ABSTRACT

WWW (W3) is defined as a vast collection of interconnected hyperlinked documents. It is a service that operates on Internet. Web directories, Search engines and Meta search engines are the tools used to retrieve information from W3. Search Engines maintain their own repository of web documents and return thousands of results in response to a user query. Statistics have shown that 37% of users always get too many results. Generally, users are not interested in all the results. Another survey has indicated that 29% of the users never get the information on the first result’s page, while 34% users seldom get the information on very first result page. This leads to the problem of information overkill. This problem further aggravates in case of inexperienced users, trying to search the information from the web. Moreover, only 40% of users try to search the results up to third page, while 28.8% of the users go up to second page and 17.2% are reluctant to go further beyond the first page.

It is a challenge for every search engine to identify more relevant documents and then to display them earlier at top position in response to a user query. The solution to this problem came from focused search engine. Focused search engines concentrate more on the quality of the information rather than the quantity. It has been observed that existing focused search engines decide the relevance of a web page based on their popularity, hyperlinks in the web page, text surrounding the hyperlinks and number of in-links pointed to a web page. They also use query term occurrence, user search history and profile, query logs, user behaviour and documents stored on user’s desktop, as some other heuristics. But these approaches have failed to resolve the ambiguity in query keywords due to the Polysemy property of words i.e. they do not consider query words to have multiple meanings. As a result, sometimes focused search engines also provide irrelevant results at top positions that are different from the user’s desired sense. Thus, there is a need to identify the context of the user query keywords as well as the words present in web pages. Then, use this contextual information to search relevant documents in the desired context. The issue of handling contextual information is addressed in this entire research work.

A novel context based focused search engine (CBFSE) has been proposed as an improvement in the existing search engine architecture. Modifications are done in the top layer at user interface and query processor level and also in the middle layer at indexer level. Context
based relevance calculator sub-module is added at query processor level to provide context score as a heuristic for relevance evaluation of web pages. In the proposed CBFSE architecture, ranking module uses context score to rank the web documents. This helps to put contextually more related documents in desired context at top positions. Back-link extractor sub-module is added in the middle layer to extract and include only contextually relevant back-links in final results. Finally, the user interface is modified to allow display of various contextual senses of query keywords to the user. User can choose a specific context, and the contextually relevant documents are displayed to them.

This thesis resulted in design and implementation of a prototype tool based on our proposed context based focused search engine architecture. The prototype was tested on 100 different query keywords having approx. 300 different contextual senses which can be scaled up further. Approximately, total 15,000 URLs have been evaluated using context based relevance evaluation mechanism.

The limitation of thesis work done is that it is valid only for the textual documents. Documents containing images and video are not considered. Moreover, the work done is dependent on WordNet dictionary for various contextual senses. The technique failed in case of proper noun where WordNet does not provide contextual senses.

The thesis work can be used by popular search engines by integrating it with their existing techniques to display more relevant as well as popular web documents to the end user at top positions. This may increase precision of their results. It can also be used at the crawler end to improve the quality of backend repository.