i et j soit minimale afin de répondre à l'inégalité triangulaire. Nous utilisons pour cela l'algorithme Dijkstra (Dijkstra, 1959) afin de déterminer la distance minimale entre deux documents i et j :

$$MC'_d : |Ref_d| \times |Ref_d|$$
$$\forall i,j \in Ref_d,\ MC'_d(i,j) = \begin{cases} \text{Dijkstra}(i,j,MC_d) & \text{si} \quad i \neq j \\ 0 & \text{sinon} \end{cases} \tag{5}$$

Après l'application de l'algorithme de Dijkstra, MC'_d définit bien un espace métrique car :
– la propriété de symétrie est satisfaite, S étant une fonction symétrique.
– S ne peut pas valoir zéro et par définition $MC'_d(i,j)$ vaut zéro quand i est égal à j.
– l'inégalité triangulaire est satisfaite.

Une fois cet algorithme appliqué, nous obtenons sur notre exemple les distances illustrées sur la figure 5.

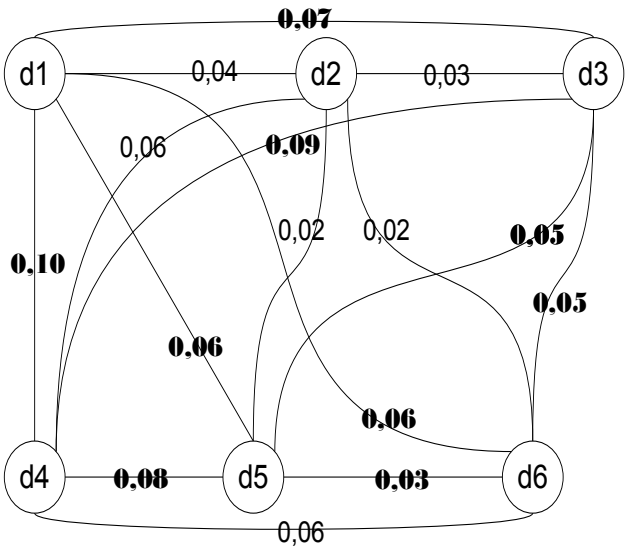

FIG. 5 – *graphe de distance après application de l'algorithme de dijkstra*

Nous avons par exemple $MC_d(1,3) > MC_d(1,2) + MC_d(2,3) = 0.07$. Nous pouvons appliquer les algorithmes de classification non supervisée, ici l'algorithme *fuzzy c-means*.

L'algorithme des c-moyennes (*fuzzy c-means*) fournit une partition de l'ensemble des données en sous-ensembles flous en optimisant la fonction 6 :

$$J = \sum_{i=1}^{n} \sum_{r=1}^{c} u_{ri}^{m} \left\| x_i - w_r \right\|^2 \tag{6}$$

avec la contrainte

$$\sum_{r=1}^{c} u_{ri}^{m} = 1 \tag{7}$$

- $X = \{x_i, i = 1....n\}$ est l'ensemble des données ;
- c le nombre de clusters cherchés, c doit être choisi ;
- w_r est le centre du cluster r ;
- u_{ri} est le degré d'appartenance de la donnée x_i au cluster r ;
- m est un hyper paramètre, fixé généralement à 2 ;

Les différentes étapes de l'algorithme *fuzzy c-means* sont les suivantes :

1. choisir le nombre de classes c ;
2. initialiser la matrice $U = [u_{ij}]$, $U^{(0)}$;
3. à la kième étape : calculer les centres $C^{(k)} = [c_j]$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m}.x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{8}$$

4. mettre à jour les degrés d'appartenance ;

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{9}$$

5. si $\left\| U^(k+1) - U^(k) \right\| < \epsilon$ alors arrêt, sinon retour à l'étape 3.

Construction des classes

En spécifiant le nombre de groupes pour l'algorithme *fuzzy c-means*, noté N_{clusters}, le résultat du regroupement est alors une matrice MG_d de dimension $|Ref_d| \times N_{\text{clusters}}$ où chaque élément $MG_d(i, j)$ représente le degré d'appartenance du document i au groupe j. On notera que la somme des degrés d'appartenance d'un document aux différents groupes vaut 1.

