PREFACE

Information Retrieval from the World Wide Web is one of the fascinating and provoking problems in web mining today due to epidemic growth in the data. Information retrieval plays a vital role in image processing, speech recognition, social network analysis, filtering the information, geographical information, etc. The main problem with the information retrieval process, as it is confided on many aspects like the number of in links, the number of out links to and from the document, changing nature of word meanings, keyword appears in the documents, the number of keywords present in the given query. The problem of obtaining more relevant results for the given query by reducing non-relevant results can be resolved by considering an efficient classification approach during the navigation of the web query with an efficient ranking algorithm.

In the recent times, information retrieval from WWW has attracted considerable research in the area of web mining. Search queries made up of keywords that are used to retrieve the documents efficiently. Based on the keywords present in the given query, search engines produce voluminous results. But the problem here is user cannot navigate through all results. Here is a real-time example, “web services and applications” query has produced 8,36,00,000 results out of which in the first page itself around three to four web pages are non-relevant. Therefore, it is essential to present results after optimization based on an efficient information retrieval process. So many researchers proposed solutions for optimizing the results for the given query. However, there was a need for further research by inflecting the intents of users, as the web of rich information and it is growing exponentially with millions of web pages added every year.

The author of this thesis focused on enhancing both the activities aforementioned. An algorithm named Simple Ordinal Classification (SOC) is proposed for B-tree based effective navigation of query results. It computes loss function of queries, and the queries are categorized into informative, navigational and transactional. Based on the category to which a query belongs, the irrelevant documents are eliminated to crop up in the B-tree, and thus effective navigation of query results is achieved besides reducing navigation cost.
Afterwards, the results of this enhancement are coupled with proposed improved distance Page Rank algorithm which employs categorization factor, visit count, Page Rank and distance among documents. The new Page Rank algorithm has resulted in further optimization of the two activities in question.

Afterwards learning to rank (LTR) with pointwise, pairwise and listwise approaches are studied. A new LTR algorithm named Scalable Listwise Online Learning Algorithm for Ranking (SLOLAR) is proposed for leveraging the two activities further. The algorithm has two-fold benefits. They are scalability in processing web query with highly relevant documents and providing a better ranking model that is dynamic and incremental as new training set arrives. The author has focused on one of the industry best practices such as “do not reinvent the wheel” by updating ranking model without rebuilding it whenever a new set of documents arrive for training. SLOLAR employs supervised learning for ranking and adds value to IR systems by making them scalable, effective in dynamic navigation and produce results whose ranking mimic’s user intent in giving web query. Thus the contributions of the author are significant and can be incorporated in LTR mechanisms being used by search engines and other enterprise applications involved in IR mechanisms.