Annotation sémantique des ressources Web : État de l'art et perspectives de recherche

Sahar Maâlej

Université de Sfax, Laboratoire MIR@CL, Pole technologique, Rte de Tunis, BP 242, CP 3021, Sfax, Tunisie.

maalej_sahar@yahoo.fr

RÉSUMÉ. *Le problème d'annotation sémantique de ressources Web intéresse des chercheurs de différentes communautés afin d'améliorer le processus de recherche d'information. En effet, l'annotation sémantique des documents et des pages Web est un travail laborieux. Une automatisation du processus de construction s'impose pour moderniser la recherche d'information dans le Web sémantique. Dans cet article, nous présentons une nouvelle approche d'annotation sémantique des ressources Web, après l'étude de certains outils d'annotation et d'extraction des métadonnées ainsi que des systèmes et approches touchant l'annotation sémantique de documents semi-structurés et Web. Notre proposition repose essentiellement sur une annotation semi-automatique des ressources Web via l'exploitation d'une ontologie et d'un outil d'extraction des métadonnées. En fait, nous voulons résoudre des problèmes liés à l'annotation dans les ressources Web, existant dans la littérature.*

ABSTRACT. *The semantic annotation problem of Web resources interest researchers from different communities to improve the information retrieval process. Indeed, semantic annotation of documents and Web pages is a hard work. An automation of construction process is essential to modernize information retrieval in the Semantic Web. In this paper, we present a new semantic annotation approach of Web resources, after the study of some annotation and metadata extraction tools, thus systems and approaches regarding semantic annotation of semi-structured documents and Web. Our proposal rest essentially on semi-automatic annotation of Web resources through the use of an ontology and a metadata extraction tool. In fact, we see solving problems related to annotation in Web resources existing in the literature.*

MOTS-CLÉS : *Annotation sémantique, Recherche d'information, Ressources Web, Métadonnées, Web sémantique.*

KEYWORDS: *Semantic annotation, Information Retrieval, Web resources, Meta-data, Semantic Web.*

1. Introduction

L'un des objectifs du Web sémantique est de construire et d'utiliser des métadonnées permettant d'annoter sémantiquement les ressources afin d'améliorer la recherche d'information. Atteindre cet objectif dépend de la prolifération des ontologies, des possibilités d'automatiser le processus d'annotation et du passage à l'échelle. Ceci constitue un axe de recherche dans différentes communautés. Par ailleurs, les ressources Web sont très hétérogènes aussi bien du point de vue de leur structuration que du point de vue du vocabulaire utilisé. Ainsi, l'annotation sémantique doit être salutaire dans un éventail d'applications intelligentes orientées contenu. Cependant, l'annotation sémantique des documents et des pages Web est une tâche laborieuse, pénible et longue. Une automatisation du processus d'annotation s'impose alors. C'est dans ce cadre que se situent nos travaux. Nous visons à résoudre les difficultés liées à l'annotation dans les ressources Web, après avoir survolé les principales orientations pour traiter ces difficultés. Pour cette raison, nous nous intéressons à la définition d'une approche d'annotation sémantique des ressources Web basée sur les ontologies, sur un modèle d'annotation, et un outil d'extraction des métadonnées. Nous nous intéressons aussi à la construction d'un système pour l'opérationnalisation de cette approche afin d'automatiser le processus d'annotation. Ce système permet la création et l'instanciation de relations contextuelles et sémantiques entre les ressources Web afin d'accélérer le processus d'interrogation.

Le papier est structuré comme suit : la deuxième section présente les outils d'annotation sémantique. Les outils d'extraction des métadonnées constituent l'objet de la troisième section. Quant à la quatrième, elle est réservée aux travaux d'annotation sémantique. Notre approche d'annotation sémantique des ressources Web est décrite dans la cinquième section.

2. Les outils d'annotation sémantique

Ces outils d'annotation se situent généralement dans le cadre du Web sémantique. Ils servent à créer et gérer les annotations du contenu des ressources Web. Ils reposent sur un modèle formel de connaissances, en général une ontologie, et exploitent de plus en plus les standards du Web sémantique. La plupart des outils d'annotation étudiés (Ontomat Annotizer (Handschuh *et al.,* 01), Video AnnEx (Lin *et al.,* 03) et Cadixe (Cadixe, 05) pour l'annotation manuelle, S-CREAM (Handschuh *et al.,* 02), KIM (Borislav *et al.,* 04) et M-OntoMat-Annotizer (Bloehdorn *et al.,* 05) pour l'annotation semi-automatique, AeroDAML (Kogut *et al.,* 01) et MnM (Vargas-Vera *et al.,* 02) pour l'annotation automatique) génèrent des annotations en utilisant RDF, XML ou SHOE, et utilisent des ontologies prédéterminées. Pour les outils d'annotation semi-automatique, un marquage manuel par l'utilisateur dans une interface graphique est nécessaire. Comme ces outils supportent des annotations manuelles, ils risquent alors de produire des erreurs

