Int J Adv Eng Sci Appl Math (2017) 9(4):195
https://doi.org/10.1007/s12572-017-0201-x

PREFACE

Preface

This special issue focuses on some aspects of the rapidly developing field of Data Science. Data science encompasses the larger areas of data analytics, machine learning and managing big data. Data analytics has become essential to glean some understanding from large data sets and convert data into actionable intelligence. With the rapid growth in the volumes of data available to enterprises, Government and on the web, automated techniques for analyzing the data have become essential. Machine learning is the backbone for identifying essential characteristics within the data, which when learnt make meaningful inferences possible. Massive amount of content on the web is in the form of natural language (text); in addition to social media and blogs, almost every major newspaper, scientific journal and government publication is available as digital content. This textual information is a veritable goldmine for individuals, enterprises and governments. Consequently, accurate and efficient extraction and analysis of textual information is an area of considerable interest to researchers in academia and industry. Efficiently storing, querying, and processing big data using distributed platforms, novel multi-dimensional frameworks and parallel algorithms is required given the increasing volumes and variety of 'big data' becoming available.

The papers in this special issue highlight the breadth of disciplines in which data science techniques have had an impact. The first paper on causal network reconstruction addresses the hard problem of recovering a network of causal effects even when data is missing. This has wide applicability in time-series modeling with constraints on data gathering. The second paper is a more specific application study on prediction of Esophageal cancer using machine learning. Similar to the first paper, this work measures the effectiveness of different approaches when not all test data is available.

The next two papers deal with online social media. The first explores using micro blog data for disaster response. The authors show that annotating the posts with the category of information they talk about increases the effectiveness of extraction actionable intelligence. The second paper, and the last in this month's issue, addresses the very important question of data exposure in micro blogs. This is a question that repeatedly arises in various aspects of data science. This paper demonstrates that even after the data is deleted, it leaves numerous traces and a good fraction of the data can be recovered from the traces. They also propose mitigation methods for such information leakage.

The fifth work, that will appear in the March issue, takes up a completely different application area that of automatic answer grading. In particular this paper studies the effectiveness of different evaluation techniques in grading of short answers and proposes a framework that learns to automatically pick the technique that best matches a question type. Together these five papers represent a good sample of exciting and cutting edge research in data science.