Annotation automatique de documents

Lylia Abrouk, Abdelkader Gouaich, Chedy Raïssi

Annotation automatique de documents

Analyse des citations

Lylia Abrouk[*,] — Abdelkader Gouaïch[*] — Chedy Raïssi[*,***]**

[*] *LIRMM, Laboratoire d'Informatique, de Robotique et de Microlectronique de Montpellier*
161 rue Ada, 34392 Montpellier Cedex 5
{abrouk, gouaich,raissi}@lirmm.fr

[**] *Système Euro-Méditerranéen d'Information sur les savoir-faire dans le Domaine de l'Eau.*
2229, route des cretes, 06560 Valbonne
l.abrouk@semide.org

[***] *EMA-LGI2P/Site EERIE*
Parc Scientifique Georges Besse
30035 Nîmes Cedex, France
raissi@ema.fr

RÉSUMÉ. *Etant impliqué dans un Système euro-méditerranéen dont le but est la diffusions d'informations multilingues, nous nous intéressons à la description de cette information afin de faciliter l'échange et la recherche de documents. Dans cet article, nous proposons une approche pour l'annotation des documents qui consiste à se baser sur les références citées afin de propager leurs annotations sur le document cible. Pour cela nous utilisons et étendons des méthodes existantes pour filtrer les références utilisées.*

ABSTRACT. *Being involved in a euro-mediterranean system which goal is to diffuse information, we focus our interest on the description of this information in order to ease the exchange and the discovery of documents. In this article, we propose an approach for the annotation of the documents based on the cited references. This is done in order to propagate their annotations on the target document. To achieve this we use existing state of the art methods to filter the references used.*

MOTS-CLÉS : *Annotation, propagation, cocitations, classification, ontologies.*

KEYWORDS: *Annotation, propagation, cocitations, classification, ontologies.*

1. Introduction

Pour de nombreux domaines, le Web et ses technologies associées sont devenues la plus grande source d'information actuelle. Mais la spécificité de telles sources d'informations les rend difficilement exploitables et leur évolution constante rend complexe les techniques de recherche d'informations. La raison principale est la suivante : les documents sont fragmentés, dispersés, hétérogènes et sont souvent très peu structurés. Il est donc nécessaire de proposer des méthodes et des outils permettant de partager, manipuler et rechercher au sein de tels documents.
L'annotation sémantique à partir des ontologies est actuellement la méthode la plus pertinente et la plus prometteuse pour pallier aux problèmes de volatilité et d'hétérogénéité des documents sur le Web. Les annotations permettent d'associer des informations complémentaires aux documents, de spécifier certaines parties de celui-ci et enfin de les partager dans le cadre d'un groupe de travail. Cette annotation est donc très utile pour affiner les réponses aux requêtes des utilisateurs. Néanmoins, l'annotation sémantique soulève deux problèmes principaux :

1) Le grand volume de ressources proposées : il est irréaliste d'envisager d'annoter manuellement des centaines de milliers de documents que ce soit par leurs auteurs ou par des documentalistes. Dans le contexte de l'annotation automatique, l'intérêt se porte sur l'information décrivant le contenu de la ressource. Cette information peut ainsi être vue soit comme une liste plate de mots-clés décrivant la ressource, soit comme une liste de concepts reliés par des relations. Dans la suite de cet article, nous nous intéressons principalement à ce dernier cas où les relations nous permettent d'affiner les recherches sur les ressources sans posséder le contenu intégral de celles-ci.

2) L'enrichissement d'ontologies : une ontologie correspond à un vocabulaire contrôlé et organisé. Celle-ci permet la formalisation explicite des relations créées entre les différents termes du vocabulaire. Ces ontologies peuvent évoluer en même temps que la masse d'informations. Il faut donc trouver des moyens de mise à jour automatique de ces vocabulaires contrôlés et des annotations de documents qu'ils décrivent.

Dans cet article, nous présentons une nouvelle approche pour l'annotation semi-automatique de ressources selon les liens de référencement. Cette approche permet notamment d'annoter directement une ressource *sans connaissance préalable de son contenu* selon un regroupement thématique construit à partir d'un classifieur flou non-supervisé.

