Web Semantics: Science, Services and Agents on the World Wide Web 36 (2016) 1–22

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

Review article

Semantic Web in data mining and knowledge discovery: A comprehensive survey

Petar Ristoski [*], Heiko Paulheim

Data and Web Science Group, University of Mannheim, B6, 26, 68159 Mannheim, Germany

ARTICLE INFO

Article history:
Received 28 March 2015
Received in revised form
5 November 2015
Accepted 1 January 2016
Available online 8 January 2016

Keywords:
Linked Open Data
Semantic Web
Data mining
Knowledge discovery

ABSTRACT

Data Mining and Knowledge Discovery in Databases (KDD) is a research field concerned with deriving higher-level insights from data. The tasks performed in that field are knowledge intensive and can often benefit from using additional knowledge from various sources. Therefore, many approaches have been proposed in this area that combine Semantic Web data with the data mining and knowledge discovery process. This survey article gives a comprehensive overview of those approaches in different stages of the knowledge discovery process. As an example, we show how Linked Open Data can be used at various stages for building content-based recommender systems. The survey shows that, while there are numerous interesting research works performed, the full potential of the Semantic Web and Linked Open Data for data mining and KDD is still to be unlocked.

Contents

[*] Corresponding author.
E-mail addresses: petar.ristoski@informatik.uni-mannheim.de (P. Ristoski), heiko@informatik.uni-mannheim.de (H. Paulheim).

http://dx.doi.org/10.1016/j.websem.2016.01.001
1570-8268/

1. Introduction

Data mining is defined as "a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [1], or "the analysis of (often large) observational datasets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [2]. As such, data mining and knowledge discovery are typically considered knowledge intensive tasks. Thus, knowledge plays a crucial role here. Knowledge can be (a) in the primary data itself, from where it is discovered using appropriate algorithms and tools, (b) in external data, which has to be included with the problem first (such as background statistics or master file data not yet linked to the primary data), or (c) in the data analyst's mind only.

The latter two cases are interesting opportunities to enhance the value of the knowledge discovery processes. Consider the following case: a dataset consists of countries in Europe and some economic and social indicators. There are, for sure, some interesting patterns that can be discovered in the data. However, an analyst dealing with such data on a regular basis will know that some of the countries are part of the European Union, while others are not. Thus, she may add an additional variable EU_Member to the dataset, which may lead to new insights (e.g., certain patterns holding for EU member states only).

In that example, knowledge has been added to the data from the analyst's mind, but it might equally well have been contained in some exterior source of knowledge, such as Linked Open Data.

Linked Open Data (LOD) is an open, interlinked collection of datasets in machine-interpretable form, covering multiple domains from life sciences to government data [3,4]. Thus, it should be possible to make use of that vault of knowledge in a given data mining, at various steps of the knowledge discovery process.

Many approaches have been proposed in the recent past for using LOD in data mining processes, for various purposes, such as the creation of additional variables, as in the example above. With this paper, we provide a structured survey of such approaches. Following the well-known data mining process model proposed by Fayyad et al. [1], we discuss how semantic data is exploited at the different stages of the data mining model. Furthermore, we analyze how different characteristics of Linked Open Data, such as the presence of interlinks between datasets and the usage of ontologies as schemas for the data, are exploited by the different approaches.

The rest of this paper is structured as follows. Section 2 sets the scope of this survey, and puts it in the context of other surveys in similar areas. Section 3 describes the knowledge discovery process according to Fayyad et al. In Section 4, we introduce a general model for data mining using Linked Open Data, followed by a description of approaches using Semantic Web data in the different stages of the knowledge discovery process in Sections 5 through 9. In Section 10, we give an example use-case of LOD-enabled KDD process in the domain of recommender systems. We conclude with a summary of our findings, and identify a number of promising directions for future research.

2. Scope of this survey

In the last decade, a vast amount of approaches have been proposed which combine methods from data mining and knowledge discovery with Semantic Web data. The goal of those approaches is to support different data mining tasks, or to improve the Semantic Web itself. All those approaches can be divided into three broader categories:

- Using Semantic Web based approaches, Semantic Web Technologies, and Linked Open Data to support the process of knowledge discovery.
- Using data mining techniques to mine the Semantic Web, also called *Semantic Web Mining*.
- Using machine learning techniques to create and improve Semantic Web data.

Stumme et al. [5] have provided an initial survey of all three categories, later focusing more on the second category. Dating back to 2006, this survey does not reflect recent research works and trends, such as the advent and growth of Linked Open Data. More recent surveys on the second category, i.e., Semantic Web Mining, have been published by Sridevi et al. [6], Quboa et al. [7], Sivakumar et al. [8], and Dou et al. [9].

Tresp et al. [10] give an overview of the challenges and opportunities for the third category, i.e., machine learning on the Semantic Web, and using machine learning approaches to support the Semantic Web. The work has been extended in [11].

In contrast to those surveys, the first category – i.e., the usage of Semantic Web and Linked Open Data to support and improve data mining and knowledge discovery – has not been subject of a recent survey. Thus, in this survey, we focus on that area.

The aim of this survey is to give a survey on the field as broad as possible, i.e., capturing as many different research directions as possible. As a consequence, a direct comparison of approaches is not always possible, since they may have been developed with slightly different goals, tailored towards particular use cases and/or datasets, etc. Nevertheless, we try to formulate at least coarse-grained comparisons and recommendations, wherever possible.

3. The knowledge discovery process

In their seminal paper from 1996, Fayyad et al. introduced a process model for knowledge discovery processes. The model comprises five steps, which lead from raw data to actionable knowledge and insights which are of immediate value to the user. The whole process is shown in Fig. 1. It comprises five steps:

1. *Selection* The first step is developing an understanding of the application domain, capturing relevant prior knowledge, and identifying the data mining goal from the end user's perspective. Based on that understanding, the target data used in the knowledge discovery process can be chosen, i.e., selecting proper data samples and a relevant subset of variables.
2. *Preprocessing* In this step, the selected data is processed in a way that allows for a subsequent analysis. Typical actions taken in this step include the handling of missing values, the identification (and potentially correction) of noise and errors in the data, the elimination of duplicates, as well as the matching, fusion, and conflict resolution for data taken from different sources.
3. *Transformation* The third step produces a projection of the data to a form that data mining algorithms can work on—in most cases, this means turning the data into a propositional form, where each instance is represented by a feature vector. To improve the performance of subsequent data mining