Chapter 8
Fuzzy Approaches to Clustering XML Structures

Michal Kozielski
Silesian University of Technology, Poland

ABSTRACT
Information on the hierarchical nature of XML data is essential in tasks of learning from XML document structures. Within this view, XML documents can be regarded as multi-represented data, which is the case when multiple representations of the document correspond to the generation of features at each structure level of the document separately. This chapter raises the importance of using fuzzy approaches to clustering XML document structures, since these approaches are shown to be effective in combining the information coming from different document representations that correspond to different hierarchy levels. For this purpose, we overview fuzzy encoding and similarity methods and present fuzzy clustering approaches which are particularly suited for being extended to handle XML document structures. We propose two different scenarios of fuzzy clustering of XML structures, which aim to either encode the document structure hierarchy using a fuzzy bag model or to specifically handle the multi-representation of the documents.

INTRODUCTION
XML documents are complex data objects having hierarchical structure. Therefore, it is not possible to apply conventional similarity measures directly to the analysis of XML document structures. One approach to address this issue is to analyse the structural similarity of XML documents by means of dedicated (tree-based) similarity measures (Nayak, 2008; Dalamagas, Cheng, Winkel, & Sellis, 2004; Nierman & Jagadish, 2002). Another approach consists in encoding the XML structure predominantly into the form of a feature vector and then analysing these feature vectors by means of
a conventional similarity measure. This approach includes such encoding methods as bit encoding (Lian, Cheung, Mamoulis, & Yiu, 2004; Yoon, Raghavan, Chakilam, & Kerschberg, 2001), signal encoding (Flesca, Manco, Masciari, & Pugliese, 2005), and fuzzy encoding (Ceravolo, Nocerino, & Viviani, 2004). Some of the encoding methods here mentioned will be described in detail in this chapter. Moreover, since the focus of this chapter is not to review existing clustering approaches and methods for XML clustering (the interested reader can refer to Chapter “The Role of Schema and Document Matchings in XML Source Clustering” and Chapter “XML Document Clustering: An Algorithmic Perspective”), we make just a mention here of some of those methods by distinguishing them from a clustering strategy viewpoint. In this respect, existing XML clustering methods have been developed based on the following main approaches: agglomerative hierarchical (Nayak, 2008; Nayak & Tran, 2007; Dalamagas, Cheng, Winkel, & Sellis, 2006; Nayak & Iryadi, 2006; Costa, Manco, Ortale, & Tagarelli, 2004; Nieman & Jagadish, 2002), frequent-itemset-based hierarchical (Tagarelli & Greco, 2010), partitional (Tagarelli & Greco, 2010, 2006; Antonellis, Makris, & Tsirakis, 2008; Wang, Liu, & Wang, 2005), tree-partitioning (Bordawekar & Shmueli, 2004), and self-organizing maps (Hagenbuchner, Sperduti, Tsoi, Trentini, Scarselli, & Gori, 2006), and fuzzy (Kozielski, 2007).

XML documents can be regarded as multi-represented data because each document can have several representations when we consider features at each structure level separately. To cluster multi-represented data, information having different representations may be combined at the level of similarity matrices by means of some arithmetic operations or fuzzy aggregation operations. The resulting partition (clustering solution) can also be determined as a particular combination of the partitions derived for each representation separately.

Essentially, multi-represented data clustering consists in combining different types of information during the clustering process. One algorithm of this type is the density-based method DBSCAN that was modified in order to handle the multi-represented nature of the data (Kailing, Kriegel, Pryakhin, & Schubert, 2004). Union and intersection operations were introduced into the algorithm in order to verify if the chosen data objects are dense in any or all the representations that are analysed. This method was also used in the OPTICS algorithm (Achtert, Kriegel, Pryakhin, & Schubert, 2005), which is based on the DBSCAN algorithm but is less sensitive to the values of the parameters required by the method.

There are also examples of fuzzy algorithms combining different types of information during the clustering process. Collaborative fuzzy clustering (Pedrycz, 2002), which is based on the Fuzzy C-Means algorithm was introduced in order to cluster data in a distributed environment. Another method is the Proximity-based Fuzzy Clustering (Pedrycz, Loia, & Senatore, 2004), which introduces additional information (e.g., expert knowledge) to the clustering process. Both methods were further developed to the form of Proximity-based Collaborative Clustering (Loia, Pedrycz, & Senatore, 2007; Pedrycz & Rai, 2008).

There is however an inherent difficulty in directly applying the aforementioned methods to XML data. XML structures are hierarchical and the representations of their different structure levels should not be considered as equally significant.

The structure of XML documents is usually modelled by a rooted labelled tree where parent-child relationships between the elements define a document hierarchy. It is a common assumption that the most general and therefore important information is enclosed nearby the root element concerning the structure of an XML document. The features which are placed on the levels neighbouring a root element should have a greater influence on the analysis results than the deeply nested leaf nodes. For example, if two XML documents have different root elements then all the paths starting from the root element will be different for each
20 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/chapter/fuzzy-approaches-clustering-xml-structures/60909?camid=4v1

www.igi-global.com/e-resources/library-recommendation/?id=1

Related Content

Data Mining In the Context of Business Network Research
Jukka Aaltonen, Annamari Turunen and Ilkka Kamaja (2010). Data Mining in Public and Private Sectors: Organizational and Government Applications (pp. 289-315).
www.igi-global.com/chapter/data-mining-context-business-network/44294?camid=4v1a

A Hyper-Heuristic for Descriptive Rule Induction
Tho Hoan Pham and Tu Bao Ho (2007). International Journal of Data Warehousing and Mining (pp. 54-66).
www.igi-global.com/article/hyper-heuristic-descriptive-rule-induction/1778?camid=4v1a

An Efficient Pruning and Filtering Strategy to Mine Partial Periodic Patterns from a Sequence of Event Sets
Kung-Juiuan Yang, Tzung-Pei Hong, Yuh-Min Chen and Guo-Cheng Lan (2014). International Journal of Data Warehousing and Mining (pp. 18-38).
www.igi-global.com/article/an-efficient-pruning-and-filtering-strategy-to-mine-partial-periodic-patterns-from-a-sequence-of-event-sets/110384?camid=4v1a

Efficient Top-k Keyword Search Over Multidimensional Databases
Ziqiang Yu, Xiaohui Yu and Yang Liu (2013). International Journal of Data Warehousing and Mining (pp. 1-21).
www.igi-global.com/article/efficient-top-keyword-search-over/78373?camid=4v1a