Chapter 5
Approximate Matching Between XML Documents and Schemas with Applications in XML Classification and Clustering

Guangming Xing
Western Kentucky University, USA

ABSTRACT
Classification/clustering of XML documents based on their structural information is important for many tasks related with document management. In this chapter, we present a suite of algorithms to compute the cost for approximate matching between XML documents and schemas. A framework for classifying/clustering XML documents by structure is then presented based on the computation of distances between XML documents and schemas. The backbone of the framework is the feature representation using a vector of the distances. Experimental studies were conducted on various XML data sets, suggesting the efficiency and effectiveness of our approach as a solution for structural classification/clustering of XML documents.

INTRODUCTION
The eXtensible Markup Language (XML) (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2004) has become the standard format for data exchange on the Internet, providing interoperability between different business applications. Such wide use results in large volumes of heterogeneous XML data, i.e., XML documents conforming to different schemas. XML documents are naturally tree structured, and may contain atomic and complex structures. XML documents are also semistructured as they incorporate both structural informa-
Approximate Matching Between XML Documents and Schemas with Applications

tion and content (Abiteboul, 1997). Dealing with structure information is important to XML data storage and management (Jagadish et al., 2002), and the presence of a schema can significantly simplify the processing of XML documents. For example, elements from different documents with similar structures can be stored together, resulting in reduced storage and faster query processing. However, processing XML documents based on the schemas may not be feasible in practice, since most XML documents are schema-less.

Studying the problem of approximate matching between an XML document and a schema and its applications in XML data management and mining is not only theoretically interesting, but also critical to many real-world scenarios. For instance, similarly structured XML documents can be stored in relational databases more efficiently than documents with arbitrary structures. As discussed in (Bertino, Castano, Ferrari, & Mesiti, 2004), the distance between an XML document and a DTD can be effective for classifying Web documents. When schemas are not available, meaningful schemas can be inferred from XML documents as explained in the XTRACT system (Garofalakis, Gionis, Rastogi, Seshadri, & Shim, 2000). The framework to compute edit distances between XML documents and inferred schemas provides an alternative to the traditional algorithms for tree-to-tree edit distance in evaluating structural similarity between XML documents (Nierman & Jagadish, 2002). Moreover, with the rapid growth of XML documents on the Internet, disseminating XML documents to a large group of information consumers is becoming more and more important. This typically involves applying XSLT to present related information to the end users, which requires that the XML documents are grouped based on the similarity of their tree structures. The approximate matching between XML documents and schemas is also useful for evaluating schema-based queries over sources of XML documents. Other applications in literature include selective dissemination of XML documents (Stanoi, Mihaila, & Padmanabhan, 2003), information extraction from Web documents (Reis, Golgher, Silva, & Laender, 2004), database integration (Parent & Spaccapietra, 1998), stream processing and document integration (Garofalakis & Kumar, 2005), and protection of XML documents (Bertino, Castano, Ferrari, & Mesiti, 2002).

In this chapter, we present a suite of algorithms for computing approximate matching between XML documents and schemas. Our proposed methods are close to the work in (Suzuki, 2005), which describes how to compute a sequence of edit scripts to transform an XML document to conform to a DTD with the minimum cost. In particular, the algorithm described in Suzuki (2005) is very similar to the unrestricted tree editing method which will be presented in this chapter, however our method achieves similar results with better time efficiency. Bertino, Castano, Ferrari and Mesiti (2008) also considered similarity between XML documents and DTDs by viewing a DTD as a special tree with nodes labeled with special operators and then applying traditional tree matching algorithms. The key differences between the algorithms in this chapter and other methods are the use of the edit distance between an ordered labeled tree and regular tree grammar. The regular tree grammar is more general than most other schema languages used in similar distance (Murrata, Lee, Mani, & Kawaguchi, 2005), and the edit distance provides a precise measure on how well an ordered tree (XML document) conforms to the regular tree grammar (schema).

One of the most important applications of the above mentioned algorithms is document classification and clustering, which is also essential to XML data storage, data integration, query processing, and schema matching. Traditionally, although not exclusively, the problems of XML document classification and clustering have been addressed based on some variants of tree edit distances. However, when the document sizes vary significantly, the performance of tree edit distance based algorithms becomes unpractical.
Related Content

Computing Join Aggregates Over Private Tables
[www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a](www.igi-global.com/article/computing-join-aggregates-over-private/1816?camid=4v1a)

Dynamic View Selection for OLAP
[www.igi-global.com/article/dynamic-view-selection-olap/1799?camid=4v1a](www.igi-global.com/article/dynamic-view-selection-olap/1799?camid=4v1a)

MILPRIT*: A Constraint-Based Algorithm for Mining Temporal Relational Patterns
[www.igi-global.com/article/milprit-constraint-based-algorithm-mining/1817?camid=4v1a](www.igi-global.com/article/milprit-constraint-based-algorithm-mining/1817?camid=4v1a)

Identifying and Analyzing Popular Phrases Multi-Dimensionally in Social Media Data
Zhongying Zhao, Chao Li, Yong Zhang, Joshua Zhexue Huang, Jun Luo, Shengzhong Feng and Jianping Fan (2015). *International Journal of Data Warehousing and Mining* (pp. 98-112).