Approche par réutilisation d'annotations sémantiques pour la recherche d'information sur le web

**Wiem YAICHE ELLEUCH*, Lobna JERIBI **,
Mohamed TMAR*, Abdelmajid BEN HAMADOU ***

** MIRACL (Multimedia InfoRmation system and Advanced Computing Laboratory)*
Université de Sfax. Institut Supérieur d'Informatique et de Multimédia de Sfax
Route de Tunis, Km 10, cité el ons 3021 BP 242, Tunisie
{Wiem.Yaiche, Mohamed.Tmar, Abdelmajid.BenHamadou}@isimsf.rnu.tn
*** RIADI-GDL (Génies Documentiel et Logiciel)*
Université de la Manouba. École Nationale des Sciences de l'Informatique
Campus Universitaire la Manouba
{ici@gnet.tn}

RÉSUMÉ. *Dans cet article, nous présentons une nouvelle approche d'aide à la recherche d'information sur le web. Elle a pour objectif de présenter à l'utilisateur courant des documents réponses pertinents pour sa requête et adaptés à son profil. Elle consiste à utiliser le mécanisme du Raisonnement à Partir de Cas (RàPC) pour mémoriser les sessions de recherche effectuées par les utilisateurs (profil utilisateur, requête, annotation d'un document pertinent, date) et à les réutiliser lorsqu'une session de recherche similaire se présente. La réutilisation des annotations au cours d'une session courante, permet en outre de reformuler automatiquement la requête courante en vue d'améliorer la qualité des réponses pour la session courante. Nos propositions ont été validées et testées par le développement du système SYRANNOT implémenté en java, utilisant l'infrastructure JENA (hp) et se servant de Google. Les premières évaluations expérimentales montrent une nette amélioration des résultats proposés par notre système relativement à google.*

ABSTRACT. *In this paper, we present a new approach for information retrieval assistance on the web. Its objective is to present to the current user relevant retrieved documents for his query and adapted to his profile. Its consists on using the mechanism of the Case Based Reasoning (CBR) in order to memorize research sessions carried out by users (user profile, query, semantic annotations of relevant documents, date), then to reuse them when a similar research session arises. The reuse annotations during a current session, can also automatically reformulate the current query for improving the quality of responses to the current session. Our proposals have been tested and validated by the development of the system SYRANNOT implemented in Java, using the infrastructure JENA (hp) and using Google. The first experimental evaluations show a significant improvement of results offered by our system compared to google*

MOTS-CLÉS : *RàPC, annotation sémantique, ontologie du domaine, recherche d'information, web.*

KEYWORDS: *CBR, semantic annotation, ontology domain, information retrieval, web.*

1. Introduction

Les problèmes principaux auxquels sont confrontés les utilisateurs au cours d'une session de recherche sur le web sont le bruit et le silence. Ces phénomènes sont essentiellement dus à la diversité et au volume du web, à la formulation de la requête qui ne reflète pas toujours les besoins réels en information des utilisateurs et à l'efficacité des moteurs de recherche (constitution des bases d'index, fonctions d'appariement entre la requête et la base d'index, …). Pour palier ces problèmes, plusieurs orientations de recherche ont été suivies. Nous citons notamment : (i) la prise en compte du profil utilisateur en modélisant ses intentions, ses spécificités cognitives et culturelles (Gaussier 2003), (ii) l'aide à la reformulation de la requête (Schenkel 2005) (iii) l'ajout d'annotations aux documents en vue de décrire leurs contenus sémantiques (Popov 2003) (Handschuh 2002).

Les travaux de recherche que nous proposons dans cet article combinent ces différentes orientations. Ils se basent sur les trois éléments suivants:

(i) Réutilisation d'expériences en utilisant un mécanisme de Raisonnement à partir de cas (RàPC) sur les sessions de recherche déjà effectuées. Une session de recherche est caractérisée par les éléments suivants : profil utilisateur, requête, annotations des documents réponses jugés pertinents, date de la session.
(ii) Annotation sémantique des documents du web pour décrire leurs contenus.
(iii) Reformulation automatique de la requête en se basant sur un *descripteur de pertinence (Dp)* construit automatiquement.

