

SYRANNOT: Information retrieval assistance system on the Web by semantic annotations re-use

Wiem YAICHE ELLEUCH¹, Lobna JERIBI², Abdelmajid BEN HAMADOU³,

^{1,3}LARIM, ISIMS, SFAX, TUNISIE

²RIADI GDL, ENSI, MANOUBA, TUNISIE

¹Wiem.Yaiche@isimsf.rnu.tn

²lj@gnet.tn

³Abdelmajid.BenHamadou@isimsf.rnu.tn

Abstract:

In this paper, SYRANNOT system implemented in java is presented. Relevant retrieved documents are given to the current user for his query and adapted to his profile. SYRANNOT is based on the mechanism of Case Based Reasoning (CBR). It memorizes the research sessions (user profile, query, annotation, session date) carried out by users, and re-use them when a similar research session arises. The first experimental evaluation carried out on SYRANNOT has shown very encouraging results.

1. Introduction

The Case Based Reasoning is a problem resolution approach based on the re-use by analogy of previous experiments called cases [AAM 94][KOL 93][SCH 89]. Some works of research assistance systems based on CBR were carried out: RADIX [COR 98], CABRI [SMA 99], COSYDOR [JER 01]. Our approach consists in applying CBR on the semantic annotations coming out of the semantic Web domain. The CBR has various advantages (information transfer between situations, evolutionary systems, etc). Nevertheless, its integration presents some difficulties such as the representation, memorizing, re-use and adaptation of the cases. These four key words constitute the CBR cycle and are the subject of our study. In the following, SYRANNOT system architecture is presented. It integrates the CBR on the semantic annotations. Special attention is given to knowledge modelling of the reasoning, as well as the search algorithms and the similarities calculation functions, in each stage of the cycle of the CBR.

2. SYRANNOT Architecture

We propose two scenarios of SYRANNOT use: the first is related to memorizing session research (cases) carried out by the user in RDF data base. Research sessions are RDF statements based on ontologies models in OWL language. The second concerns re-use cases by applying research algorithms and similarity functions to collect the most similar cases to the current one, and to exploit them in order to present to the current user relevant retrieved documents for its

query and adapted to its profile. In the following, both scenarios processes are detailed.

2.1 Cases memorizing scenario

A user having a given profile, memorized in the user profiles RDF data base in the form of RDF statements based on user ontology, expresses his need of information by formulating a query which he submits to the search engine. It collects and presents to the user the retrieved documents. When the current user finds a document which he considers relevant to his query, he annotates it. The annotation created is memorized in the RDF data base of the annotations in the form of RDF statements based on ontology annotation. The research session (user profile identifier, submitted query, annotation identifier, session date) is memorized in the RDF data base of cases in the form of RDF statements based on the cases ontology. Figure 1 presents the scenario proposed to memorize cases.

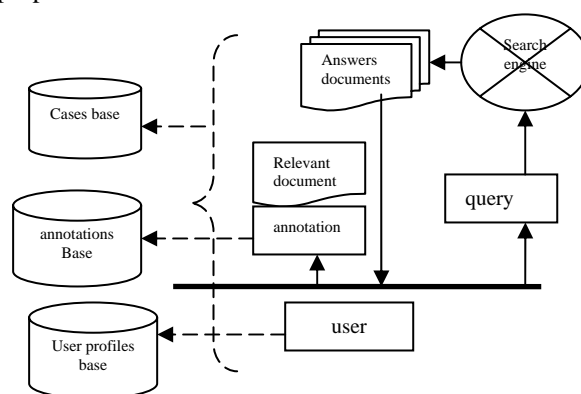


Figure 1 : Scenario of memorizing case

The case memorizing scenario is illustrated by the interfaces figures 2, 4, 5 and 6. Figure 2 shows the user new inscription interface. The user having a single identifier (PID) assigned by the system fills in his name (*yaiche*), his first name (*wiem*), his login (*wiem*), his password (******) and a set of interests which he selects from the ontology domain (*Case based reasoning, annotation*). The domain ontology is organised in a concepts tree. The user profile created is memorized in RDF data base of user profiles in the form of RDF statements, based on the user ontology

modelled in OWL (figure 3). The RDF data base of user profiles will be re-used later.