Le résultat de notre exemple donne la matrice suivante en fixant le nombre de clusters à 3 :

$$\begin{pmatrix} 0.01 & 0.01 & 0.97 \\ 0.13 & 0.79 & 0.07 \\ 0.85 & 0.09 & 0.05 \\ 0.38 & 0.33 & 0.28 \\ 0.15 & 0.75 & 0.08 \\ 0.07 & 0.87 & 0.04 \end{pmatrix}$$

Ceci donne un regroupement des références illustré dans la figure 6. La première colonne de la matrice définit le degré d'appartenance au cluster 1 et la deuxième au cluster 2, etc. Les documents d_2 et d_5 avec des degrés d'appartenance à C_2 respectivement 0.79 et 0.75 sont dans le deuxième cluster.

Les regroupements des documents par thème effectués, nous annotons le document d. La dernière étape consiste à l'importation des annotations des références choisies.

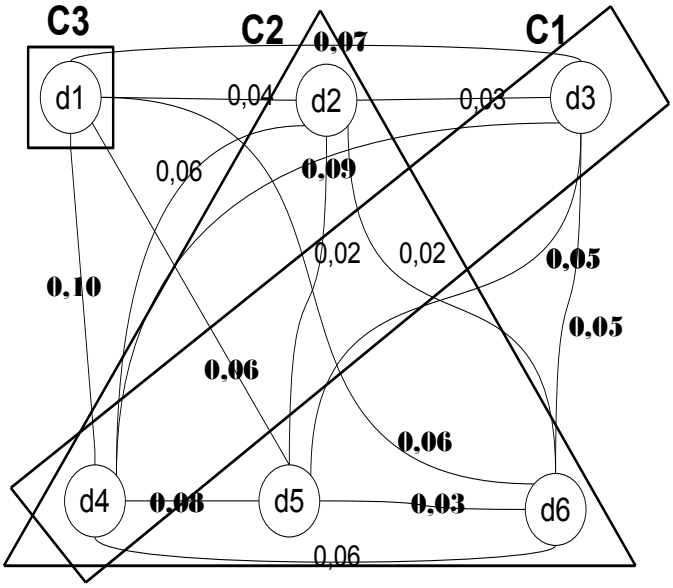

FIG. 6 – *Le résultat du regroupement des références après application de l'algorithme fuzzy c-means*

3.2 Importation et fusion des annotations

Le but dans cette étape est d'importer et d'ordonner les annotations des documents cités par un document d. La présentation des annotations importées est faite en définissant un choix multi-critères pour sélectionner des annotations à utiliser dans la phase suivante.

Dans un premier temps, nous définissons la liste des annotations importées pour le document d comme suit :

$$\textit{annotation\_list}_d = \bigcup_{i \in Ref_d} \textit{Annotation}(i) \tag{10}$$

La fonction *Annotation*(i) regroupe l'ensemble des annotations du document i, ces annotations correspondent aux annotations des références classées dans l'étape précédente. Il faut rappeler que les éléments de l'annotation sont des concepts définis dans l'ontologie globale d'annotation. Il faut considérer $annotation\_list_d$ comme un multi-ensemble, c'est-à-dire un ensemble avec une répétition des éléments, car deux références peuvent avoir des concepts communs.

L'ensemble des annotations du document d est alors défini comme suit :

$$\textit{annotation\_set}_d = \bigcup_{x \in \textit{annotation\_list}_d} \{x\} \tag{11}$$

Il s'agit maintenant de doter cet ensemble d'une fonction d'ordre totale en se basant sur plusieurs critères. Nous retenons les critères suivants pour ordonner les annotations dans l'ensemble $Annotations_d$:

1. l'importance du cluster contenant le document d'où l'annotation a été importée. Si l'annotation apparaît dans plusieurs documents, on considérera l'importance du cluster le plus grand (maximal) ;
2. le degré d'appartenance du document au cluster d'où l'annotation a été importée. Si l'annotation apparaît dans plusieurs documents, on ne considérera que le document qui a un degré d'appartenance du cluster maximale.
3. le nombre de fois où l'annotation apparaît dans *annotation_list*$_d$.

