The dawn of a new age: a new discipline digging deep into big data for big value

John Wang

Department of Information and Operations Management,
Montclair State University,
1 Normal Ave., Montclair, NJ 07043, USA
Email: j.john.wang@gmail.com

Ruiliang Yan

Department of Marketing & Business Analytics,
College of Business,
Texas A&M University – Commerce,
Commerce, TX 75428, USA
Email: ruiliangy@gmail.com

Biographical notes: John Wang is a Professor in the Department of Information and Operations Management at Montclair State University, USA. Having received a scholarship award, he came to the USA and completed his PhD in Operations Research from Temple University. He has published over 100 refereed papers and six books. He has also developed several computer software programs based on his research findings. He has served as a guest editor and referee for many highly prestigious journals. His long-term research goal is on the synergy of operations research, data mining, and cybernetics.

Ruiliang Yan is an Assistant Professor of Marketing at Texas A&M University – Commerce. His research interests include marketing strategy, theoretical and empirical modelling, retailing, and e-commerce. His works have appeared in *Journal of the Academy of Marketing Science, Journal of Business Research, Industrial Marketing Management, Omega*, and many other journals.

With the Age of Big Data upon us, should we be drowned in a struggling swirl in a flood tide of digital data? Or could we be lifted to a lofty level by a reservoir of meaningful flowing information? Big data spans five dimensions (volume, variety, velocity, volatility, and veracity) and is generally steered towards one critical destination – value. Big data has now become a critical part of the business world and daily life. Containing big information and big knowledge, big data does indeed have big value. *IJDS* confronts the challenges of extracting a fountain of knowledge from 'mountains' of big data.

The main objective of *International Journal of Data Science* (*IJDS*) is to employ an interdisciplinary approach and bridge the gap between different disciplines, including computer science, OR/MS, statistics, data mining, DSS, graphic design and human-computer interaction. The process of knowledge creation therefore can include multiple components and perspectives (Davenport and Harris, 2007; Lohr, 2012). By

adopting such a diverse set of tools/techniques while employing the synergies involved, companies and organisations can make faster (real-time), frequent and fact-based decisions.

There are six regular papers, plus another short one for the Practitioner Column, in this glorious inaugural issue that are written by a select number of our Associate Editors and other distinguished scholars, who are experts in the related areas. In the first insightful paper, Michael A. Walker, President of the Data Science Association, suggests data science is establishing basic foundations to become a profession. Like the professionalisation of law and medicine in the past hundred years, he asserts the data science field is at the very beginning of becoming a profession – with competency standards, a code of professional conduct, specialised graduate level curriculums, certification and licensure, and self-regulation.

Walker's hypothesis is the data science community can follow a roadmap for how data science can be professionalised by reviewing the history of the medical and legal professions. He provides an outline consisting of seven (7) process steps for the professionalisation of data science:

1 the full-time occupation of data science is identified

2 educational programs are established where specialised knowledge and skills are identified and incorporated into a data science curriculum

3 a professional association is established to help define standards of the data science profession

4 a professional code of ethics is created

5 certifications and licenses are developed to distinguish qualified from unqualified practitioners

6 a professional association defines entry requirements and disciplinary procedures

7 gaining the support of law for self-autonomy and self-regulation.

Walker persuasively posits that data science ought to become a profession for the same reasons medicine and law became professions: each requires highly specialised education, a code of conduct and self-regulation by knowledgeable professionals to assure competency and protect the public.

Walker provocatively offers evidence that scientific misconduct is common in both the hard and soft sciences. He cites Dr. John Ioannidis's 2005 paper "Why most published research findings are false" to argue a code of professional conduct is needed for data science to maintain credibility and protect the public. Walker reminds us that raw datasets are not objective – they contain human and structural biases in how they are selected, collected, filtered, structured and analysed presenting serious risks for decision makers – the consumers of data science. As a result, Walker convincingly argues it is crucial for data scientists to follow a code of conduct and prudent data science processes to prevent and mitigate damage, and ensure the public of competency to uphold the reputation and maintain credibility of data science.

Finally, Walker calls on all members of the data science community to collaborate in developing and professionalising data science and suggests data science can become a profession like medicine and law in about 10 years.

The second paper focuses on Computerised Maintenance Management Systems (CMMS). Business information systems have been used more and more in recent decades, and in particular, Enterprise Resource Planning (ERP). This phenomenon is associated with a significant increase in contributions in the literature on the choice of programs of this type, but also in showing problems in their introduction which can be encountered, examples of successful installations, etc.

