Pre-matching: Large XML Schemas Decomposition Approach

Sana Sellami, Aïcha-Nabila Benharkat, and Youssef Amghar

University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
{Sana.Sellami,Nabila.benharkat,Youssef.Amghar}@insa-lyon.fr

1 XML Schemas Decomposition Approach

We propose a decomposition approach, as a pre-matching phase, which break down large XML schemas into smaller sub-schemas to improve the quality of large schema matching. Our approach identifies and extracts common structures between and within XML schemas (inter and intra-schemas) and finds the sub-schemas candidates for matching.

As illustrated in Fig.1, our proposed approach is composed of three phases:

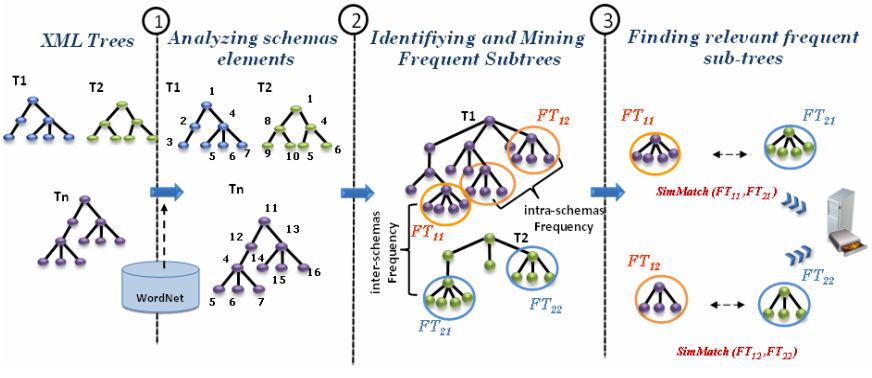

Fig. 1. Decomposition approach

(1) Converting XML schemas in trees: The goal of this initial phase is to transform XML schemas into trees and to find linguistic relations between elements. This aims at improving decomposition with considering not only exactly the same labels of elements but also the linguistic similar elements. We firstly need to parse the XML schemas and transforming them into trees. The main feature of these large schemas is that they contain referential constraints. Then parsing these schemas becomes a difficult exercise. To cope with these constraints, we duplicate the segment which they refer to resolve their multiple contexts. We notice that most previous match systems focused on simple schemas without referential elements.

(2) Identifying and mining frequent sub-trees: The main goal of this phase is to decompose the input schemas into smaller ones. To this end, we identify and extract

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 9–10, 2009.

the common sub-structures from XML schemas describing the same domains. We propose to use tree mining techniques to identify these structures. More precisely, we use the algorithm proposed in [2]. Tree mining is a classical pattern mining problem (an important class of data mining problem) which aims at discovering automatically sub-trees that appear frequently in a set of trees.

(3) Finding relevant frequent sub-trees: The focus of this phase is to identify the sub-trees candidates for matching. This aims at reducing match effort by only matching relevant parts from the other schemas. These sub-schemas are then selected for matching. This pre-matching phase includes two main steps: selection of maximal sub-trees and finding the most similar ones.

2 Experimental Evaluation

We conducted our experiments on real XML schemas (XCBL[1] and OAGIS[2]). We have implemented the decomposition approach in our PLASMA (Platform for LArge Schema MAtching) prototype. We compared our decomposition results with those of fragmentation COMA++ approach [1]. Our results (fig.2) show that decomposition approach provides a better quality of matching in comparison to the fragmentation approach in COMA++.

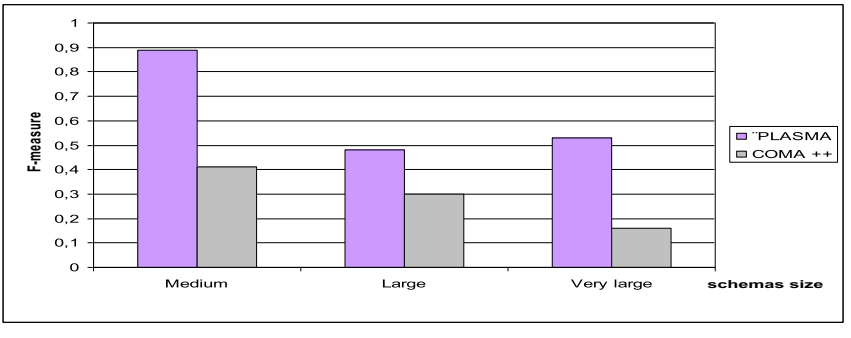

Fig. 2. F-measure obtained by decomposition approach in PLASMA and fragmentation approach in COMA++

References

1. Do, H.H., Rahm, E.: Matching large schemas: Approaches and evaluation. Journal of Information Systems, 857–885 (2007)
2. Termier, A., Rousset, M.A., Sebag, M.: DRYADE: a new approach for discovering closed frequent trees in heterogeneous tree databases. In: Proceedings of the 4th IEEE International Conference on Data Mining (ICDM), pp. 543–546 (2004)

[1] www.xcbl.org
[2] www.oagi.org