Concernant le premier critère d'ordre, nous partons de l'hypothèse que les groupements importants définissent la thématique du document d. Nous utilisons la matrice d'appartenance aux clusters GR_d. En effet, nous déterminons l'appartenance d'un document à un groupe en utilisant le maximum des degrés d'appartenances. Le cardinal de chaque cluster est le nombre des documents contenus dans celui-ci. Le cardinal des groupes nous indique un premier critère afin de déterminer quels sont les groupes les plus importants. Dans la figure 6, le groupe le plus important est celui du cluster $C2$, avec un cardinal égal à 4.

Dans notre exemple nous sélectionnons les documents d_2, d_4, d_5 et d_6, avec les annotations suivantes (figure 7).

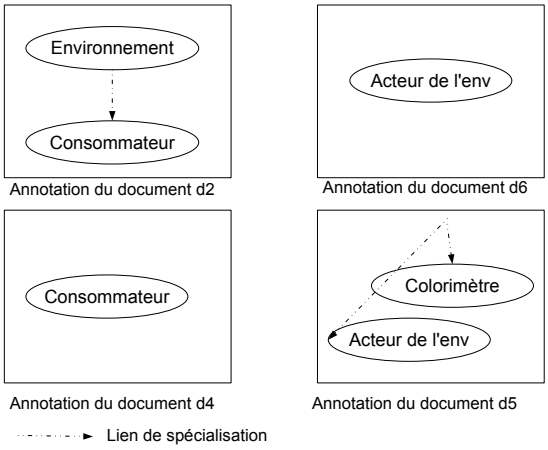

FIG. 7 – *Annotation des documents cibles*

En ce qui concerne le deuxième critère d'ordre, le degré d'appartenance d'un document aux différents groupes est déjà calculé dans la matrice GR_d. En ce qui concerne les documents de la figure 6, nous nous réfèrons à la matrice calculée lors de l'étape de regroupement, le

degré d'appartenance des documents d_2, d_4, d_5 et d_6 est le suivant :

$$\begin{pmatrix} 0.79 \\ 0.33 \\ 0.75 \\ 0.87 \end{pmatrix}$$

Pour le troisième critère d'ordre, il s'agit simplement de calculer la répétition de l'annotation dans la liste *annotation_list*$_d$. Nous remarquons que dans notre exemple *consommateur* est dans d_2 et d_4, alors que *environnement* n'est que dans d_2.

La fonction d'ordre est un ordre total et tous les éléments de l'ensemble *annotation_set*$_d$ peuvent être ordonnés. Nous trouvons ainsi les annotations importantes en fonction de l'ordre suivant :

1. les annotations qui proviennent des documents situés dans des clusters importants ;
2. au sein des annotations qui proviennent d'un même cluster ou bien de clusters qui ont la même importance. Les annotations importantes sont celles qui proviennent des documents qui ont un degré important d'appartenance au cluster. Ainsi, l'importance de l'annotation d'un document dépend de l'importance du document dans le cluster. Ici, les documents sont classés par ordre décroissant comme suit : d_6, d_2, d_5 et enfin d_4 ;
3. si les annotations proviennent d'un même document ou de documents qui ont le même degré d'appartenance au même cluster, alors nous considérons comme importantes les annotations redondantes.

Dans la deuxième étape, les annotations sont ordonnées de la façon suivante :
- « acteur de l'environnement »issu de deux documents d_6 et d_5, d_6 est le document qui a le plus fort degré d'appartenance au cluster,
- « consommateur »issu des documents d_2 et d_4,
- « environnement »issu du document d_2,
- « colorimètre »issu du document d_5,

Dans la dernière partie de cet article, nous présentons l'outil implémenté nommé *Reference Annotation System (RAS)*, ainsi que notre expérimentation.

4 Expérimentation

Notre expérimentation s'est déroulée en deux étapes :
- construction d'un outil d'annotation RAS ;
- évaluation du résultat de l'annotation sur un gros corpus de documents.

Chaque étape et les choix associés sont détaillés tout au long de cette section.