Le reste de l'article est organisé de la manière suivante. La section 2 présente le contexte de ce travail et les objectifs d'annotation de documents dans le cadre du SEMIDE. Dans la section 3 nous proposons un survol des travaux liés et nos motivations pour une nouvelle approche. La section 4 présente notre approche, et la section 5 les résultats de nos expérimentations. Enfin, une conclusion est proposée dans la section 6.

2. Contexte

L'objectif de cet article est de décrire un système d'annotation dans le cadre spécifique de partage et de diffusion de ressources dans le projet SEMIDE[1] (Système Euro-Méditerranéen d'Information sur les savoir-faire dans le Domaine de l'Eau). Ce système d'information est un instrument pour l'échange de connaissances en matière d'eau entre tous les pays du Partenariat Euro-Méditerranéen. Il s'agit d'un système réparti où l'information réside chez les fournisseurs et qui est donc par définition très variée et difficile d'accès. Un tel contexte impose une double annotation des ressources :

1) selon le contexte de création : nom des auteurs, date de création, lieu etc.

2) selon la sémantique du contenu

Le SEMIDE offre la possibilité de partager un grand volume de données et de ressources sur un ensemble d'individus ou d'organisations. Ces informations disponibles n'existent cependant que de façons segmentées et hétérogènes. C'est pourquoi il est apparu nécessaire d'engager un effort de rationalisation et de lisibilité de l'information. Les travaux présentés dans cet article sont réalisés dans ce contexte et nous définissons une nouvelle approche permettant aux experts d'annoter semi-automatiquement un grand volume de ressources en se basant sur les liens de références et de citations existants. Les informations du projet SEMIDE n'étant pas actuellement disponibles, nous avons choisi d'illustrer notre proposition à l'aide de données ayant des caractéristiques similaires. Notre choix s'est donc porté sur la base CiteSeer décrite section 5.

3. Etat de l'art

De façon générale, deux types d'annotations peuvent être définies *(i)* l'annotation de document en utilisant son contenu, *(ii)*l'annotation des documents en utilisant le contexte. C'est à ce dernier type d'annotation que nous nous intéressons 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.

3.1. *Annotation des documents en utilisant le contexte de citation*

La bibliométrie [LAU 97] est l'application de méthodes statistiques ou mathématiques sur des ensembles de références bibliographiques. La bibliométrie est donc une mesure auquel on fait appel pour aider à la comparaison et à la compréhension d'un ensemble d'éléments bibliographiques.
Le domaine de la bibliométrie s'intéresse à la citation entre les documents : *l'analyse des citations*. L'analyse des citations examine les relations entre les auteurs et les pu-

1. http ://www.semide.org

blications, ainsi qu'à d'autres types de relations qui peuvent exister, la co-citation et la co-référence.
La figure 1 illustre les différentes relations de citation 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 ou cet auteur a été cité ;

– couplage bibliographique : Kessler a eu en premier l'idée d'utiliser les citations comme relation entre les documents scientifiques. Kessler [KES 65] a utilisé l'analyse des citations de manière évoluée en élaborant une méthode d'analyse bibliométrique par association bibliographique. Le principe est que si deux articles qui citent un ou plusieurs documents communs, alors ces articles ont une relation significative avec une force d'association traduite par le nombre d'articles en commun. Les articles sont couplés s'ils partagent au moins une référence bibliographique en commun ;

– La méthode de co-citation [GAR 93], 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 dates quelconques, fréquemment citées ensemble, ont une parité thématique.

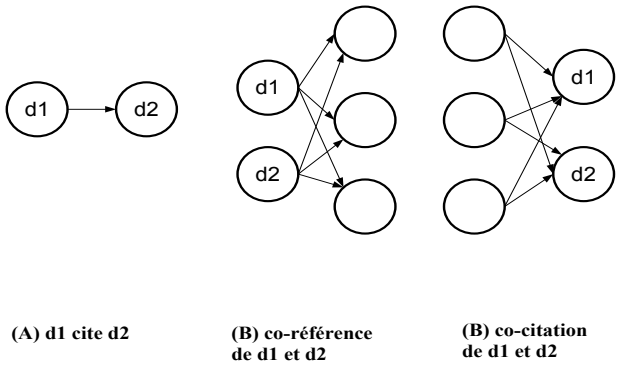

Figure 1. *Les relations entre les documents*

3.2. *Méthode de Marchiori*

Marchiori [MAR 98], a été le premier à s'intéresser à la propagation des métadonnées[2], son approche permet de propager des mots clés en utilisant la structure du Web. Ces mots clés sont pondérés par un coefficient entre 0 et 1.

