Chapter XIII

Clustering Similar Schema Elements Across Heterogeneous Databases: A First Step in Database Integration

Huimin Zhao, University of Wisconsin-Milwaukee, USA
Sudha Ram, University of Arizona, USA

ABSTRACT

Interschema relationship identification (IRI), that is, determining the relationships among schema elements in heterogeneous data sources, is an important first step in integrating the data sources. This chapter proposes a cluster analysis-based approach to semi-automating the IRI process, which is typically very time-consuming and requires extensive human interaction. We apply multiple clustering techniques, including K-means, hierarchical clustering, and self-organizing map (SOM) neural network, to identify similar schema elements from heterogeneous data sources, based on multiple types of features, such as naming similarity, document similarity, schema specification, data patterns, and usage patterns. We describe an SOM prototype we have developed that provides users with a visualization tool for displaying clustering results and for incremental evaluation of potentially similar elements. We also report on some empirical results demonstrating the utility of the proposed approach.
INTRODUCTION

In today’s technological environment, organizations and users are constantly faced with the challenge of integrating heterogeneous data sources. Most organizations have developed a variety of information systems for operational purposes over time. Having an integrated data source, however, is a prerequisite for decision-support applications, such as On-Line Analytical Processing (OLAP) and data mining, which require simultaneous and transparent access to data from the underlying operational systems. Business mergers and acquisitions further amplify the emergence of heterogeneous data environments and the need for data integration. Cooperating enterprises and business partners also need to share or exchange data across system boundaries for applications such as supply chain management.

The information systems that need to be integrated are typically heterogeneous in several aspects, such as hardware, operating systems, data models, database management systems (DBMS), application programming languages, structural formats, and data semantics. Many technologies are already available for bridging the syntactic differences across heterogeneous information systems. Some examples are heterogeneous DBMS, connectivity middleware (e.g., open database connectivity [ODBC], object linking and embedding for databases [OLE DB], and Java database connectivity [JDBC]), and the emerging Web services technology (Hansen, Madnick, & Siegel, 2002). However, resolving the heterogeneities in data semantics across systems is still a resource-consuming process and demands automated support.

A particularly critical step in semantic integration of heterogeneous data sources is to identify semantically corresponding schema elements, that is, tables that represent the same entity type in the real world and attributes that represent the same property of an entity type, from the data sources (Seligman, Rosenthal, Lehner, & Smith, 2002). This problem has been referred to as interschema relationship identification (IRI) (Ram & Venkataraman, 1999). IRI has been shown to be a very complex and time-consuming task in integrating large data sources due to various kinds of semantic heterogeneities among the data sources. For example, Clifton, Houseman, and Rosenthal (1997) reported on a project performed by the MITRE Corporation over a period of several years to integrate the information systems that had been developed semi-independently over decades for the U.S. Air Force. They found that tremendous effort was required from the investigator, local database administrators (DBAs), and domain experts to determine attribute correspondences across systems.

While completely automating the IRI process is generally infeasible, it is possible to semi-automate the process using techniques to reduce the amount of human interaction. We propose a cluster analysis-based approach to semi-automating the IRI process. We apply multiple clustering techniques, including K-means, hierarchical clustering, and self-organizing map (SOM) neural network, to identify similar schema elements from heterogeneous data sources, based on multiple types of features, such as naming similarity, document similarity, schema specification, data patterns, and usage patterns. An SOM prototype we have developed provides a visualization tool for users to display clustering results and for incremental evaluation of candidate solutions. We have empirically evaluated our approach using real-world heterogeneous data sources and report on some encouraging results in this chapter.
Related Content

Providing Approximate Answers Using a Knowledge Abstraction Database
www.igi-global.com/article/providing-approximate-answers-using-knowledge/3261?camid=4v1a

INDUSTRY AND PRACTICE: Research Issues in Multimedia Databases
www.igi-global.com/article/industry-practice-research-issues-multimedia/51171?camid=4v1a

Content-Based Video Semantic Analysis
Shuqiang Jiang, Yonghong Tian, Qingming Huang, Tiejun Huang and Wen Gao (2009). *Semantic Mining Technologies for Multimedia Databases* (pp. 211-235).
www.igi-global.com/chapter/content-based-video-semantic-analysis/28835?camid=4v1a
Web Content Management and Dynamic Web Pages-A Tutorial
[www.igi-global.com/chapter/web-content-management-dynamic-web/31424?camid=4v1a](www.igi-global.com/chapter/web-content-management-dynamic-web/31424?camid=4v1a)