Dans ce papier, nous commençons par présenter un état de l'art sur les travaux utilisant les annotations sémantiques et le RàPC en RI. Dans la section 3, nous présentons l'approche proposée à partir de la quelle nous détaillons les modélisations des connaissances utilisées (section 4). Ensuite, nous présentons le processus de réutilisation d'annotations (section 5) et la reformulation de la requête (section 6). Enfin, nous présentons le prototype SyRANNOT et les premiers tests.

2. Bref aperçu de l'état de l'art

Le RàPC est une approche de résolution de problèmes basée sur la réutilisation par analogie d'expériences passées appelées **cas** (Aamodt 1994) (Kolodner 1993) (Schank 1989) (Lieber 1999). De nombreux systèmes de RI utilisant le RàPC ont été élaborés : RADIX (Corvaisier 1998), COSYDOR (Jéribi 2001), etc.

Par ailleurs, l'utilisation des annotations pour améliorer la qualité d'un système de recherche d'information a fait l'objet de plusieurs travaux et de réalisations. Nous citons en particulier les systèmes Annotea (Koivunen 2001), SHOE Knowledge annotator (Heflin 2000), CREAM (Handschuh 2002) et KIM (Popov 2003), etc. Ces systèmes offrent des architectures pour le stockage, la recherche, la consultation des annotations et la recherche de documents en se basant sur celles-ci. Cependant, l'intégration des annotations dans un système de RàPC n'a pas été, à notre

connaissance, suffisamment explorée dans le cadre de la recherche d'informations. L'approche que nous proposons s'inscrit dans cette perspective.

3. Présentation générale de l'approche

Nous décrivons l'approche proposée selon les deux scénarios : (i) alimentation de la base de cas, (ii) réutilisation de la base et reformulation de la requête (figure 1).

L'alimentation de la base de cas correspond à la mémorisation des annotations des documents pertinents des sessions de recherche. Le processus d'annotation se réfère à une *ontologie du domaine.* La réutilisation de la base de cas est effectuée pour la session de recherche courante. Elle consiste à chercher, dans la base, les sessions antérieures les « plus proches » de la courante (profils et requêtes similaires). Les résultats obtenus constituent une *première* liste de documents potentiellement pertinents pour l'utilisateur courant.

Les annotations relatives aux documents obtenus permettent de créer *un descripteur de pertinence (Dp)* constitué par l'union des concepts des annotations. Il servira pour reformuler automatiquement une nouvelle requête. Celle ci est soumise au moteur de recherche qui collecte de nouveaux documents réponse à partir du web pour enrichir la première liste et dans une étape ultérieure d'enrichir la base des cas.

L'approche proposée permet ainsi de combiner deux types de recherche pour une session courante : (i) une recherche dans un système fermé (recherche dans la base des cas) (ii) une recherche dans un système ouvert (recherche dans le web). Cette ouverture permet d'une part d'actualiser la base des cas d'une manière systématique et d'autre part d'améliorer le rappel. La figure 1 détaille cette approche.

