Abstract of the Ph.D. Work

Name of the Research Scholar : **Rafeeq Ahmed**

Supervisor : Prof. Tanvir Ahmad

Department : Department of Computer Engineering,
Faculty of Engineering and Technology,
Jamia Millia Islamia, New Delhi

Title : Text Mining using Big Data Analytics - A Fuzzy Approach

Increasing unstructured data on the internet by academics, digital documents, and social media leads to the process of mining and representation of the latent knowledge contained in it. In the Big Data age, a large quantity of data has been produced in various areas, from online social networking portals like Facebook, Twitter, etc., to news articles unstructured data, from health services to genomic functionalities. As an important field of research, text mining has many ground-level challenges because of no common platform, especially for free-flowing text or natural languages. Extracting hidden necessary information in the text is a major challenge. While handling an issue, we generally need to provide numerous different datasets.

The human brain can astutely confront any issue in any domain, in any event. Humans learn and disambiguate the essence of a passage by its context. Polysemic words, many morphologically similar terms, difficulty in human languages makes the task much difficult for the machines. Unfortunately, with an exponential increase in data, the process of information extraction becomes difficult. For text data, this information is represented in the form of context vectors. However, the generation of context vectors is limited by the memory heap and RAM of traditional systems. This

study aims to examine and propose a framework for computing context vectors of large dimensions over Big Data, trying to overcome traditional systems' bottleneck. The proposed framework is based on a set of mappers and reducers, implemented on Apache Hadoop. With an increase in the input dataset's size, the dimensions of the related concepts (in the form of a resultant matrix) increase beyond the capacity of a single system. In the form of unstructured data, the input falls under the "variety" part of Big Data.

This work focusses on one of the characteristics of Big Data i.e., "Variety" which includes all types of data including structured, semi-structured, and unstructured data. Unstructured data includes images, videos, text files, etc. The textual files are eBooks, ppt, pdf, news articles, blogs, emails, scholarly articles. Since Scholarly articles are a great source of knowledge, so learning from them like E-learning requires automatic approaches to build concept-maps, learning paths, etc. These sources are monotonically increasing and Big too. These sources have multi-domain, variety, huge volumes, which is, in fact, the characteristics of Big Data. The extraction of the concept is taken from key phrase extraction in information retrieval and the area of text mining. These forms for knowledge representation techniques like concept-map, ontology, etc.

Taxonomies or Concept hierarchies' plays an important role in any knowledge representation system like online learning or E-learning. The application of concept extraction is to create concept maps, a knowledge representation technique, for domains like e-Learning. Concepts are mined with their fuzzy context. The fuzzy concept means an important term in the document with its context vector. A concept map is a graphical representation of different concepts with nodes as concept and weights of edges as strength of their relationships. Taking out the concept, we have reflected the fuzzy distance between concepts. We have implemented a concept-map learning system to help users for better visualization and get a better learning path. Also, relationships between nodes or predicates, and the common concept or predicate among the concepts have been computed to help reduce the user's cognitive load. The concept-map generated is used for obtaining the Learning path

based on the Semantic relatedness of concepts, we can say nodes with maximum edges have been traversed and displayed.

Semantic Relatedness computation has been a fundamental and essential step for domains like Information Retrieval, Natural Language Processing, Semantic Web, etc. Many techniques for Semantic Relatedness calculations have been used but for a single domain. These techniques give inappropriate results for the multi-domain massive dataset because they give some relation between concepts across the different domains, although they are not related anyway, and their similarity should be minimum. A novel method, Modified Balanced Mutual Information (MBMI), has been proposed to handle data from multiple sources belonging to multiple domains and minimize semantic relatedness across multi-domain data. After the extraction of concepts, it is followed by a fuzzy vector from the given corpus to get semantic relatedness. A comparison of the proposed method with prominent techniques has been made. The dataset taken was for the e-learning domain, and it was unstructured. The dataset was research articles from medical and computer background articles. Scholarly articles from different domains were taken for our experimental work, and we tested on the BBC data set with 100 and 650 documents. Since the number of domains is known, and concepts domain is stored, we have done both clustering as well as the classification for testing our fuzzy-based semantic system. In classification, we have used logistics regression, Support Vector Machine (SVM) with Linear kernel, Polynomial kernel, Radial Basis Function (RBF) Kernel, Sigmoid kernel to obtain maximum accuracy up to 94% to 96% for all data sets. In clustering, using K-Means, we obtained precision up to 93%. This system can generate adaptive learning paths, concept map extraction, and Big Data-based E-Learning portals.