The field of Maintenance also has specific information systems called Computerised Maintenance Management Systems (CMMS) whose use makes all the data on machines, facilities and resources available online, which makes effective decision making on maintenance easier. Despite the importance that this decision making has in the areas of Production, Quality and Safety, the scientific contributions dealing with selection, installation and control of CMMS's may be considered almost non-existent; this is the case, even though some authors state that up to 90% of CMMS installations can be considered a failure or do not achieve the established goals.

In this study, then, María Carmen Carnero contributes to an area where research is seriously deficient; the real innovation in this research, however, is the development of a model to control the introduction of CMMS's, a topic in which there is no precedent for describing models or methodologies. Although there are a few prior contributions to do with CMMS selection, they do not analyse the key matter of how to control the introduction of a CMMS. With the model described here, maintenance managers can, for the first time, audit the progress of CMMS installation regularly, detect anomalies and take corrective action to achieve a successful introduction, as part of a process of continuous improvement. The usefulness in practice of this research should thus be underlined, as is shown in by the case study in which the model accompanying the research is applied.

She has used fuzzy multi-criteria techniques to construct the model, a vision which breaks with the tradition in maintenance of taking decisions based solely on experience, without the backing of models of techniques. Therefore, this research provides a highly innovative model in a field which urgently requires this type of contribution.

Drawing on a series of case studies, Franco Caron, Politecnico di Milano explore the improvement of the project planning and control process by using all the data potentially available to the project team. At each Time-Now (TN) throughout the project life cycle, a part of the work is completed (WC) and a part of the work is the Work Remaining (WR) that is still to be done. Based on Earned Value Management System (EVMS), the two components of the estimate at completion (EAC), i.e., the overall final cost of the project, are given by Actual Cost (AC) of the WC plus the Estimate To Complete (ETC) of the WR. Similar considerations may be applied to the estimate of Time at Completion (TAC). In the project control process the role of ETC is critical, since the information drawn from the ETC, may highlight the need for and the type of corrective actions changing the project plan. This approach corresponds to a *feed-forward* type control loop, since analysis of the future informs present-day decisions.

An effective process of forecasting/planning depends on utilising all the available knowledge – data records and experts' judgement – in particular when facing a high level of project uncertainty and complexity. Note that lack of information may be a very practical issue (information missing, documentation not completed, documentation unclear, documentation delayed, reviews not performed, contractual provisions unclear, plans unclear or missing, governance framework unclear or missing, etc.).

The Bayes Theorem represents a rigorous and formal approach allowing for an update of a prior distribution, which expresses the experts' preliminary opinion, by means of the data records gathered in the field. For instance, the project team may assume a prior estimate of the final budget overrun, based on subjective expectations about the development of the current project, and this prior estimate may be updated based on the actual performance of the current project at Time Now.

In particular, the contribution given by tacit knowledge, i.e., by the project stakeholders, about the future development of the project, may regard:

- the impact deriving from drivers which explain the project development during the WC, and also presumably affecting the WR
- possible behaviour of the stakeholders involved in the project
- certain/uncertain events or conditions affecting project performance during the WR which may originate both internally and externally to the project
- weak signals, i.e., risk triggers, indicating emerging situations which could possibly affect project performance.

Since stakeholders are the main sources of knowledge about the project, their early engagement may increase significantly the effectiveness of the planning process. As a matter of fact, forecasting becomes a participatory forecasting, based on interactive and participative methods involving a wide variety of stakeholders, increasing legitimacy of forecasting results and a shared sense of commitment. The goal of the stakeholder management process is to foster a knowledge sharing progress among the project stakeholders.

The fourth paper by Tomoko Saiki delves into patent data searches. It presents the paths for patent information management in patent searches by intellectual property (IP) management departments of firms. The importance of IP for firms has been growing and firms have to manage their IP appropriately in order to maximise their market value. As information on IP, patents contain a large amount of information relating to a technology. According to the increasing number of patent applications, the amount of patent data in the world has become immense. It is becoming challenging to identify the relevant information in patents. Accurate, relevant and timely information on patent is crucial for effective decision making of business and patent management. Generally, large-scaled firms have their IP management departments. IP management departments of firms are required to identify and analyse information effectively to support proper decision making by conducting patent data searches.

This paper discusses what the information quality of patent data in the form of search results means from the practical viewpoint of the IP management departments of firms. The different types of patent searches are performed during the business process and these patent searches have varied purposes. The IP management departments have to provide the highest quality search results in accordance with these purposes, while being as inexpensive and time efficient as possible. The IP management departments of firms are the user of patent databases and the results of patent data searches. This paper provides a discussion to better understand the perspective of patent data search policy in IP management departments of firms.

In this paper, the dimensions of information quality of patent data in patent searches are grouped and practical assessment methods for each dimension are presented in brief.