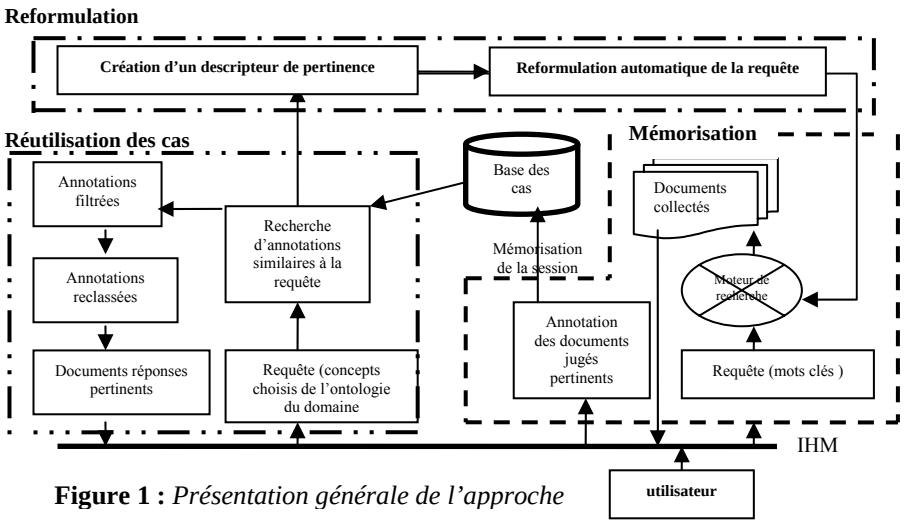

Figure 1 : *Présentation générale de l'approche*

4. Les connaissances utilisées

4.1. Ontologie du domaine

L'ontologie du domaine que nous proposons est constituée par des concepts qui représentent un domaine particulier. Ces concepts sont reliés entre eux par des relations de subsomption, formant un arbre de concepts. Notons que pour l'expérimentation de l'approche proposée, nous avons choisi *le domaine du web sémantique* (37 concepts). L'ontologie du domaine est utilisée comme référence pour : (i) annoter les documents jugés pertinents par l'utilisateur, (ii) formuler une requête de recherche dans la base des cas, (iii) calculer la similarité entre des sessions en vue d'identifier celles réutilisables pour une session courante.

4.2 *Profil utilisateur*

Le profil utilisateur permet de modéliser les connaissances relatives à un utilisateur. Un utilisateur possède un identifiant unique (PID), un nom, un prénom, un login et un mot de passe, et un ensemble de mots clés représentant le domaine qui l'intéresse, sélectionnés à partir de l'ontologie du domaine.

Une instance du profil utilisateur est représentée par un ensemble de déclarations RDF mémorisées dans une base de données dédiée aux profils utilisateur. Cette base sera utilisée pour le calcul de similarité entre deux profils utilisateurs.

4.3. Schéma d'annotation

Le schéma d'annotation regroupe les propriétés standardisées du Dublin Core tels que auteur, langue, URL etc, et des concepts descripteurs du document sélectionnés par l'utilisateur à partir de l'ontologie du domaine (figure 2).

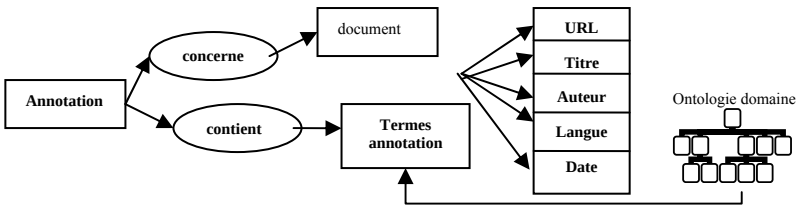

Figure 2: *Schéma des annotations sémantiques*

Une instance du schéma d'annotation pour un document donné est aussi représentée par un ensemble de déclarations RDF mémorisées dans une base de données dédiée aux annotations. Cette base permet de retrouver les cas réutilisables pour une session courante.

4.4 *Calcul de similarité sémantique dans l'ontologie du domaine*

Pour calculer la similarité entre deux ensembles de concepts dans l'ontologie du domaine, nous avons retenu la mesure de Wu Palmer qui s'applique bien au type de

représentation que nous avons choisi. La similarité est définie par rapport à la distance qui sépare deux ensembles de concepts dans la hiérarchie et par la distance qui les sépare du concept racine :

$$Sim(P1,P2)=\frac{1}{2}\left(\frac{1}{|P1|}\sum_{Ai\in P1}\max_{Bi\in P2}\left(ConSim(Ai,Bi)\right)+\frac{1}{|P2|}\sum_{Bi\in P2}\max_{Ai\in P1}\left(ConSim(Ai,Bi)\right)\right)$$

- ConSim est la fonction de Wu palmer de calcul de similarité entre deux concepts.
- P1 est un ensemble n concepts Ai ; P1={A1, A2,..,An}
- P2 est un ensemble p concepts Bi ; P2={B1, B2,..,Bp}

5. Processus de réutilisation d'annotation

5.1. *Principe du Raisonnement à Partir de Cas (RàPC)*

Le RàPC est une approche de résolution de problèmes basée sur la réutilisation par analogie d'expériences antérieures similaires mémorisées (Aamodt 1994) (Kolodner 1993). Dans ce qui suit, nous présentons les quatres étapes du cycle du RàPC (représentation, mémorisation, réutilisation et adaptation). L'étape de représentation d'un cas constitué du profil utilisateur et requête (partie problème) et de l'annotation d'un document pertinent (PartieSolution), a été déjà présentée.

