Chapter 5

Approaches for Semantic Association Mining and Hidden Entities Extraction in Knowledge Base

Thabet Slimani
ISG of Tunis, Tunisia

Boutheina Ben Yaghlane
IHEC of Carthage, Tunisia

Khaled Mellouli
IHEC of Carthage, Tunisia

ABSTRACT

Due to the rapidly increasing use of information and communications technology, Semantic Web technology is being increasingly applied in a large spectrum of applications in which domain knowledge is represented by means of an ontology in order to support reasoning performed by a machine. A semantic association (SA) is a set of relationships between two entities in knowledge base represented as graph paths consisting of a sequence of links. Because the number of relationships between entities in a knowledge base might be much greater than the number of entities, it is recommended to develop tools and invent methods to discover new unexpected links and relevant semantic associations in the large store of the preliminary extracted semantic association. Semantic association mining is a rapidly growing field of research, which studies these issues in order to create efficient methods and tools to help us filter the overwhelming flow of information and extract the knowledge that reflect the user need. The authors present, in this work, an approach which allows the extraction of association rules (SWARM: Semantic Web Association Rule Mining) from a structured semantic association store. Then, present a new method which allows the discovery of relevant semantic associations between a preliminary extracted SA and predefined features, specified by user, with the use of Hyperclique Pattern (HP) approach. In addition, the authors present an approach which allows the extraction of hidden entities in knowledge base. The experimental results applied to synthetic and real world data show the benefit of the proposed methods and demonstrate their promising effectiveness.

INTRODUCTION

As described by Tim Berners-Lee, with the intense activity of Semantic Web (SW) in industry and academia (Berners-Lee, T. et al, 2001), it is reasonable to expect that increasingly more metadata describing domain information about resources on the Web will become available. The characteristic of distributed information, in the Semantic Web allows anyone to link anything to anything (linked data). The massive growth of the linked data stored in the Web is managed by a network of interrelated data sources (Semantic Link Network: SLN) which contains several types of entities (persons, companies, domain knowledge, general knowledge, scientific publications, books, etc.). SLN is a network containing semantic nodes and semantic links. A semantic node (Semantic Node: SN) can be a concept, an instance of concept, an URI, an entity, a particular form of resources in knowledge base (Zhuge H., 2004). A semantic link reflects a kind of relational knowledge represented as a pointer with a tag describing such semantic relations as *causeEffect*, *implication*, *subtype*, *similar*, *instance*, *sequence*, *reference* and *equal* (Zhuge H., 2007). Thus, SW is not a Web of documents, but a Web of semantic relations between entities denoting real world objects such as people, places and events.

The core idea of Semantic Web is to represent entities and relationships between them using ontologies for the purpose of the interoperability between machines themselves or the interoperability between machines and humans.

Consequently, the use of ontologies has been proven to be a good choice for knowledge and various information representations in a human understandable and machine-readable format consisting of entities, attributes and relationships. The direct relationship between two entities refers to us as semantic relations and indirect relationship as semantic association.

Semantic association (SA) discovery and mining is an important issue for applications including networked data. The approach of semantic association discovery initially appeared with the theories and the methods coming from the research of LSDIS laboratory at the University of Georgia[1]. A semantic association is essentially a graph-theoretic based approach that represents, discovers and interprets complex relationships between entities contained in RDF graph (Aleman-Meza et al, 2003) (Anyanwu K. et al, 2005) (Aleman-Meza B. et al, 2006) (Ning X. et al, 2006). A formal definition is presented in (Anyanwu & Sheth, 2002): "Semantic Associations capture complex relationships between entities involving sequences of predicates, and sets of predicate sequences that interact in complex ways". Since the predicates are semantic metadata extracted from various multi-source documents, this is an attempt to discover complex relationships between objects described in those documents. Detecting such associations is crucial for many research and analytical activities that are important to applications in national security, business intelligence and bioinformatics. The datasets that semantic associations operate over are RDF/RDFS graphs.

Thus, for a better exploitation of the extracted semantic association in several applications, it is essential to extract significant patterns and relevant information from preliminary extracted associations.

Semantic Web Mining aims at combining the two areas "Semantic Web" and "Web Mining". The Semantic Web addresses the first part of the new challenge posed by the great success of the current WWW aiming to make the data (also) machine-understandable, while Web Mining addresses the second part by (semi-) automatically extracting the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions. The term Semantic Web Mining can be interpreted as *Semantic* (*Web Mining*) and as (*Semantic Web*) *mining*.

In this chapter, we concentrate on mining approaches that refer to the explicit structure included in semantic association.

Related Content

Two Rough Set-based Software Tools for Analyzing Non-Deterministic Data

Mao Wu, Michinori Nakataand Hiroshi Sakai (2014). *International Journal of Rough Sets and Data Analysis (pp. 32-47).*

www.irma-international.org/article/two-rough-set-based-software-tools-for-analyzing-non-deterministic-data/111311

Financial Fraud, Technology Disruption, and Cyber-Governance

Yves Ekoué Amaïzo (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 1526-1538).*

www.irma-international.org/chapter/financial-fraud-technology-disruption-and-cyber-governance/112556

New Perspectives of Pattern Recognition for Automatic Credit Card Fraud Detection

Addisson Salazar, Gonzalo Safont, Alberto Rodriguezand Luis Vergara (2018). *Encyclopedia of Information Science and Technology, Fourth Edition (pp. 4937-4950).*

www.irma-international.org/chapter/new-perspectives-of-pattern-recognition-for-automatic-credit-card-fraud-detection/184197

A Comparison of Data Exchange Mechanisms for Real-Time Communication

Mohit Chawla, Siba Mishra, Kriti Singhand Chiranjeev Kumar (2017). *International Journal of Rough Sets and Data Analysis (pp. 66-81).*

www.irma-international.org/article/a-comparison-of-data-exchange-mechanisms-for-real-time-communication/186859

Handling Imprecise Data in Geographic Databases

Cyril de Runz, Herman Akdagand Asma Zoghlami (2015). *Encyclopedia of Information Science and Technology, Third Edition (pp. 1785-1799).*

www.irma-international.org/chapter/handling-imprecise-data-in-geographic-databases/112584