IRM PRESS
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.irm-press.com
ITB9197

Chapter IV

Managing Document Taxonomies in Relational Databases

Ido Millet
Penn State Erie, USA

ABSTRACT

This chapter addresses the challenge of applying relational database technologies to manage taxonomies, which are commonly used to classify documents, knowledge and websites into a hierarchy of topics. It first describes how denormalizing the data model can facilitate data retrieval from such topic hierarchies. It then shows how the typical data maintenance difficulties associated with denormalized data models can be solved using database triggers.

INTRODUCTION

The need to maintain classification and retrieval mechanisms that rely on concept hierarchies is as old as language itself. Familiar examples include the Dewey decimal classification system used in libraries and the system for classifying life forms developed in the 1700s by Carolus Linnaeus. A more recent example is Yahoo's subject taxonomy.

Information technology has led to an explosive growth in digital documents, records, multi-media files and websites. To facilitate end-user access to these resources, topic hierarchies are frequently maintained to allow intuitive navigation and searching for resources related to specific categories. This chapter deals with the challenges of using relational database technology to maintain and facilitate queries against such topic hierarchies.

This chapter appears in the book, *Effective Databases for Text & Document Management* by Shirley A. Becker. Copyright © 2003, IRM Press, an imprint of Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

In a chapter written for another book (Millet, 2001), I discuss the generic issue of managing hierarchies in relational databases. This chapter focuses on applying these techniques to the specific needs of managing document, website and knowledge management taxonomies. For example, this domain is complicated by the typical need to classify a single document or website under multiple topics.

Relational databases and the current SQL standard are poorly suited to retrieval of hierarchical data. After demonstrating the problem, this chapter describes how two approaches to data denormalization can facilitate hierarchical data retrieval. Both approaches solve the problem of data retrieval, but as expected, come at the cost of difficult and potentially inconsistent data updates. This chapter then describes how we can address these update-related shortcomings via database triggers. Using a combination of denormalized data structure and triggers, we can have the best of both worlds: easy data retrieval and simple, consistent data updates.

THE CHALLENGE

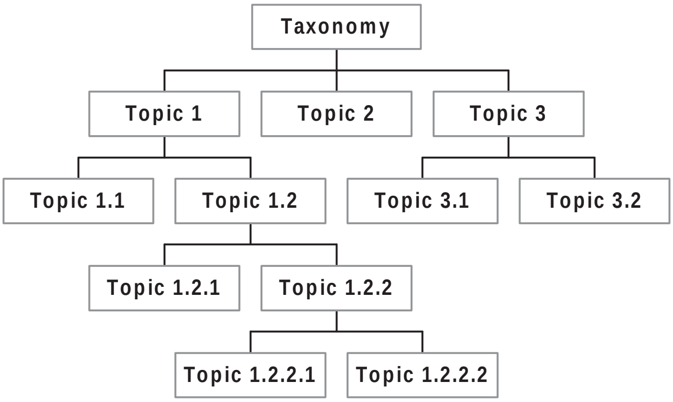
To demonstrate the data retrieval difficulties associated with topic hierarchies, consider a document database where each document is classified into a hierarchy of topics shown in Figure 1.

First, let us discuss how the topic hierarchy itself is stored in a relational database. Since each subtopic has at most one parent topic, we can implement this hierarchy via a recursive relationship. This means that each topic record maintains a foreign key pointing to the topic record above it. Figure 2 shows a data model for this situation. Note that the classify table allows us to assign a single document to multiple topics.

To demonstrate the difficulty of hierarchical data retrieval against the normalized data model in Figure 2, consider the following requests:

- Show a list of all Topics (at all levels) under Topic 1
- Show a list of all Documents (at all levels) under Topic 1
- Show how many Documents (at all levels) are classified under each Topic at Level 1 of the hierarchy

Figure 1. A Topic Hierarchy



Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.

15 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/managing-document-taxonomies-relational-databases/9204

Related Content

Object-Oriented Publish/Subscribe for Modeling and Processing Imperfect Information

Haifeng Liuand Hans Arno Jacobsen (2005). *Advances in Fuzzy Object-Oriented Databases: Modeling and Applications (pp. 301-332).*

www.irma-international.org/chapter/object-oriented-publish-subscribe-modeling/4815

Database High Availability: An Extended Survey

Moh'd A. Radaidehand Hayder Al-Ameed (2009). *Selected Readings on Database Technologies and Applications (pp. 21-43).*

www.irma-international.org/chapter/database-high-availability/28571

Issues in Mobile Electronic Commerce

Asuman Dogacand Arif Tumer (2002). *Journal of Database Management (pp. 36-42).*

www.irma-international.org/article/issues-mobile-electronic-commerce/3275

A Requirements Engineering Framework for Software Startup Companies

Sudhaman Parthasarathyand Maya Daneva (2021). *Journal of Database Management (pp. 69-94).*

www.irma-international.org/article/a-requirements-engineering-framework-for-software-startup-companies/282445

Overview of Internet of Medical Things Security Based on Blockchain Access Control

Yikai Liu, Fenglan Ju, Qunwei Zhang, Meng Zhang, Zezhong Ma, Mingduo Li, Aimin Yangand Fengchun Liu (2023). *Journal of Database Management (pp. 1-20).*

www.irma-international.org/article/overview-of-internet-of-medical-things-security-based-on-blockchain-access-control/321545