SUMMARY

This research is concerned with the study and analysis of search engines, storage mechanism of the search engine repository and the algorithms used for search and it proposes a traffic adaptive optimum updating scheme for updating the search engine repository. It also proposes a semantic search process to improve precision in search results. Under conventional crawling scheme, web crawlers use polling technique to update search engine repository. Crawlers make HTTP requests to determine whether a page has undergone a change or not. If there is no change, then the HTTP requests were needless and the web server could have been better spent servicing users. Moreover, the web crawlers impose extra load on network bandwidth and web servers.

The proposed updating scheme aims to eliminate the needless requests of web crawlers in updating the search engine repository by using a top down approach. A self-managing autonomic computing architecture is proposed to regulate the load on network bandwidth and web servers. Clients and servers are designed as AC nodes and AC servers. The task knowledge of AC nodes and AC servers are designed to contain the representation of actual system behavior, the system itself and the environment perceived by the system. Delta extraction, web negotiator and request analyser algorithms use the task knowledge to verify the load on the network and web server before issuing and accepting the requests. The load on network bandwidth and web server is further reduced by using delta transmission of the updated document.

The proposed updating scheme is tested for the freshness of the search engine repository and is compared with the page refresh policies used by web crawlers. The load on network bandwidth and web servers is tested for effective resource utilization and is compared with the one consumed during crawler updation. Web workload model is used in testing the architecture. The workload parameters are validated for their characteristics when used to generate results.

The thesis also proposes a semantic search process to improve the relevance of results produced by search engines. Search process operates in phases to transform the user query to search results. Query formulation, query processing, repository indexing, searching and ranking are the phases involved in the search process. At present the results for a given query are identical and independent of the users or the context in which the user made the request. The proposed semantic search process intends to introduce semantics of the web document in the search process itself. Context and category of the web document are included using additional entity headers with the HTTP requests. The index of the search engine is extended to accommodate the semantics. Semantic pagerank of the web document is calculated and it is used to rank the search results. Taking into account the effect of additional HTTP headers on network bandwidth, the precision of the search results has been tested using user relevancy reports. The test results show that the proposed updating algorithm and semantic search process algorithm improve the effectiveness of the search results.