L'annotation sémantique de documents

Directrices de thèse : Chantal Reynaud et Brigitte Safar

L'explosion du nombre de sources d'information disponibles via le Web multiplie les besoins de techniques permettant d'accéder au contenu sémantique de ces sources et de les interroger plus finement qu'avec une simple recherche de mots clés. Le paradigme le plus puissant proposé actuellement pour résoudre cette tâche est celui établi dans la vision d'un web sémantique : développement d'ontologies décrivant les concepts et les relations mis en jeu dans un domaine particulier, puis annotation des documents des sources avec les éléments de l'ontologie et enfin interrogation des sources par le biais du vocabulaire exprimé dans l'ontologie. L'un des aspects clés de ce paradigme est la phase d'annotation des documents, i.e. l'accrochage d'un élément de l'ontologie à un fragment de document, qui implique de reconnaître (ou de déduire) la présence de l'élément ontologique dans la forme de surface du document.

Différents outils d'annotation ont été proposés pour des domaines spécifiques. Ils exploitent des techniques variées s'appuyant sur des patrons lexico-syntaxiques génériques ou définis par un expert du domaine ([2], [4]), des ressources lexicales pré-établies ([3],[4]), éventuellement la structure du document [1], pour reconnaître la forme de surface d'un type d'élément ontologique particulier (concept, instance de concept, Entité Nommée, relation). Le traitement des éléments pris individuellement conduit toutefois souvent à de mauvaises interprétations. Comme l'ont montré les travaux précédents, il faut être capable, pour établir une annotation, de prendre davantage en considération l'aspect sémantique. Cela signifie interpréter la présence simultanée de différents termes linguistiques et l'absence d'autres, comprendre le contexte défini par différents éléments reconnus dans le document et mis en relation, savoir interpréter l'absence d'autres éléments, inférer certaines connaissances.

Il s'agira dans cette thèse de définir une approche permettant d'exploiter les différents composants d'une ontologie (définition et classification des concepts, relations, contraintes) afin d'établir différents formes d'annotations plus complexes que de simples concepts, c'est-à-dire des représentations formelles structurées, interprétables à l'aide d'une ontologie, et sur lesquelles il est possible de raisonner pour produire de réelles annotations sémantiques. Ces annotations devront pouvoir être établies en raisonnant sur le contexte défini par les différents éléments reconnus dans le document, ce qui suppose une représentation formelle de ce contexte. Dans un second temps, il s'agira de définir des mécanismes de raisonnement portant sur les annotations pour faciliter la recherche d'information et produire des réponses à des requêtes les plus pertinentes possibles.

Bibliographie :

[1] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak & S. Hellmann. *DBpedia : a crystallization point for the web of data.* Web Semantics: Science, Services and Agents on the World Wide Web 7 (3), 154-165

[2] P. Cimiano, G. Ladwig & S. Staab. *Gimme The Context : Context-driven semantic annotation with C-PANKOW*. In proc. of WWW'05, pp. 332-341, 2005.

[3] P.N. Mendes, M. Jakob, A. Garcia Silva & C. Bizer. *DBpedia spotlight: shedding light on the web of documents.* In proc. of I-Semantics'11, pp. 1-8, 2011.

[4] F. M. Suchanek & G. Weikum. *SOFIE: a Self-Organizing Framework for Information Extraction.* In proc. of WWW'09, pp. 631-640, 2009.