ABSTRACT

In the contemporary era, Information Retrieval (IR) from World Wide Web (WWW) or other repositories has gained unprecedented significance. The rationale behind this, there is an exponential growth of web documents over WWW from different sources and information systems. People from all walks of life get benefited while surfing through search engines that produce millions of results from WWW. From the user’s point of view, there are some problems with results provided by the search engines for a given query. From the user's point of view, search engines produce a massive number of web pages as results. Sometimes, the results may not reflect the user intent. The user needs to navigate further or refine the query to get intended results. When the search engine does not provide expected results at the top of the results page, and it leads to user dissatisfaction. From the search engine point of view, it is essential to understand the intention of the user behind giving a web query. Moreover, users might not have the skill to frame search queries properly. In other words, many users do not know how to formulate queries meaningfully. It creates a problem for the search engines in producing useful results. Nevertheless, it is a very challenging problem to understand, to process based on keywords and to provide highly relevant results for a given query efficiently.

Existing research on information retrieval from WWW found in the literature showed many insights. Due to a massive increase in the document corpus, retrieval of highly relevant and user intended results became a challenging task. Information retrieval in this context depends on various factors like the meaning of query words, number of in-links, number of out-links, the frequency of query words in documents, and classification of web query and so on. Efficient classification of web queries can help in the better interpretation of queries and represent significant results for presentation rather than presenting all web pages to the user against the given query. Search engines use ranking algorithms to display results for a given query from highly relevant documents to least relevant documents. Learning to Rank (LTR) is an essential application of machine learning and computational intelligence which paves the way for rendering high-quality
results. LTR is a standard algorithm used with search engines, recommender systems and question-answer kind of applications for presenting results.

This research aims to propose various algorithms for retrieving highly relevant, and user intended results while surfing the web. Information retrieval from World Wide Web primarily includes two steps, retrieving relevant web documents and sorting these retrieved documents from highly relevant to least relevant using ranking algorithms. As a first step to retrieve relevant documents, this thesis proposes efficient navigation of query results using simple ordinal classification algorithm. This algorithm uses B-tree for classification of queries to improve the relevancy of results. This algorithm reduces most of the non-relevant results and retrieves relevant documents as intermediate results. As a second step, an improved Page Rank algorithm is proposed to rank the retrieved documents. This algorithm is based on structure, content, and distance between the web pages. It also includes session logs to reflect the intent of users. This algorithm presents better results with high precision compared to the existing Page Rank algorithms.

Afterwards, dynamic navigation of B-tree using improved Page Rank algorithm is proposed. This algorithm is based on user search session logs. As the simple ordinal classification approach is based on regression models, time complexity is more. So the proposed algorithm uses categorization factor to calculate the relevancy of a query to the specific category. The usage of B-tree for dynamic navigation of web queries lessens the navigation time while processing a query. The usage of improved distance Page Rank algorithm sorts the above-resulted documents efficiently. Our proposed approach for categorization and ranking of web queries is conversant, and it yields more relevant web pages regarding ranking and content in comparison to present ranking and categorization techniques.

The main problem with improved Page Rank algorithm is scalability, and it cannot handle large real-time data sets. To overcome these issues, a new algorithm named Scalable List wise Online Learning Algorithm for Ranking (SLOLAR) is proposed. Most of the
Page Rank and learning to rank algorithms build their training data model from laceration. The proposed SLOLAR algorithm overcomes the scalability problem of existing ranking approaches. It results in better performance over its counterparts using benchmark datasets. The algorithms proposed in this research can be used for improving the performance of search engines. These algorithms provide highly relevant results by reflecting user intent, while processing user queries related to textual or web documents retrieval.