Chapter LXXI
XML Document Clustering

Andrea Tagarelli
University of Calabria, Italy

INTRODUCTION

The ability of providing a "standardized, extensible means of coupling semantic information within documents describing semistructured data" (Chaudhri, Rashid, & Zicari, 2003) has led to a steady growth of XML (extensible markup language) data sources, so that XML is touted as the driving force for representing and exchanging data on the Web.

The motivation behind any clustering problem is to find an inherent structure of relationships in the data and expose this structure as a set of clusters where the objects within the same cluster are each to other highly similar but very dissimilar from objects in different clusters.

The clustering problem finds in text databases a fruitful research area. Since today semistructured text data has become more prevalent on the Web, and XML is the de facto standard for such data, clustering XML documents has increasingly attracted great attention. Any application domain that needs organization of complex document structures (e.g., hierarchical structures with unbounded nesting, object-oriented hierarchies) as well as data containing a few structured fields together with some largely unstructured text components can be profitably assisted by an XML document clustering task.

In principle, the availability of schemas for XML data may be useful to drive or simplify a clustering task. For instance, in case of structural classification, XML documents with different element values but similar schemas could be grouped together. An XML DTD defines the document schema by means of constraints (element content models) that specify the element types, hierarchical relationships between elements, and other properties such as multiple occurrences of elements (operator +), optional elements (operator ?), and alternate elements (operator |). XML documents available from real data sources tend to have such characteristics.

However, exploiting XML schemas profitably for classification purposes is not always feasible in practice. On one hand, most XML sources provide documents that are schema-less, that is, documents without an explicitly associated element type definition. On the other hand, XML documents available from the same data source may have quite different size and structure mainly due to nesting and repetition of elements, although they conform to a unique DTD. Also, XML documents with different schemas may have similar contents, or, in a more complicated case, XML documents coming from heterogeneous sources may represent semantically related data even if they use different markup tags that refer to different schemas. The above are main reasons to conceive the task of clustering XML documents without assuming the availability of predefined XML schemas.

Within this view, we aim to provide a broad overview of the state of the art and a guide to recent advances and emerging challenges in the research field of clustering XML documents.

BACKGROUND

Several approaches and methodologies have been proposed in the last years to address the problem of clustering XML documents, and major differences can be found in how the following issues have been settled:

- **Representation model:** A natural interpretation of an XML document is a labeled rooted tree or, if references between elements are included, a labeled, rooted, directed graph. The element and attribute names are mapped to inner nodes (or, alternatively, to edges) of the tree (or graph), and the text sequences enclosed by the innermost elements are assigned to leaf nodes.
- **Features:** An XML feature is a component of the XML representation model and expresses what and how information available in an XML document is considered and modeled as its attribute. Structural information (e.g., element tag names, element location in the tree hierarchy), content information (e.g., textual elements, attribute values), or both information can be involved in XML features.
- **Related tasks**
  - **Pattern matching:** Any mechanism of information retrieval needs a method of locating substructures of a larger structure, the target, by comparing them against a given form called the pattern. In the XML domain, a common problem is tree matching, which is concerned with finding the instances, or matches, of a given pattern tree in a given target tree.
  - **Similarity detection:** Defining a suitable distance or similarity measure between XML documents requires one to consider at least the features upon which documents are identified as similar and the method of pattern matching that is used according to the model chosen for representing an XML document.
  - **Summarization:** The capability of summarizing an XML document or sets of XML documents has a likely impact on the performance of the similarity computation in a clustering task. In particular, XML summaries can help in estimating the selectivity of path expressions and devising indexing techniques for clusters to finally improve the construction of query plans.

In the following we describe major aspects, merits, and shortcomings of significant solutions existing in the current research literature concerning representation and summarization models, feature extraction, similarity computation, and clustering schemes for XML documents. In particular, the focus here is on approaches and methods conceived to address the following main problems: the detection of structural similarities among XML documents, clustering of XML documents by structure, summarization of XML documents, and clustering of XML documents by content.

CLUSTERING XML DOCUMENTS BY STRUCTURE

Clustering XML documents has its roots in the problem of comparing semistructured documents, which originally arises from several applications in the management of semistructured data, such