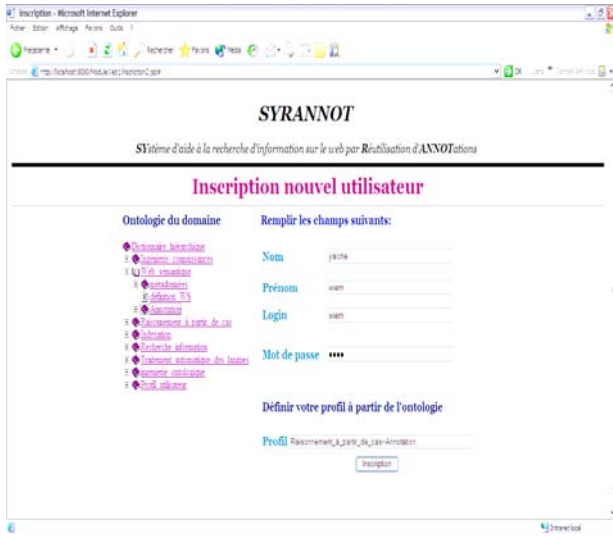


Figure 2: User new inscription interface

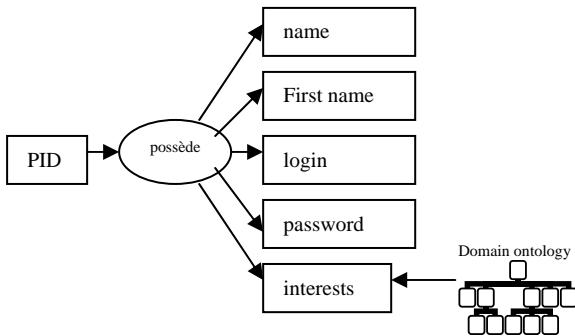


Figure 3 : User ontology diagram

Figure 4 shows the SYRANNOT home page. The enrichment of the cases data base consists in submitting queries (*semantic Web*) on the google search engine, collecting retrieved documents, and annotating the relevant documents for the query.



Figure 4: Scenario choice interface

Figure 5 shows the answers collected by google for the submitted query. It is a list of URL, each one is

preceded by an icon. When the user finds a document which he considers relevant to his query, he annotates it by clicking on the icon.

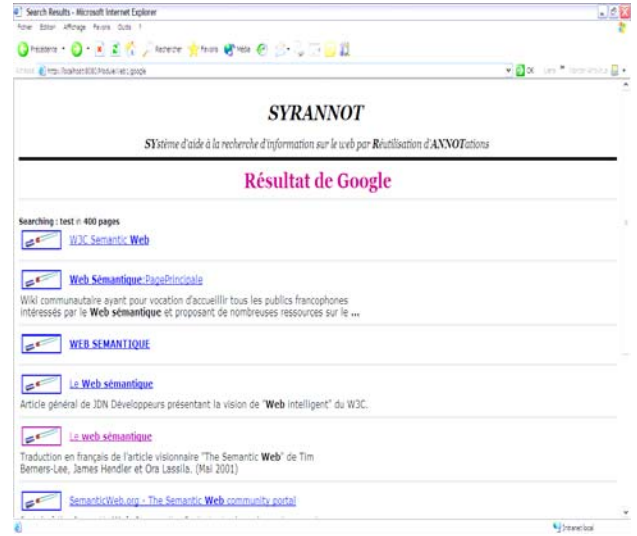


Figure 5: Google retrieved documents for a query

Figure 6 shows the interface which permits to annotate a document considered by the user to be relevant to his query. The annotation consists on the one hand in determining the standardized properties of Dublin Core such as URL (*http://www.scientificamerican...*), the title (*the semantic Web*), the author (*Tim Berners-Lee, James Hendler, Ora Lassila*), the date (*May 2001*) and the language of the document (*English*), and on the other hand to select a set of concepts from the ontology domain in order to describe the document according to the user point of view (*semantic Web definition, ontology definition, annotation definition*).