La figure 8 illustre le protocole d'expérimentation qui se déroule en deux phases : l'annotation et l'évaluation.

La phase d'annotation se compose des étapes suivantes :
- constitution d'un corpus qui répond à nos besoins, en l'occurrence des documents qui s'inter-référencent ;
- utilisation d'une ontologie pour l'annotation des documents ;
- annotation avec l'outil RAS.

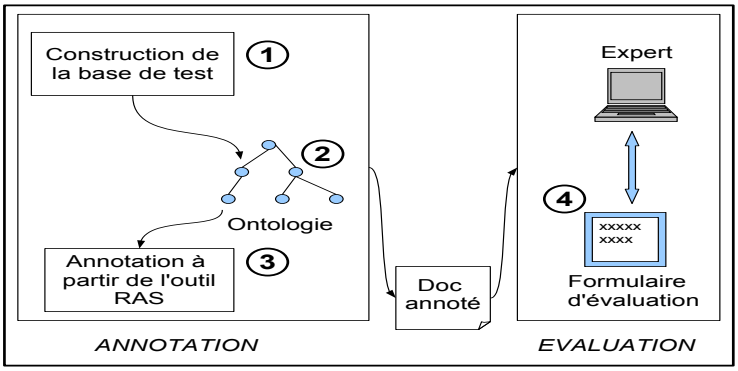

FIG. 8 – *Le protocole d'expérimentation*

Le résultat de cette phase est un ensemble de documents annotés.

L'évaluation du résultat de l'annotation des documents est basée sur deux méthodes : la comparaison avec une annotation existante, faite généralement par l'auteur du document et l'évaluation par des experts du domaine. La deuxième méthode est réalisée à partir d'un formulaire d'évaluation.

4.1 Constitution du corpus de tests

La base du SEMIDE ne constitue pas pour le moment une base d'une taille suffisamment importante pour que l'on puisse tester notre approche. Pour cela notre choix s'est porté sur une base composée d'un nombre important de documents qui s'inter-référencent.

4.1.1 Collection Citeseer

Nous avons choisi comme collection de tests la base de Citeseer[2] (Stribling et al., 2005), (Ghita et al., 2005) qui est une bibliothèque numérique sur la littérature scientifique. Citeseer localise les articles scientifiques sur le web, extrait différentes informations telles que les citations, le titre des articles. Cette collection a été choisie pour deux raisons : *(i)* le nombre importants de documents ; *(ii)* l'inter-référencement des articles, ce qui convient exactement à nos expérimentations.

La bibliothèque nous a permis de construire une base de plus de 550 000 documents qui s'inter-référencent. La figure 9 illustre le schéma de la base :

- une publication a un titre, une URL et un ou plusieurs auteurs ;
- une publication est citée par une autre publication ;
- une publication cite une autre publication ;
- une publication est co-citée avec une autre publication.

[2]http ://citeseer.ist.psu.edu/

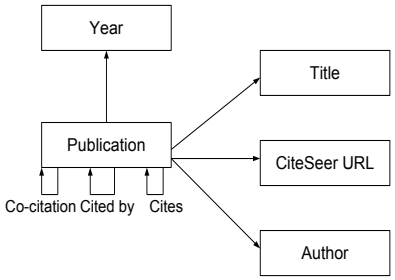

FIG. 9 – *Le schéma de la base Citeseer*

4.1.2 Utilisation d'une ontologie

La description des documents de Citeseer ne peut pas s'utiliser directement. Effectivement, Citeseer utilise un vocabulaire général pour décrire les documents. Or dans notre cas, nous nous intéressons uniquement à la description des documents utilisant une ontologie. Nous avons utilisé l'ontologie DMOZ afin de décrire les documents. Notre approche a consisté à utiliser le moteur de recherche de DMOZ[3] afin de créer la correspondance entre les mots clés de Citeseer et l'ontologie du domaine.

L'outil développé au cours de notre travail, RAS (Reference Annotation System), basé sur notre approche d'annotation est présenté dans la section suivante.

