Speeding Up the Internet in Big Data Era:
Exploiting Historical User Request Patterns for Web Caching to Reduce User Delays

Chetan (Chet) Kumar
California State University – San Marcos, USA

INTRODUCTION

The Internet has witnessed a tremendous growth in the amount of available information, and this trend of increasing traffic is likely to continue. The rapid rise of Big Data across the technology world has lead to an explosion of data. According to McAfee and Brynjolfsson (2012) the key characteristic of Big Data that separates it from analytics of the past is the volume, velocity, and variety of data. They quantify that more data now cross the Internet every second than were stored in the entire Internet 20 years ago. A few excerpts from McAfee and Brynjolfsson (2012) on volume, velocity, and variety of Big Data are as follows:

- ***Volume:*** *As of 2012, about 2.5 exabytes of data are created each day, and that number is doubling every 40 months or so...This gives companies an opportunity to work with many petabyes of data in a single data set—and not just from the internet. For instance, it is estimated that Walmart collects more than 2.5 petabytes of data every hour from its customer transactions.*
- ***Velocity:*** *For many applications, the speed of data creation is even more important than the volume. Real-time or nearly real-time information makes it possible for a company to be much more agile than its competitors. For instance,…a group at MIT Media Lab used location data from mobile phones to infer how many people were in Macy's parking lots on Black Friday…This made it possible to estimate the retailer's sales on that critical day even before Macy's itself had recorded those sales.*
- ***Variety:*** *Big data takes the form of messages, updates, and images posted to social networks; readings from sensors; GPS signals from cell phones, and more. Many of the most important sources of big data are relatively new. The huge amounts of information from social networks, for example, are only as old as the networks themselves; Facebook was launched in 2004, Twitter in 2006. The same holds for smartphones and the other mobile devices that now provide enormous streams of data tied to people, activities, and locations. (McAfee and Brynjolfsson 2012)*

According to IDC Worldwide Big Data Technology and Services 2014–2018 Forecast, the trend of increasing Big Data applications online is to continue (IDC Forecast Report 2014). The Report states: "IDC expects the Big Data technology and services market to grow at a 26.24% compound annual growth rate through 2018 to reach $41.52 billion." Despite technological advances this traffic increase can lead to significant user delays in web access (Sorn & Tsuyoshi 2013, Zhao & Wu 2013, Kumar 2010, Hosanagar & Tan 2004, Datta et al. 2003).

Web caching is one approach to reduce such delays. Caching involves temporary storage of web object copies at locations that are relatively close to the end user. As a result user requests can be served faster

DOI: 10.4018/978-1-4666-9787-4.ch062

than if they were served directly from the origin web server (Davison, 2013; Ali et al,. 2012; Hosanagar & Tan, 2004).

Caching can be performed at different levels in a computer network. Proxy caches are situated at computer network access points for web users (Davison, 2013). Other locations where caching may be performed include browser and web-server levels (Ali et al. 2012; Kumar, 2010; Kumar & Norris, 2008; Davison, 2001). Proxy caches can store copies of web objects and directly serve requests for them in the network, consequently avoiding repeated requests to origin web servers. As a result there is reduced network traffic, load on web servers, and average delays experienced by web users (Ali et al. 2012; Datta et al. 2003; Cao & Irani, 1997). Kumar (2009) illustrate the benefit of a network of proxy caches using an example of the IRCache network (www.ircache.net). Figure 1 shows how a network of proxy caches with nodes at three locations can reduce user delays. If the U.K. node has requests for web pages chrysler.com, ford.com, and mercedes-benz.com, that it has not cached, then these requests can be satisfied from the U.S. and Germany nodes. Therefore the U.K. node need not go to the origin web server to satisfy requests for objects it does not hold itself but are held by neighbor caches. Since origin server requests typically have the longest waiting times, by reducing them proxy caches can significantly reduce network delays (Ali et al. 2012; Kumar, 2009). Proxy caching is widely used by computer network administrators and technology providers (Davison, 2013). Examples include proxy caching solution providers such as Oracle (www.oracle.com/technology/products/ias/web_cache/index.html), content delivery network (CDN) firms such as Akamai (www.akamai.com), and Internet service providers (ISP) such as AT&T (www.att.com). The following are two illustrations, adapted from Davison (2013), of how some firms may practically benefit from caching. In one case a company such as Intel may employ a proxy cache near its network gateway to serve its many users (e.g., clients within Intel) with cached objects from many servers. As a result Intel reduces the bandwidth required over expensive dedicated Internet connections. In another scenario a content provider such as Yahoo can place a proxy cache directly in front of a particular server to reduce the number of requests that the server must handle. This service to speed up content delivery, also called reverse caching as a proxy node may cache objects for many clients but from usually only one server, is professionally provided by CDN firms such as Akamai. In both scenarios access delays are reduced thereby benefitting all Internet users (Davison 2013). Of course in choosing caching solutions, as in any IT investment decision, firms have to evaluate costs of an implementation versus its benefit, before deciding on the appropriate caching service. In this article we discuss some proxy caching approaches that exploit historical user request patterns to reduce user request delays (Irani & Lam, 2015; Kumar, 2010; Kumar & Norris, 2008; Zeng et al., 2004).

RELATED LITERATURE AND BACKGROUND

There is a growing interest in caching due to its application in reducing user delays while accessing the increasingly congested Internet (Sorn &Tsuyoshi, 2013; Davison, 2013; Kumar, 2010; Datta et al. 2003). Zhang et al. (2103), Ali et al. (2012), Zeng et al. (2004), and Podlipnig and Boszormenyi (2003) provide an extensive survey of numerous caching techniques. These include popular cache replacement strategies such as least recently used (LRU), where the least recently requested object is evicted from the cache to make space for a new one, and their many extensions. While most caching studies focus on improving performance on measures such as user latency and bandwidth reduction, there have been relatively fewer studies that consider a data or model driven approach for managing caches. Ali et al. (2012) discuss web proxy caching approaches based on machine learning techniques. Zhao and Wu