Chapter VI

Caching on the Web

Mehregan Mahdavi, The University of New South Wales, Australia
Boualem Benatallah, The University of New South Wales, Australia

Abstract

The World Wide Web provides a means for sharing data and applications among users. However, its performance, in particular providing fast response time, is still an issue. Caching is a key technique that addresses some of the performance issues in today’s Web-enabled applications. Deploying dynamic data, especially in an emerging class of Web applications, called Web portals, makes caching even more interesting. In this chapter, we study Web caching techniques with focus on dynamic content. We also discuss the limitations of caching in Web portals and study a solution that addresses these limitations. The solution is based on the collaboration between the portal and its providers.

Introduction

The World Wide Web has influenced many aspects of life such as communication, education, business, shopping, and entertainment. There are many resources on the Internet; some provide data for being used and shared among users, while others are designed to provide services. For example, some Web sites such as Web sites of universities, people’s home pages, and yellow and white pages provide data. Examples of Web sites that provide services include
those used for online shopping, booking and buying flight tickets, and banking. Users find and access appropriate data or services through Web browsers.

Performance is one of the major issues in today’s Web-enabled applications. Previous research has shown that the abandonment of Web sites dramatically increases with an increase in response time (Zona Research Inc., 2001), resulting in loss of revenue by businesses. In other words, providing fast response time is one of the critical issues that today’s Web applications must deal with. Nowadays, many Web sites employ dynamic Web pages by accessing a back-end database and formatting the result into HTML (hypertext markup language) pages. Accessing the database and assembling the final result on the fly is an expensive process and a contributive factor to the performance issue. Server workload or failure, and network traffic are some other contributing factors for slow response times.

Caching is a key technique that addresses some of the performance issues of Web-enabled applications. Caching can improve the response time. As a result, customer satisfaction is increased and better revenue for the business is generated. In addition, network traffic and the workload on the providers’ servers are considerably reduced. This in turn improves throughput and scalability, and reduces hardware and software costs.

In this chapter, we study caching solutions for Web applications with focus on dynamic content. We discuss the limitations of existing caching solutions in Web portals as an emerging class of Web applications, and we introduce a caching strategy based on the collaboration between the portal and its providers. Providers trace their logs, extract information to identify good candidates for caching, and notify the portal. Caching at the portal is decided based on the scores calculated by providers and is associated with objects. We also address the issue of heterogeneous scoring policies by different providers and introduce mechanisms to regulate them.

Web Caching Overview

A Web cache is located somewhere between the Web browser and the origin content provider and stores Web objects for future requests. Candidate Web objects for caching include HTML (Hyper Text Markup Language) pages, images, audio and video files, XML (eXtensible Markup Language) pages or fragments, query results (e.g., SQL [Structured Query Language]), results of dynamic Web pages (e.g., JSP [Java Server Pages] or Servlet, ASP [Active Server Pages], PHP [Personal Home Page Hypertext Preprocessor]), and programs (e.g., Java applets). When the cache server receives a request, it checks the cache to see whether the request can be answered locally or not. If it can, then the result is sent to the client. Otherwise, the request will be forwarded to the content provider. A Web cache can result in one or more of the following.

• Reducing network traffic and therefore reducing network costs for both content providers and consumers
• Reducing user-perceived delay

Copyright © 2007, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
Watching the Web: An Ontological and Epistemological Critique of Web-Traffic Measurement
www.igi-global.com/chapter/watching-web-ontological-epistemological-critique/21996?camid=4v1a