ABSTRACT

Creating text summaries from large volume of unstructured text like customer reviews, web log posts and social media posts is an important task in text mining applications. These summaries reveal the useful information which portrays the entire document or reviews. Summarization task could be performed using two major approaches: Extractive and abstractive approach. Extractive approach identifies the significant portion of the document that exposes the entire content and extracts it to form summary. Abstractive approach creates summary based on keywords and semantics from the document, which makes it difficult when compared to the other approach. The extractive summarization task is complex due to redundancy, large volume of text, variability and semantics of natural language in the text. The wide applicability as well as challenging nature of the task has inspired active research in the domain by both academic and industry experts. This research focuses on the design of machine learning based systems for two core areas related to text mining, namely, Feature based text summarization and text similarity detection.

Existing feature ranking and summarization systems employ a variety of methods including latent semantic indexing; Naïve Bayes' and other semantics based approaches. Due to the complexity of the task there is a need for developing efficient systems. In this thesis, a feature ranking system based on customer preferences have been developed. Three different machine learning approaches have been adopted for feature based micro level extractive text summary formation. In the feature ranking system, features are extracted using standard dependency parsing algorithm. Then these features are annotated in the domain ontology tree. A novel aspect scoring including significant score and author preference score along with ontology annotation has been proposed. This is efficient in ranking the features when compared with statistical and semantic approaches

because author preferences are included. Pair wise ranking algorithm is developed to compute scores for these features including customer preferences and ranked accordingly. The system is evaluated using customer reviews from two domains: Movie and hotel domain measured using true ranking, mean squared error and mean absolute error metrics.

The top features from these systems are used for summary generation from customer reviews. Distributed clustering algorithm is constructed with underlying the principles of K-means algorithm. Map reduce framework had been used for deployment and this makes it efficient for working with large volume of customer reviews. Outlier detection in the clustering algorithm in mapper and reducer improves the quality of the summary and also reduces noisy sentences from the reviews. Proposed Combiner optimization based on silhouette coefficient is efficient for removal of redundant sentences. This improves the quality of the summary by removing redundant sentences. Next particle swarm optimization with two different text modelling is used. An efficient diversity improved PSO with discrete and continuous versions has been recommended for optimizing the swarms. A new multi objective fitness function has been modeled for text summarization. This helps in enhancing the summary quality by minimizing similarity and maximizing the relevance score. The system showed an improved performance when continuous modelling is used. The third method involves in-node optimizations using Map Reduce. An optimized mapper for sentence extraction using hash map and partitioner module with different feature terms are used for extracting summarized sentences. Feature based partitions are generated using reducer with summary sentences. This displays an enhanced efficiency in removing noisy sentences thereby improving the quality of the summary.

The final impact in the thesis is on text similarity detection. This involves developing a similarity detection system using graph data bases. This includes verbal intents in identifying similarities using grammatical

linkages. Grammatical linkages used aids to identify similarity based on context, thus enhancing the performance. This helps to find similarities between two documents with improved accuracy. The system exhibits an upgraded performance in identifying text similarities between two text documents using similarity metrics like jaccard coefficient, Dice coefficient and cosine similarity index.

The efficiency of text summarization can be improved using machine learning techniques with optimization. Text similarity detection can be enhanced by using semantic modelling with graph databases.