Chapter III

Building an Active Content Warehouse

Serge Abiteboul, INRIA-Futurs, Orsay, France
Benjamin Nguyen, Université de Versailles Saint-Quentin, France
Gabriela Ruberg, Federal University of Rio de Janeiro, Brazil

Abstract

Non-quantitative content represents a large part of the information available nowadays, such as Web pages, e-mails, metadata about photos, etc. In order to manage this new type of information, we introduce the concept of content warehousing, the management of loosely structured data. The construction and maintenance of a content warehouse is an intricate task, involving many aspects such as feeding, cleaning and enriching semi-structured data. In this chapter, we introduce the Acware (for active content warehouse) specification language, whose goal is to help all sorts of users to organize content in a simple manner. The problem we are faced with is the following: The data are semi-structured, and the operations to be executed on this data may be of any sort. Therefore, we base our approach on XML to represent the data, and Web Services, as generic
components that can be tailored to specific applicative needs. In particular, we discuss the specification of mappings between the warehouse data and the parameters/results of services that are used to acquire and enrich the content. From the implementation point of view, an Acware specification of a content warehouse is compiled into a set of Active XML documents, i.e., XML documents with embedded Web service calls. These Active XML documents are then used to build and maintain the warehouse using the Active XML runtime environment. We illustrate the approach with a particular application drawn from microbiology and developed in the context of the French RNTL e.dot project.

Introduction

The management of non-quantitative content (e-mails, legal documents, customer relationship management information, news articles, analysis reports, meta-data about audio/video/photos, and many more) that originates from various heterogeneous environments is a key issue for companies. In particular, with the broadening use of information retrieved from the World Wide Web by many applications, the integration of unstructured and unsure information with structured, proprietary data has sparked a lot of research these last years. This chapter deals with integration difficulties, and proposes a model and system to simplify related tasks. More precisely, we propose a model to design a (Web) content warehouse and tools to build and maintain it.

The focus of this chapter is non-quantitative information, since the field of data warehousing is vast. Therefore the processing of quantitative information with traditional warehousing techniques is beyond the scope of this paper. Let us introduce the term content warehouse (CW) (Abiteboul, 2003). We call content warehouse any (large) amount of non-quantitative information. The information need not be structured, nor originate from a well established source. This contrasts with traditional data warehouses (Buneman, Davidson, Fan, Hara, & Tan, 2003; Hammer, Garcia-Molina, Widom, Labio, & Zhuge, 1995), whose sources are known, have a regular format and whose data model usually is either relational or multidimensional. Many issues arise when building a content warehouse. First, it is necessary to identify and capture relevant content. This is achieved using standard services provided by ETL (extract, transform and load) tools, search engines or Web crawlers that return flows of documents. The content is extremely diverse, ranging from structured (e.g., relational and Excel tables) to very unstructured data (e.g., plain text and HTML pages), and the data need to be cleaned, classified, transformed and enriched in order to be properly...
Potential Power and Problems in Sentiment Mining of Social Media
www.igi-global.com/article/potential-power-problems-sentiment-mining/78345?camid=4v1a

A Knowledge Worker Desktop Model (KWDM) Applied to Decision Support System
www.igi-global.com/chapter/knowledge-worker-desktop-model-kwdm/11298?camid=4v1a