S&T In-Depth

The latest in science and technology research literature

FEATURE TOPIC: BIG DATA

TABLE OF CONTENTS

ANALYTICS - PREDICTIVE ANALYSIS

- Advanced predictive-analysis-based decision support for collaborative logistics networks (Hungary) 2015
- Big Data and Predictive analysis: What's New? (USA) 2014
- Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control (USA) 2016
- Big-Data or Slim-Data: Predictive analysis Will Rule with World (USA) 2016
- Data quality for data science, Predictive analysis, and big data in supply chain management: An introduction to the problem and suggestions for research and applications (USA) 2014
- The future is in the numbers: the power of predictive analysis in the biomedical educational environment (USA) 2016
- Predictive analysis, Data Mining and Big Data: Myths, Misconceptions and Methods (USA) 2015
- Special Issue on Big Data and Predictive analysis Application in Supply Chain Management

SUPPORT TECHNOLOGIES (Acquistion, Storage, Computation)

- Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks (Spain) 2014
- The design of data storage system based on Lustre for EAST (China) 2016
- A distributed frequent itemset mining algorithm using Spark for Big Data analytics (China) 2015
- The Gamma Matrix to Summarize Dense and Sparse Data Sets for Big Data Analytics (USA) 2016
- Handling big data: research challenges and future directions (Greece) 2016
- Knowledge discovery from social media using big data-provided sentiment analysis (SoMABiT) (Germany) 2015
- Optical storage arrays: a perspective for future big data storage (Australia) 2014
- Parallel Processing Systems for Big Data: A Survey (China) 2016
- The rise of "big data" on cloud computing: Review and open research issues (Malaysia) 2015
- Social big data: Recent achievements and new challenges (Spain) 2016
- A Software Framework for Efficient IoT Contexts Acquisition and Big Data Analytics (South Korea) 2014
- A View on Fuzzy Systems for Big Data: Progress and Opportunities (Spain) 2016

Review Articles

Big data, open science and the brain: lessons learned from genomics (Canada) 2014

Choudhury, S (Choudhury, Suparna); Fishman, JR (Fishman, Jennifer R.); McGowan, ML (McGowan, Michelle L.); Juengst, ET (Juengst, Eric T.)

ABSTRACT: The BRAIN Initiative aims to break new ground in the scale and speed of data collection in neuroscience, requiring tools to handle data in the magnitude of yottabytes (10(24)). The scale, investment and organization of it are being compared to the Human Genome Project (HGP), which has exemplified "big science" for biology. In line with the trend towards Big Data in genomic research, the promise of the BRAIN Initiative, as well as the European Human Brain Project, rests on the possibility to amass vast quantities of data to model the complex interactions between the brain and behavior and inform the diagnosis and prevention of neurological disorders and psychiatric disease. Advocates of this "data driven" paradigm in neuroscience argue that harnessing the large quantities of data generated across laboratories worldwide has numerous methodological, ethical and economic advantages, but it requires the neuroscience community to adopt a culture of data sharing and open access to benefit from them. In this article, we examine the rationale for data sharing among advocates and briefly exemplify these in terms of new "open neuroscience" projects. Then, drawing on the frequently invoked model of data sharing in genomics, we go on to demonstrate the complexities of data sharing, shedding light on the sociological and ethical challenges within the realms of institutions, researchers and participants, namely dilemmas around public/private interests in data, (lack of) motivation to share in the academic community, and potential loss of participant anonymity. Our paper serves to highlight some foreseeable tensions around data sharing relevant to the emergent "open neuroscience" movement.

Author(s) affiliation: [Choudhury, Suparna] McGill Univ, Div Social & Transcultural Psychiat, Montreal, PQ H3A 1A1, Canada. [Choudhury, Suparna] Jewish Gen Hosp, Lady Davis Inst, Montreal, PQ H3A 1A1, Canada.
[Fishman, Jennifer R.] McGill Univ, Social Studies Med Dept, Biomed Eth Unit, Montreal, PQ, Canada.
[McGowan, Michelle L.] Case Western Reserve Univ, Sch Med, Dept Bioeth, Cleveland, OH USA.
[Juengst, Eric T.] Univ N Carolina, Ctr Bioeth, Chapel Hill, NC USA.
Reprint Address: Choudhury, S (reprint author), McGill Univ, Div Social & Transcultural Psychiat, Montreal, PQ H3A 1A1, Canada.
E-mail Author(s) affiliation: suparna.choudhury@mcgill.ca

Times Cited: 19
Number of references: 79

Tags: Big data - Review article

Big Data: Survey, Technologies, Opportunities, and Challenges (Malaysia) 2014

Khan, N (Khan, Nawsher); Yaqoob, I (Yaqoob, Ibrar); Hashem, IAT (Hashem, Ibrahim Abaker Targio); Inayat, Z (Inayat, Zakira); Ali, WK (Ali, Waleed KamaleldinMahmoud); Alam, M (Alam, Muhammad); Shiraz, M (Shiraz, Muhammad); Gani, A (Gani, Abdullah)

ABSTRACT: Big Data has gained much attention from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. Currently, over 2 billion people worldwide are connected to the Internet, and over 5 billion individuals own mobile phones. By 2020, 50 billion devices are expected to be connected to the Internet. At this point, predicted data production will be 44 times greater than that in 2009. As information is transferred and shared at light speed on optic fiber and wireless networks, the volume of data and the speed of market growth increase. However, the fast growth rate of such large data generates numerous challenges, such as the rapid growth of data, transfer speed, diverse data, and security. Nonetheless, Big Data is still in its infancy stage, and the domain has not been reviewed in general. Hence, this study comprehensively surveys and classifies the various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined based on opportunities and several open issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal techniques to address Big Data.

Author(s) affiliation: [Khan, Nawsher; Yaqoob, Ibrar; Hashem, Ibrahim Abaker Targio; Inayat, Zakira; Ali, Waleed KamaleldinMahmoud; Shiraz, Muhammad; Gani, Abdullah] Univ Malaya, Fac Comp Sci & Informat Technol, Mobile Cloud Comp Res Lab, Kuala Lumpur 50603, Malaysia.
[Khan, Nawsher] Abdul Wali Khan Univ Mardan, Dept Comp Sci, Mardan 23200, Pakistan.
[Inayat, Zakira] Univ Engn & Technol Peshawar, Dept Comp Sci, Peshawar 2500, Pakistan.

[Alam, Muhammad] Saudi Elect Univ, Riyadh, Saudi Arabia.
[Alam, Muhammad] Univ Kuala Lumpur, Kuala Lumpur 50603, Malaysia.
Reprint Address: Khan, N (reprint author), Univ Malaya, Fac Comp Sci & Informat Technol, Mobile Cloud Comp Res Lab, Kuala Lumpur 50603, Malaysia.
E-mail Author(s) affiliation: nawsherkhan@gmail.com
Times Cited: 5
Cited Refernce Count: 107

Tags: Big data - Review article

Challenges of Big Data analysis (USA) 2014

Fan, JQ (Fan, Jianqing); Han, F (Han, Fang); Liu, H (Liu, Han)

ABSTRACT: Big Data bring new opportunities to modern society and challenges to data scientists. On the one hand, Big Data hold great promises for discovering subtle population patterns and heterogeneities that are not possible with small-scale data. On the other hand, the massive sample size and high dimensionality of Big Data introduce unique computational and statistical challenges, including scalability and storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity and measurement errors. These challenges are distinguished and require new computational and statistical paradigm. This paper gives overviews on the salient features of Big Data and how these features impact on paradigm change on statistical and computational methods as well as computing architectures. We also provide various new perspectives on the Big Data analysis and computation. In particular, we emphasize on the viability of the sparsest solution in high-confidence set and point out that exogenous assumptions in most statistical methods for Big Data cannot be validated due to incidental endogeneity. They can lead to wrong statistical inferences and consequently wrong scientific conclusions.