5.2. *Mémorisation d'un cas (scénario 1)*

Un utilisateur, ayant un profil donné soumet une requête à un moteur de recherche. Lorsque celui-ci trouve un document qu'il juge pertinent il l'annote et crée un nouveau cas constitué par les éléments suivants : l'identifiant du profil, la requête soumise, l'identifiant de l'annotation et la date de la session. L'ensemble des cas d'une session de recherche est mémorisé dans la base de cas (figure 3).

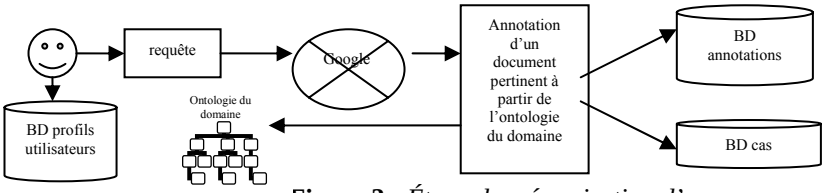

Figure 3 : *Étape de mémorisation d'un cas*

5.3. *Recherche des cas utiles à la réutilisation (scénario 2)*

Le système parcourt la base d'annotations et collecte toutes celles contenant dans le champ terme d'annotation, les concepts de la requête ou leurs pères ou leurs fils, et ce en utilisant l'ontologie du domaine. Les annotations retenues sont les *annotations candidates à la réutilisation*. Le système filtre ensuite ces annotations en appliquant la fonction de similarité de Wu Palmer entre la requête courante et les termes d'annotations de chaque annotation. Le filtrage des annotations candidates à

la réutilisation permet de retenir les plus pertinentes : *les annotations utiles à la réutilisation* (figure 4).

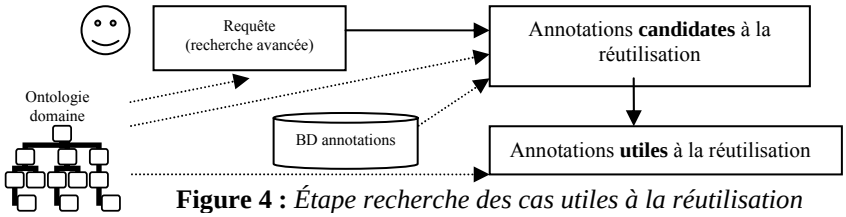

Figure 4 : *Étape recherche des cas utiles à la réutilisation*

5.4. *Réutilisation et Adaptation des cas utiles à la réutilisation*

Pour chacune des annotations utiles à la réutilisation, le système parcourt la base de cas pour déterminer l'identifiant du créateur de l'annotation. Ensuite, il va parcourir la base des profils, et à partir de chaque identifiant, il va déterminer les intérêts de chaque créateur d'annotation. Le système calcule ensuite la similarité entre le profil courant et ces profils antérieurs en vue de reclasser les annotations utiles à la réutilisation. Le système parcourt la base d'annotations et extrait les URLs à partir des annotations qu'il présente à l'utilisateur courant (figure 5).