syntaxiques ou des références incorrectes. Nous remarquons, en plus, que les outils d'annotation automatique risquent de produire les mêmes erreurs puisqu'ils utilisent les outils d'extraction d'informations. Nous pouvons conclure alors que ces outils d'annotation sémantique présentent plusieurs inconvénients au niveau du traitement séparé des composants d'une page Web, et au niveau de l'utilisation des outils d'extraction d'information. Par rapport à notre objectif, ces outils ne s'intéressent pas à la description de la totalité d'une ressource Web, et plus précisément à la mise en valeur des liens sémantiques entre les ressources. Cette description est décrite par les métadonnées d'annotation RDF, FOAF, etc. Nous allons présenter, dans ce qui suit, les outils étudiés d'extraction des métadonnées à partir des pages Web.

3. Les outils d'extraction des métadonnées

Les outils d'extraction des métadonnées permettent d'extraire des métadonnées (sous formats XML ou RDF) à partir de différents types de fichiers (documents Microsoft Office, HTML, etc.). Dans nos travaux, nous nous concentrons sur les outils d'extraction des métadonnées à partir des pages Web. Parmi ces outils, on trouve ceux qui s'intéressent à l'extraction des métadonnées Dublin Core, et ceux qui concernent les données sémantiques (SIOC, FOAF, DOAP, RDFa, etc.) existant dans une page. C'est sur ces derniers que se base notre réflexion (cf. Tableau 1). Ces outils permettent d'extraire, automatiquement, la sémantique caractérisant une page Web aussi que ses relations avec d'autres pages. Cette sémantique est reconnue et valorisée dans plusieurs standards, dont les plus connus sont FOAF (Friend-of-a-friend), SIOC (Semantically-Interlinked Online Communities), DOAP (Description Of A Project) et RDFa (Resource Description Framework in attributes) (SemanticR, 09). De ce fait, nous pouvons utiliser l'un de ces outils pour annoter sémantiquement les ressources Web. Dans ce cadre, nous avons réalisé une étude qui nous a permis de choisir l'outil le plus adéquat à notre besoin. Il s'agit du plug-in Semantic Radar (cf. Tableau 1), puisqu'il permet d'extraire automatiquement les données sémantiques existant dans une page Web et ses liens. En sus, nous proposons d'utiliser des ontologies de domaines et d'applications afin de pouvoir valider le modèle d'annotation pertinente envisageable. À cet effet, nous avons aussi étudié les travaux utilisant le standard XML-RDF pour l'annotation et la génération automatique de descriptions sémantiques de documents semi-structurés et Web, ainsi que les travaux utilisant le standard Topic Maps (XTM : XML Topic Maps) pour l'extraction et la gestion des ontologies dans différents contextes d'applications. Nous présentons et discutons ces travaux pour préciser la démarche à suivre pour atteindre notre premier objectif ; atteindre une annotation semi-automatique des ressources Web ?

	Critères de comparaison	
Outils	**Résultat d'extraction**	**Existence des métadonnées dans le résultat d'extraction**
Semantic Radar (SemanticR, 09)	Données du Web sémantique RDF (FOAF, SIOC, DOAP, RDFa, etc.)	Extraction de toutes les données du Web sémantique existant dans la page Web
RDFa distiller (RDFa D, 10)	RDF (uniquement RDFa du XHTML)	Extraction très limitée : uniquement les descripteurs RDFa existants dans la page Web
RDF Distiller (RDF D, 11)	RDF (RDFa, RDFXml)	Extraction très limitée : uniquement les descripteurs RDFa et RDFXml existants dans la page Web

Tableau 1. ***Études des différents outils d'extraction des métadonnées sémantiques***

