Designing similarity measures for XML

Ismael Sanz, María Pérez, and Rafael Berlanga

Universitat Jaume I de Castelló, Spain
{isanz,maria.perez,berlanga}@uji.es

Abstract. In this demonstration we will show a series of tools that support a methodology [1] for the design of complex similarity functions in the context of heterogenous XML systems.

1 Introduction

The existence of highly complex, publicly available XML-based databases has motivated research into multi-similarity XML applications, which support multiple notions of similarity to tailor queries to users with diverse information needs. Such applications arise e.g. in the integration and merging of highly heterogeneous XML databases, and in systems handling objects with complex structures such as protein data, music retrieval systems, or shape databases. Until now, little attention has been paid to the problem of designing suitable similarity measures for such applications.

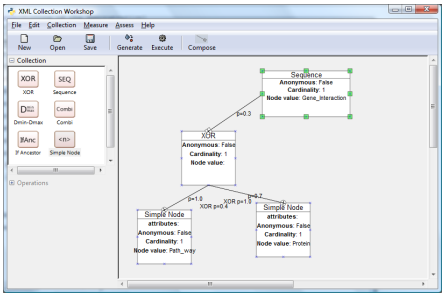
[1] introduces a methodology to support the design of similarity functions for heterogeneous XML-based information systems, based on The following four steps: (1) Characterize a set of relevant XML collections and queries that describe the information needs of users. (2) Establish a candidate set of similarity measures. (3) Evaluate the suitability of the candidate measures. According to the result of this assessment it may be the case that (*i*) the measure may need some adjustment, which implies a change in the measure and the re-evaluation of the candidate set; or that (*ii*) deficiencies in the specification of the information needs are detected, which may cause the candidate set to change completely. (4) Finally, the final set of measures is obtained, and the indexing requirements are established and targeted for physical implementation and optimization

2 Outline of the demonstration

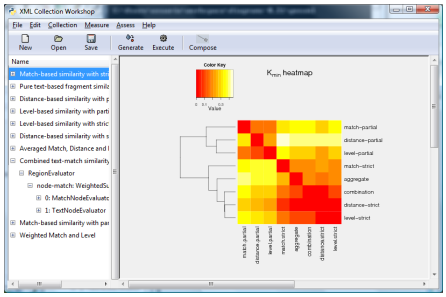
We have implemented a set of tools, the *XML Collection Workshop*, which uses techniques that help in each of the steps of the previously sketched methodology. We will use two different collections: The ASSAM [1] highly heterogeneous collection, whose documents span several different domains, and a collection of large, publicly-avaliable XML databases containing Bioinformatics-related data

The demonstration will proceed through the following steps:

[1] http://moguntia.ucd.ie/repository/datasets/



(a) Part of a generated model opened for editing



(b) Correlation between candidate measures, shown as a heatmap and a corresponding hierarchical clustering

Fig. 1. Screenshots of the XML Collection Workshop

1. *Characterization of XML collections.* Using the ASSAM collection as a case study, we will demonstrate how to create a simplified, probabilistic model of a highly complex XML collection using a probabilistic model described in [2].
2. *Design of test collections and associated queries.* We will show how to use a GUI-based tool, depicted in Figure 1(a), to display the model generated by the previous step, and edit it interactively to generate an XML test collection and a a set of queries which are suitable for testing candidate measures.
3. *Semi-automatic selection of measures.* The selected candidate measures will include a representative set of features: structural matching, text retrieval approaches, and Bioinformatics-specific algorithms. Using the Bioinformatics-based use case, we will show how to select appropriate measures using several assessment criteria. First, we will use a correlation measure to prune redundant candidates; for example, Figure 1(b) graphically displays a clustering of a set of candidate measures based on the the K_{min} distance [3], after performing a run of experiments on the collection and queries generated in the previous step. Then, we will apply standard techniques such as the F_1-measure to study the quality of the remaining candidates.

References

1. Sanz, I., Pérez, M., Berlanga, R.: Measure selection in multi-similarity XML applications. In: Third International Workshop on Flexible Database and Information System Technology (FlexDBIST-08). (2008)
2. Sanz, I., Mesiti, M., Guerrini, G., Berlanga, R.: Fragment-based approximate retrieval in highly heterogeneous XML collections. Data & Knowledge Engineering **64**(1) (January 2008) 266–293
3. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top-k lists. SIAM Journal on Discrete Mathematics **17**(1) (2003) 134–160