

Name of the Scholar : Iqbal Hasan
Name of the Supervisor : Syed Afzal Murtaza Rizvi
Name of the Department : Department of Computer Science
Topic of Research : A Data Analytics Framework for Knowledge Discovery from Heterogeneous Data Sources (DAFDKD)

Keywords: Data Analytics Framework, Knowledge Discovery, Heterogeneous Data, Machine learning, Data mining process

Research Findings and Key Contributions of the Thesis

The key contributions of thesis can be summarised as follows:

Proposing an AI techniques driven cognitive computing framework: Cognitive Computing Framework is a combination of technologies performing tasks that facilitate systems to emulate human intelligence. The AI techniques augments human thinking to solve the real-world problems focusing on to reflect the reality and improving the performance. In this work, we have presented how AI techniques can work with cognitive systems and help such systems to improve the quality and have human like thinking and intelligence. The modern decision support system can leverage simultaneously the benefits of next-gen technologies, AI techniques, and cognitive intelligence. In future work, we would present a system using the model based on this work.

Design of a knowledge-based data analytics approach for the next-gen e-governance:

.For effective service delivery, organized initiatives and knowledge management as a valuable resource are required. Our work discusses next-generation technologies and their effective application through knowledge management for the long-term sustainability and resilience of e-governance and the efficient delivery of public services. A strong next-gen technological infrastructure, combined with effective knowledge management and sharing, could play a critical role in next-generation e-governance.

Development of an AI-driven chatbot framework for citizen-centric delivery of e-governance services: In this work, we have discussed an AI-enabled chatbot developed by our group at National Informatics Centre to streamline the interaction between government and citizens and better citizen's experience in terms of quality of services delivered with improved operational efficiency and reduced time. The chatbot has been deployed in more

than fifty citizen centric IT solutions, We are already working on Next-gen conversational chatbot with better NLP and NLU capabilities.

Development of an approach for fraud detection in e-commerce transaction data: In this work we have discussed the mechanisms to detect and prevent online fraud using suitable machine learning and AI techniques and reduce the risks of the existing payment systems. Moreover, the work presents fraud detection in eCommerce transactions along with the pros and cons of various online payment systems and types of e-frauds. Finally, it discussed the security challenges associated with online payment systems and various controlling and mitigation strategies to reduce the risks of the existing payment systems.

Design of a multi-criteria decision-making technique for employee's performance evaluation (MCDM-EPE): This work introduces a novel EPE methodology by applying MCDM techniques AHP and TOPSIS with Machine learning. The AHP and TOPSIS method set the scores of the employees. The Random Forest, Naïve Baye, KNN, and Support Vector Machine algorithms used for building ML classification model was trained using 12 given features set. The Random Forest algorithm with gave 93% accuracy was most suitable, followed by NB with 77.77%, SVM with 77.69% and KNN with 75.91% respectively. The results gave 95% CI and provides an adequate level of AUROC and 10-fold cross-validation values. The proposed method gives significant improvement over existing models. This approach can be considered an efficient smart HR tool for predicting and classifying the performance of the candidates. The future work may include testing the model with bigger datasets to increase the accuracy of the model.

Development of a scalable framework to analyze data from heterogeneous sources at different levels of granularity: We have introduced DafkdSQL framework for achieving the Semantic Data Lake, for querying diverse and huge data sources using Semantic Web approaches. It lets users to describe changes that enable joinability on-the-fly at query time, and it conducts distributed cross-source join operations. DafkdSQL is designed with Spark and Elasticsearch, two cutting-edge Big Data technologies. It removes users from handcrafting wrappers by relying on the latter's connectors to wrap the data – a fundamental bottleneck in supporting data variation across the literature.

Design of a novel clustering algorithm based on data spread patterns using AR spread index: Through present research work, a novel clustering algorithm is proposed and to draw comparison between proposed and existing algorithm, the same datasets used in each algorithm testing. Performance of the existing algorithms is compared with proposed

algorithm which clearly performed better than all of clustering algorithms under study. As illustrated, these algorithms are implemented in MATLAB. The experimental results reveal the fact that the clustering algorithm proposed in this research, performs much better than the other algorithms tested using same datasets.

Development of a Covid-19 digital response framework: This work presents a digital response framework for surveillance and control of COVID-19 in India and various other technological initiatives to fight against this global crisis. It discusses the roles, essential supports, and technical solutions under various governing bodies at different levels of governance. Moreover, the work also presents the analysis of COVID-19 cases in various parts of India. Implemented many apps, digital interfaces, and dashboards to combat this pandemic and limit its spread.

A data analytics and knowledge management approach for Covid-19 surveillance and control:

The extensive data analytics and knowledge management with integrated control and command centre (iCCC) guided the authorities in making informed decisions for comprehensive management of the COVID-19 pandemic. The iCCC dynamic dashboard based on the Delhi Covid Data Management (DCDM) data repository and other time-series data sources is being used for current and predictive data analytics. Data related to hospitals, vaccination, oxygen, recoveries/deaths, rehabilitation, and other aspects of COVID management are accumulated, classified, and analyzed on a real-time basis. Based on the given time series datasets the trends have been plotted for daily active positive cases, the tests performed, the positive number of cases, the trend of oxygen received, demanded, and consumed, the trend of daily reported deaths, a trend of hospitalization against active cases. Moreover, the contribution implemented ARIMA model for forecasting the active number of cases, positive patients' rate, and death rates of all the COVID-19 virus infected patients. The Box-Ljung statistic was applied for finding the autocorrelation issues within the developed models. We have found that only Death rate cases have autocorrelation issue, however, Active Cases and Positive Patients rate does not have autocorrelation issue. The data was forecasted for the next 25 weeks' time period in order to provide better surveillance and control of COVID-19 patients in Delhi state. The adequate level of accuracy measure is provided by the respective ARIMA model while forecasting the active cases, positive patients rate, and death cases.