4.2 L'outil RAS

Un outil nommé RAS a été réalisé afin de montrer la faisabilité de notre approche. C'est un outil d'annotation de documents basée sur les liens de citation. L'outil est disponible en ligne à l'adresse suivante : *http ://www.lirmm.fr/annotation*

Afin d'annoter un document, l'outil RAS regroupe les différentes étapes de notre approche, qui sont les suivantes :

- calcul des indices de co-citations
- calcul de la similarité thématique des références
- génération des classes de thèmes
- fusion des annotations

Les classes de thèmes sont construites en appliquant l'algorithme flou des c-moyennes. Dans notre travail, nous avons posé comme hypothèse qu'un document ne peut pas couvrir plus de 3 thèmes, d'où le choix du nombre de classes à 3.

A partir du résultat de regroupement, le système utilise l'ordre décrit dans notre approche afin d'annoter le nouveau document :

- l'importance du cluster ;

[3] http ://www.dmoz.org/

– le degré d'appartenance au cluster ;
– le nombre de fois où l'annotation apparaît.

3000 documents ayant des références déjà annotées ont été sélectionnés de manière aléatoire dans la base et annotés avec RAS. Les méthodes d'évaluation utilisées et les résultats obtenus sont présentés dans la suite.

4.3 Evaluation

Nous avons utilisé deux méthodes d'évaluation de notre approche :
– une comparaison avec l'annotation de la base Citeseer ;
– un formulaire pour les experts.

Le nombre de documents déjà annotés et annotés par RAS n'étant pas assez important, la première méthode à été complétée par l'avis d'experts à travers un formulaire d'évaluation.

4.3.1 Comparaison avec l'annotation existante

La première méthode d'évaluation consiste à comparer le résultat avec une annotation existante. Parmi les 3000 documents annotés avec RAS, nous avons sélectionné ceux déjà indexés dans Citeseer, généralement par l'auteur du document.

Les documents sélectionnés ici devaient répondre aux contraintes suivantes :
– les documents dans la base doivent être annotés afin de pouvoir les comparer avec le résultat de notre approche ;
– les documents doivent avoir au moins trois références annotées, et ceci afin de pouvoir appliquer notre approche de propagation d'annotation. Nous avons remarqué de façon empirique que pour moins de 3 références, la propagation n'est pas pertinente. Dans ce cas, une annotation est effectuée par l'expert.

Le tableau 1 illustre le nombre de documents utilisés :
– le nombre de documents annotés est de 17417. Notre base contient 550000 documents, mais nous n'avons pas pu obtenir tous les mots clés associés aux documents. Les documents de Citeseer sont décrits avec un ensemble de mots clés, un passage à l'annotation avec l'ontologie a été effectué ;
– afin de propager les annotations, nous avons sélectionné les documents ayant plus de 3 références annotées ;
– l'intersection des deux premiers ensembles de documents nous donne 66 documents. Cet ensemble sert de comparaison avec l'annotation générée par notre approche. La première méthode a donc été effectuée sur un petit corpus de documents.

Documents dans la base	550 000
Documents annotés par les experts	17 417
Documents ayant plus de 3 références annotées	940
Documents déjà annotés et ayant plus de 3 références annotées	66
Nombre moyen de termes annotant par document dans Citeseer	6,77

TAB. 1 – *comparaison des annotations*

Concernant les 66 documents évalués nous obtenons :
– une moyenne de $6,77$ concepts par document ;
– une moyenne de $4,96$ concepts de documents correspondent à l'annotation existante.

Chaque document est annoté avec une moyenne de $6,77$ concepts. Notons qu'un index a été fourni par l'administrateur de la base Citeseer sous forme de mots clés associés à un ensemble de documents. Afin de pouvoir comparer avec le résultat de notre approche, nous avons transformé ces mots clés en une annotation en utilisant les concepts de *Dmoz*. Parmi les résultats de notre approche $4,96$ concepts correspondent à l'annotation existante, ce qui correspond à 73%.