Author(s) affiliation: [Fan, Jianqing; Liu, Han] Princeton Univ, Dept Operat Res & Financial Engn, Princeton, NJ 08544 USA.
[Han, Fang] Johns Hopkins Univ, Dept Biostat, Baltimore, MD 21205 USA.
Reprint Address: Fan, JQ (reprint author), Princeton Univ, Dept Operat Res & Financial Engn, Princeton, NJ 08544 USA.
E-mail Author(s) affiliation: jqfan@princeton.edu
Times Cited: 44
Number of references: 121

Tags: Big data - Review article

Data discretization: taxonomy and big data challenge (Spain) 2016

Author(s): Ramirez-Gallego, S (Ramirez-Gallego, Sergio); Garcia, S (Garcia, Salvador); Mourino-Talin, H (Mourino-Talin, Hector); Martinez-Rego, D (Martinez-Rego, David); Bolon-Canedo, V (Bolon-Canedo, Veronica); Alonso-Betanzos, A (Alonso-Betanzos, Amparo); Benitez, JM (Manuel Benitez, Jose); Herrera, F (Herrera, Francisco)

ABSTRACT: Discretization of numerical data is one of the most influential data preprocessing tasks in knowledge discovery and data mining. The purpose of attribute discretization is to find concise data representations as categories which are adequate for the learning task retaining as much information in the original continuous attribute as possible. In this article, we present an updated overview of discretization techniques in conjunction with a complete taxonomy of the leading discretizers. Despite the great impact of discretization as data preprocessing technique, few elementary approaches have been developed in the literature for Big Data. The purpose of this article is twofold: a comprehensive taxonomy of discretization techniques to help the practitioners in the use of the algorithms is presented; the article aims is to demonstrate that standard discretization methods can be parallelized in Big Data platforms such as Apache Spark, boosting both performance and accuracy. We thus propose a distributed implementation of one of the most well-known discretizers based on Information Theory, obtaining better results than the one produced by: the entropy minimization discretizer proposed by Fayyad and Irani. Our scheme goes beyond a simple parallelization and it is intended to be the first to face the Big Data challenge. (C) 2015 John Wiley & Sons, Ltd

Author(s) affiliation: [Ramirez-Gallego, Sergio; Garcia, Salvador; Manuel Benitez, Jose; Herrera, Francisco] Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain.
[Mourino-Talin, Hector; Martinez-Rego, David; Bolon-Canedo, Veronica; Alonso-Betanzos, Amparo] Univ A Coruna, Dept Comp Sci, La Coruna, Spain.
[Martinez-Rego, David] UCL, Dept Comp Sci, London, England.
Reprint Address: Garcia, S (reprint author), Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain.
E-mail Author(s) affiliation: salvagl@decsai.ugr.es
Times Cited: 0
Number of references: 70

Tags: Big data - Review article

Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications (Spain) 2016

Author(s): Pastur-Romay, LA (Anton Pastur-Romay, Lucas); Cedron, F (Cedron, Francisco); Pazos, A (Pazos, Alejandro); Porto-Pazos, AB (Belen Porto-Pazos, Ana)

ABSTRACT: Over the past decade, Deep Artificial Neural Networks (DNNs) have become the state-of-the-art algorithms in Machine Learning (ML), speech recognition, computer vision, natural language processing and many other tasks. This was made possible by the advancement in Big Data, Deep Learning (DL) and drastically increased chip processing abilities, especially general-purpose graphical processing units (GPGPUs). All this has created a growing interest in making the most of the potential offered by DNNs in almost every field. An overview of the main architectures of DNNs, and their usefulness in Pharmacology and Bioinformatics are presented in this work. The featured applications are: drug design, virtual screening (VS), Quantitative Structure-Activity Relationship (QSAR) research, protein structure prediction and genomics (and other omics) data mining. The future need of neuromorphic hardware for DNNs is also discussed, and the two most advanced chips are reviewed: IBM TrueNorth and SpiNNaker. In addition, this review points out the importance of considering not only neurons, as DNNs and neuromorphic chips should also include glial cells, given the proven importance of astrocytes, a type of glial cell which contributes to information processing in the brain. The Deep Artificial Neuron-Astrocyte Networks (DANAN) could overcome the difficulties in architecture design, learning process and scalability of the current ML methods.

Author(s) affiliation: [Anton Pastur-Romay, Lucas; Cedron, Francisco; Pazos, Alejandro; Belen Porto-Pazos, Ana] Univ A Coruna, Dept Informat & Commun Technol, La Coruna 15071, Spain.
[Pazos, Alejandro; Belen Porto-Pazos, Ana] Complexo Hosp Univ A Coruna CHUAC, Inst Invest Biomed A Coruna INIBIC, La Coruna 15006, Spain.
Reprint Address: Porto-Pazos, AB (reprint author), Univ A Coruna, Dept Informat & Commun Technol, La Coruna 15071, Spain.
Porto-Pazos, AB (reprint author), Complexo Hosp Univ A Coruna CHUAC, Inst Invest Biomed A Coruna INIBIC, La Coruna 15006, Spain.
E-mail Author(s) affiliation: pastur90@gmail.com; flanciskinho@gmail.com; apazos@udc.es; ana.portop@udc.es

Times Cited: 0
Number of references: 140

Tags: Big data - Review article

Emerging trends and technologies in big data processing (Spain) 2015

Author(s): Casado, R (Casado, Ruben); Younas, M (Younas, Muhammad)

ABSTRACT: Big Data encompasses large volume of complex structured, semi-structured, and unstructured data, which is beyond the processing capabilities of conventional databases. The processing and analysis of Big Data now play a central role in decision making, forecasting, business analysis, product development, customer experience, and loyalty, to name but a few. In this paper, we examine the distinguishing characteristics of Big Data along the lines of the 3Vs: variety, volume, and velocity. Accordingly, the paper provides an insight into the main processing paradigms in relation to the 3Vs. It defines a lifecycle for Big Data processing and classifies various available tools and technologies in terms of the lifecycle phases of Big Data, which include data acquisition, data storage, data analysis, and data exploitation of the results. This paper is first of its kind that reviews and analyzes current trends and technologies in relation to the characteristics, evolution, and processing of Big Data. Copyright (c) 2014 John Wiley & Sons, Ltd.

Author(s) affiliation: [Casado, Ruben] Treelogic, Dept Res & Innovat, Asturias, Spain.
[Younas, Muhammad] Oxford Brookes Univ, Dept Comp & Commun Technol, Oxford OX3 0BP, England.
Reprint Address: Casado, R (reprint author), Treelogic, Dept Res & Innovat, Asturias, Spain.
E-mail Author(s) affiliation: ruben.casado@treelogic.com

Times Cited: 6
Number of references: 37

Tags: Big data - Review article

Emergent Technologies in Big Data Sensing: A Survey (USA) 2015

Author(s): Zhu, T (Zhu, Ting); Xiao, S (Xiao, Sheng); Zhang, QQ (Zhang, Qingquan); Gu, Y (Gu, Yu); Yi, P (Yi, Ping); Li, YH (Li, Yanhua)

ABSTRACT: When the number of data generating sensors increases and the amount of sensing data grows to a scale that traditional methods cannot handle, big data methods are needed for sensing applications. However, big data is a fuzzy data science concept and there is no existing research architecture for it nor a generic application structure in the field of sensing. In this survey, we explore many scattered results that have been achieved by combining big data techniques with sensing and present our vision of big data in sensing. Firstly, we outline the application categories to generally summarize existing research achievements. Then we discuss the techniques proposed in these studies to demonstrate challenges and opportunities in this field. Finally, we present research trends and list some directions of big data in future sensing. Overall, mobile sensing and its related studies are hot topics, but other large-scale sensing researches are flourishing too. Although there are no "big data" techniques acting as research platforms or infrastructures to support various applications, multiple data science technologies, such as data mining, crowd sensing, and cloud computing, serve as foundations and bases of big data in the world of sensing.

Author(s) affiliation: [Zhu, Ting; Zhang, Qingquan] Univ Maryland, Baltimore, MD 21250 USA.
[Xiao, Sheng] Hunan Univ, Changsha 410082, Hunan, Peoples R China.
[Gu, Yu] IBM Res, Austin, TX 78758 USA.
[Yi, Ping] Shanghai Jiao Tong Univ, Shanghai 200240, Peoples R China.
[Li, Yanhua] Univ Minnesota Twin Cities, Minneapolis, MN 55416 USA.
Reprint Address: Zhu, T (reprint author), Univ Maryland, Baltimore, MD 21250 USA.
E-mail Author(s) affiliation: zt@umbc.edu

Times Cited: 0
Number of references: 69
Tags: Big data - Review article

A general perspective of Big Data: applications, tools, challenges and trends (Mexico) 2016

Author(s): Rodriguez-Mazahua, L (Rodriguez-Mazahua, Lisbeth); Rodriguez-Enriquez, CA (Rodriguez-Enriquez, Cristian-Aaron); Sanchez-Cervantes, JL (Luis Sanchez-Cervantes, Jose); Cervantes, J (Cervantes, Jair); Garcia-Alcaraz, JL (Luis Garcia-Alcaraz, Jorge); Alor-Hernandez, G (Alor-Hernandez, Giner)

ABSTRACT: Big Data has become a very popular term. It refers to the enormous amount of structured, semi-structured and unstructured data that are exponentially generated by high-performance applications in many domains: biochemistry, genetics, molecular biology, physics, astronomy, business, to mention a few. Since the literature of Big Data has increased significantly in recent years, it becomes necessary to develop an overview of the state-of-the-art in Big Data. This paper aims to provide a comprehensive review of Big Data literature of the last 4 years, to identify the main challenges, areas of application, tools and emergent trends of Big Data. To meet this objective, we have analyzed and classified 457 papers concerning Big Data. This review gives relevant information to practitioners and researchers about the main trends in research and application of Big Data in different technical domains, as well as a reference overview of Big Data tools.