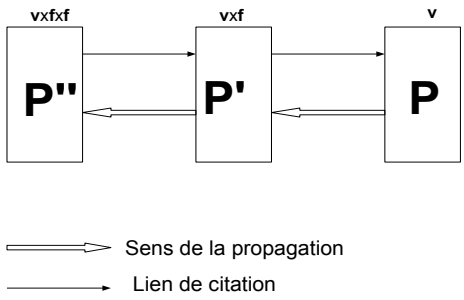

Figure 2. *Propagation de métadonnées (Marchiori)*

L'hypothése de Marchiori est la suivante : si une ressource R du Web posséde des métadonnées (mots clés) associées $A : v$, avec 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 est également accessible par R', car il existe un lien entre les deux ressources.

La pertinence de R' pour le mot clé A n'est pas identique à $R(v)$, puisque l'information pour R est seulement potentiellement accessible de R', mais n'est pas directement contenue dans R'. Pour résoudre ce problème, il suffit d'atténuer 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 $A : v \times f$. Si il existe une autre ressource R'' avec un lien vers R', nous pouvons propager les métadonnées $A : v \times f$ exactement de la même manière ; l'ensemble de métadonnées associé sera alors $A : v \times f \times f$.

3.3. *Méthode de Prime*

L'affectation des métadonnées aux pages est une tâche difficile, c'est pour cela que [PRI 04] se sont intéressés à l'attribution de ces métadonnées en les propageant dans

2. Les métadonnées sont des "données sur des données". Il s'agit d'un ensemble standard et structuré d'informations, traitable par un logiciel, décrivant une ressource sur support électronique ou papier.

le graphe du Web. Tout comme Marchiori, ils se basent sur l'hypothèse que si une page P contient un lien vers une autre page P' alors ces pages partagent des métadonnées communes. Contrairement à Marchiori, l'approche ne s'applique pas sur le graphe du Web mais sur un graphe construit avec la méthode de *co-sitation*. La relation de *co-sitation* est établie entre deux pages p_1 et p_2 si elles sont citées par une autre page sur un site différent de p_1 et p_2. Pour cela, les auteurs considérant deux étapes principales :

– La construction d'un corpus (un corpus est un ensemble de documents, regroupés dans une but précis) avec la méthode de *co-sitation* afin d'obtenir des classes partageant des propriétés. Dans [PRI 04] le corpus est mono langue et mono thématique. ensuite un travail manuel d'attribution de métadonnées est effectué sur un ensemble de documents ;

– la propagation des métadonnées dans les sous corpus

La construction du corpus se fait par : *(i)*le calcul de la matrice de *co-sitation* et la similarité entre les pages à partir de la matrice, *(ii).* le regroupement des pages en classes avec une méthode de classification hiérarchique ascendante.
Pour la propagation des annotations, deux méthodes sont utilisées : *(i)* la première méthode consiste à choisir les deux pages les plus éloignées dans la classe et partageant les mêmes métadonnées aux autres pages, *(ii)* la deuxième méthode consiste à prendre la page la plus proche de toutes les autres et à propager ses métadonnées sur les autres pages. Dans ce cas, cette page est considérée comme la page représentative de la classe.

3.4. *Méthode de Propagation de signatures lexicales*

Cette méthode [M.B 06] a pour but de propager des signatures lexicales dans le graphe du Web. *(i),* la *signature lexicale* $S(p)$ est un ensemble de termes pondérés décrivant une page, *(ii),* la *signature interne* $I(p)$ d'une page p est la signature lexicale que souhaite donner l'auteur à la page, *(iii),* la *signature externe* (p) d'une page p est la signature lexicale perçue par les auteurs des pages qui pointent la page p.
L'idée ce cette approche est de calculer la signature externe d'une page à partir des précédentes signatures internes des pages qui la pointent sur cette page (*propagation avant*). De la même façon, le calcul de la signature interne d'une page p est fait à partir de son contenu et des précédentes signatures externes des pages qu'elle pointe (*propagation arrière*).