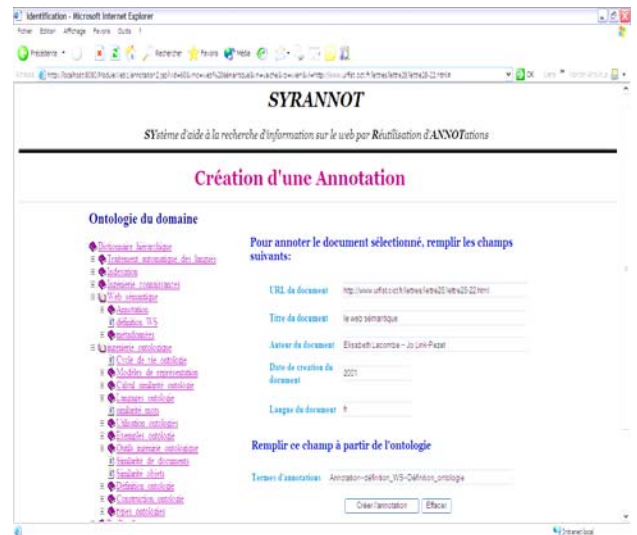


Figure 6: Annotation creation interface

The annotation created has a single identifier (AID) assigned by the system, and is memorized in the RDF data base of annotations, in the form of RDF statements based on ontology annotation (figure 7).

The RDF data base of annotations will be re-used later.

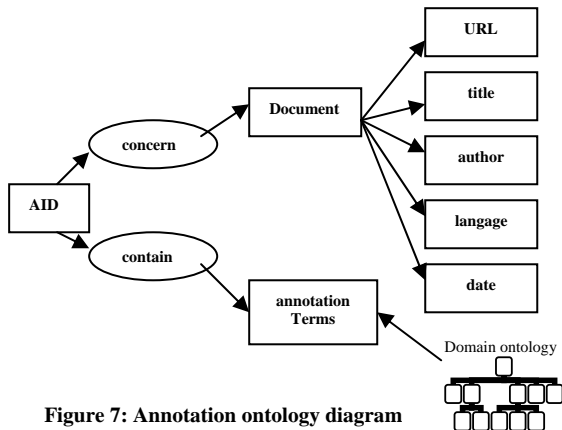


Figure 7: Annotation ontology diagram

The research session (the user identifier, the submitted query, the annotation identifier, session date) is memorized in the RDF data base of cases, in the form of RDF statements based on the ontology cases (figure 8).

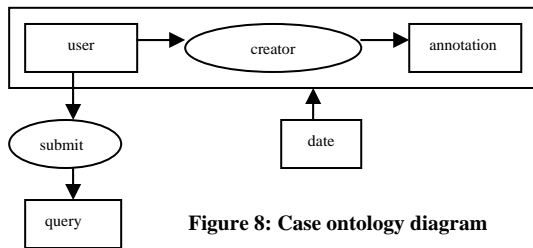


Figure 8: Case ontology diagram

The scenario presented above corresponds to the stages of representation and memorizing of the CBR cycle.

2.2 Cases re-use Scenario

The current user, having a given profile memorized in the RDF data base of user profiles, formulates his query by selecting one or more concepts from the domain ontology. The system scans the RDF data base of annotations and collects those having at least one concept of the query in the annotations terms field. The system filters these annotations by calculating the similarity between the current query and the annotation terms of each annotation in order to retain the 20 most relevant annotations. Then, the system reclassifies them by calculating the similarity between the current user profile and the profile of the user who has created the annotation. Finally, the system extracts and presents to the current user some information about the relevant documents (URL, author, date, etc). Figure 9 illustrates the cases re-use scenario.