Author(s) affiliation: [Rodriguez-Mazahua, Lisbeth; Rodriguez-Enriquez, Cristian-Aaron; Luis Sanchez-Cervantes, Jose; Alor-Hernandez, Giner] Inst Tecnol Orizaba, Div Res & Postgrad Studies, Ave Oriente 9 852 Col Emiliano Zapata, Orizaba 94320, Mexico.
[Cervantes, Jair] Univ Autonoma Estado Mexico, Ctr Univ UAEM Texcoco, Ave Jardin Zumpango S-N, Texcoco, Estado De Mexic, Mexico.
[Luis Garcia-Alcaraz, Jorge] Univ Autnoma Ciudad Juarez, Inst Ingn & Tecnol, Dept Ingn Ind & Manufactura, Ciudad Juarez, Mexico.
Reprint Address: Alor-Hernandez, G (reprint author), Inst Tecnol Orizaba, Div Res & Postgrad Studies, Ave Oriente 9 852 Col Emiliano Zapata, Orizaba 94320, Mexico.
E-mail Author(s) affiliation: lrodriguez@itorizaba.edu.mx; crodriguezen@gmail.com; isc.jolu@gmail.com; chazarra17@gmail.com; jorge.garcia@uacj.mx; galor@itorizaba.edu.mx

Times Cited: 0
Number of references: 168
Tags: Big data - Review article

Iterative big data clustering algorithms: a review (Malaysia) 2016

Author(s): Mohebi, A (Mohebi, Amin); Aghabozorgi, S (Aghabozorgi, Saeed); Teh Ying Wah (Teh Ying Wah); Herawan, T (Herawan, Tutut); Yahyapour, R (Yahyapour, Ramin)

ABSTRACT: Enterprises today are dealing with the massive size of data, which have been explosively increasing. The key requirements to address this challenge are to extract, analyze, and process data in a timely manner. Clustering is an essential data mining tool that plays an important role for analyzing big data. However, largescale data clustering has become a challenging task because of the large amount of information that emerges from technological progress in many areas, including finance and business informatics. Accordingly, researchers have dealt with parallel clustering algorithms using parallel programming models to address this issue. MapReduce is one of the most famous frameworks, and it has attracted great attention because of its flexibility, ease of programming, and fault tolerance. However, the framework has evident performance limitations, especially for iterative programs. This study will first review the proposed iterative frameworks that extended MapReduce to support iterative algorithms. We summarize these techniques, discuss their uniqueness and limitations, and explain how they address the challenging issues of iterative programs. We also perform an in-depth review to understand the problems and the solving techniques for parallel clustering algorithms. Hence, we believe that no well-rounded review provides a significant comparison among parallel clustering algorithms using MapReduce. This work aims to serve as a stepping stone for researchers who are studying big data clustering algorithms. Copyright (c) 2015 John Wiley & Sons, Ltd.

Author(s) affiliation: [Mohebi, Amin; Aghabozorgi, Saeed; Teh Ying Wah; Herawan, Tutut] Univ Malaya, Fac Comp Sci & Informat Technol, Kuala Lumpur, Malaysia.
[Yahyapour, Ramin] Gesell Wissensch Datenverarbeitung mbH Gottingen, Gottingen, Germany.
Reprint Address: Mohebi, A (reprint author), Univ Malaya, Fac Comp Sci & Informat Technol, Kuala Lumpur, Malaysia.
E-mail Author(s) affiliation: aminmohebbi@siswa.um.edu.my

Times Cited: 1
Number of references: 76
Tags: Big data - Review article

Rethinking big data: A review on the data quality and usage issues (China) 2016

Author(s): Liu, JZ (Liu, Jianzheng); Li, J (Li, Jie); Li, WF (Li, Weifeng); Wu, JZ (Wu, Jianzheng)

ABSTRACT: The recent explosive publications of big data studies have well documented the rise of big data and its ongoing prevalence. Different types of "big data" have emerged and have greatly enriched spatial information sciences and related fields in terms of breadth and granularity. Studies that were difficult to conduct in the past time due to data availability can now be carried out. However, big data brings lots of "big errors" in data quality and data usage, which cannot be used as a substitute for sound research design and solid theories. We indicated and summarized the problems faced by current big data studies with regard to data collection, processing and analysis: inauthentic data collection, information incompleteness and noise of big data, unrepresentativeness, consistency and reliability, and ethical issues. Cases of empirical studies are provided as evidences for each problem. We propose that big data research should closely follow good scientific practice to provide reliable and scientific "stories", as well as explore and develop techniques and methods to mitigate or rectify those 'big-errors' brought by big data. (C) 2015 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

Author(s) affiliation: [Liu, Jianzheng; Li, Jie; Li, Weifeng] Univ Hong Kong, Dept Urban Planning & Design, Fac Architecture, Knowles Bldg,Pokfulam Rd, Hong Kong, Hong Kong, Peoples R China.
[Wu, Jianzheng] Peking Univ, Shenzhen Grad Sch, Key Lab Human Environm Sci & Technol, Room E318, Shenzhen 518055, Peoples R China.
[Wu, Jianzheng] Peking Univ, Key Lab Earth Surface Proc, Coll Urban & Environm Sci, Beijing 100871, Peoples R China.
Reprint Address: Li, WF (reprint author), Univ Hong Kong, Dept Urban Planning & Design, Fac Architecture, Knowles Bldg,Pokfulam Rd, Hong Kong, Hong Kong, Peoples R China.
E-mail Author(s) affiliation: jzliu@hku.hk; Jessieleepku@hotmail.com; wfli@hku.hk; wujs@pkusz.edu.cn

Times Cited: 2
Number of references: 55
Tags: Big data - Review article

Analytics - Algorithms

Big Data Analytics for Emergency Communication Networks: A Survey (Japan) 2016

Author(s): Wang, JB (Wang, Junbo); Wu, YL (Wu, Yilang); Yen, N (Yen, Neil); Guo, S (Guo, Song); Cheng, ZX (Cheng, Zixue)

ABSTRACT: In this paper, we survey existing ECNs and big data analytics from both the content and the spatial points of view. From the content point of view, we survey existing data mining and analysis techniques, and further survey and analyze applications and the possibilities to enhance ECNs. From the spatial point of view, we survey and discuss the most popular methods and further discuss the possibility to enhance ECNs. Finally, we highlight the remaining challenging problems after a systematic survey and studies of the possibilities.

Author(s) affiliation: [Wang, Junbo; Wu, Yilang; Yen, Neil; Guo, Song; Cheng, Zixue] Univ Aizu, Grad Sch Comp Sci & Engn, Aizu Wakamatsu, Fukushima 9658580, Japan.
Reprint Address: Wang, JB (reprint author), Univ Aizu, Grad Sch Comp Sci & Engn, Aizu Wakamatsu, Fukushima 9658580, Japan.
E-mail Author(s) affiliation: j-wang@u-aizu.ac.jp

Times Cited: 0
Number of references: 154

Tags: Analytics - Algorithms, Big data

Big data analytics with swarm intelligence (China) 2016

Author(s): Cheng, S (Cheng, Shi); Zhang, QY (Zhang, Qingyu); Qin, QD (Qin, Quande)

ABSTRACT: Purpose - In this paper, the potential application of SI in big data analytics is analyzed. The correspondence and association between big data analytics and SI techniques are discussed. As an example of the application of the SI algorithms in the big data processing, a commodity routing system in a port in China is introduced. Another example is the economic load dispatch problem in the planning of a modern power system. Findings - The characteristics of big data include volume, variety, velocity, veracity, and value. In the SI algorithms, these features can be, respectively, represented as large scale, high dimensions, dynamical, noise/surrogates, and fitness/objective problems, which have been effectively solved. Research limitations/implications - In current research, the example problem of the port is formulated but not solved yet given the ongoing nature of the project. The example could be understood as advanced IT or data processing technology, however, its underlying mechanism could be the SI algorithms. This paper is the first step in the research to utilize the SI algorithm to a big data analytics problem. The future research will compare the performance of the method and fit it in a dynamic real system. Originality/value - Based on the combination of SI and data mining techniques, the authors can have a better understanding of the big data analytics problems, and design more effective algorithms to solve real-world big data analytical problems.

Author(s) affiliation: [Cheng, Shi] Univ Nottingham, Div Comp Sci, Ningbo, Zhejiang, Peoples R China.
[Zhang, Qingyu; Qin, Quande] Shenzhen Univ, Dept Management Sci, Shenzhen, Peoples R China.
[Zhang, Qingyu; Qin, Quande] Res Inst Business Analyt & Supply Chain Managemen, Shenzhen, Peoples R China.
[Qin, Quande] Beijing Inst Technol, Ctr Energy & Environm Policy Res, Beijing, Peoples R China.
Reprint Address: Qin, QD (reprint author), Shenzhen Univ, Dept Management Sci, Shenzhen, Peoples R China.
Qin, QD (reprint author), Res Inst Business Analyt & Supply Chain Managemen, Shenzhen, Peoples R China.
Qin, QD (reprint author), Beijing Inst Technol, Ctr Energy & Environm Policy Res, Beijing, Peoples R China.
E-mail Author(s) affiliation: q.yu.zhang@gmail.com

Times Cited: 0
Number of references: 71

Tags: Analytics - Algorithms, Big data

Big Data, Big Results: Knowledge Discovery in Output from Large-Scale Analytics (USA) 2014

Author(s): McCormick, TH (McCormick, Tyler H.); Ferrell, R (Ferrell, Rebecca); Karr, AF (Karr, Alan F.); Ryan, PB (Ryan, Patrick B.)