3.5. *Autres utilisations des liens dans le Web*

– La classification des pages Web est un exemple de l'utilisation de l'analyse des liens afin de retrouver les pages les plus importantes, les deux algorithmes les plus connus de classement de pages sont l'algorithme *Page Rank* [ARA 01], [BRI 98], [MIH 04] et l'algorithme *HITS*[KLE 99].

– on peut aussi citer d'autres travaux utilisant les liens comme : *(i)* la catégorisation des pages, *(ii)* la portée géographique d'un document[THI 04], [BUY 99], *(iii)* la découverte de pages similaires, par exemple dans [PHE 02]. Ces auteurs utilisent les liens afin de calculer la similarité entre deux pages hypertextes, ainsi que *(iv)* la découverte de communauté dans le Web[GIB 98], [KUM 99], [VAN 04].

4. Approche : Propagation d'annotations

Rappelons que le but de notre travail est de présenter une méthode pour l'annotation automatique des articles en utilisant les relations de référencement. Dans cette section, nous détaillons les différentes étapes de cette approche. Nous supposons disposer d'une base qui contient des articles ainsi que leurs relations de référencement. Nous supposons également qu'une partie seulement de ces articles sont déjà annotés. Le problème consiste alors à annoter un nouveau document qui doit être ajouté à cette base.

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 dans un ensemble noté Ref_d ;

2) regrouper thématiquement 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.

Dans la suite nous développons plus en détail ces différentes étapes afin d'obtenir une annotation automatique des documents sans accéder directement à leur contenu.

4.1. *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 [GAR 93]. Par la suite nous utilisons cette caractéristique pour déterminer le thème d'un document sans avoir accès à son contenu mais en utilisant simplement les références de celui-ci.

Toutes les références d'un document ne sont cependant pas pertinentes pour la détermination du thème du document citant. En effet, un document peut aborder certains aspects mineurs et les références qui sont utilisées dans ces aspects mineurs ne devront pas être prises en compte pour l'importation des annotations. Dans ce contexte, il est alors important de retrouver les sujets les plus importants abordés par le document et d'ignorer les sujets les moins importants.

Afin de déterminer les thèmes les plus importants dans l'ensemble des références d'un document, nous utilisons l'hypothèse de la co-citation qui est une mesure importante dans le domaine de la bibliométrie [H.R 96]. Selon cette hypothèse, si deux documents sont souvent citées ensemble alors ils sont thématiquement proches. Par exemple, Prime [PRI 04] expose 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-citation 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 à l'origine pour normaliser la fraction et ainsi rendre le résultat dans l'intervalle $[0, 1]$. En effet, les documents i et j sont indépendants et peuvent par exemple apparaître à des périodes différentes (si on suppose que le document i est plus ancien que j). Dans ce cas, le document i peut être cité plusieurs fois et on aura 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 on suppose qu'à chaque fois que le document j est cité, le document i est cité dans le même document, on s'attend une proximité thématique. 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.

Afin de résoudre ce problème nous proposons d'utiliser une autre définition pour la fonction de distance. Cette définition utilise l'indice de co-citation :

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

L'équation 2 prend en compte simplement l'indice de co-citation entre deux documents afin de déterminer leur proximité thématique. Ainsi, plus deux documents sont cités ensembles, plus la distance $S_{(i,j)}$ sera proche du zéro.