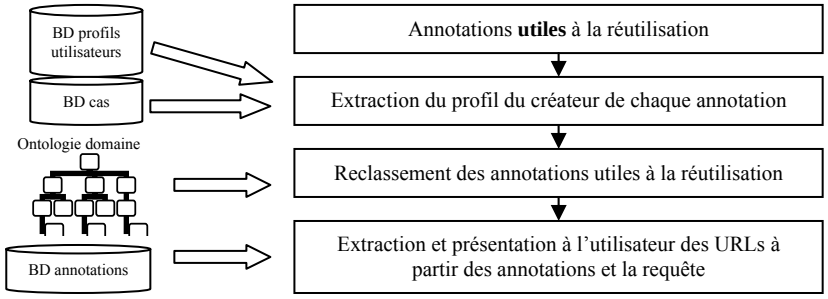

Figure 5 : *Étape réutilisation et adaptation des cas utiles à la réutilisation*

6. Reformulation automatique de la requête

Cette étape permet d'enrichir la liste des documents pertinents proposés à partir de la base de cas. Elle consiste à reformuler la requête de l'utilisateur en se basant sur le résultat obtenu. Ainsi, les annotations des documents pertinents fournis à l'utilisateur sont fusionnées pour constituer le *Descripteur de pertinence Dp*.

$$Dp = \bigcup_{\substack{i=1..n \\ j=1..p}} Cij$$

n: Nb d'annotations utiles, **p**: Nb de concepts de l'annotation i et **Cij**: le j^{ème} concept. Ce descripteur de pertinence constitue les éléments de la requête qui sera automatiquement soumise à un moteur de recherche (Figure 6). Les documents collectés par le moteur de recherche sont présentés à l'utilisateur courant pour choisir les pertinents, qu'il va annoter et mémoriser ces nouveaux cas selon le scénario 1 décrit précédemment (voir figure 3).

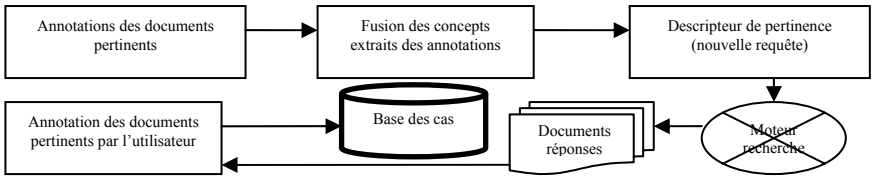

Figure 6 : *Processus de reformulation automatique de la requête*

7. Mise en œuvre et évaluation : le système SYRANNOT

Pour valider et expérimenter notre approche, nous avons développé un système baptisé SYRANNOT (**SY**stème d'aide à la recherche d'informations sur le web par **R**éutilisation d'**ANNOT**ations) (Yaiche 2005). Nous avons créé une cinquantaine de profils utilisateurs différents, une centaine d'annotations de documents scientifiques PDF extraits des actes de la conférence ISWC (2006 et 2007) et une centaine de cas.

La démarche d'évaluation consiste à comparer les résultats de la recherche fournis par SyRANNOT et Google pour le même besoin en information. Pour Google, la requête est librement formulée, alors que pour SyRANNOT elle est formulée à partir des concepts de l'ontologie du domaine et reformulée automatiquement en utilisant le Dp. L'étape suivante consiste à calculer séparément, pour les 100 premières réponses proposées par Google et SyRANNOT le nombre de documents pertinents et le nombre de documents non pertinents. Notre objectif est de déterminer pour chacun d'eux le *taux de précision* et *le taux de rappel*.

Les tests effectués nous ont permis de constater que les résultats fournis par SyRANNOT sont globalement de meilleure qualité que ceux fournis par Google. L'allure de la courbe de précision-rappel obtenue sur la base de 5 sessions de recherche le montre (figure 7). En effet, la moyenne des taux précision-rappel est de 0,51 pour SyRANNOT alors qu'elle est de 0,32 pour Google.

Nous remarquons que pour les quarante premiers documents de la liste, les taux rappel-précision se rapprochent de la valeur 1. Ceci s'explique par le fait qu'ils sont principalement restitués de la base des cas. Aussi, plus le nombre de documents considérés se rapproche de cent, plus les taux rappel-précision s'éloignent de la valeur 1. Ceci est du à l'ouverture progressive de SyRANNOT sur le web.

