Aggregative and Neighboring Approximations to Query Semi-Structured Documents

Y. Mrabet* N. Pernelle* N. Bennacer** M. Thiam*

*4, Rue J. Monod, Parc Club Orsay université, 91483 Orsay Cedex
first.last@lri.fr,
**Supelec, F-91192 Gif-sur-Yvette Cedex
nacera.bennacer@supelec.fr

Abstract. Structures heterogeneity in Web resources is a constant concern in element retrieval (i.e. tag retrieval in semi-structured documents). In this paper we present the *SHIRI* [1] querying approach which allows to reach more or less structured document parts without an a priori knowledge on their structuring.

1 Approximate Queries According to Document Structuring

To retrieve the most suited tagged element according to a user query, classical approaches tend to use a statistical indexing of the tagged zones. But, while such indexing has shown to be very efficient for document retrieval, it remains unsatisfying for element retrieval. Cases where the query is composed of many terms, which are not necessarily localized in the same parts of the documents, are not well covered. Furthermore, even if the neighboring tags are taken into account through an in-document distance, the ranking of the retrieved parts does not embed any notion of structuring (e.g. a document node talking only about a conference A, may have the same rank as a node talking about three different conferences).

We propose a semantic solution to cope with structures heterogeneity by making explicit the structuring levels [Thiam et al. (2008)]. A document node is so said to be a part of speech (i.e annotated by the *PartOfSpeech* metadata) if it contains many instances of different concepts. Another node containing only one single instance of a given concept is annotated as being an instance of that concept and respectively for the *SetOf* case, where a node contains a set of instances of the same type. Furthermore the structural imbrication between document nodes is used to infer semantic relations between the annotated instances. E.g. if the node '$< ul >$' is annotated as an instance of the '$Article$' concept and the next '$< li >$' node is annotated as an instance of the $Person$ concept, the relation $< ul,\ authored\_by,\ li >$ is created. Referring to the above annotation model, we propose two approximation types. The first, called *aggregative approximation*, uses the aggregate metadata defined in the ontology extension (*PartOfSpeech* and *SetOfConcepti*) to look for less structured document parts if no better structuring is found. The second approximation, called *neighboring approximation*, is used to cover cases where we look for semantic relations that are not retrieved in the annotation base (i.e. there is no imbrication between two document nodes which are annotated

[1] SHIRI : Digiteo labs project (LRI, SUPELEC)

respectively by the domain and range of the relation). The queries we consider are in RDF triple pattern (e.g. SPARQL queries). For instance, the query ($<?x\ type\ Article> \ AND \ <?x\ authored_by\ ?y> \ AND \ <?y\ hasName\ "Victor\ Vianu"\ >$) should return parts of documents which are imbricated in an appropriate manner and annotated by the appropriate concept types according to the query. However, this structuring is highly binding. Our first approximation type allows to search for document parts annotated by aggregates instead of the initial concepts specified by the user query. A first approximation of the query above can so be (<?x type Article> AND <?x authored_by_Set ?y> AND <?y type *SetOfAuthors*>). A second can be to search for a PartOfSpeech metadata indexed by the $Article$ and $Author$ concepts. From another hand, the neighboring approximation allows to replace the wished imbrication by a more general structural neighboring. In example, the query above could be approximated by replacing the triple $<?x\ authored\_by\ ?y>$ by $<?x\ ParentOf[N]\ ?y>$. The $ParentOf[N]$ relation represents the fact that the structural unit $?x$ and $?y$ are at a distance N of each other (e.g. <table/tr[1]/td[3]> is at a distance 3 from <table/tr[2]/td[1]>). In our first experimentations we proposed a dynamic algorithm which combines the two approximations types and returns results of the original query and its ranked approximations. We tested this approach on three distinct annotated data sources (bibliographic references from DBLP, HAL and INRIA server)! . The results we obtained after applying a set of more or less complicated queries showed that our approach makes it possible to reach heterogeneous data structures without an a priori knowledge. We had an overall recall of 28% with the original query, 19,9% with the aggregative approximation, 52.3% with the neighboring approximations and 72% with the combined approximations. Overlapping put away, the total recall of three approximations is of 100%. The precision was also of 100% due to structures regularity of the corpus we used. Contextually, our approach doesn't replace ontology-based approximations proposed in [Hurtado et al. (2006); Corby et al. (2006)]). It applies as an independent layer to cope with structures heterogeneity. Future works will encompass the enhancement of our ranking criteria, the experimentation with irregular document structures and an unsupervised annotation process.

References

Corby, O., R. Dieng-Kuntz, F. Gandon, and C. Faron-Zucker (2006). Searching the semantic web : Approximate query processing based on ontologies. *IEEE intelligent systems Journal*.

Hurtado, C.-A., A. Poulovassilis, and P.-T. Wood (2006). A relaxed approach to rdf querying. *International Semantic Web Conference, ISWC*.

Thiam, M., N. Pernelle, and N. Bennacer (2008). Contextual and Metadata-based Approach for the Semantic Annotation of Heterogeneous Documents. *ESWC-SeMMA*.

Résumé

L'hétérogéneité des structures dans les documents Web est un souci constant en recherche d'éléments. Dans ce papier, nous présentons l'approche d'interrogation sémantique SHIRI qui permet d'atteindre des parties de document plus ou moins structurées sans connaissances a priori sur leur structuration.