ABSTRACT

The goal of the thesis is to understand the contributions of concrete areas such as framework efficiency variability factors and query variability factors to overall retrieval variability. This brought together four different Information Retrieval (IR) models and set them to improve effectiveness performance. Then for each designated model, a detailed analysis of each framework with its retrieved documents was done and observations have yielded hypotheses for individual model and are tested in experiments.

The major work of the thesis is organized into 2 Chapters. Chapter 1 starts with model formulations of efficient IR techniques spanning sections 1-4. Section 1 presents a novel method called DETS, a hybrid Differential Evolution Algorithm (DE) with TABU Search (TS) which yields accurate results for document retrieval. The performance of DETS is compared with the TS and K-means with TS. Section 2 implements TABU Annealing, a heuristic approach which is a combination of TABU and simulated annealing with clustering approach. The results of the proposed approach are superior to simulated annealing and TABU. Section 3 aims at text mining framework consist of four distinct stages: 1.Text preprocessing 2.Dimensionality Reduction using Latent Semantic Indexing 3.Clustering based on Hybrid combination of Particle Swarm Optimization and K-means Algorithm
4. Information Retrieval Process using Simulated Annealing. This framework provides more relevant documents to the user and reduces the irrelevant documents. Section 4 proposes a novel clustering method based on Harmony Search (HS) optimization, a Meta heuristic algorithm that mimics the improvisation of musicians. By modeling Retrieval as an optimization problem, we hybridize Harmony clustering approach with TABU to achieve better retrieval. Experimental results reveal that the proposed algorithm can find better results when compared to HS and TABU methods. Finally, Feedback Language models are used to further improve the retrieved results.

As a part of the investigation, the following observations had been made: 1) Ranking search results are essential for an IR and Web search. 2) Websites recommended by the search engines have no guarantee for information correctness. 3) Faceted interfaces represent a new powerful paradigm which has been proven to be a successful complement to keyword searching. 4) Newly-emerging domains like comment e-rulemaking, in which many of the comments are “form letters” and modified copies of form letters (Near Duplicates). Spotting exact, near duplicates and their unique component(s) is a more challenging task. 5) It is necessary to investigate a problem of information seekers satisfaction in Question answering communities.

Chapter 2, the next part of the thesis spanning sections 1-5 covers more advanced IR concepts from the above observations. Section 1 focuses
on document ranking used by IR frameworks to rank matching documents according to their relevance to a given search query. We describe a 0/1 knapsack procedure for automatically selecting features to use within Generalization model for Ranking. We propose an approach for Relevance Feedback using Expectation Maximization method and experimental results show that our feature selection algorithm produces models that are either significantly more or equally effective as, models such as Markov random field model, Correlation Co-efficient and Count Difference method.

Section 2 introduces a Web recommender system called Truth Discovery based Content and Collaborative RECommender that helps users make decisions in the complex information space and to provide the user with further reading material by combining content and collaborative filtering. Along with this Truth finder is incorporated to present trustworthy websites to the users which outperforms the existing methods like Learning automata and Weighted Association Rule. Results suggest that the proposed system is better in predicting the next request of a web user. Section 3 presents a domain independent automatic Facet Generation Framework to extract the facets for efficient document retrieval. We also discover an efficiency improving semantically related feature sets with WordNet that reflects the contents of the target information collection. Experiments show that our approach can effectively generate multi-faceted arbitrary topics; and are comparable with traditional approaches like Baseline, Greedy and Feedback Language models. Section 4 deal with a framework for duplicate document detection problem
that uses a dynamic program called All Pairs Shortest Path in the text
collection to minimize the impact of duplicates on search results. We
compared our solution to the state of the art and found that our method has
produced promising results in improved accuracy of exact duplicate detection,
and also detected partial and neighbor replica.

Finally, Section 5 investigates the problem of predicting
information seeker satisfaction in yahoo question answering communities,
whether an asker is satisfied with the answers submitted by the community
participants. We explore automatic ranking, creating an abstract from
retrieved answers, and history updation for user satisfaction. Experimental
results, obtained from an evaluation over thousands of real questions and user
ratings, demonstrate the feasibility of modeling and predicting asker
satisfaction.