Name of Scholar : **Jahiruddin**
Name of Supervisor : **Dr. Muhammad Abulaish** (Associate Prof., D/o Computer Sc., JMI, Delhi, India)
Name of Co-supervisor : **Dr. Lipika Dey** (Principal Scientist, Innovation Labs, TCS, Gurgaon, India)
Department : **Computer Science**
Title of Thesis : **Text Mining for Knowledge Discovery from Domain-Specific Text**

Abstract

Due to easy and cheap accessibility of Web and growing research activities in the field of biomedical science, the number of text documents disseminating biomedical knowledge has gone up manifolds and the challenge is increasingly becoming not the lack of information, but the excess of it. It is difficult to use and to integrate the knowledge embedded within biomedical texts with other biological data sources. Consequently, there is an increasing demand for automatic curation schemes to extract knowledge from scientific documents and store them in a structured form without which the assimilation of knowledge from this vast repository is becoming practically impossible. A number of techniques including information retrieval, information extraction, document classification, document clustering, etc. have been developed to ease extraction and understanding of information embedded within text documents. However, knowledge that is embedded in natural language texts is difficult to extract using simple pattern matching techniques and most of these methods do not help users directly understand key-concepts and their semantic relationships in document corpora, which are critical for capturing their conceptual structures. The problem arises due to the fact that most of the information is embedded within unstructured or semi-structured texts that computers can't interpret easily. Word-based searches for relevant information retrieve a huge number of documents and burden the users with pile of documents resulting in *"information overload"* problem.

Text mining has the potential to fuse knowledge embedded within the biomedical literature. The information fusion would enable biologists to exploit knowledge more effectively by processing complex texts and bio-inference sentences. Text mining systems enriched with knowledge

visualization techniques can be used to extract only relevant snippets of texts in response to user's query and present them in a comprehendible form. The visualization graph can assist users to navigate through the extracted knowledge at different levels of specificity without exploring the pile of text documents.

While various interesting applications are developed centered around biomedical text mining, some of the core research issues have been directed towards different aspects of knowledge extraction from biomedical literature. In this thesis, we have attempted to address a fundamental area that needs serious consideration to design and implement knowledge extraction and representation system to extract important information components consisting of biomedical concepts and their inter-relationships for conceptualization of underlying text corpora and contextual query processing over them. To tackle various aspects of these problems, we have proposed the design of a novel Biomedical Knowledge Extraction, Visualization and Query Answering (**BioKEViQA**) framework to identify key information components from biomedical text documents that are centered on key-concepts (a.k.a. keyphrases). **BioKEViQA** applies linguistic analysis and statistical techniques to identify key-concepts. The information component extraction principle is based on natural language processing techniques and semantic-based analysis which is further analyzed using statistical techniques and co-occurrence based analysis to identify feasible biomedical relations and their morphological variants. We have also presented a method for collating information extracted from multiple sources to generate semantic network which provides distinct user perspectives and allows navigation over documents with similar information components and is also used to provide a comprehensive view of the collection. The system stores the extracted information components in structured repositories that are integrated with a query-processing module to answer contextual biomedical queries over text documents. We have also proposed a document ranking mechanism to present retrieved documents along with information components in order of their relevance to user's query.