ABSTRACT: Observational healthcare data, such as electronic health records and administrative claims databases, provide longitudinal clinical information at the individual level. These data cover tens of millions of patients and present unprecedented opportunities to address such issues as post-market safety of medical products. Analyzing patient-level databases yields

population-level inferences, or 'results', such as the strength of association between medical product exposure and subsequent outcomes, often with thousands of drugs and outcomes. In this article, by contrast, we study 'big results', which are the product of applying thousands of alternative analysis strategies to five large patient databases. These results were produced by the Observational Medical Outcomes Partnership. All together, there are more than 6 million results, comprising risk assessments for 399 medical product-outcome pairs analyzed across five observational databases using seven statistical methods, each of which has between a few dozen and a few hundred variants representing parameters or 'tuning variables'. We focus on the value of knowledge discovery methods and the challenges in extracting clinically relevant knowledge from big results. We believe our analyses are both scientifically and methodologically valuable as they reveal information about how methods/algorithms perform under various circumstances, as well as provide a basis for comparison of these methods. (C) 2014 Wiley Periodicals, Inc.

Author(s) affiliation: [McCormick, Tyler H.; Ferrell, Rebecca] Univ Washington, Dept Stat, Seattle, WA 98195 USA.
[Karr, Alan F.] Natl Inst Stat Sci, Res Triangle Pk, NC 27709 USA.
[Ryan, Patrick B.] Janssen Res & Dev, Titusville, NJ 08560 USA.
[Ryan, Patrick B.] OMOP, Titusville, NJ 08560 USA.
Reprint Address: McCormick, TH (reprint author), Univ Washington, Dept Stat, Seattle, WA 98195 USA.
E-mail Author(s) affiliation: tylermc@uw.edu
Times Cited: 3
Number of references: 14
Tags: Analytics - Algorithms, Big data

Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives (China) 2014

Author(s): Zhou, ZH (Zhou, Zhi-Hua); Chawla, NV (Chawla, Nitesh V.); Jin, YC (Jin, Yaochu); Williams, GJ (Williams, Graham J.)

ABSTRACT: The authors bring together diverse perspectives, coming from different geographical locations with different core research expertise and different affiliations and work experiences. The aim of this paper is to evoke discussion rather than to provide a comprehensive survey of big data research.

Author(s) affiliation: [Zhou, Zhi-Hua] Nanjing Univ, Natl Key Lab Novel Software Technol, Nanjing 210023, Jiangsu, Peoples R China.
[Chawla, Nitesh V.] Univ Notre Dame, Dept Comp Sci & Engn, Notre Dame, IN 46556 USA.
[Chawla, Nitesh V.] Univ Notre Dame, Interdisciplinary Ctr Network Sci, Notre Dame, IN 46556 USA.
[Jin, Yaochu] Univ Surrey, Dept Comp, Guildford GU2 7XH, Surrey, England.
[Williams, Graham J.] Australian Taxat Off, Canberra, ACT, Australia.
Reprint Address: Zhou, ZH (reprint author), Nanjing Univ, Natl Key Lab Novel Software Technol, Nanjing 210023, Jiangsu, Peoples R China.
Times Cited: 23
Number of references: 62
Tags: Analytics - Algorithms, Big data

Critical analysis of Big Data challenges and analytical methods (England) 2017

Author(s): Sivarajah, U (Sivarajah, Uthayasankar); Kamal, MM (Kamal, Muhammad Mustafa); Irani, Z (Irani, Zahir); Weerakkody, V (Weerakkody, Vishanth)

ABSTRACT: Given the significant nature of the BD and BDA, this paper presents a state-of-the-art review that presents a holistic view of the BD challenges and BDA methods theorized/proposed/ employed by organizations to help others understand this landscape with the objective of making robust investment decisions. In doing so, systematically analysing and synthesizing the extant research published on BD and BDA area. More specifically, the authors seek to answer the following two principal questions: Q1 - What are the different types of BD challenges theorized/proposed/confronted by organizations? and Q2 - What are the different types of BDA methods theorized/proposed/employed to overcome BD challenges?. This systematic literature review (SLR) is carried out through observing and understanding the past trends and extant patterns/themes in the BDA research area, evaluating contributions, summarizing knowledge, thereby identifying limitations, implications and potential further research avenues to support the academic community in exploring research themes/patterns. Thus, to trace the implementation of BD strategies, a profiling method is employed to analyze articles (published in English-speaking peer-reviewed journals between 1996 and 2015) extracted from the Scopus database. The analysis presented in this paper has identified relevant BD research studies that have contributed both conceptually and empirically to the expansion and accrual of intellectual-wealth to the BDA in technology

and organizational resource management discipline. (C) 2016 The Authors. Published by Elsevier Inc.

Author(s) affiliation: [Sivarajah, Uthayasankar; Kamal, Muhammad Mustafa; Irani, Zahir; Weerakkody, Vishanth] Brunel Univ London, Brunel Business Sch, Uxbridge UB8 3PH, Middx, England.
Reprint Address: Sivarajah, U (reprint author), Brunel Univ London, Coll Business Arts & Social Sci, Brunel Business Sch, Uxbridge UB8 3PH, Middx, England.
E-mail Author(s) affiliation: Sankar.Sivarajah@brunel.ac.uk; Muhammad.Kamal@brunel.ac.uk; Zahir.lrani@brunel.ac.uk; Vishanth. Weerakkody@bmnel.ac.uk

Times Cited: 0
Number of references: 110
Tags: Analytics - Algorithms, Big data

Differential evolution framework for big data optimization (Australia) 2016

Author(s): Elsayed, S (Elsayed, Saber); Sarker, R (Sarker, Ruhul)

ABSTRACT: In this paper, a general differential evolution framework is proposed, in which the most suitable differential evolution algorithm for a problem on hand is adaptively configured. A local search is also employed to increase the exploitation capability of the proposed algorithm. The algorithm is tested on the 2015 big data optimization competition problems (six single objective problems and six multi-objective problems). The results show the superiority of the proposed algorithm to several state-of-the-art algorithms.

Author(s) affiliation: [Elsayed, Saber; Sarker, Ruhul] Univ New S Wales, Sch Engn & Informat Technol, Canberra, ACT, Australia.
[Elsayed, Saber] Zagazig Univ, Fac Comp & Informat, Zagazig, Egypt.
Reprint Address: Elsayed, S (reprint author), Univ New S Wales, Sch Engn & Informat Technol, Canberra, ACT, Australia.
Elsayed, S (reprint author), Zagazig Univ, Fac Comp & Informat, Zagazig, Egypt.
E-mail Author(s) affiliation: s.elsayed@adfa.edu.au; r.sarker@adfa.edu.au

Times Cited: 0
Number of references: 38
Tags: Analytics - Algorithms, Big data

A High-Order CFS Algorithm for Clustering Big Data (China) 2016

Author(s): Bu, FY (Bu, Fanyu); Chen, ZK (Chen, Zhikui); Li, P (Li, Peng); Tang, T (Tang, Tong); Zhang, Y (Zhang, Ying)

ABSTRACT: In this paper, we propose a high-order CFS algorithm (HOCFS) to cluster heterogeneous data by combining the CFS clustering algorithm and the dropout deep learning model, whose functionality rests on three pillars: (i) an adaptive dropout deep learning model to learn features from each type of data, (ii) a feature tensor model to capture the correlations of heterogeneous data, and (iii) a tensor distance-based high-order CFS algorithm to cluster heterogeneous data. Furthermore, we verify our proposed algorithm on different datasets, by comparison with other two clustering schemes, that is, HOPCM and CFS. Results confirm the effectiveness of the proposed algorithm in clustering heterogeneous data.

