Chapter 7
Web Usage Mining: Improving the Performance of Web-Based Application through Web Mining

Sathiyamoorthi V
Sona College of Technology, India

ABSTRACT

In recent days, Internet technology has provided a lot of services for sharing and distributing information across the world. Among all the services, World Wide Web (WWW) plays a significant role. The slow retrieval of Web pages may lessen the interest of users from accessing them. To deal with this problem, Web caching and Web pre-fetching are the two techniques used. Web proxy caching plays a key role in improving Web performance by keeping Web objects that are likely to be used in the near future in the proxy server which is closer to the end user. It helps in reducing user perceived latency, network bandwidth utilization, and alleviating loads on the Web servers. Thus, it improves the efficiency and scalability of Web based system. This chapter gives an overview of Web usage mining and its application on Web and discusses various approaches for improving the performance of Web.

INTRODUCTION

It is generally observed throughout the world that in the last two decades, while the average speed of computers has almost doubled in a span of around eighteen months, the average speed of the network has doubled merely in a span of just eight months! In order to improve the performance, more and more researchers are focusing their research in the field of computers and its related technologies. Internet is one such technology that plays a major role in simplifying the information sharing and retrieval. World Wide Web (WWW) is one such service provided by the Internet. It acts as a medium for sharing of information. As a result, millions of applications run on the Internet and cause increased network traffic and put a great demand on the available network infrastructure. The rapid growth of the WWW and Web development has been the result of many innovative advances in Web technology. Web works with arrays of technologies for better communication with the Internet user, but the inconveniences to users still persistent among the users. A possible solution for this problem is, to add a new resource and distribute

DOI: 10.4018/978-1-5225-1877-8.ch007

the network traffic across one or more resources. Web caching is one such method which is widely used to reduce the network traffic by storing Webpages to a location nearer to the client (Pallis et al 2008). A proxy server is responsible for Web caching which acts as a mediator between the Web server and the Web client and thus reduces latency in retrieving the pages. This proxy-based Web caching system can still be improved to control the performance of the Web. This chapter thus focuses on a methodology for improving the proxy-based Web caching system. It uses Web Usage Mining (WUM) to optimize the performance of the Web based system through Web caching and pre-fetching.

Basic Terminologies

The word cache means, fastest memory. Caching refers to the storage of recently or frequently retrieved information for future access (Wessels & Duane 2001). It reduces latency in accessing Webpages and also improves the performance of Web-based systems. The most important terms used in cache memory references are: i) cache hit and ii) cache miss. If the user requested object is not present in the cache, then it is called a cache miss else if it is present then it is called a cache hit. The hit rate also known as hit ratio (HR) is the percentage of user's requests served from the cache. The byte hit rate (BHR) also known as byte hit ratio is the percentage of bytes served from the cache. Thus, caching saves bandwidth utilization and byte hit rate used to measures it. Byte hit rate is used to measure network performance whereas hit rate used to measure the user satisfaction. Client is the program that makes a request for some resources or objects from the server. Server is a service provider for the client. It is the storage of multiple heterogeneous resources that are accessed by multiple clients. Each server has a unique name or identifier through which a client can refer the server and make requests for some resources. This unique identifier is called URL. Usually, communication can take place only between clients which initiates the communication by sending a request and the server which processes the user request and sends response. Web-based system is an example for client server system environment. In this, the most commonly used client is Web browser. A proxy server lies in between clients and servers and reduces latency. It acts as a client when interacting with a Web server and acts as a server when interacting with a browser. Web documents might be classified into either dynamic or static documents. Dynamic documents are generated by the server when a request arrives and is dependent on time of user's request whereas static documents are produced independent of any user's request. These documents are identified based on file extensions such as.jsp,.asp,.html,.xml and so on. The interaction between client and server is initiated by the protocol called Hypertext Transfer Protocol (HTTP). The response header of HTTP protocol contains information that is originally requested by the user and control information which includes size, type and various cache control directives. Response header also contains status code which tells about the success or failure of the user request. The commonly exchanged status codes are 200 (OK), 404 (Not Found) (Wessels & Duane 2001). The other codes exchanged between proxy server and Web server is 304 (Freshness). Cache has limited space. So when a new object arrives and no space has been left for the incoming object, then it must remove some objects from the cache. To do this, a cache replacement policy has been designed. It assigns a priority value to each object in the cache by using some heuristic technique and removes the least expensive objects based on priority value. It depends on the cache replacement policies used for replacement. Different heuristic techniques have been adapted by various replacement policies. Some of these standard policies include LRU, LFU, and FIFO. A user session is interaction between a client and a server during particular time period or visits. During this session, user might have accessed large number of Web pages. Web pages are piece of information present in a