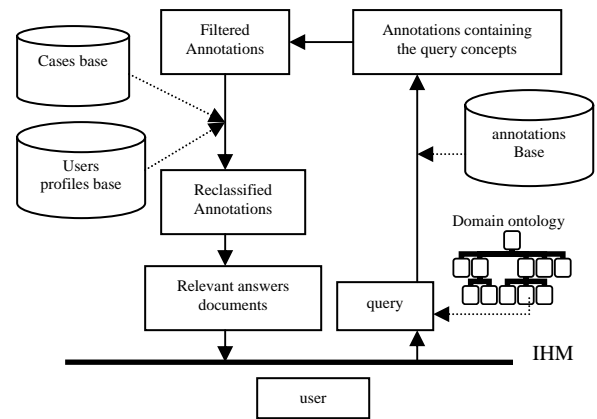


Figure 9: Cases re-use scenario

In figure 4, the link *recherche sur SYRANNOT* permits the current user to have retrieved documents from previous similar experiments (similar profiles, similar queries).

Figure 10 presents the query formulation interface which allows the user to interrogate the memorized cases via SYRANNOT. The user expresses his need of information by selecting concepts from the domain ontology (*semantic Web definition*). He can also make an advanced research on the author, the date or the language of the document.

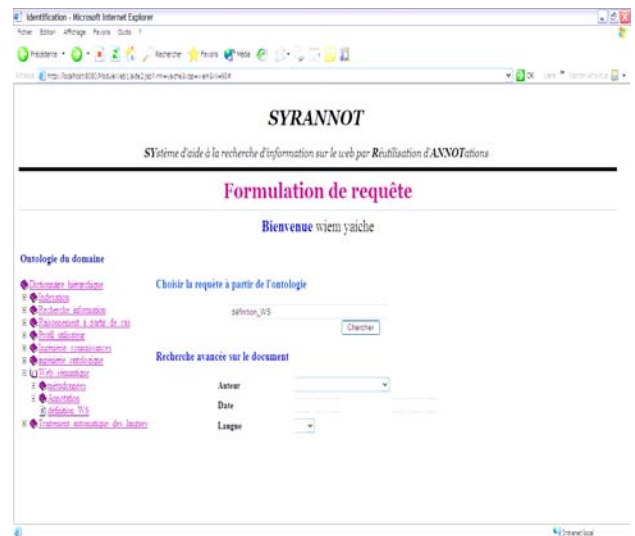


Figure 10 : Query formulation interface

SYRANNOT scans the RDF data base of annotations and collects all the annotations containing at least one element of the query in the annotations terms field.

SYRANNOT then filters these annotations in order to retain the 20 most relevant annotations by using API JENA [JENA] developed by the HP company (the objective of JENA is to develop applications for the semantic Web) and by calculating the similarity between the concepts of the current query and the

concepts of the field terms of annotations corresponding to each annotation.

JENA is used to carry out inferences on the ontologies and on the RDF data bases.

The similarity calculation of two sets of concepts is carried out by using the Wu Palmer formula:

$$Sim(A, B) = \frac{1}{2} \left(\frac{1}{|A|} \sum_{A_i \in P1} \max(ConSim(A_i, B_i)) + \frac{1}{|B|} \sum_{B_i \in P2} \max(ConSim(A_i, B_i)) \right)$$

With

A: set of concepts {A_i}, |A| cardinal of A

B: set of concepts {B_i}, |B| cardinal of B

ConSim(C1, C2): similarity calculation function between two concepts C1 and C2, in a concepts tree.

$$ConSim(C1, C2) = 2 * \text{depth}(C) / (\text{depth}(C1) + \text{depth}(C2))$$

With

C is the smallest generalizing of C1 and C2 in arcs number, depth (C) is the number of arcs which separates C from the root.

The system then reclassifies the 20 relevant annotations by calculating the similarity between the current user profile and the profile of the user who has created the annotation by using JENA and the Wu Palmer formula. The system extracts from each annotation and presents to the user the URL, the title, the author, the language of the document, as well as the query submitted to google for a possible reformulation of the user query (figure 11).