Dés que les références d'un document d sont récupérées, nous construisons le graphe de citation GC_d :

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

Tel que décrit dans l'équation 4, le graphe de citation est un graphe complet où les nœuds représentent les documents cités dans d, et 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 citation, MC, définie comme suit :

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

Á partir de cette matrice, nous pouvons rechercher les groupements (clusters) de documents proches. Pour cela nous utilisons un algorithme de groupement 'fuzzy c-means' [DUN 74] qui utilise la théorie des ensembles flous . Pour utiliser cette matrice comme entrée à l'algorithme 'fuzzy c-means' il faut veiller à ce que les valuations des liens définissent bien une distance au sens mathématique. En effet, comme les documents sont indépendants et que le calcul de $S_{(i,j)}$ ne prend en compte que l'indice de cocitation de ces documents, la spécification d'une distance au sens mathématique peut ne pas être satisfaite. En effet, on peut très bien avoir un graphe de citation et une matrice de citation correspondante 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. Dans ce cas, le graphe de citation ne présente pas une distance mathématique et l'utilisation de l'algorithme 'fuzzy c-means' ne sera pas appropriée. Pour résoudre ce problème nous transformons la matrice de citation pour que la distance entre deux documents i et j soit minimale. On utilise pour cela l'algorithme Dijkstra [DIJ 59] afin de déterminer la distance minimale entre deux document i et j :

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

Dans ce cas, MC'_d définit bien un espace métrique car :

– la propriété de symétrie est satisfaite, car S est une fonction symétrique.

– S ne peut pas valoir zéro et par définition $MC'_d(i,j)$ vaut zéro quand $i \neq j$.

– l'inégalité triangulaire est satisfaite en utilisant Dijkstra.

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.

4.2. *Importation des annotations*

Le but dans cette partie est d'importer et de présenter de façon pertinente les annotations des documents cités par un document d. La présentation des annotations importées est faite en définissant un choix multi-critéres pour sélectionner des annotations à utiliser dans la phase suivante.

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

$$\text{annotation\_list}_d = \bigcup_{i \in Ref_d} \text{Annotation}(i) \tag{6}$$

La fonction Annotation(i) récupère l'ensemble des annotations du document i. 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 une liste ou bien comme un multi-ensemble, c'est à dire, un ensemble avec une répétition des éléments.

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

$$\text{annotation\_set}_d = \bigcup_{x \in \text{annotation\_list}_d} \{x\} \tag{7}$$

Il s'agit maintenant de doter cet ensemble d'un ordre multi-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 document, on considérera l'importance du cluster maximale.

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 la liste 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, on détermine 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.

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.

Pour le troisième critère d'ordre, il s'agit simplement de calculer la répétition de l'annotation dans la liste annotation_list$_d$.

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

1) Les annotations qui proviennent des documents qui sont 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. Ce qui revient à dire que l'importance de l'annotation d'un document dépend de l'importance du document dans le cluster ;

3) Si, les annotations proviennent d'un même document, ou bien de documents qui ont le même degré d'appartenance au même cluster alors nous considérons comme importantes les annotations redondantes.

5. Expérimentations et résultats

Nous sommes partis du constat que les documents du SEMIDE sont techniques et référencent trés souvent d'autres documents de la base du SEMIDE. N'ayant pas encore un corpus complet, nous avons utilisé un autre corpus de tests qui représente notre problématique.

Dans le cadre de nos expérimentations nous avons choisi comme collection de tests la base de CiteSeer[3]. CiteSeer[STR 05],[GHI 05] est une bibliothèque numérique sur la littérature scientifique. CiteSeer localise les articles scientifiques sur le Web, extrait différentes informations tels que les citations , le titre des articles ,etc. 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. Nous avons construit une base qui contient plus de 550 000 documents.

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 un vocabulaire contrôlé ou une ontologie. Nous avons utilisé l'ontologie DMOZ afin de décrire les documents. Notre approche a consisté à nous servir du moteur de recherche de DMOZ[4] afin de créer la correspondance entre les mots clés de Citesser et l'ontologie du domaine.

L'approche présentée a été implémentée et nous a servi à annoter des documents à partir de documents déjà annotés. L'expérimentation nous a montré que l'on pouvait annoter des documents sans leur contenu et elle peut aussi servir à affiner une indexation déjà faite sur les documents. Aussi, pour un concept x, qui a été sélectionné lors

3. http ://citeseer.ist.psu.edu/
4. http ://www.dmoz.org/

de l'annotation, ses ancêtres dans l'arbre de l'ontologie seront aussi ajoutés à l'annotation à améliorer la recherche.

Afin de mieux expliquer nos différentes étapes, nous illustrons notre approche par un exemple, dans la suite de cette section. La première étape de nos expérimentations a consisté à calculer la matrice de cocitations de tous les documents de la base. Cette étape a duré plus de 30 jours.

