

Databases on the Web

Raghu Ramakrishnan
Yahoo! Research
Santa Clara, CA 95054, USA

ABSTRACT

What role will database management play in the next generation of the Web? I believe that a number of trends signal a growing and central role for the ideas and techniques that have emerged over the past three decades of database research. However, we will have to re-examine some basic issues and build a new generation of data management infrastructures in order to address many of the new demands made by Web data management.

Categories and Subject Descriptors

H.3.4 [Information Systems]: Information Storage and Retrieval – *Systems and Software*

General Terms: Management, Performance, Design

Keywords: Web search, Web data management

1. INTRODUCTION

Data is central to Yahoo!’s business. Some common workload patterns, made significantly more challenging by the sheer magnitude of the datasets involved, are:

Nugget storage/retrieval: User-facing properties constantly store and retrieve data nuggets. In some cases all that is required is key-based get/put, e.g., store/retrieve Bob’s stock portfolio. Increasingly, more complex retrieval is required, e.g. get Bob’s friends’ favorite restaurants in cities on Bob’s travel itinerary.

Canned large-scale processing: Many properties rely on large-scale preprocessing of massive data sets. The obvious example is search, which does link analysis over the many-billion page Web graph. However, many other Yahoo! properties also rely on large-scale classification, clustering, entity extraction or wrapper induction over large corpora of Web pages, news or advertisements.

Ad-hoc large-scale processing: Analysts look for patterns to inform the design of new products (and tune existing ones). Data sets of interest include Web crawls, click streams, advertisement entries, and various derived data. These “R&D queries” tend to be ad-hoc, and make heavy use of user-defined transformations, aggregation, and in some cases joins.

In addition, the following trends are creating significant new challenges and opportunities for Web data management:

The Emergence of Structure: The Web contains a variety of data, ranging from html documents to query logs, user profiles, advertisements, user interactions, and multi-media. Capturing and exploiting metadata describing common structure is a key to next-generation Web applications, including search and advertising.

The use of many different types of large-scale data could benefit from additional, integrated structure. Existing content repositories often do not expose structure explicitly (e.g.,

individuals’ home pages often contain contact information, but embedded in free text), and even when they do, rarely conform to uniform metadata standards (e.g. *price* on one shopping site might correspond to *cost* on another; a well-known data integration problem). While structure extraction technology ranging from entity extraction to site wrappers is currently applied only to a limited degree; it is likely this will change.

Online Communities and User-Generated Content: Online communities are a fundamental and increasingly important part of the Web. We routinely see bulletin boards with presence and real-time chat, tagging and folksonomy data arrangement, rating and reputation systems, and other such capabilities, on a wide range of sites. Hosted application providers for such capabilities, with potential advantages such as single authentication, integrated payment systems, and shared personalization, are on the horizon.

User-contributed content is becoming a key resource. In many cases, it conforms to a large and growing set of schemas meant to incorporate specifics of either data type (question/answer pairs, new articles, blog posts, Web pages), or domain data (auto repair, medical information, etc). As a variant of user-contributed content, we see user-contributed metadata in the form of reviews, tags, bookmarks, ratings, or even clicks.

The Web as a Delivery Channel: The Web has become a powerful and ubiquitous means of delivering a range of end-user (e.g., email), collaborative (e.g., IM, Y! Answers), and enterprise (e.g., Salesforce.com) applications to hundreds of millions of users. As online application development moves in the direction of “mash-ups” of online APIs (to a wide range of capabilities that may be combined into an application), the requirements on the backend change substantially.

This has created a radically different approach to developing and distributing applications, disrupting the traditional software distribution model. In turn, it has challenged us to develop new types of service-oriented software platforms, new kinds of customizable application environments, and forced us to think about massively distributed systems with novel quality of service guarantees, fail-over mechanisms, and the ability to manage massive numbers of application instances.

2. ACKNOWLEDGMENTS

This talk is based on joint work with many people at Yahoo! Research; in particular, Sihem Amer-Yahia, Phil Bohannon, Brian Cooper, Ravi Kumar, Chris Olston, Andrew Tomkins, Jai Shanmugasundaram and members of the Community Systems group. While I focus on data management related issues in this talk, see [1] for a broader view.

3. REFERENCE

- [1] The Yahoo! Research Team: Content, Metadata, and Behavioral Information: Directions for Yahoo! Research. *IEEE Data Engineering Bulletin*, (Dec. 2006), 10-18.