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)