Dans notre cas, nous voudrions annoter le document d qui a le titre "Optimizing ML with Run-Time Code Generation". L'étape suivante de l'approche est de sélectionner les références du document d, et calculer la matrice de distances à partir de la matrice de cocitation. L'algorithme de classification est ensuite appliqué sur les références à partir de la matrice des distances.

La figure 3 illustre le résultat de l'étape de classification sur le document d. 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 clusters à 3.

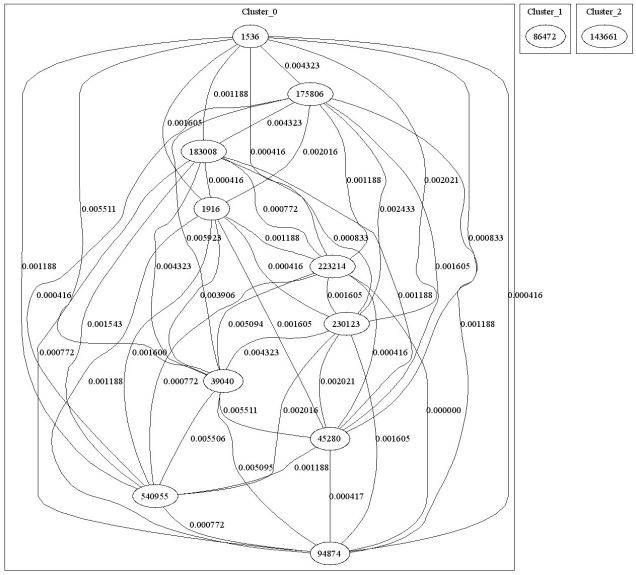

Figure 3. *Résultat de l'étape de classification des références*

On remarque ici qu'il existe deux références qui sont classées à part. Ces deux documents ne serviront pas à l'annotation du document d. Cette étape consiste au regroupement thématique des références.

Enfin, la dernière étape consiste à importer des annotations des références. La figure 4 illustre le résultat par notre système de l'annotation du document d. Le rank

correspond au rang des annotations des documents références basés sur les critères décrits dans la section 4.2.

Rank	Annotation	Include?
10	/Programming/Languages/ML	☑
10	/Programming/Languages/POP-11	☑
10	/Programming/Languages/Oz	☑
10	/Programming/Languages/Functional	☑
10	/Programming/Languages/Ada/Books	☑
10	/Programming/Languages/Erlang/News_and_Media/Theses	☑
10	/Programming/Languages/C++/Books	☑
10	/Algorithms/Computational_Algebra/Research_Groups	☑
10	/Programming/Languages/Java/Extensions	☑
10	/Programming/Methodologies/Patterns_and_Anti-Patterns	☑
10	/Programming/Languages/Miranda	☑
10	/Programming/Metaprogramming	☑
10	/Programming/Languages/Lisp/Scheme/Software	☑
10	/Artificial_Intelligence/Machine_Learning/Software	☑
10	/Programming/Languages/Forth/Implementations	☑
1	/Parallel_Computing/Programming/Languages	☐

Figure 4. *Résultat de l'annotation*

Les premiers résultats sont apparus satisfaisants. Pour le moment, notre méthode d'évaluation est basée sur le jugement d'experts du domaine, en comparant l'annotation de notre système avec celle de l'expert.

6. Conclusion et perspectives

Cet article a décrit notre système d'annotation de documents dans un domaine spécialisé (SEMIDE). L'annotation manuelle de documents est une tâche difficile voir impossible à faire compte tenu du temps que cela nécessite. Nous avons utilisé la relation de citation d'un document avec les autres documents de la base afin de l'annoter en se basant sur une ontologie du domaine sans avoir besoin de son contenu. 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 cocitations. Les expérimentations ont été effectuées sur la base CiteSeer. Ce choix a été motivé par la taille de la base et les similarités structurelle avec la base du SEMIDE (des documents qui s'inter référencent). Les premiers résultats ba-

sés sur l'évaluation d'experts sont trés satisfaisants. Les perspectives associées à ce travail sont les suivantes. Nous pensons tout d'abord étendre notre évaluation en utilisant notre annotation pour la recherche de documents et puis analyser la pertinence du document résultat correspondant à l'annotation.

