Database Technologies for E-Commerce

Rakesh Agrawal

Abstract
Beginning with a general survey of the new requirements placed on the database technology by the e-commerce applications, we discuss four specific problems as concrete examples and their solutions.

1. The search provided by most e-Commerce sites is rather primitive and is based on the traditional database metaphor of submitting an SQL query and packaging the response as an HTML page. The Eureka parametric search engine replaced the above submit/response metaphor with a continuous querying metaphor that seamlessly integrates querying with result browsing. We discuss this new metaphor and present the required computational techniques and data structures.

2. The problem of integrating documents from different sources into a master catalog is pervasive in e-Commerce websites, marketplaces and portals. The conventional technology for automating this process consisted of building a classifier that uses the categorization of documents in the master catalog to construct a model for predicting the category of unknown documents. However, many of the data sources have their own categorization, and classification accuracy can be improved by factoring in the implicit information in these source categorizations. Empirical evaluation of the catalog integrator built using this insight showed very large improvement in accuracy. We present this new integration technology.

3. Another hard problem in creating searchable E-Commerce catalogs is that the data extraction to obtain attribute-value pairs to be able to do parametric searches is difficult and error-prone. We take the audacious approach of not requiring the relationship between attributes and values to be accurately established and simply use only numbers during the searches. We present this new technology and characterize its applicability.

4. e-Commerce catalogs require data schemas that are constantly evolving and sparsely populated. The conventional horizontal row representation used in database systems fails to meet functional and manageability requirements for such data. This problem is addressed by representing objects in a vertical format, storing an object as a set of tuples. However, writing queries against this format becomes cumbersome and error prone and the existing tools no longer work. We developed techniques for defining a logical horizontal view and transforming the horizontal view queries into queries against the vertical format. By using clever rewriting, we were able to obtain very high query performance of the vertical representation on DB2. We discuss the transformations and present performance results.

We will conclude by pointing out some open research problems.

Rakesh Agrawal (Ph.D., University of Wisconsin, 1983) researches at the IBM Almaden Research Center in San Jose, California. He has received IBM’s Corporate Award, Outstanding Innovation Awards, Research Division Award, and several Invention Achievement and Patent Awards. His current research interests include data mining, text and web mining, OLAP, and electronic commerce. He has published extensively in technical journals and conferences and served on program committees on several database conferences.