4.3.2 Formulaire d'évaluation

Une deuxième méthode d'évaluation a consisté à se baser sur des experts à travers un formulaire d'évaluation. Le formulaire consultable sur le web[4], contient les questions suivantes :

1. L'annotation correspond-elle au titre du document ?
2. L'annotation correspond-elle à la description du document ?
3. Les termes correspondent-ils bien au document ou pas ?
4. Avis général sur l'annotation (très satisfaisant, satisfaisant, peu satisfaisant, pas satisfaisant)

Les articles de Citeseer relèvent de plusieurs sous domaines, le choix des évaluateurs a été fait en utilisant différentes listes de diffusion de la communauté du domaine informatique. Ces évaluations ont été faites sur 322 documents. Cela a duré trois mois.

La figure 10 illustre l'avis sur l'annotation en général.

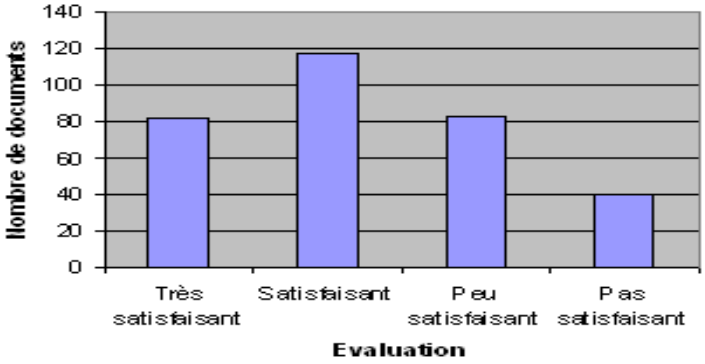

FIG. 10 – *Avis sur l'annotation sur un échantillon de 315 documents*

Afin d'estimer la moyenne des annotations incohérentes pour un document, l'évaluateur saisi le nombre d'annotations qui ne correspondent pas au document. Nous avons obtenu pour une moyenne de $10,88$ concepts par document, $2,65$ concepts incorrects dans l'ensemble de l'annotation.

[4]http ://www.lirmm.fr/annotation/evaluation

4.4 Analyse

Dans cette section, nous tentons d'analyser les mauvais résultats et ceci en nous basant sur les résultats et sur les commentaires saisis par les évaluateurs insatisfaits du résultat de l'annotation. Nous avons catégorisé les commentaires par critère, qui sont les suivants :

1. le nombre des concepts de l'annotation est insuffisant ;
2. le dernier niveau des concepts est trop spécialisé ;
3. les concepts spécifiques au document n'apparaissent pas (annotation correcte mais pas assez spécialisée) ;
4. incohérence entre les concepts, due au passage de mots clés vers l'ontologie Dmoz.

L'annotation d'un document est basée sur des annotations existantes des documents cités, le regroupement de ces références est complètement indépendant de leurs annotations étant donné que ce regroupement est basé sur les co-citations.

Cette propagation nécessite alors une base déjà annotée, ce qui n'est pas toujours le cas. Quelques documents évalués avaient un petit nombre de références annotées, ce qui a généré dans certains cas une annotation se composant de un ou deux concepts et dans ce cas un nombre de concepts insuffisant.

La deuxième raison de la mauvaise annotation est la propagation des concepts de références spécifiques au document référencé. Prenons par exemple une référence qui traite du SGBD "MySQL", ce document est une définition des bases de données. Ne faisant pas la distinction entre la nature des références (dans une définition ou dans le corps du document), nous nous basons sur la proximité thématique des documents, dans ce cas nous propageons le concept "MySQL" qui représente un niveau trop spécifique pour le document à annoter. Une solution serait de propager les concepts d'un niveau plus haut dans la hiérarchie du concept (ex : SGBD).

Le contraire peut générer une annotation insuffisante en n'ayant pas de concepts spécifiques au document à annoter. En propageant des références proches thématiquement, qui sont dans le même domaine que le document à annoter mais qui ne traitent pas forcément du même sujet (de façon détaillée), nous générons une annotation correcte et cohérente mais pas assez fine (spécifique).

Enfin, lors de nos expérimentations, quelques incohérences sont apparues et ceci lors du passage de mots clés vers l'ontologie Dmoz. Un traitement manuel de filtrage a été effectué afin d'éliminer les concepts n'appartenant au domaine.