Author(s) affiliation: [Bu, Fanyu; Chen, Zhikui; Li, Peng] Dalian Univ Technol, Sch Software Technol, Dalian 116620, Peoples R China.
[Bu, Fanyu] Inner Mongolia Univ Finance & Econ, Sch Comp Informat Management, Hohhot 010070, Peoples R China.
[Tang, Tong] Southwest Univ, Dept Student Work, Chongqing 400715, Peoples R China.
[Zhang, Ying] Dalian Univ Finance & Econ, Coll Business Adm, Dalian 116622, Peoples R China.
Reprint Address: Bu, FY (reprint author), Dalian Univ Technol, Sch Software Technol, Dalian 116620, Peoples R China.
Bu, FY (reprint author), Inner Mongolia Univ Finance & Econ, Sch Comp Informat Management, Hohhot 010070, Peoples R China.
E-mail Author(s) affiliation: bufanyu@imufe.edu.cn

Times Cited: 0
Number of references: 9
Tags: Analytics - Algorithms, Big data

A Hybrid Approach to Clustering in Big Data (Australia) 2016

Author(s): Kumar, D (Kumar, Dheeraj); Bezdek, JC (Bezdek, James C.); Palaniswami, M (Palaniswami, Marimuthu); Rajasegarar, S (Rajasegarar, Sutharshan); Leckie, C (Leckie, Christopher); Havens, TC (Havens, Timothy Craig)

ABSTRACT: Clustering of big data has received much attention recently. In this paper, we present a new clusiVAT algorithm and compare it with four other popular data clustering algorithms. Three of the four comparison methods are based on the well known, classical batch k-means model. Specifically, we use k-means, single pass k-means, online k-means, and clustering using representatives (CURE) for numerical comparisons. clusiVAT is based on sampling the data, imaging the reordered distance

matrix to estimate the number of clusters in the data visually, clustering the samples using a relative of single linkage (SL), and then noniteratively extending the labels to the rest of the data-set using the nearest prototype rule. Previous work has established that clusiVAT produces true SL clusters in compact-separated data. We have performed experiments to show that k-means and its modified algorithms suffer from initialization issues that cause many failures. On the other hand, clusiVAT needs no initialization, and almost always finds partitions that accurately match ground truth labels in labeled data. CURE also finds SL type partitions but is much slower than the other four algorithms. In our experiments, clusiVAT proves to be the fastest and most accurate of the five algorithms; e.g., it recovers 97% of the ground truth labels in the real world KDD-99 cup data (4 292 637 samples in 41 dimensions) in 76 s.

Author(s) affiliation: [Kumar, Dheeraj; Palaniswami, Marimuthu; Rajasegarar, Sutharshan] Univ Melbourne, Dept Elect & Elect Engn, Melbourne, Vic 3010, Australia.
[Bezdek, James C.; Leckie, Christopher] Univ Melbourne, Dept Comp & Informat Syst, Melbourne, Vic 3010, Australia.
[Havens, Timothy Craig] Michigan Technol Univ, Dept Elect & Comp Engn, Houghton, MI 49931 USA.
Reprint Address: Kumar, D (reprint author), Univ Melbourne, Dept Elect & Elect Engn, Melbourne, Vic 3010, Australia.
E-mail Author(s) affiliation: dheerajk@student.unimelb.edu.au; jcbezdek@gmail.com; palani@unimelb.edu.au; sraja@unimelb.edu.au; caleckie@unimelb.edu.au; thavens@mtu.edu
Times Cited: 0
Number of references: 53
Tags: Analytics - Algorithms, Big data

Modeling and Optimization for Big Data Analytics (Greece) 2014

Author(s): Slavakis, K (Slavakis, Konstantinos); Giannakis, GB (Giannakis, Georgios B.); Mateos, G (Mateos, Gonzalo)

ABSTRACT: This article contributes to the ongoing cross-disciplinary efforts in data science by putting forth encompassing models capturing a wide range of SP-relevant data analytic tasks, such as principal component analysis (PCA), dictionary learning (DL), compressive sampling (CS), and subspace clustering. It offers scalable architectures and optimization algorithms for decentralized and online learning problems, while revealing fundamental insights into the various analytic and implementation tradeoffs involved. Extensions of the encompassing models to timely data-sketching, tensor- and kernel-based learning tasks are also provided. Finally, the close connections of the presented framework with several big data tasks, such as network visualization, decentralized and dynamic estimation, prediction, and imputation of network link load traffic, as well as imputation in tensor-based medical imaging are highlighted.

Author(s) affiliation: [Slavakis, Konstantinos] TokyoTech, Tokyo, Japan.
Author(s) affiliation: [Slavakis, Konstantinos] Univ Athens, Dept Informat & Telecommun, GR-10679 Athens, Greece.
[Slavakis, Konstantinos] Univ Peloponnese, Dept Telecommun & Informat, Tripolis, Greece.
[Giannakis, Georgios B.] Univ Minnesota, Minneapolis, MN 55455 USA.
[Slavakis, Konstantinos; Giannakis, Georgios B.] Univ Minnesota, Dept Elect & Comp Engn, Minneapolis, MN 55455 USA.
[Giannakis, Georgios B.] Univ Minnesota, Digital Technol Ctr, Minneapolis, MN 55455 USA.
[Mateos, Gonzalo] Univ Rochester, Dept Elect & Comp Engn, Rochester, NY 14627 USA.
Reprint Address: Slavakis, K (reprint author), Univ Minnesota, Dept Elect & Comp Engn, Minneapolis, MN 55455 USA.
E-mail Author(s) affiliation: kslavaki@umn.edu; georgios@umn.edu; mate0058@umn.edu
Times Cited: 27
Number of references: 76
Tags: Analytics - Algorithms, Big data

SIMD parallel MCMC sampling with applications for big-data Bayesian analytics (USA) 2015

Author(s): Mahani, AS (Mahani, Alireza S.); Sharabiani, MTA (Sharabiani, Mansour T. A.)

ABSTRACT: Two opportunities for Single-Instruction Multiple-Data (SIMD) parallelization of MCMC sampling for probabilistic graphical models are presented. In exchangeable models with many observations such as Bayesian Generalized Linear Models (GLMs), child-node contributions to the conditional posterior of each node can be calculated concurrently. In undirected graphs with discrete-value nodes, concurrent sampling of conditionally-independent nodes can be transformed into a SIMD form. High-performance libraries with multi-threading and vectorization capabilities can be readily applied to such SIMD opportunities to gain decent speedup, while a series of high-level source-code and runtime modifications provide further performance boost by reducing parallelization overhead and increasing data locality for Non-Uniform Memory Access architectures. For big-data Bayesian GLM graphs, the end-result is a routine for evaluating the conditional posterior and its gradient vector that is 5 times

faster than a naive implementation using (built-in) multi-threaded Intel MKL BLAS, and reaches within the striking distance of the memory-bandwidth-induced hardware limit. Using multi-threading for cache-friendly, fine-grained parallelization can outperform coarse-grained alternatives which are often less cache-friendly, a likely scenario in modern predictive analytics workflow such as Hierarchical Bayesian GLM, variable selection, and ensemble regression and classification. The proposed optimization strategies improve the scaling of performance with number of cores and width of vector units (applicable to many-core SIMD processors such as Intel Xeon Phi and Graphic Processing Units), resulting in cost-effectiveness, energy efficiency ('green computing'), and higher speed on multi-core x86 processors. (C) 2015 Elsevier B.V. All rights reserved.

Author(s) affiliation: [Mahani, Alireza S.] Sentrana Inc, Sci Comp Grp, Washington, DC 20006 USA.
[Sharabiani, Mansour T. A.] Univ London Imperial Coll Sci Technol & Med, Natl Heart & Lung Inst, London SW7 2AZ, England.
Reprint Address: Mahani, AS (reprint author), Sentrana Inc, 1725 I St NW Suite 900, Washington, DC 20006 USA.
E-mail Author(s) affiliation: alireza.s.mahani@gmail.com
Times Cited: 2
Number of references: 82
Tags: Analytics - Algorithms, Big data

Analytics - Deep Learning

Big Data Deep Learning: Challenges and Perspectives (USA) 2014

Author(s): Chen, XW (Chen, Xue-Wen); Lin, XT (Lin, Xiaotong)
Source: IEEE ACCESS Volume: 2 Pages: 514-525 DOI: 10.1109/ACCESS.2014.2325029 Published: 2014

ABSTRACT: Deep learning is currently an extremely active research area in machine learning and pattern recognition society. It has gained huge successes in a broad area of applications such as speech recognition, computer vision, and natural language processing. With the sheer size of data available today, big data brings big opportunities and transformative potential for various sectors; on the other hand, it also presents unprecedented challenges to harnessing data and information. As the data keeps getting bigger, deep learning is coming to play a key role in providing big data predictive analytics solutions. In this paper, we provide a brief overview of deep learning, and highlight current research efforts and the challenges to big data, as well as the future trends.

Author(s) affiliation: [Chen, Xue-Wen] Wayne State Univ, Dept Comp Sci, Detroit, MI 48404 USA.
[Lin, Xiaotong] Oakland Univ, Dept Comp Sci & Engn, Rochester, MI 48309 USA.
Reprint Address: Chen, XW (reprint author), Wayne State Univ, Dept Comp Sci, Detroit, MI 48404 USA.
E-mail Author(s) affiliation: xwen.chen@gmail.com
Times Cited: 27
Number of references: 104
Tags: Analytics - Predictive analysis, Big data

Cross-Domain Feature Learning in Multimedia (China) 2015

Author(s): Yang, XS (Yang, Xiaoshan); Zhang, TZ (Zhang, Tianzhu); Xu, CS (Xu, Changsheng)
Source: IEEE TRANSACTIONS ON MULTIMEDIA Volume: 17 Issue: 1 Pages: 64-78 DOI: 10.1109/TMM.2014.2375793 Published: JAN 2015

ABSTRACT: We propose a novel cross-domain feature learning (CDFL) algorithm based on stacked denoising auto-encoders. By introducing the modal correlation constraint and the cross-domain constraint in conventional auto-encoder, our CDFL can maximize the correlations among different modalities and extract domain invariant semantic features simultaneously. To evaluate our CDFL algorithm, we apply it to three important applications: sentiment classification, spam filtering, and event classification. Comprehensive evaluations demonstrate the encouraging performance of the proposed approach.

