Semantic Data Mining

Protima Banerjee
Drexel University, USA

Xiaohua Hu
Drexel University, USA

Illhio Yoo
Drexel University, USA

INTRODUCTION

Over the past few decades, data mining has emerged as a field of research critical to understanding and assimilating the large stores of data accumulated by corporations, government agencies, and laboratories. Early on, mining algorithms and techniques were limited to relational data sets coming directly from On-Line Transaction Processing (OLTP) systems, or from a consolidated enterprise data warehouse. However, recent work has begun to extend the limits of data mining strategies to include "semi-structured data such as HTML and XML texts, symbolic sequences, ordered trees and relations represented by advanced logics." (Washio and Motoda, 2003)

The goal of any data mining endeavor is to detect and extract patterns in the data sets being examined. Semantic data mining is a novel approach that makes use of graph topology, one of the most fundamental and generic mathematical constructs, and semantic meaning, to scan semi-structured data for patterns. This technique has the potential to be especially powerful as graph data representation can capture so many types of semantic relationships. Current research efforts in this field are focused on utilizing graph-structured semantic information to derive complex and meaningful relationships in a wide variety of application areas-- national security and web mining being foremost among these.

In this article, we review significant segments of recent data mining research that feed into semantic data mining and describe some promising application areas.

BACKGROUND

In mathematics, a graph is viewed as a collection of vertices or nodes and a set of edges which connect pairs of those nodes; graphs may be partitioned into sub-graphs to expedite and/or simplify the mining process. A tree is defined as an acyclic sub-graph, and trees may be ordered or unordered, depending on whether or not the edges are labeled to specify precedence. If a sub-graph does not include any branches, it is called a path.

The two pioneering works in graph-based data mining, the algorithmic precursor to semantic data mining, take an approach based on greedy search. The first of these, SUBDUE, deals with conceptual graphs and is based on the Minimum Description Length (MDL) principle. (Cook and Holder, 1994) SUBDUE is designed to discover individual concepts within the graph by starting with a single vertex, which represents a potential concept, and then incrementally adding nodes to it. At each iteration, a more "abstract" concept is evaluated against the structure of the original graph, until the algorithm reaches a stopping point which is defined by the MDL heuristic. (Cook and Holder, 2000)

The second of the seminal graph mining works is called Graph Based Induction (GBI), and like SUBDUE, it is also designed to extract concepts from data sets. (Yoshida, Motoda, and Inokuchi, 1994) The GBI algorithm repeatedly compresses a graph by replacing each found sub-graph or concept with a single vertex. To avoid compressing the graph down to a single vertex, an empirical graph size definition is set to establish the size of the extracted patterns, as well as the size of the compressed graph.

Later researchers have applied several other approaches to the graph mining problem. Notable among these are the Apriori-based approach for finding frequent sub-graphs (Inokuchi, Washio, and Motoda, 2000; Kuramochi and Karypis, 2002), Inductive Logic Processing (ILP), which allows background knowledge to be incorporated in to the mining process; Inductive Database approaches which have the advantage of practical computational efficiency; and the Kernel Function approach, which uses the mathematical kernel function measure to compute similarity between two graphs. (Washio and Motoda, 2003)

Semantic data mining expands the scope of graph-based data mining from being primarily algorithmic, to include ontologies and other types of semantic information. These methods enhance the ability to systematically extract and/or construct domain specific features in data.

MAIN THRUST OF CHAPTER

Defining Semantics

The effectiveness of semantic data mining is predicated on the definition of a domain-specific structure that captures semantic meaning. Recent research suggests three possible methods of capturing this type of domain knowledge:

- Ontologies
- Semantic Associations
- Semantic Metadata

In this section, we will explore each of these in depth.

An ontology is a formal specification in a structured format, such as XML or RDF, of the concepts that exist within a given area of interest and the semantic relationships among those concepts. The most useful aspects of feature extraction and document classification, two fundamental data mining methods, are heavily dependent on semantic relationships. (Phillips and Buchanan, 2003) For example, a news document that describes "a car that ran into a gasoline station and exploded like a bomb" might not be classified as a terrorist act, while "a car bomb that exploded in a gasoline station" probably should be. (Gruenwald, McNutt and Mercier, 2003) Relational databases and flat documents alone do not have the required semantic knowledge to intelligently guide mining processes. While databases may store constraints between attributes, this is not the same as describing relationships among the attributes themselves. Ontologies are uniquely suited to characterize this semantic meta-knowledge. (Phillips and Buchanan, 2003)

In the past, ontologies have proved to be valuable in enhancing the document clustering process. (Hotho, Staab, and Strumme, 2003) While older methods of text clustering were only able to relate documents that used identical terminology, semantic clustering methods were able to take into account the conceptual similarity of terms such as might be defined in terminological resources or thesauri. Beneficial effects can be achieved for text document clustering by integrating an explicit conceptual account of terms found in ontologies such as WordNet. For example, documents containing the terms "beef" and "chicken" are found to be similar, because "beef" and "chicken" are both sub-concepts of "meat" and, at a higher level, "food". However, at a more granular clustering level, "beef" may be more similar to "pork" than "chicken" because both can be grouped together under the sub-heading of "red meat". (Hotho, Staab, and Strumme, 2003)

Ontologies have also been used to augment the knowledge discovery and knowledge sharing processes. (Phillips and Buchanan, 2003) While in the past, prior knowledge had been specified separately for each new problem, with the use of an ontology prior knowledge found to be useful for one problem area can be reused in another domain. Thus, shared knowledge can be stored even in a relatively simple ontology, and collections of ontologies can be consolidated together at later points in time to form a more comprehensive knowledge base.

At this point it should be noted that that the issues associated with ontology construction and maintenance are a research area in and of themselves. Some discussion of potential issues is presented in (Gruenwald, McNutt and Mercier, 2003) and (Phillips and Buchanan, 2003), but an extensive examination of this topic is beyond the scope of the current article.