4. Les travaux d'annotation sémantique des documents semi-structurés et Web

Nous avons étudié les approches récentes d'annotation sémantique de documents semi-structurés et Web, telle que celle présentée dans (Thiam, 10), ainsi que les systèmes de génération automatique de descriptions sémantiques tels que le framework WebCat (Martins *et al.,* 05) et le système ALLRIGHT (Shchekotykhin *et al.,* 07). Ces travaux utilisent le standard XML-RDF pour l'annotation. Par rapport à notre objectif, ces derniers se limitent à une partie d'une page Web (textes, images, etc.) pour l'extraction des métadonnées RDF. En plus, ils ne traitent pas les liens sémantiques entre les ressources Web lors de cette extraction. Nous avons étudié aussi des approches d'extraction et de gestion des ontologies se basant sur le standard Topic Maps, telles que celles proposées dans (Gaëlle *et al.,* 06) et (Steven *et al.,* 07). Nous déduisons les mêmes limites définies au niveau de RDF. Par rapport à notre objectif, ces derniers ignorent le profil utilisateur lors d'une interrogation Web. D'après ces études, nous voyons que les résultats de ces travaux sont insuffisants par rapport à notre objectif d'annotation sémantique des ressources Web après une interrogation du Web par l'utilisateur. Nous récapitulons ces insuffisances dans le tableau qui suit.

	Avantages	**Inconvénients**
Les travaux qui utilisent le standard RDF	Extraction des métadonnées RDF à partir de documents-structurés ou Web.	- Limitée soit à un corpus de données, soit à une partie d'une page Web. - Ne traite pas les liens existants entre les pages Web. - Risque de produire de faux termes lors de l'utilisation des outils de TALN (Traitement Automatique du Langage Naturel).
	Génération d'annotation sémantique par rapport à une ontologie.	Nécessite l'intervention de l'utilisateur (choix des concepts de l'ontologie pour l'annotation).

Les travaux qui utilisent le standard Topic Maps	Extraction d'une structure ontologique en XTM d'un corpus de données ou d'une page Web.	- L'extraction est soit manuelle, soit semi-automatique. - L'extraction est limitée à un corpus de données ou bien à une partie de la structure d'une page Web, sans tenir compte les liens entre les pages.
	Utilisation d'une ontologie pour générer le XTM des ressources.	- Les concepts de l'ontologie représentent les mots clés de la recherche ; absence d'un langage d'interrogation. - Ignorance du profil utilisateur lors d'une interrogation des ressources Web.

Tableau 2. *Les avantages et les inconvénients des travaux d'annotation*

L'objectif de ces travaux est alors de générer des métadonnées pour les ressources semi-structurées ou Web. Ces métadonnées représentent l'annotation sémantique des ressources. La plupart de ces travaux se basent sur une ontologie. Nous allons suivre la même démarche que ces travaux : annoter les ressources en reposant sur une ontologie. Mais, ces derniers présentent plusieurs limites par rapport à notre objectif. En effet, nous voulons arriver à assurer une annotation sémantique des ressources Web en tenant compte des liens existants entre les pages Web. Cette annotation sera plus pertinente dans le cas d'utilisation d'une ontologie. Nous suggérons aussi que l'interrogation de ces ressources annotées soit faite par l'extension d'un langage d'interrogation supportant les dimensions sémantiques des métadonnées d'annotation. Pour atteindre notre objectif, nous proposons la création d'une nouvelle approche d'annotation sémantique des ressources Web. Nous présentons dans la section suivante la démarche de notre proposition et nos premières idées concernant le modèle d'annotation.

5. Approche d'annotation sémantique des ressources Web

5.1. Démarche à suivre pour la définition de l'approche (cf. Figure 1)

Nous avons indiqué précédemment que nous utilisons le plug-in Semantic Radar intégré à Firefox Mozilla pour l'extraction des données du Web sémantique. Mais, cet outil ne suffit pas pour arriver à assurer une annotation sémantique semi-automatique des pages Web. Nous proposons alors de le coupler avec les ontologies afin de pouvoir offrir l'annotation sémantique la plus pertinente. À cet effet, nous proposons de créer une nouvelle approche d'annotation sémantique des ressources Web. Nous présentons ci-après la démarche générale de l'approche proposée.

(1) Une interrogation des ressources Web par un langage d'interrogation, en se basant sur les concepts de l'ontologie. Cette interrogation est lancée depuis une interface graphique que nous envisageons de construire. Le langage utilisé doit tenir compte des dimensions sémantiques des métadonnées des ressources Web.

(2) La définition d'une méthode de filtrage des pages Web en se basant sur un seuil de pertinence choisi : les pages Web retournées après interrogation passent par le processus de filtrage afin d'utiliser les plus pertinentes.