Mots clé	URL	Titre	Auteur	Date langue	Profil Utilisateur	Similarité entre profils
profil conceptuel	www.iaa.fr/rapports/iaa-2004-2005/2004-2005-01.pdf	Formulation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies: applications à la description de documents audiovisuels	Raphaël Tricot	fr	OT1 - Annotations - modèle_graphes_conceptuels - Construction_sémantique - Définition_sémantique - RDF - RDF/S	0.6537142857142857
RDF/S	www.iaa.fr/rapports/iaa-2004-2005/2004-2005-02.pdf	Vers le web sémantique	Philippe Leclerc	fr	OT1 - RDF - RDF/S - Définition_WS 0.6	0.6537142857142857
rd	http://fr.wikipedia.org/wiki/Portail:Publiweb/Portail:Publiweb/2004	Web sémantique et pratiques documentaires	Jérôme Euzenat Raphaël Tricot	fr 2004	RDFS - Définition_WS - RDF - Dublin_Core - OT1 - Sémantique	0.5813492063492063
RDF Jena	http://www.perthnet.com.au/~jhenr45602/2004/	Jena: RDF au Jena	Jena	fr	RDF - Jena	0.4484285714285714
RDF	http://www.iaa.fr/rapports/iaa-2004-2005/2004-2005-03.pdf	Introductions à RDF	Philippe Leclerc	fr 2004	RDF	0.4484285714285714
RDF Jena	http://www.iaa.fr/rapports/iaa-2004-2005/2004-2005-04.pdf	Jena Documentation	Jena	fr	RDF - Jena	0.4484285714285714
RDF	http://www.iaa.fr/rapports/iaa-2004-2005/2004-2005-05.pdf	RDF Tutorial	Pierre-Antoine Champin	fr 05-04-2001	RDF	0.4484285714285714

Figure 11: SYRANNOT Results

The scenario presented above corresponds to the stages of re-use and adaptation of the CBR cycle.

3. SYRANNOT tests and evaluations

To evaluate the contribution of SYRANNOT, we have initialized the data bases of cases, profiles, and annotations by simulating research sessions. Thus, we have built a corpus including a hundred PDF scientific documents annotated using the domain ontology. First evaluations showed that the fact that the concepts used for the annotations are elements of the current query permits to SYRANNOT to present a significant assistance to the current user. Our current research tasks focus on the comparison of the performances of SYRANNOT to other existing systems based on annotations.

4. Conclusion

In this paper, we presented the SYRANNOT system architecture which assists a user in the information retrieval session by presenting relevant retrieved documents for his query and adapted to his profile. SYRANNOT integrates the CBR mechanism in the semantic annotations coming out of the semantic Web field. Ontological models were presented, as well as the research algorithms and the similarity calculation functions proposed in each stage of the CBR cycle. Experimental evaluations have shown very encouraging results in particular when the data base of cases is important and diversified.

References

[AAM 94] AAMODT, A., PLAZA, E. Case-Based Reasoning : Foundational Issues, Methodological Variations and System Approaches. March 1994, AI Communications, the European journal on AI, 1994, Vol 7, N°1, p. 39-59.

[COR 98] 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.

[JENA] jena.sourceforge.net/

[KOL 93] KOLODNER, J. Case based reasoning. San Mateo, CA: Morgan Kaufman, 1993.

[JER 01] JÉRIBI, L. Improving Information Retrieval Performance by Experience Reuse. Fifth International ICCS/IFIP conference on Electronic Publishing: '2001 in the Digital Publishing Odyssey' ELPUB2001. Canterbury, United Kingdom, 5-7 July 2001, p.78-92.

[SCH 89] SCHANK, R. C., RIESBECK, C. K. Inside Case Based Reasoning. Hillsdale, New Jersey, Usa : Lawrence Erlbaum Associates Publishers, 1989, 423 p.

[SMA 99] SMAÏL, M. Recherche de régularités dans une mémoire de sessions de recherche d'information documentaire", InforSID'99, actes des conférences, XVIIème congrès, La Garde, Toulon, 2-4 juin 1999, p. 289-304.