Chapter III
Semantic Integration and Knowledge Discovery for Environmental Research

Zhiyuan Chen
University of Maryland, Baltimore County (UMBC), USA

Aryya Gangopadhyay
University of Maryland, Baltimore County (UMBC), USA

George Karabatis
University of Maryland, Baltimore County (UMBC), USA

Michael McGuire
University of Maryland, Baltimore County (UMBC), USA

Claire Welty
University of Maryland, Baltimore County (UMBC), USA

ABSTRACT

Environmental research and knowledge discovery both require extensive use of data stored in various sources and created in different ways for diverse purposes. We describe a new metadata approach to elicit semantic information from environmental data and implement semantics-based techniques to assist users in integrating, navigating, and mining multiple environmental data sources. Our system contains specifications of various environmental data sources and the relationships that are formed among them. User requests are augmented with semantically related data sources and automatically presented as a visual semantic network. In addition, we present a methodology for data navigation and pattern discovery using multi-resolution browsing and data mining. The data semantics are captured and utilized in terms of their patterns and trends at multiple levels of resolution. We present the efficacy of our methodology through experimental results.

INTRODUCTION

The urban environment is formed by complex interactions between natural and human systems. Studying the urban environment requires the collection and analysis of very large datasets that span many disciplines, have semantic (including spatial and temporal) differences and interdependencies, are collected and managed by multiple organizations, and are stored in varying formats. Scientific knowledge discovery is often hindered because of challenges in the integration and navigation of these disparate data. Furthermore, as the number of dimensions in the data increases, novel approaches for pattern discovery are needed.

Environmental data are collected in a variety of units (metric or SI), time increments (minutes, hours, or even days), map projections (e.g., UTM or State Plane) and spatial densities. The data are stored in numerous formats, multiple locations, and are not centralized into a single repository for easy access. To help users (mostly environmental researchers) identify data sets of interest, we use a metadata approach to extract semantically related data sources and present them to the researchers as a semantic network. Starting with an initial search (query) submitted by a researcher, we exploit stored relationships (metadata) among actual data sources to enhance the search result with additional semantically related information. Although domain experts need to manually construct the initial semantic network, which may only include a small number of sources, we introduce an algorithm to let the network expand and evolve automatically based on usage patterns. Then, we present the semantic network to the user as a visual display of a hyperbolic tree; we claim that semantic networks provide an elegant and compact technique to visualize considerable amounts of semantically relevant data sources in a simple yet powerful manner.

Once users have finalized a set of environmental data sources, based on semantic networks, they can access the actual sources to extract data and perform techniques for knowledge discovery. We introduce a new approach to integrate urban environmental data and provide scientists with semantic techniques to navigate and discover patterns in very large environmental datasets.

Our system provides access to a multitude of heterogeneous and autonomous data repositories and assists the user to navigate through the abundance of diverse data sources as if they were a single homogeneous source. More specifically, our contributions are:

1. **Recommendation of Additional and Relevant Data Sources:** We present our approach to recommend data sources that are potentially relevant to the user's search interests. Currently, it is tedious and impractical for users to locate relevant information sources by themselves. We provide a methodology that addresses this problem and automatically supplies users with additional and potentially relevant data sources that they might not be aware of. In order to discover these additional recommendations, we exploit semantic relationships between data sources. We define *semantic networks* for interrelated data sources and present an algorithm to automatically refine, augment, and expand an initial and relatively small semantic network with additional and relevant data sources; we also exploit *user profiles* to tailor resulting data sources to specific user preferences.
2. **Visualization and Navigation of Relevant Data Sources:** The semantic network with the additional sources is shown to the user as a visual hyperbolic tree improving usability by showing the semantic relationships among relevant data sources in a visual way. After the user has decided on the choice of relevant data sources of interest (based on our metadata approach) and has accessed the actual data, we also assist the user in navigating through the plethora of environmental