Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 160 (2005) 353–364
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

Data warehouse clustering on the web

Aristides Triantafillakis *, Panagiotis Kanellis, Drakoulis Martakos

Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, University Campus, Athens 15771, Greece

Received 5 June 2002; accepted 24 July 2003
Available online 9 December 2003

Abstract

In collaborative e-commerce environments, interoperation is a prerequisite for data warehouses that are physically scattered along the value chain. Adopting system and information quality as success variables, we argue that what is required for data warehouse refreshment in this context is inherently more complex than the materialized view maintenance problem and we offer an approach that addresses refreshment in a federation of data warehouses. Defining a special kind of materialized views, we propose an open multi-agent architecture for their incremental maintenance while considering referential integrity constraints on source data.

Keywords: Decision support systems; Multi-agent systems; Data warehouse refreshment; Materialized view maintenance

1. Introduction

A base requirement for the success of a data warehouse is the ability to provide decision makers with both accurate and timely information (information quality) as well as fast query response times (system quality) [3]. Common methods that are used in practice for achieving those are largely focused on the concept of the materialized views (MVs) where a business question (i.e. query) is more quickly answered against the MV than accessing directly the base data sources [10], which may potentially involve time-demanding operations such as large-table scans and joins. However, inevitable updates at the data sources introduce the information quality problem—how to keep the MVs at a certain level of consistency with the source data when update transactions take place [13,16]. In short, system and information quality are bound together – MVs can improve query performance if we can manage to update the views consistently [4].

Collaborative electronic commerce (Ce-commerce), simply augments this challenge because these data sources are not only internal, as they largely were a mere few years ago. For example, the emergence of business communities in the form of Business-to-Business (B2B) exchanges means that the boundaries of organizations are more fluid than they used to be [14]. Ce-commerce poses new challenges to the MV maintenance problem, as 'extended enterprises' have to integrate far more

* Corresponding author.
E-mail address: triant@di.uoa.gr (A. Triantafillakis).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2003.07.012

data originating outside the organization into a single repository. A recent paper [7] by Hammer examines this trend by felicitously pointing out that the streamlining of cross-company processes "... is the next great frontier for reducing costs, enhancing quality, and speeding operations" (p. 83). In this way multiple enterprises within a shared market segment can collaboratively plan and manage the flow of goods, services, and information along the supply chain in a way that increases customer-perceived value and optimizes at the same time internal efficiencies [14].

What this means for data warehouse information and system quality is that we should start viewing Data Warehouse Refreshment (DWR) as an operational process that must provide explicit support for cross-enterprise collaboration in terms of changing business constraints. Therefore, DWR should not only be limited to MV maintenance in the context of a single warehouse (as we are accustomed to), but also support the refreshment in a federation of data warehouses. This, in turn, translates to providing a new set of algorithms and techniques to materialize views from source data that may reside in remote sites, and to incrementally maintain these views. One should also note that separate DWR processes in separate data warehouses augments this challenge, because there might exist different maintenance policies on the MVs of interest for each DWR process. To the best of our knowledge, previous work on view maintenance has mainly considered the case of SPJ (Select-Project-Join) views in a single warehouse, while not providing insights for a data warehousing solution in such an environment. In short, the problem we address in this paper can be stated as follows: "how to maintain MVs in environments where a data warehouse utilizes data from other data warehouses".

Considering system and information quality as success variables [5], we argue in the next section that for collaborative e-commerce there is more to DWR than the MV maintenance problem. In Section 3 we present an agent-based framework based on the eXtended Markup Language (XML) standard for the incremental maintenance of a special kind of MVs and in section 4 we clarify the limitations of the empirical part of our research, providing at the same time some directions for future work.

2. Issues and challenges for data warehouse refreshment in collaborative e-commerce environments

In the case of a single data warehouse, DWR can be viewed as a process that involves a hierarchy of data stores accommodating a range from source data to highly aggregated data [2,16]. The Operational Data Store (ODS) stores source data in a uniform and clean representation whilst the Corporate Data Warehouse (CDW) stores aggregate views of interest, or in other words, MVs. This hierarchy of data stores is a logical way to represent the flow of information, which goes from the data sources to the MVs.

DWR is a complex process composed of both asynchronous and parallel tasks of which the main task is to capture the differential changes in the sources and to propagate them through this hierarchy of data stores in the data warehouse. Fig. 1 presents a generic workflow for the refreshment process, where for simplicity we have not considered archiving.

The information flow begins with the loading sub-process, which 'feeds' the data warehouse and is composed of the data extraction and data cleansing steps. Other sub-processes include data integration, update propagation and customization, which propagate the summarized data to the data marts. In the remaining section we will demonstrate why DWR in a federation of data

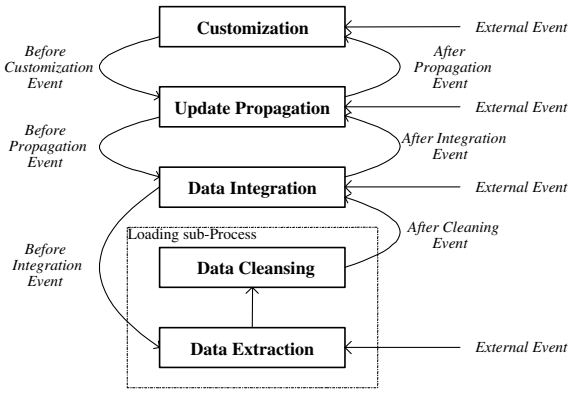

Fig. 1. A generic workflow for DWR (adapted from [2]).