Chapter 31
XML Mining for Semantic Web

Rafael Berlanga
Universitat Jaume I, Spain

Victoria Nebot
Universitat Jaume I, Spain

ABSTRACT

This chapter describes the convergence of two influential technologies in the last decade, namely data mining (DM) and the Semantic Web (SW). The wide acceptance of new SW formats for describing semantics-aware and semistructured contents have spurred on the massive generation of semantic annotations and large-scale domain ontologies for conceptualizing their concepts. As a result, a huge amount of both knowledge and semantic-annotated data is available in the web. DM methods have been very successful in discovering interesting patterns which are hidden in very large amounts of data. However, DM methods have been largely based on simple and flat data formats which are far from those available in the SW. This chapter reviews and discusses the main DM approaches proposed so far to mine SW data as well as those that have taken into account the SW resources and tools to define semantics-aware methods.

INTRODUCTION

XML (Bray, Paoli, Sperberg-McQueen, & Maler, 2000) has been extensively used to represent and publish semistructured data across the Web both in the academic and business communities as it provides inter-operability and a well-defined, extensible and machine-readable format. The widespread adoption of XML as the de-facto standard has prompted the development of new techniques that address the problem of XML management and knowledge discovery. Many research efforts have been directed towards mining the structure of XML documents as a way to integrate data sources based on structure similarity. As a step forward, content features borrowed from the text mining field have been introduced to enrich the process of XML mining. However, the increase in volume and heterogeneity of XML-based applications demands new analysis techniques that consider semantic features in the process of knowledge discovery so that more meaningful analysis can be performed.

On the other hand, the Web of Data is currently coming into existence, as opposed to the classical Web of documents, through the Linked

DOI: 10.4018/978-1-4666-2455-9.ch031

Copyright © 2013, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Data effort (Bizer, Heath, & Berners-Lee, 2009). The general idea is to extend the Web by creating typed entities and links between data resources in a way that is machine-readable and the meaning (i.e., semantics) is explicitly defined. This new data model, whose representation formats rely on XML, opens a new range of challenges and opportunities in the data mining and knowledge discovery area.

The aim of this chapter is to review the literature and discuss how semantic features have been incorporated and dealt with in the process of mining complex structured and semistructured data. From the data viewpoint, we provide a state-of-the-art review on approaches focused both on mining complex semistructured data (i.e., XML sources) and SW data. We conceive SW data as both formal knowledge resources that have been created with clear and well-defined semantics (e.g., an ontology conceptualizing the human anatomy) and also structured, semistructured or unstructured data that has been a posteriori enriched with semantics (i.e., linked to a semantic knowledge resource as claimed in the Linked Data effort) through the process of semantic annotation.

We believe the integration of heterogeneous data sources into a common semantic formalism, as is OWL-DL, provides a great asset for enhancing the knowledge discovery process. We discuss all the benefits provided by ontologies and knowledge representation formalisms (e.g., OWL-DL) and claim that semantics should be taken into account during the whole mining process.

Semantics-aware mining is a very young and novel field of research. The aim of this chapter is to show how well known statistics-based techniques from artificial intelligence (e.g., clustering, association rules, etc.) can benefit from inferred information coming from logic-based approaches followed in the Semantic Web. We provide a state-of-the-art review structured according to the mining phase in which semantics is incorporated.

The chapter is organized as follows. First we introduce the motivation of integrating knowledge resources and data mining algorithms. Afterwards, we introduce the semantic web scenario which serves as the technological platform for all the semantics-aware mining methods. Taking into account this scenario, we organize and discuss the existing literature according to the mining phase in which semantics is incorporated. Finally, we give some future trends and conclusions.

INCORPORATING BACKGROUND KNOWLEDGE TO DATA MINING PROCESSES

Data Mining (DM) processes are aimed at discovering interesting patterns (data regularities) from huge amounts of data, which can be helpful for decision-making tasks such as classification, prediction and summarization. Basically, a DM algorithm takes as input a set of objects described with a fixed set of features, which are usually derived from the data structures (e.g., database schema). Most DM algorithms require input datasets to be clean, ho mogeneous in format and semantics, noise-free, non-redundant and complete (i.e., with no missing data). As a result, DM algorithms aimed at very large-scale scenarios, are usually implemented on top of a data warehouse, which homogeneously stores data facts under a simple well-defined schema (e.g., multidimensional schemas). In (Han & Kamber, 2006), a generic architecture for analysis over large databases is presented. In the Web scenario, the concept of data warehouse is still an open issue, for Web data (e.g., XML documents) present very heterogeneous structures and contents (Pérez, Berlanga, Aramburu, & Pedersen, 2008). In the SW scenario the problem of semantic heterogeneity is alleviated thanks to the existence of ontologies for vocabulary control, although SW data is also semistructured, dynamic and incomplete. The definition of data warehouses for SW data has been recently discussed in (Nebot, Berlanga, Pérez, Aramburu, & Pedersen, 2009) and (Nebot & Berlanga, 2010b).

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/xml-mining-semantic-web/73461

Related Content

A Boosting-Aided Adaptive Cluster-Based Undersampling Approach for Treatment of Class Imbalance Problem
Debashree Devi, Suyel Namasudraand Seifedine Kadry (2020). *International Journal of Data Warehousing and Mining (pp. 60-86).*
www.irma-international.org/article/a-boosting-aided-adaptive-cluster-based-undersampling-approach-for-treatment-of-class-imbalance-problem/256163

Connectionist and Evolutionary Models for Learning, Discovering and Forecasting Software Effort
Parag C. Pendharkarand Girish Subramanian (2003). *Managing Data Mining Technologies in Organizations: Techniques and Applications (pp. 250-276).*
www.irma-international.org/chapter/connectionist-evolutionary-models-learning-discovering/25770

An Immune Systems Approach for Classifying Mobile Phone Usage
Hanny Yulius Limanto, Tay Joc Cingand Andrew Watkins (2007). *International Journal of Data Warehousing and Mining (pp. 54-66).*
www.irma-international.org/article/immune-systems-approach-classifying-mobile/1784

Graph-Based Modelling of Concurrent Sequential Patterns
Jing Lu, Weiru Chenand Malcolm Keech (2010). *International Journal of Data Warehousing and Mining (pp. 41-58).*
www.irma-international.org/article/graph-based-modelling-concurrent-sequential/42151

Detecting Trends in Social Bookmarking Systems: A del.icio.us Endeavor
Robert Wetzker, Carsten Zimmermannand Christian Bauckhage (2010). *International Journal of Data Warehousing and Mining (pp. 38-57).*
www.irma-international.org/article/detecting-trends-social-bookmarking-systems/38953