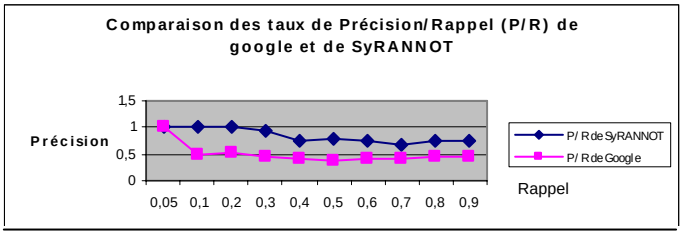

Figure 7 : *Courbes de Précision/Rappel de Google et de SyRANNOT*

Ainsi, ces études expérimentales nous ont permis de prouver l'impact de la tâche d'annotation et de la taille de la base de cas sur les performances de notre approche.

8. Conclusion et perspectives

Dans ce papier, nous avons présenté une nouvelle approche incrémentale d'aide à la recherche d'information sur le web. Elle présente à l'utilisateur courant des documents réponses pertinents et adaptés à son profil. Elle combine deux types de recherche : une recherche *fermée* par réutilisation des résultats de sessions antérieures et une autre *ouverte* sur le web qui enrichit, à chaque fois, la liste des documents pertinents et incrémente la base des cas de manière systématique.

Pour valider et expérimenter notre approche, nous avons développé le système SYRANNOT. Les premiers résultats obtenus montrent une amélioration encourageante de la qualité des réponses proposées à l'utilisateur. Les perspectives d'amélioration se situent au niveau de l'ontologie du domaine en introduisant de nouvelles relations comme la disjonction et l'équivalence et de nouvelles mesures de similarité.

9. Bibliographie

Aamodt, A., Plaza, E. "Case-Based Reasoning : Foundational Issues, Methodological Variations and System Approaches". March 1994, AI Communications, the Europeen journal on AI, 1994, Vol 7, N°1, p. 39-59.

Corvaisier F., Mille A., Pinon J.M. Radix 2, assistance à la recherche d'information documentaire sur le web. In IC'98, Ingénierie des Connaissances, Pont-à-Mousson, France, INRIA-LORIA, Nancy, 1998, p. 153-163.

Gaussier E., Stefanini MH., Assistante Intelligente à la recherche d'informations, Hermes Science, ISBN 2-7462-0726-5, 2003.

Handschuh S., Staab S.. "Authoring and Annotation of Web Pages in CREAM" WWW 2002

Heflin J., Hendler James."Searching the Web with SHOE", AAAI-2000 Workshop.

Jéribi, L. "Improving Information Retrieval Performance by Experience Reuse". Digital Publishing Odyssey' ELPUB2001. Canterbury, United Kingdom, 5-7 July 2001, p.78-92.

Kolodner, J. "Case based reasoning". San Mateo, CA: Morgan Kaufman, 1993.

Koivunen M., Swick R. "Metadata Based Annotation Infrastructure offers Flexibility and Extensibility for Collaborative Applications and Beyond", workshop on knowledge markup & semantic annotation. KCAP 2001

Lieber J., Napoli A., Raisonnement à partir de cas et résolution de problèmes dans une représentation par objets. Revue Intelligence Artificielle 13 : 9-35, 1999.

Popov B., Kiryakov A., Kirilov A., Manov D., Ognyanoff D., Goranov M. "KIM – Semantic Annotation Platform", 2nd International Semantic Web Conference (ISWC2003), 20-23 October 2003, Florida, USA. LNAI Vol. 2870, pp. 834-849, Springer-Verlag.

Schank R. C., Riesbeck C. K. "Inside Case Based Reasoning". Hillsdale, New Jersey, USA : Lawrence Erlbaum Associates Publishers, 1989, 423 p.

Schenkel R., Theobald M., « Relevance Feedback for Structural Query Expansion », *INEX 2005 Workshop Pre-Proceedings*, Germany, November 2005, p. 260,272.

Wu Z. & Palmer M. (1994) Verb Semantics and Lexical Selection, *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133-138.

Yaiche W. Jeribi L., Ben Hamadou A. "SYRANNOT: Information retrieval assistance system on the Web by semantic annotations re-use", First International Workshop On Open Source Web Information Retrieval OSWIR 2005.