ABSTRACT

Searching in the web is an interesting application. Information retrieval using web is an ever green field for research and is always the choice of researchers due to its importance in any real time application. Every day the search engines processes billions of queries. The increase in web users and web applications has led to increase in latency, network congestion and server overloading. Many research activities are going on to improve the search performance. One well known traditional methodology is web caching. When web caching is deployed, it caches popular objects at the locations close to the clients.

Cache can be deployed at three levels - client level, proxy level and original server level. Client level local cache is a part of hard disk in local browser, allocated for web object caching. But the current browser cache cannot share their contents with other clients. Each browser cache is isolated and its resources are not utilized completely. A co-operative browser level web cache sharing system based on chord is suggested for performance improvement. A proxy server is residing between client machines and the web server. It can service request from a number of clients. For each user request, the system first checks the result in the local cache or in the authenticated cooperative caches. In both the cases if a miss occurs, the proxy checks whether it has a valid retrieved result. If cache hit occurs, it sends the requested result to the client on behalf of origin server. If a miss occurs, it forwards the request to the origin server. The retrieved results are stored in the proxy cache for future purpose.
Web prefetching is an additional technique to improve the web caching mechanism. Predicted objects are prefetched from the origin server during the browser idle time and stored in the client’s cache. Prefetching increases the cache hits and reduces the user-perceived latency. The client side prediction helps in keeping track of the patterns of a single user across the various web servers. The proxy based prefetching helps in keeping track of the patterns of a group of users accessing many web servers.

The objective of the present work is to improve the search performance using an integrated approach. Accordingly, the main contributions of this work include the following:

1. An Access Log Manager (ALM) is used for providing classified training data set to proxy server. Data without classification is huge and includes many irrelevancies. A Support Vector Machine (SVM) classifier is utilized to avoid irrelevancy.

   Once the sample data set is obtained, the proxy server will perform pattern identification.

2. The Proxy Cache Manager (PCM) will perform pattern identification through clustering. A transparent proxy cache server software is implemented along with web prefetching technique for achieving a better performance.

   The main advantage is that it prevents bandwidth underutilization and reduces the user perceived latency and reduces cost of downloading.
3. An authenticated co-operative client network is created for web object sharing with the help of Dynamic Hash Table (DHT) chord protocol.

4. Real time implementation of an integrated web object sharing with cluster based prefetching scheme.

The effectiveness of the integrated system is increased by sharing the web objects and increasing the cache hit by 23%, minimizing network congestion by 35%, and reduces user perceived latency. There is a clear distinction between the values of metrics before and after classification and clustering. The result obtained shows that the developed system has a higher efficiency. Therefore the integrated web cache system with SVM based classification and cluster based prefetching is found to be a viable solution for effective information retrieval.