Author(s) affiliation: [Yang, Xiaoshan; Zhang, Tianzhu; Xu, Changsheng] Chinese Acad Sci, Inst Automat, Natl Lab Pattern Recognit, Beijing 100190, Peoples R China.
[Yang, Xiaoshan; Zhang, Tianzhu; Xu, Changsheng] China Singapore Inst Digital Media, Singapore 119613, Singapore.
Reprint Address: Yang, XS (reprint author), Chinese Acad Sci, Inst Automat, Natl Lab Pattern Recognit, Beijing 100190, Peoples R China.
E-mail Author(s) affiliation: xiaoshang.yang@nlpr.ia.ac.cn; tzzhang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn
Times Cited: 9
Number of references: 59
Tags: Analytics - Deep learning, Big data

Deep Computation Model for Unsupervised Feature Learning on Big Data (China) 2016

Author(s): Zhang, QC (Zhang, Qingchen); Yang, LT (Yang, Laurence T.); Chen, ZK (Chen, Zhikui)

ABSTRACT: This paper proposes a deep computation model for feature learning on big data, which uses a tensor to model the complex correlations of heterogeneous data. To fully learn the underlying data distribution, the proposed model uses the tensor distance as the average sum-of-squares error term of the reconstruction error in the output layer. To train the parameters of the proposed model, the paper designs a high-order back-propagation algorithm (HBP) by extending the conventional back-propagation algorithm from the vector space to the high-order tensor space. To evaluate the performance of the proposed model, we carried out the experiments on four representative datasets by comparison with stacking auto-encoders and multimodal deep learning models. Experimental results clearly demonstrate that the proposed model is efficient to perform feature learning when evaluated using the STL-10, CUAVE, SANE and INEX datasets.

Author(s) affiliation: [Zhang, Qingchen; Chen, Zhikui] Dalian Univ Technol, Sch Software Technol, Dalian 116620, Peoples R China. [Yang, Laurence T.] St Francis Xavier Univ, Dept Math Stat & Comp Sci, Antigonish, NS B2G 1C0, Canada.
Reprint Address: Zhang, QC; Chen, ZK (reprint author), Dalian Univ Technol, Sch Software Technol, Dalian 116620, Peoples R China. Yang, LT (reprint author), St Francis Xavier Univ, Dept Math Stat & Comp Sci, Antigonish, NS B2G 1C0, Canada.
E-mail Author(s) affiliation: qingchen@mail.dlut.edu.cn; ltyang@gmail.com; zkchen@mail.dlut.edu.cn

Times Cited: 1
Number of references: 33
Tags: Analytics - Deep learning, Big data

An Energy-Efficient and Scalable Deep Learning/Inference Processor With Tetra-Parallel MIMD Architecture for Big Data Applications (South Korea) 2015

Author(s): Park, SW (Park, Seong-Wook); Park, J (Park, Junyoung); Bong, K (Bong, Kyeongryeol); Shin, D (Shin, Dongjoo); Lee, J (Lee, Jinmook); Choi, S (Choi, Sungpill); Yoo, HJ (Yoo, Hoi-Jun)

ABSTRACT: This paper presents a SoC implementation to enable deep learning applications to run with low cost platforms such as mobile or portable devices. Different from conventional works which have adopted massively-parallel architecture, this work adopts task-flexible architecture and exploits multiple parallelism to cover complex functions of convolutional deep belief network which is one of popular deep learning/inference algorithms. In this paper, we implement the most energy-efficient deep learning and inference processor for wearable system. The implemented 2.5 mm x4.0 mm deep learning/inference processor is fabricated using 65 nm 8-metal CMOS technology for a battery-powered platform with real-time deep inference and deep learning operation. It consumes 185 mW average power, and 213.1 mW peak power at 200 MHz operating frequency and 1.2 V supply voltage. It achieves 411.3 GOPS peak performance and 1.93 TOPS/W energy efficiency, which is 2.07x higher than the state-of-the-art.

Author(s) affiliation: [Park, Seong-Wook; Park, Junyoung; Bong, Kyeongryeol; Shin, Dongjoo; Lee, Jinmook; Choi, Sungpill; Yoo, Hoi-Jun] Korea Adv Inst Sci & Technol, Taejon 305701, South Korea.
Reprint Address: Park, SW (reprint author), Korea Adv Inst Sci & Technol, Taejon 305701, South Korea.
E-mail Author(s) affiliation: seongwook.park@kaist.ac.kr

Times Cited: 0
Number of references: 25
Tags: Analytics - Deep learning, Big data

An Intelligent Fault Diagnosis Method Using Unsupervised Feature Learning Towards Mechanical Big Data (China) 2016

Author(s): Lei, YG (Lei, Yaguo); Jia, F (Jia, Feng); Lin, J (Lin, Jing); Xing, SB (Xing, Saibo); Ding, SX (Ding, Steven X.)

ABSTRACT: Inspired by the idea of unsupervised feature learning that uses artificial intelligence techniques to learn features from raw data, a two-stage learning method is proposed for intelligent diagnosis of machines. In the first learning stage of the method, sparse filtering, an unsupervised two-layer neural network, is used to directly learn features from mechanical vibration signals. In the second stage, softmax regression is employed to classify the health conditions based on the learned features. The proposed method is validated by a motor bearing dataset and a locomotive bearing dataset, respectively. The results show that the proposed method obtains fairly high diagnosis accuracies and is superior to the existing methods for the motor bearing dataset. Because of learning features adaptively, the proposed method reduces the need of human labor and makes intelligent fault diagnosis handle big data more easily.

Author(s) affiliation: [Lei, Yaguo; Jia, Feng; Lin, Jing; Xing, Saibo] Xi An Jiao Tong Univ, State Key Lab Mfg Syst Engn, Xian 710049, Peoples R China.
[Ding, Steven X.] Univ Duisburg Essen, Inst Automat Control & Complex Syst, D-47057 Duisburg, Germany.
Reprint Address: Lei, YG; Jia, F; Lin, J; Xing, SB (reprint author), Xi An Jiao Tong Univ, State Key Lab Mfg Syst Engn, Xian 710049, Peoples R China.
Ding, SX (reprint author), Univ Duisburg Essen, Inst Automat Control & Complex Syst, D-47057 Duisburg, Germany.
E-mail Author(s) affiliation: yaguolei@mail.xjtu.edu.cn; jiafeng1237@sina.com; jinglin@mail.xjtu.edu.cn; xingsaibo@stu.xjtu.edu.cn; steven.ding@uni-due.de
Times Cited: 2
Number of references: 51
Tags: Analytics - Deep learning, Big data

Learning methodologies for wireless big data networks: A Markovian game-theoretic perspective (China) 2016

Author(s): Yang, CG (Yang, Chungang)

ABSTRACT: Motivated by the idea of divide-and-conquer, in this article, we first present a multiple cognitive agent-based divide-and-conquer network management and control architecture. Furthermore, a Markovian game-theoretic modeling framework is proposed to model the state big data-based decision-making problem. Then, we investigate various learning methodologies with respect to different kinds of the state information, in particular, we concentrate on the construction of state space, the state transition computation, and the convergence of parallel Q-learning technique. This work provides a suitable network management architecture, an effective modeling tool, and various learning techniques for wireless big data networks. (C) 2015 Elsevier B.V. All rights reserved.

Author(s) affiliation: [Yang, Chungang] Xidian Univ, ISN, State Key Lab, Xian 710071, Shaanxi, Peoples R China.
[Yang, Chungang] Southeast Univ, Natl Mobile Commun Res Lab, Nanjing, Jiangsu, Peoples R China.
Reprint Address: Yang, CG (reprint author), Xidian Univ, PB Box 102,2 South Taibai Rd, Xian 710071, Shaanxi, Peoples R China.
E-mail Author(s) affiliation: cgyang@mail.xidian.edu.cn
Times Cited: 0
Number of references: 18
Tags: Analytics - Deep learning, Big data

Mobile Big Data Analytics Using Deep Learning and Apache Spark (Singapore) 2016

Author(s): Abu Alsheikh, M (Abu Alsheikh, Mohammad); Niyato, D (Niyato, Dusit); Lin, SW (Lin, Shaowei); Tan, HP (Tan, Hwee-Pink); Han, Z (Han, Zhu)

ABSTRACT: This article presents an overview and brief tutorial on deep learning in mobile big data analytics and discusses a scalable learning framework over Apache Spark. Specifically, distributed deep learning is executed as an iterative MapReduce computing on many Spark workers. Each Spark worker learns a partial deep model on a partition of the overall mobile, and a master deep model is then built by averaging the parameters of all partial models. This Spark-based framework speeds up the learning of deep models consisting of many hidden layers and millions of parameters. We use a context-aware activity recognition application with a real-world dataset containing millions of samples to validate our framework and assess its speedup effectiveness.