5 Conclusion et perspectives

Cet article a décrit notre système d'annotation de documents RAS. L'annotation manuelle de documents est une tâche difficile voire impossible à faire, compte tenu du temps que cela nécessite. Nous avons utilisé la relation de citation d'un document avec d'autres documents afin de déduire les concepts de l'annotation sans accéder au contenu du document. Les concepts sont issus d'une ontologie du domaine. Ce type d'annotation, contrairement à l'indexation

classique avec une liste plate de mots clés, améliorera le processus de recherche de documents et résoudra le problème de multilinguisme puisqu'un terme dans différentes langues est associé à un seul concept. Notre approche est basée sur un regroupement thématique des citations en utilisant l'algorithme *fuzzy c-means*, le rapprochement thématique est construit à partir de la méthode de co-citations.

La base documentaire du SEMIDE n'étant pas suffisante, les expérimentations ont été effectuées sur la base Citeseer. Ce choix a été motivé par la taille de la base et l'inter-référencement des documents. L'évaluation du résultat de l'annotation a été réalisée par une comparaison avec des annotations existantes, ainsi que par un formulaire pour les experts. Les résultats obtenus sont satisfaisants ; néanmoins il en ressort quelques mauvaises annotations.

Nous avons constaté qu'une grande partie des mauvais résultats est due au passage des mots clés vers l'ontologie DMOZ. Le moteur de recherche donnant quelques fois plusieurs hiérarchies avec le même nombre d'occurrences. Dans ce cas, il est impossible de choisir de manière automatique la hiérarchie correspondante au mot.

Les perspectives associées à ce travail seraient d'étendre notre évaluation en proposant pour chaque document l'annotation faite par le système RAS aux auteurs de l'article et d'utiliser notre annotation pour la recherche de documents et analyser la pertinence des documents retournés en résultat.

Références

Charlet, J., P. Laublet, et C. Reynaud (2003). In *Action spécifique 32 Web sémantique : Rapport final*.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 269–271.

Dunn, J. C. (1974). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics 3*, 32–57.

Enguehard, C., P. Malvache, et P. Trigano (1992). Indexation de textes : l'apprentissage des concepts. In *Proceedings of 14th conference on Computational linguistics*, Morristown, NJ, USA, pp. 1197–1202. Association for Computational Linguistics.

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).

Ghita, S., N. Henze, et W. Nejdl (2005). Task specific semantic views : Extracting and integrating contextual metadata from the web. In *Proceedings of the International Semantic Web Conference Workshop on The Semantic Desktop - Next Generation Personal Information Management and Collaboration Infrastructure, ISWC*, Galway, Ireland.

Kessler, M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American documentation 14*, 10–15.

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

Prime-Claverie, C. (2004). *Vers une prise en compte de plusieurs aspects des besoins d'information dans les modèles de la recherche documentaire : Propagation de métadonnées sur le World Wide Web*. Ph. D. thesis, Ecole supérieure des Mines de Saint-Etienne.

Rijsbergen, V. (1979). *Information Retrieval, 2nd edition*. London : Dept. of Computer Science, University of Glasgow.

Rostaing, H. (1996). *La bibliométrie et ses techniques*. Sciences de la Société Collection.

Salton, G. (1971). A comparaison between manual and automatic indexing methods. In *Proceedings of Journal of American documentation*.

Stribling, J., I. Councill, J. Li, M. F. Kaashoek, D. Karger, R. Morris, et S. Shenker (2005). Overcite : A cooperative digital research library. In *Proceedings of International Workshop on Peer-to-Peer Systems*.

Summary

The semantic annotation of resources is at the heart of the semantic Web vision. In fact, the annotation of resources by meta-data, makes it possible for both human users and software agents to find the appropriate resources despite being in a large and distributed medium, such as the Web. However, the annotation process should not be considered as granted, especially with the growing number of resources when it is almost impossible to assume a manual annotation process of resources. We propose in this paper, an approach for automatic and semi-automatic annotation of resources based on the context of citations. The approach presented in this paper has been experimented and evaluated with the Citeseer database of scientific documents. The evaluation shows the benefit of using not only the content of resources for the annotation but also using the citation relationships as context to describe the subjects and themes of resources.