(3) L'application d'une analyse par Semantic Radar pour chaque page Web filtrée. Cette analyse retourne un RDF pour la page contenant des descripteurs, comme FOAF, SIOC, etc., décrivant les données sémantiques.

(4) La définition d'une méthode d'annotation des pages Web. Cette méthode aide à produire des métadonnées RDF à chaque ressource Web, en se basant sur des règles d'équivalence et sur un modèle d'annotation que nous envisageons proposer. Ces règles sont des résultats d'une recherche d'équivalence entre le RDF généré et les ontologies de domaine. Les métadonnées RDF vont être publiées sur le Web, associées à leurs ressources. Elles sont mises à jour à chaque nouvelle interrogation.

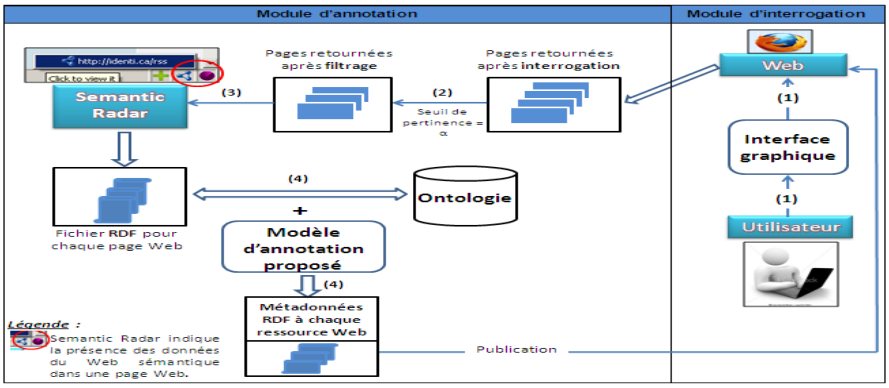
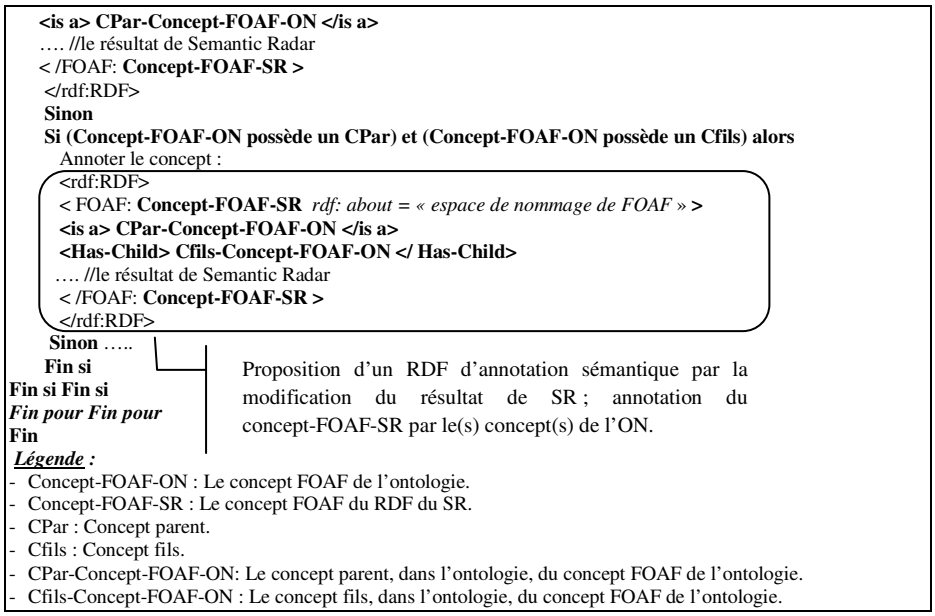


Figure 1. *Démarche de notre approche*

5.2. Le modèle d'annotation proposé

Nous avons entrepris la construction du modèle d'annotation que nous proposons. Ce modèle aide à générer un RDF d'annotation sémantique des ressources Web et ses liens. Cette annotation est le résultat d'une recherche d'équivalence entre le RDF généré par Semantic Radar (SR) pour chaque page Web et les ontologies de domaine (ON) après une interrogation du Web par l'utilisateur. Nous nous limitons dans ce qui suit à la présentation de quelques propositions d'annotation (pour les concepts FOAF), de façon algorithmique.

