Abstract

In Today's world, a large percentage of internet users rely on search engines to find relevant information rapidly. Most internet search engines use classical information retrieval techniques for retrieving documents in response to a natural language query. People commonly use popular search engines such as Google, Yahoo, DuckDuckGo, Baidu, etc. Before the advent of WWW(World Wide Web), retrieval systems were used to manage information in libraries, government organizations, and archiving. In these cases, the professional indexers acted as an intermediary for the common users and they could identify the customer needs through an interactive dialogue with the customers because they knew the collection, and document representation, and they were experts in formulating Boolean search operators to retrieve the documents. However, information retrieval systems used by the various modern search engines are designed for naive users unfamiliar with the collection, document representation, and formulation of queries. Most existing IR systems were designed for retrieving English documents in response to an English query.

Nowadays non-English content is rapidly increasing on the Internet. Therefore, the major search engines are trying to add support for the Indian languages including Bengali. Although most search engines use basic IR techniques, the application of IR techniques designed to focus on a certain target language can help in improving retrieval performance. Compared to English, the Bengali language has a more complex morphology. Bengali is a language in which verbs and nouns have many inflected forms. In the Bengali language, proper nouns can appear in various forms, for example, Ratan-i, Ratan-o, and Ratan-er. Proper nouns can also have more complex morphology where multiple atomic suffixes are added to proper nouns, Ratan-der-ke-o. The singular and plural forms of nouns can have inflection like aam-guli, jamider-er. Compared to nouns, verbs are more inflected in Bengali. Compound words are very frequently used in the Bengali language. Compound words containing verbs, adjectives, and nouns can have many morphological variations. Besides this morphological variation, like in English, new words can be formed by derived morphology. Due to the rich morphology of the Bengali language, the term mismatching problem becomes more critical while using Bengali queries for searching.

Although research on Bengali information retrieval had started in the long past, the research progress had been a little bit slow. Most earlier studies on Bengali IR focused on morphological analysis of the terms and finding roots or stems of the words using a lemmatizer or stemmer for effective indexing of the Bengali texts and integrating it with the traditional IR models such as the TF-IDF model, Okapi BM25 model. In recent times, an effective Deep Learning technique known as transfer learning has transformed and revolutionized the field of natural language processing. Several transfer learning techniques and architectures have been developed to advance the state-of-the-art on a variety of NLP tasks. The pre-trained word embeddings have been more popular due to the widespread availability and simplicity of integration with a new task. The use of pre-trained word embeddings in word and document representation can be effective for dealing with the word mismatch problem which is a critical problem for Bengali IR caused by morphological variants. To date, much research work has not been done for Bengali information retrieval to date and research on Indian languages is not at par with the English language also. Motivated by this, in this thesis, we investigate novel algorithms for Bengali IR. We have proposed several techniques for improving the Bengali IR model in various ways (1) Enhancing the smoothing process for a document language model, (2) semantic matching-based retrieval using large language model(LLM)-based word or document representation, (3) query expansion using pre-trained word embedding, pseudo-relevance feedback, or user feedback or hybrid methods, and (4) combining multiple IR models. We have evaluated the proposed approaches using benchmark IR datasets for Bengali. The experimental results and analyses establish that our proposed methods are effective for Bengali information retrieval.

Since an IR model returns a large number of documents, to aid the users in spotting the relevant information quickly, we propose a method for clustering the search results and producing summaries of the clusters so that the user can quickly identify the cluster containing the most relevant document by reading the summaries. We have proposed a new method for multi-document summarization which is used for creating a summary for each cluster. For clustering search results, we have used a histogram-based clustering method enhanced using word embedding. The multi-document summarization method has also been tested on a benchmark summarization dataset for English and a Bengali dataset developed by us. The experimental results and analyses establish that the proposed multi-document summarization method is effective for creating cluster summaries. It is also established that clustering and summarization of search results are useful for quickening to find the relevant information.

Keywords: Bengali Language, Information Retrieval, Smoothing, Document Language Model, Semantic Matching-based Retrieval, Word Embedding, Query Expansion, Pseudo-relevance Feedback, Hybrid Query Expansion, Model Combination, Clustering, Multi-document Summarization