7. Bibliographie

[ARA 01] ARASU A., NOVAK J., TOMKINS A., TOMLIN J., « PageRank Computation and the Structure of the Web : Experiments and Algorithms », 2001.

[BRI 98] BRIN S., PAGE L., « The anatomy of a large-scale hypertextual Web search engine », *Proceedings of the seventh international conference on World Wide Web 7*, Australia, 1998, p. 107-117.

[BUY 99] BUYUKKOKTEN O., CHO J., GARCIA-MOLINA H., GRAVANO L., SHIVAKUMAR N., « Exploiting Geographical Location Information of Web Pages », *WebDB (Informal Proceedings)*, 1999.

[DIJ 59] DIJKSTRA E. W., « A Note on Two Problems in Connexion with Graphs. », *Numerische Mathematik*, vol. 1, 1959, p. 269–271.

[DUN 74] DUNN J. C., « A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters », *Journal of Cybernetics*, vol. 3, 1974, p. 32-57.

[GAR 93] GARFIELD E., « 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*, vol. 15, n^o 19, 1993.

[GHI 05] GHITA S., HENZE N., NEJDL W., « Task specific semantic views : Extracting and integrating contextual metadata from the web », *In Submitted for publication, L3S Technical Report*, 2005.

[GIB 98] GIBSON D., KLEINBERG J., RAGHAVAN P., « Inferring Web Communities from Link Topology », *UK Conference on Hypertext*, 1998.

[H.R 96] H.ROSTAING, Ed., *La bibliométrie et ses techniques*, Sciences de la Société , Collection, 1996.

[KES 65] KESSLER M., « Comparison of the results of bibliographic coupling and analytic subject indexing », *American documentation*, vol. 14, 1965, p. 10–15.

[KLE 99] KLEINBERG J. M., « Authoritative Sources in a Hyperlinked Environment », *Journal of the ACM*, 1999, p. 139-146.

[KUM 99] KUMAR R., RAGHAVAN P., RAJAGOPALAN S., TOMKINS A., « Trawling the Web for Emerging Cyber-Communities », *Computer Networks*, Amsterdam, Netherlands, 1999.

[LAU 97] LAURI P., « The bibliometrics, a trend indicator », *International Journal Information Sciences for Decision Making*, 1997, p. 28-36.

[MAR 98] MARCHIORI M., « The limits of Web metadata, and beyond », *Proceedings of the Seventh International World Wide Web Conference*, Australia, 1998, p. 1–9.

[M.B 06] M.BOUKLIT, M.LAFOURCADE, « Propagation de signatures lexicales dans le graphe du Web », *RFIA 2006, 15e congrès francophone AFRIF-AFIA, Reconnaissance des Formes et Intelligence Artificielle*, Janvier 2006.

[MIH 04] MIHALCEA R., TARAU P., FIGA E., « PageRank on Semantic Networks, with Application toWord Sense Disambiguation », *Proceedings of the 20th international conference on computational linguistics (COLING2004)*, Geneva, Switzerland, 2004.

[PHE 02] PHELAN D., KUSHMERICK N., « A descendant-based link analysis algorithm for Web search », 2002.

[PRI 04] PRIME-CLAVERIE C., « 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 », PhD thesis, Ecole supérieure des Mines de Saint-Etienne, 2004.

[STR 05] STRIBLING J., COUNCILL I. G., LI J., KAASHOEK M. F., KARGER D. R., MORRIS R., SHENKER S., « OverCite : A Cooperative Digital Research Library », *International Workshop on Peer-to-Peer Systems*, 2005.

[THI 04] THILLIEZ M., DELOT T., « Evaluation de requêtes dépendantes de la localisation dans les réseaux mobiles », *Premières Journées Francophones : Mobilité et Ubiquité 2004*, Nice, Juin 2004.

[VAN 04] VANDAELE V., FRANCQ P., DELCHAMBRE A., « Analyse d'hyperliens en vue d'une meilleure description des profils », *Proceedings of JADT 2004, 7es Journées internationales d'Analyse statistique de Données Textuelles*, 2004.