Author(s) affiliation: [Abu Alsheikh, Mohammad; Niyato, Dusit] Nanyang Technol Univ, Sch Comp Engn, Singapore, Singapore.
[Abu Alsheikh, Mohammad] Inst Infocomm Res, Singapore, Singapore.
[Lin, Shaowei] Singapore Univ Technol & Design, Singapore, Singapore.
[Tan, Hwee-Pink] Singapore Management Univ, Informat Syst Practice, Singapore 178902, Singapore.
[Han, Zhu] Univ Houston, Elect & Comp Engn Dept, Houston, TX 77004 USA.
[Han, Zhu] Univ Houston, Dept Comp Sci, Houston, TX 77004 USA.
Reprint Address: Abu Alsheikh, M (reprint author), Nanyang Technol Univ, Sch Comp Engn, Singapore, Singapore.
Abu Alsheikh, M (reprint author), Inst Infocomm Res, Singapore, Singapore.
E-mail Author(s) affiliation: stumyhaa@i2r.a-star.edu.sg; dniyato@ntu.edu.sg; shaowei_lin@sutd.edu.sg; hptan@smu.edu.sg; zhan2@uh.edu
Times Cited: 0
Number of references: 15
Tags: Analytics - Deep learning, Big data

A survey of machine learning for big data processing (China) 2016

Author(s): Qiu, JF (Qiu, Junfei); Wu, QH (Wu, Qihui); Ding, GR (Ding, Guoru); Xu, YH (Xu, Yuhua); Feng, S (Feng, Shuo)

ABSTRACT: In this paper, we present a literature survey of the latest advances in researches on machine learning for big data processing. First, we review the machine learning techniques and highlight some promising learning methods in recent studies, such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. Next, we focus on the analysis and discussions about the challenges and possible solutions of machine learning for big data. Following that, we investigate the close connections of machine learning with signal processing techniques for big data processing. Finally, we outline several open issues and research trends.

Author(s) affiliation: [Qiu, Junfei; Wu, Qihui; Ding, Guoru; Xu, Yuhua; Feng, Shuo] PLA Univ Sci & Technol, Coll Commun Engn, Nanjing 210007, Jiangsu, Peoples R China.
Reprint Address: Ding, GR (reprint author), PLA Univ Sci & Technol, Coll Commun Engn, Nanjing 210007, Jiangsu, Peoples R China.
E-mail Author(s) affiliation: dingguoru@gmail.com
Times Cited: 0
Number of references: 129
Tags: Analytics - Deep learning, Big data

Traffic Flow Prediction With Big Data: A Deep Learning Approach (China) 2015

Author(s): Lv, YS (Lv, Yisheng); Duan, YJ (Duan, Yanjie); Kang, WW (Kang, Wenwen); Li, ZX (Li, Zhengxi); Wang, FY (Wang, Fei-Yue)

ABSTRACT: In this paper, a novel deep-learning-based traffic flow prediction method is proposed, which considers the spatial and temporal correlations inherently. A stacked autoencoder model is used to learn generic traffic flow features, and it is trained in a greedy layerwise fashion. To the best of our knowledge, this is the first time that a deep architecture model is applied using autoencoders as building blocks to represent traffic flow features for prediction. Moreover, experiments demonstrate that the proposed method for traffic flow prediction has superior performance.

Author(s) affiliation: [Lv, Yisheng; Duan, Yanjie; Kang, Wenwen; Wang, Fei-Yue] Chinese Acad Sci, Inst Automat, State Key Lab Management & Control Complex Syst, Beijing 100190, Peoples R China.
[Li, Zhengxi] North China Univ Technol, Beijing 100144, Peoples R China.
Reprint Address: Lv, YS (reprint author), Chinese Acad Sci, Inst Automat, State Key Lab Management & Control Complex Syst, Beijing 100190, Peoples R China.
E-mail Author(s) affiliation: yisheng.lv@ia.ac.cn; duanyanjie2012@ia.ac.cn; kangwenwen2012@ia.ac.cn; lzx@ncut.edu.cn; feiyue@ieee.org
Times Cited: 15
Number of references: 61
Tags: Analytics - Deep learning, Big data

Analytics - Predictive Analysis

Advanced predictive-analysis-based decision support for collaborative logistics networks (Hungary) 2015

Author(s): Ilie-Zudor, E (Ilie-Zudor, Elisabeth); Ekart, A (Ekart, Aniko); Kemeny, Z (Kemeny, Zsolt); Buckingham, C (Buckingham, Christopher); Welch, P (Welch, Philip); Monostori, L (Monostori, Laszlo)

ABSTRACT: The paper establishes an overview of challenges and opportunities of current significance in the area of big data, specifically in the context of transparency and processes in heterogeneous enterprise networks. Within this context, the paper presents how existing components and purpose-driven research were combined for a solution implemented in a nationwide network for less-than-truckload consignments. Findings - Aside from providing an extended overview of today's big data situation, the findings have shown that technical means and methods available today can comprise a feasible process transparency solution in a large heterogeneous network where legacy practices, reporting lags and incomplete data exist, yet processes are sensitive to inadequate policy changes. Practical implications - The means introduced in the paper were found to be of utility value in improving process efficiency, transparency and planning in logistics networks. The particular system design choices in the presented solution allow an incremental introduction or evolution of resource handling practices, incorporating existing fragmentary, unstructured or tacit

knowledge of experienced personnel into the theoretically founded overall concept. Originality/value - The paper extends previous high-level view on the potential of big data, and presents new applied research and development results in a logistics application.

Author(s) affiliation: [Ilie-Zudor, Elisabeth; Kemeny, Zsolt; Monostori, Laszlo] Hungarian Acad Sci, Res Lab Engn & Management Intelligence, Budapest, Hungary.
[Ekart, Aniko; Buckingham, Christopher; Welch, Philip] Aston Univ, Dept Comp Sci, Birmingham B4 7ET, W Midlands, England.
Reprint Address: Ilie-Zudor, E (reprint author), Hungarian Acad Sci, Res Lab Engn & Management Intelligence, Budapest, Hungary.
Total Times Cited: 0
Number of references: 84
Tags: Analytics - Predictive analysis, Big data

Big Data and Predictive analysis: What's New? (USA) 2014

Author(s): Earley, S (Earley, Seth)

ABSTRACT: There's a lot of noise and nonsense about so-called big data, especially about its role in the new and exciting field of "predictive analytics." But many industries have been using data and analytics for decades. What's truly new in terms of predictive analytics, and what does this mean for the IT industry?

Author(s) affiliation: Earley & Associates Www Earley Com, Carlisle, MA 01741 USA.
Reprint Address: Earley, S (reprint author), Earley & Associates Www Earley Com, Carlisle, MA 01741 USA.
E-mail Author(s) affiliation: seth@earley.com
Times Cited: 0
Number of references: 1
Tags: Analytics - Predictive analysis, Big data

Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control (USA) 2016

Author(s): Moyne, J (Moyne, James); Samantaray, J (Samantaray, Jamini); Armacost, M (Armacost, Michael)

ABSTRACT: As requirements on data volumes, rates, quality, merging, and analytics increase exponentially in the digital universe, semiconductor manufacturers are faced with a need for new approaches to data management and use across the Fab. These are often termed "big data" challenges. In our industry big data solutions will be key to scaling advanced process control (APC) solutions to finer levels of control and diagnostics. However the main impact will be to better enable more effective predictive technologies such as predictive maintenance (PdM), virtual metrology and yield prediction, all of which utilize data from traditional APC capabilities that include fault detection and classification and run-to-run control. PdM represents one area where big data solutions are generating significant benefits across a variety of process types. Moving to big data solutions involves addressing the aforementioned requirements either with enhancements of existing systems or moving to more big data friendly platforms. Big data friendly platforms applied to APC systems provide quantifiable cost-of-ownership and speed improvements, thereby better enabling high quality prediction solutions. Initially, big data solutions will largely be delegated to off-line and on-time critical tasks; over the longer term these big data solutions will increasingly be leveraged for time critical and real-time capabilities.

