A Scalable Algorithm for One-to-One, Onto, and Partial Schema Matching with Uninterpreted Column Names and Column Values

Boris Rabinovich, Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel

Mark Last, Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer Sheva, Israel

ABSTRACT

In this paper, the authors propose a five-step approach to the problem of identifying semantic correspondences between attributes of two database schemas. It is one of the key challenges in many database applications such as data integration and data warehousing. The authors’ research is focused on uninterpreted schema matching, where the column names and column values are uninterpreted or unreliable. The approach implements Bayesian networks, Pearson’s correlation and mutual information to identify inter-attribute dependencies. Additionally, the authors propose an extension to their algorithm that allows the user to manually enter the known mappings to improve the automated matching results. The five-step approach also allows data privacy preservation. The authors’ evaluation experiments show that the proposed approach enhances the current set of schema matching techniques.

Keywords: Attribute Dependency, Data Mining, Data Modeling, Database Semantics, Graph Matching, Schema Matching

DOI: 10.4018/JDM.2014100101
INTRODUCTION

Manual solutions to schema matching involve examining all tables in the integrated databases and identifying semantically corresponding attributes in each pair of a source table and a target table. These manual processes are prone to errors, limited in their scalability and often too slow to meet the requirement of fast information integration. The goal of automatic and semi-automatic schema matching techniques is to cope with these difficulties in such database applications as schema integration, data warehousing, etc.

There are two main approaches to automatic schema matching, namely the schema-based and the instance-based approach. The schema-based matching considers the schema information like attribute names, data types, relationship types, integrity constraints, and data descriptions. When the column names and values are uninterpreted, the instance-based schema matching approach (Liang, 2008) can identify the semantic correspondence of schema attributes by analyzing the instance data without any schema-level information. A deep insight into the content of schema elements such as data range, data domain and statistics is very helpful for uncovering semantic correspondences. To identify matching columns, which use different encodings of logically similar domains, one can also utilize the inter-attribute dependencies in each table.

The following three types of cardinality constraints should be considered in schema matching:

- **One-to-One Mapping ([1,1] – [1,1], in UML Notation):** Each attribute in the source table A has a unique match in the target table B and vice versa. This applies to mapping between tables having the same number of attributes. Consequently, the problem simply involves finding a correspondence between the attributes;
- **Onto Mapping ([0,1] – [1,1]):** Each attribute in A has a unique match in B while each attribute in B either has a unique match in A or remains unmatched. This corresponds to cases where we know that table A’s attributes are a subset of table B’s. Hence, we have to discover this subset and then to map its attributes;
- **Partial Mapping ([0,1] – [0,1]):** Each attribute in A either has a unique match in B or remains unmatched and vice versa. This corresponds to the most general and difficult cases, where we do not know which attributes of A map to B or even how many attributes of A map to B. In this case, we need to find the best subset of attributes of A to be mapped to B as well as the best mapping for this subset.

Zhao (2010) uses information-theoretic metrics to identify matching attributes of the most common types across two databases. The method application scope
Shortest Path in Transportation Network and Weighted Subdivisions
Radwa Elshawi and Joachim Gudmundsson (2012). Graph Data Management: Techniques and Applications (pp. 463-474).
www.igi-global.com/chapter/shortest-path-transportation-network-weighted/58623?camid=4v1a

Dealing with Dangerous Data: Part-Whole Validation for Low Incident, High Risk Data
www.igi-global.com/article/dealing-with-dangerous-data-part-whole-validation-for-low-incident-high-risk-data/160350?camid=4v1a