Brzeziński, Dariusz; **Leśniewska, Anna**; **Morzy, Tadeusz**; **Piernik, Maciej**
XCleaner: a new method for clustering XML documents by structure. (English)
Zbl 1318.68137
Control Cybern. 40, No. 3, 877-891 (2011).

Summary: With the vastly growing data resources on the Internet, XML is one of the most important standards for document management. Not only does it provide enhancements to document exchange and storage, but it is also helpful in a variety of information retrieval tasks. Document clustering is one of the most interesting research areas that utilize semi-structural nature of XML. In this paper, we put forward a new XML clustering algorithm that relies solely on document structure. We propose the use of maximal frequent subtrees and an operator called Satisfy/Violate to divide documents into groups. The algorithm is experimentally evaluated on real and synthetic data sets with promising results.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
68T30 Knowledge representation

Keywords:

XML; clustering; patterns

Software:

ToXgene; Xproj; XCleaner