Reducing Retrieval Latencies in the Web: the Past, the Present, and the Future

Themistoklis Palpanas
Department of Computer Science
University of Toronto
10 King's College Road, Toronto
Ontario, M5S 3G4, CANADA
themis@cs.toronto.edu

Balachander Krishnamurthy
AT&T Labs-Research
180 Park Ave
Florham Park, NJ 07932 USA
bala@research.att.com

Graduate Department of Computer Science
University of Toronto
Technical Report CSRG-378

Abstract

One of the main directions of research in the Web is to reduce the time latencies users experience when navigating through Web sites. The work in this direction originated from relevant research in operating systems (e.g. caching), where the aim is to reduce file system latencies. Caching is already being used in the Web domain. However, recent studies indicate that the benefits from this technique are rather limited. Thus, another method that was first applied in operating systems, prefetching, is now being studied in the Web context. In addition to prefetching, there is a number of other techniques that are being examined in the Web domain and are specifically targeted to this environment. These techniques include compression, and piggybacking of information between the servers and the clients. The main contributions of this survey paper are: A taxonomy of the caching and prefetching techniques, according to the underlying principles, and the algorithms used; a critical survey of the methods that have been proposed so far for dealing with the retrieval latencies problem, and how each one of those fits in the taxonomy; a discussion of the results of the past and ongoing work, and the identification of future research directions.