Chapter VIII

On the Usage of Structural Distance Metrics for Mining Hierarchical Structures

Theodore Dalamagas, National Technical University of Athens, Greece

Tao Cheng, University of Illinois at Urbana-Champaign, USA

Timos Sellis, National Technical University of Athens, Greece

Abstract

The recent proliferation of XML-based standards and technologies demonstrates the need for effective management of hierarchical structures. Such structures are used, for example, to organize data in product catalogs, taxonomies of thematic categories, concept hierarchies, etc. Since the XML language has become the standard data exchange format on the Web, organizing data in hierarchical structures has been vastly established. Even if data are not stored natively in such structures, export mechanisms make data publicly available in hierarchical structures to enable its automatic processing by programs, scripts and agents. Processing data
encoded in hierarchical structures has been a popular research issue, resulting in the design of effective query languages. However, the inherent structural aspect of such encodings has not received strong attention till lately, when the requirement for mining tasks, like clustering/classification methods, similarity ranking, etc., on hierarchical structures has been raised. The key point to perform such tasks is the design of a structural distance metric to quantify the structural similarity between hierarchical structures. The chapter will study distance metrics that capture the structural similarity between hierarchical structures and approaches that exploit structural distance metrics to perform mining tasks on them.

Introduction

Hierarchical structures have been widely used in the past in the form of tree or graph structures to organize data in thematic categories, taxonomies, catalogs, SGML files, etc. Since the XML language is becoming the standard data exchange format on the Web, the idea of organizing data in hierarchical structures to enable its automatic processing by programs, Web scripts and agents has been re-visited. Vast amounts of data from many knowledge domains are available or processed to be available on the Web, encoded in hierarchical structures under the XML format.

While the processing and management of data encoded in such structures have been extensively studied (Abiteboul, Buneman, & Suciu, 2000), operations based on the structural aspect of such an encoding have received strong attention only lately. Structural distance metrics is a key issue for such operations. A structural distance metric can quantify the structural similarity between hierarchical structures. Thus, it is a tool that can support mining tasks for such structures. Examples of these tasks are clustering methods, classification methods and similarity ranking. For instance, a clustering task can identify sets of structures such that each set includes similar structures (in terms of their form or the way that they organize data). A similarity ranking mechanism can detect hierarchical structures which are similar to a hierarchical structure given as a test pattern, and also quantify such a similarity.

In this chapter we will first study distance metrics that capture the structural similarity between hierarchical structures. Then we will present approaches that exploit structural distance metrics to perform mining tasks on hierarchical structures. We concentrate on hierarchical structures encoded as XML documents due to the proliferation of the XML language for encoding data on the Web.
A Novel Methodology for Configuring a Job Satisfaction Matrix
www.igi-global.com/article/novel-methodology-configuring-job-satisfaction/74356?camid=4v1a

An Application of the Integrated IBA-TOPSIS Model in Supplier Selection
www.igi-global.com/article/an-application-of-the-integrated-iba-topsis-model-in-supplier-selection/125884?camid=4v1a