Additionally, the framework to determine and assess information quality of search results was developed and proposed from the practical viewpoint of IP management department. The dimensions presented in this paper would help the IP management department to standardise the criteria of patent search results and it would be useful for the IP management department to provide objectively evaluated search results in accordance with the purposes of patent searches. The paper presents a conceptual framework of patent information management in patent searches that is expected to promote patent search abilities within the IP management department.

The regretted professor George Isac discovered and introduced, especially for the study of Pareto type efficiency in locally convex spaces, the notion of 'nuclear cone' in 1981, published it in 1983 and called later on 'supernormal cone', since it appears stronger than the usual concept of 'normal cone' considered the most appropriate for the investigations of the ordered topological vector spaces. For the first time, Vasile Postolică named these convex cones in separated locally convex spaces as 'Isac's Cones' in 2009, after the previous, long ago, acceptance on professor Isac's part. This research work is devoted to the study of Isac's cones in comparison with varied topologies on the same linear space, in order to continue the investigations given in Encyclopaedia of Business Analytics and Optimisation (Wang, 2014). Thus, it is well known that the concept of normal cone is essential in the theory of ordered topological linear spaces and its applications. So, much more its generalisation represented by Isac's (nuclear or supernormal) cones in Hausdorff locally convex spaces, for the efficiency. In this context, Vasile Postolică presents the noticeable importance of Isac's cones in the infinite dimensional ordered vector spaces concerning the general efficiency, following his recent results in this field and significant scientific links. Until now, the concept of Isac's cone is the best for the study of the existence of the efficient points in separated locally convex spaces and fundamental in the theory of multi-objective optimisation, but not only. Thus, it has be considered to several kinds of problems in: optimisation, potential theory and applications, the best approximation theory, the fixed point theory, the study of vector optimisation for multi-valued functions, conically bounded sets, Grothendieck's nuclearity of topological vector spaces, absolute summability, C^*-algebras, geometrical aspects of Ekeland's principle, multiple scalarisations and so on. The contribution of Vasile Postolică is absolutely new concerning the behaviour of Isac's cones under dualities, with strong applications and implications at least in the following pertinent themes of this first edition of the *IJDS*: Mathematical Optimisation and Mathematics of Decision Science, Optimisation, Performance Measurement, Management Science, Social Sciences and Statistics.

In the sixth paper, Bharti Joshi and Rustom D. Morena proposed a novel approach for queries on nested objects. Object oriented programming has achieved a significant recognition in the design and functioning of complex applications, which need more compound structure for objects, new data types for storing images and/or large textual items. Taking advantage of nested objects, object oriented database can support these complex applications. However, optimisation of cost of execution of query on nested object has been a challenge in traditional methods. The new approach could reduce the cost of execution of nested objects queries. A conversion formula has been derived via observing specific pattern of object storage for execution of query on backward access. Their approach also opens up new vistas for its effective use in new technologies such as Service Oriented Systems and can further help in building composite algorithms for best backward path from basic service to composite service.

In addition to publishing regular research articles, *IJDS* has also designated 'Practitioner Column', a special territory for professional data science practitioners. *IJDS* welcomes innovative ideas, research notes, practice comments, reviews or surveys, technical or management reports, case studies, commentaries, and big news.

IJDS provides a professional forum for examining the processes and results associated with obtaining data, as well as munging, scrubbing, exploring, modelling, interpreting, communicating and visualising data. As a new discipline, Data Science takes data everywhere, including cyberspace, as a research object. The goal is an integrated and interconnected process designed to form a common ground from which a knowledge-based system can be built, shared and supported by professionals from different disciplines.

Hopefully, IJDS, along with *Int. J. of Applied Management Science (IJAMS)*, *Int. J. of Data Mining, Modelling and Management (IJDMMM)*, *Int. J. of Information and Decision Sciences (IJIDS)*, *Int. J. of Data Analysis Techniques and Strategies (IJDATS)* will be able to ameliorate a manager's burdens, meet a practitioner's challenges, explore an executive's opportunities, and realise an entrepreneur's dreams.

Together, let us celebrate the *birth* of *IJDS*, nurture its *growth*, contribute to its *strength* and protect its *health*.

References

Davenport, T.H. and Harris, J.G. (2007) *Competing on Analytics*, Harvard Business School Press, Boston.

Lohr, S. (2012) 'The age of big data', *New York Times*, SR. 1, February 12, Retrieved from http://search.proquest.com/docview/921038884?accountid=12536

Wang, J. (Ed.) (2014) *Encyclopedia of Business Analytics and Optimization*, 1st ed., IGI Global, Hershey, PA.