Début
Pour tout **Concept-FOAF-ON**
Pour tout **Concept-FOAF-SR**
Si (Concept-FOAF-ON = Concept-FOAF-SR) alors
Si (Concept-FOAF-ON possède un CPar) et (Concept-FOAF-ON ne possède pas un Cfils) alors
Annoter le concept :
<rdf:RDF>
< FOAF: **Concept-FOAF-SR** *rdf: about = « espace de nommage de FOAF »* **>**



7. Conclusion et perspectives

Dans cet article, nous nous sommes intéressés à l'étude des outils d'annotation sémantique, des outils d'extraction des métadonnées ainsi que les approches et les systèmes d'annotation sémantique des documents semi-structurés et Web. D'après ces recherches, nous nous sommes proposés de définir une nouvelle approche d'annotation sémantique des ressources Web. Cette approche d'annotation permettra d'améliorer la recherche d'information dans un environnement Web sémantique. Comme perspectives à court terme, nous envisageons la validation de cette première proposition de modèle d'annotation, par l'automatisation l'élicitation et l'instanciation RDF d'annotation. En sus, nous envisageons de finaliser la méthode d'annotation des pages Web (les étapes 3 et 4 de l'approche d'annotation). Comme perspectives à long terme, nous étudions l'utilisation et l'extension du langage d'interrogation SPARQL[1] dédié aux ressources RDF. Nous souhaitons enfin trouver une solution pour filtrer les pages Web après interrogation (les étapes 1 et 2 de l'approche).

8. Bibliographie

Bloehdorn S., Petridis K., Saathoff C., Simou N., Tzouvaras V., Avrithis Y., Handschuh S., Kompatsiaris Y., Staab S., G. Strintzis M., « Semantic annotation of images and videos

1. www.w3.org

for multimedia analysis », T*he 2nd European Semantic Web Conference,* 29 May–1 June 2005, Heraklion, Crete, Greece, p. 592-607.

Borislav P., Atanas K., Damyan O., Dimitar M., Atanas ., « KIM – a semantic annotation platform for information extraction and retrieval », *Natural Language Engineering*, vol. 10, n° 3-4, 2004, p. 375-392.

Cadixe., « Le projet Caderige », *Catégorisation Automatique de Documents pour l'Extraction de Réseaux d'Interactions Geniques,* 2005. http://www-leibniz.imag.fr/SICLAD/Caderige/Cadixe/index.html.

Lortal G., Todirascu A., Lewkowicz M., « Soutenir la coopération par l'indexation semi-automatique d'annotations », *Actes d'IC 2006*, Nantes-France, p. 61-70.

Handschuh S., Staab S., Maedche A., « CREAM - Creating relational metadata with a component-based, ontology-driven annotation framework », *The Knowledge Capture Conference*, 2001, Banff-Canada, p. 76-83.

Handschuh S., Staab S., Ciravegna F., « S-CREAM–Semiautomatic CREAtion of Metadata », *The 13th International Conference on Knowledge Engineering and Knowledge Management Ontologies and the Semantic Web*, vol. 2473, 2002, p. 358-372.

Kogut P., Holmes W., « AeroDAML: applying information extraction to generate DAML annotations from web pages *», The First International Conference on Knowledge Capture K-CAP*, 2001, Victoria, British Columbia.

Lin C-Y, Tseng B. L., Smith J. R., « VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning », *IEEE Intl. Conf. on Multimedia & Expo (ICME)*, July 2003, Baltimore, MD.

Martins B., Silva M.J., « The WebCAT Framework - Automatic Generation of Meta-Data for Web Resources », *Web Intelligence 2005*, Compiegne- France, p. 236-242.

RDFa D., « RDFa Distiller », 2010. http://www.w3.org/2007/08/pyRdfa/.

RDF D., « RDF Distiller », 2011. http://rdf.kellogg-assoc.com/distiller.

SemanticR., « Semantic Radar », 2009. https://addons.mozilla.org/en-US/firefox/addon/semantic-radar/.

Shchekotykhin K. M., Jannach D., Friedrich G., Kozeruk O., « AllRight: Automatic Ontology Instantiation from Tabular Web Documents », *The 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, 2007, Busan-Korea, p. 466-479.

RobersonS., Dicheva D., « Semi-automatic ontology extraction to create draft topic maps », *ACM Southeast Regional Conference*, 2007, Winston-Salem, North Carolina, USA, p. 100-105.

Thiam M., Annotation Sémantique de Documents Semi-structurés pour la Recherche d'Information, Thèse de doctorat, Université Paris Sud-Paris XI, France, 2010.

Vargas-vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F., « MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup », *Knowledge Engineering and Management*, 2002, Spain, p. 379-391.