

Chapter 5

Semantic Integration of Structured and Unstructured Data in Data Warehousing and Knowledge Management Systems

Liane Haak

University of Oldenburg, Germany

ABSTRACT

Nowadays, increasing information in enterprises demands new ways of searching and connecting the existing information systems. This chapter describes an approach for the integration of structured and unstructured data focusing on the application to Data Warehousing (DW) and Knowledge Management (KM). Semantic integration is used to improve the interoperability between two well-known and established information systems in the business context of nowadays enterprises. The objective is to introduce a semantic solution in the field of Business Intelligence based on ontology integration. The main focus of this chapter is not to provide a complete literature review of all existing approaches or just to point out the motivation for such an approach. In fact, it presents, under consideration of the most important research approaches, a solution for how a Semantic Integration could be technically achieved in this specific application area. After pointing out the motivation, a short introduction to Semantic Integration, the problems and challenges occurring from it, and the application area of Knowledge Management and Data Warehousing are given. Besides the basic ideas of ontologies and ontology integration are introduced. The approach itself starts with a short overview on the determined requirements, followed by a concept for generating an ontology from a Data Warehouse System (DWS) to be finally integrated with Knowledge Management Systems (KMS) ontology. Finally SENAGATOR, an exemplarily system for semantic navigation based on integrated ontologies, is shortly introduced.

DOI: 10.4018/978-1-60960-126-3.ch005

INTRODUCTION AND MOTIVATION

The amount of information and the demand is continuously increasing. Due to this fact, information systems become a critical success factor in today's business. Every year, companies invest a large amount of money in their system landscape and infrastructure to retrieve relevant information supporting their decisions. Numerous information systems, inside and outside the company, offer a huge amount of information. Developments in this area vary, ranging from industry software solutions up to standard software often represented as stand-alone solutions. Some of these solutions offer a high adaptation degree towards the business processes and requirements of the enterprise. One example are ERP systems which are also a kind of integrated Business Information Systems. Nevertheless, most of the existing information systems have a deficit regarding their incapability to collaborate with other information systems, especially cross-company systems. This applies especially for systems with different data types and structures and in particular for those with unstructured data and structured data. The data in these kinds of systems is often partially redundant and inconsistent. Users need to have a consolidated knowledge of all systems that contain possibly relevant information. Because of alteration and permanent technological changes of the IT landscape in today's companies, it is challenging for the user to find the proper and accurate information. Therefore employees need effective and efficient ways to find relevant information. Data structures in heterogeneous information systems pose many challenges achieving this objective.

Two exemplars of heterogeneous information systems are Data Warehouse Systems (DWS) and Knowledge Management Systems (KMS). Resulting from different data sources, and because of unequal data stored in these systems, especially DWS and KMS are often not linked to each other (Dittmer & Gluchowski, 2002; Klesse et al. 2003). The main reason for the gap between such

systems is the different kind of data managed by those systems. There is predominantly structured data in Data Warehouses, which means there is an identifiable regularity within the amount of data and hence a dependency in the data set itself. In contrast, the data in the knowledge base of a Knowledge Management System is mainly unstructured and explicit in documents in different formats (for example *.pdf, *.docx or *.txt). This unstructured data eventually has a manually given structure (like an index or similar) but not an obviously identifiable dependency inside the data itself. Normally these documents are manually categorized or indexed for better searching, but this is less fine granular than the data structures used in DWS.

In large and medium sized companies, these two different kinds of systems usually exist side by side. Semantic Integration offers new possibilities to get information in context. Generally, an employee must know how to find the relevant information in the both different systems. In this case, the quality of the result is depending on his specific knowledge about the information background and about the domain he is searching in. A main benefit from Semantic Integration is that relevant content could be provided automatically. Ontologies could be used to bridge the gap between these two classes of information systems and their different data sets. Ontologies are widely applied in the area of Knowledge Management Systems but rarely used in the area of Data Warehouse Systems.

The approach introduced in this chapter focuses on Semantic Integration of Knowledge Management Systems and Data Warehouse Systems using ontologies. The objective is to show how Semantic Integration can be used to integrate different data sources under consideration of the context (semantics) to allow an overlapping search in the integrated information. The focus is not restricted to a 'pure' technical solution based on a component or layer integration of two different application architectures, for instance done by middleware

23 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/semantic-integration-structured-unstructured-data/60057

Related Content

General Adaption Framework: Enabling Interoperability for Industrial Web Resources

Olena Kaykova, Oleksiy Khriyenko, Dmytro Kovtun, Anton Naumenko, Vagan Terziyanand Andriy Zharko (2005). *International Journal on Semantic Web and Information Systems* (pp. 31-63).

www.irma-international.org/article/general-adaption-framework/2810

Research on Intelligent Landscape Design Based on Distributed Integrated Model

Xihui Tang (2023). *International Journal on Semantic Web and Information Systems* (pp. 1-19).

www.irma-international.org/article/research-on-intelligent-landscape-design-based-on-distributed-integrated-model/325002

Dynamic Knowledge Discovery in Open, Distributed and Multi-Ontology Systems: Techniques and Applications

Silvana Castano, Alfio Ferraraand Stefano Montanelli (2006). *Web Semantics & Ontology* (pp. 226-258).

www.irma-international.org/chapter/dynamic-knowledge-discovery-open-distributed/31203

Overview of MERA: An Architecture to Perform Record Linkage in Music-Related Databases

Daniel Fernández-Álvarez, José Emilio Labra Gayo, Daniel Gayo-Avelloand Patricia Ordoñez de Pablos (2019). *Semantic Web Science and Real-World Applications* (pp. 219-245).

www.irma-international.org/chapter/overview-of-mera/215067

Building Virtual Communities through a De-Marginalized View of Knowledge Networking

Kam Hou Vat (2009). *Social Web Evolution: Integrating Semantic Applications and Web 2.0 Technologies* (pp. 233-248).

www.irma-international.org/chapter/building-virtual-communities-through-marginalized/29300