Semantic Data Mining

Protima Banerjee
Drexel University, USA

Xiaohua Hu
Drexel University, USA

Illhoi 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 Online 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 & 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 paper, 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 that 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 & 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 & 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, & 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, & Motoda, 2000; Kuramochi & 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 & 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 informa-

tion. These methods enhance the ability to systematically extract and/or construct domain specific features in data.

MAIN THRUST

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 & 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, & 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 & Buchanan, 2003).

In the past, ontologies have proved to be valuable in enhancing the document clustering process (Hotho, Staab, & 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, & Strumme, 2003).

Ontologies have also been used to augment the knowledge discovery and knowledge sharing processes (Phillips & 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, & Mercier (2003) and Phillips & Buchanan (2003), but an extensive examination of this topic is beyond the scope of the current paper.

In addition to ontologies, another important tool in extracting and understanding meaning is semantic associations. "Semantic associations lend meaning to information, making it understandable and actionable, and provide new and possibly unexpected insights" (Aleman-Meza, et al., 2003). Looking at the Internet as a prime example, it becomes apparent that entities can be connected in multiple ways to other entities by types of relationships that cannot be known or established a priori. For example, a "student" can be related to a "university," "professors," "courses," and "grades;" but she can also be related to other entities by different relations like financial ties, familial ties, neighborhood, and etcetera. "In the Semantic Web vision, the RDF data model provides a mechanism to capture the meaning of an entity or resource by specifying how it relates to other entities or classes of resources" (Aleman-Meza et. al., 2003) – each of these relationships between entities is a "semantic association" and users can formulate queries against them. For example, semantic association queries in the port security domain may include the following:

1. Are any passengers on a ship coming into dock in the United States known to be related by blood to one or more persons on the watch list?
2. Does the cargo on that ship contain any volatile or explosive materials, and are there any passengers on board that have specialized knowledge about the usage of those materials?

Semantic associations that span several entities and these constructs are very important in domains such as national security because they may enable analysts to uncover non-obvious connections between disparate people, places and events.

In conjunction with semantic associations, semantic metadata is an important tool in understanding the mean-