Author(s) affiliation: [Moyne, James; Samantaray, Jamini; Armacost, Michael] Appl Mat Inc, Santa Clara, CA 95054 USA.
Reprint Address: Moyne, J (reprint author), Appl Mat Inc, Santa Clara, CA 95054 USA.
E-mail Author(s) affiliation: moyne@umich.edu
Times Cited: 0
Number of references: 20
Tags: Analytics - Predictive analysis, Big data

Big-Data or Slim-Data: Predictive analysis Will Rule with World (USA) 2016

Author(s): Combs, D (Combs, Daniel); Shetty, S (Shetty, Safal); Parthasarathy, S (Parthasarathy, Sairam)

ABSTRACT: In this issue of the Journal, Ustun and colleagues set us in such a deliberate direction by applying a new machine learning method known as SLIM (Supersparse Linear Integer Models). They tested the hypothesis that a diagnostic screening tool based on routinely available medical information would be superior to one based solely on patient-reported sleep-related symptoms. Their rationale was that the application of such technology—in an automated manner—to data residing in electronic medical records can assist with large-scale screening for obstructive sleep apnea (OSA).

Author(s) affiliation: [Combs, Daniel] Univ Arizona, Dept Pediat, Tucson, AZ 85721 USA.

[Combs, Daniel; Shetty, Safal; Parthasarathy, Sairam] Univ Arizona, Ctr Sleep Disorders, Tucson, AZ USA.
[Shetty, Safal; Parthasarathy, Sairam] Univ Arizona, Dept Med, Tucson, AZ USA.
Reprint Address: Parthasarathy, S (reprint author), Univ Arizona, 1501 N Campbell Ave,AHSC Rm 2342D, Tucson, AZ 85724 USA.
E-mail Author(s) affiliation: sparthasarathy@deptofmed.arizona.edu
Times Cited: 0
Number of references: 22
Tags: Analytics - Predictive analysis, Big data

Data quality for data science, Predictive analysis, and big data in supply chain management: An introduction to the problem and suggestions for research and applications (USA) 2014

Author(s): Hazen, BT (Hazen, Benjamin T.); Boone, CA (Boone, Christopher A.); Ezell, JD (Ezell, Jeremy D.); Jones-Farmer, LA (Jones-Farmer, L. Allison)

ABSTRACT: Today's supply chain professionals are inundated with data, motivating new ways of thinking about how data are produced, organized, and analyzed. This has provided an impetus for organizations to adopt and perfect data analytic functions (e.g. data science, Predictive analysis, and big data) in order to enhance supply chain processes and, ultimately, performance. However, management decisions informed by the use of these data analytic methods are only as good as the data on which they are based. In this paper, we introduce the data quality problem in the context of supply chain management (SCM) and propose methods for monitoring and controlling data quality. In addition to advocating for the importance of addressing data quality in supply chain research and practice, we also highlight interdisciplinary research topics based on complementary theory. (C) 2014 Elsevier B.V. All rights reserved.

Author(s) affiliation: [Hazen, Benjamin T.] Univ Tennessee, Coll Business Adm, Dept Mkt & Supply Chain Management, Stokely Management Ctr 310, Knoxville, TN 37996 USA.
[Boone, Christopher A.] Georgia So Univ, Coll Business Adm, Dept Mkt & Logist, Statesboro, GA 30458 USA.
[Ezell, Jeremy D.; Jones-Farmer, L. Allison] Auburn Univ, Dept Supply Chain & Informat Syst Management, Harbert Coll Business, Auburn, AL 36849 USA.
Reprint Address: Hazen, BT (reprint author), Univ Tennessee, Coll Business Adm, Dept Mkt & Supply Chain Management, Stokely Management Ctr 310, Knoxville, TN 37996 USA.
E-mail Author(s) affiliation: hazen@utk.edu; caboone@georgiasouthern.edu; jde0009@tigermail.auburn.edu; joneall@auburn.edu
Times Cited: 23
Number of references: 98
Tags: Analytics - Predictive analysis, Big data

The future is in the numbers: the power of predictive analysis in the biomedical educational environment (USA) 2016

Author(s): Gullo, CA (Gullo, Charles A.)

ABSTRACT: Biomedical programs have a potential treasure trove of data they can mine to assist admissions committees in identification of students who are likely to do well and help educational committees in the identification of students who are likely to do poorly on standardized national exams and who may need remediation. In this article, we provide a step-by-step approach that schools can utilize to generate data that are useful when predicting the future performance of current students in any given program. We discuss the use of linear regression analysis as the means of generating that data and highlight some of the limitations. Finally, we lament on how the combination of these institution-specific data sets are not being fully utilized at the national level where these data could greatly assist programs at large.

Author(s) affiliation: [Gullo, Charles A.] Marshall Univ, Joan C Edwards Sch Med, Off Med Educ, 1600 Med Ctr Dr,Suite 3411, Huntington, WV 25701 USA.
Reprint Address: Gullo, CA (reprint author), Marshall Univ, Joan C Edwards Sch Med, Off Med Educ, 1600 Med Ctr Dr,Suite 3411, Huntington, WV 25701 USA.
Total Times Cited: 0
Number of references: 8
Tags: Analytics - Predictive analysis, Big data

Predictive analysis, Data Mining and Big Data: Myths, Misconceptions and Methods (USA) 2015

Author(s): Shen, WJ (Shen, Wenjing)

ABSTRACT: Retailers, banks, governments, social networking sites, credit reference agencies and télécoms companies, amongst others, hold vast amounts of information about us. They know where we live, what we spend our money on, who our friends and family are, our likes and dislikes, our lifestyles a nd our opinions. Every year the amount of electronic information about us grows as we increasingly use internet services, social media and smart devices to move more and more of our lives into the online environment.

Author(s) affiliation: [Shen, Wenjing] Drexel Univ, Dept Decis Sci & MIS, LeBow Coll Business, Philadelphia, PA 19104 USA.
Reprint Address: Shen, WJ (reprint author), Drexel Univ, Dept Decis Sci & MIS, LeBow Coll Business, Philadelphia, PA 19104 USA.
E-mail Author(s) affiliation: wenjings@drexel.edu

Times Cited: 0
Number of references: 1
Tags: Analytics - Predictive analysis, Big data

Special Issue on Big Data and Predictive analysis Application in Supply Chain Management

Author(s): Gunasekaran, A (Gunasekaran, Angappa); Tiwari, MK (Tiwari, Manoj Kumar); Dubey, R (Dubey, Rameshwar); Wamba, SF (Wamba, Samuel Fosso)

ABSTRACT: The aim of this special issue of Computers and Industrial Engineering (CAIE) is to attract manuscripts which are firmly grounded in supply chain theories, using big data and predictive analytics to take the current supply chain theory and practice to the next level of excellence in terms of supporting suitable supply design and operations in the 21st century organizational competitiveness. These deliverables will re-align the supply chain into a more coordinated and integrated network with collaborative, efficient and secure service systems

Author(s) affiliation: agunasekaran@umassd.edu; mktiwari9@iem.iitko.ernet.in; rameshwardubey@gmail.com; SamuelFOSSO. WAMBA@neoma-bs.fr

Times Cited: 0
Number of references: 3
Tags: Analytics - Predictive analysis, Big data

Support Technologies (Acquistion, Storage, Computation)

Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks (Spain) 2014

Author(s): Fernandez, A (Fernandez, Alberto); del Rio, S (del Rio, Sara); Lopez, V (Lopez, Victoria); Bawakid, A (Bawakid, Abdullah); del Jesus, MJ (del Jesus, Maria J.); Benitez, JM (Benitez, Jose M.); Herrera, F (Herrera, Francisco)

ABSTRACT: In this article, we provide an overview on the topic of Big Data, and how the current problem can be addressed from the perspective of Cloud Computing and its programming frameworks. In particular, we focus on those systems for large-scale analytics based on the MapReduce scheme and Hadoop, its open-source implementation. We identify several libraries and software projects that have been developed for aiding practitioners to address this new programming model. We also analyze the advantages and disadvantages of MapReduce, in contrast to the classical solutions in this field. Finally, we present a number of programming frameworks that have been proposed as an alternative to MapReduce, developed under the premise of solving the shortcomings of this model in certain scenarios and platforms. WIREs Data Mining Knowl Discov 2014, 4:380-409. doi: 10.1002/widm.1134 For further resources related to this article, please visit the . Conflict of interest: The authors have declared no conflicts of interest for this article.

Author(s) affiliation: [Fernandez, Alberto; del Jesus, Maria J.] Univ Jaen, Dept Comp Sci, Jaen, Spain.
[del Rio, Sara; Lopez, Victoria; Benitez, Jose M.; Herrera, Francisco] Univ Granada, Dept Comp Sci & Artificial Intelligence, Granada, Spain.
[Bawakid, Abdullah; Herrera, Francisco] King Abdulaziz Univ, Fac Comp & Informat Technol North Jeddah, Jeddah 21413, Saudi Arabia.

Reprint Address: Fernandez, A (reprint author), Univ Jaen, Dept Comp Sci, Jaen, Spain.
E-mail Author(s) affiliation: alberto.fernandez@ujaen.es
Times Cited: 16
Number of references: 162
Tags: Big data - Supporting technologies, Big data

The design of data storage system based on Lustre for EAST (China) 2016

Author(s): Wang, F (Wang, Feng); Chen, Y (Chen, Ying); Li, S (Li, Shi); Yang, F (Yang, Fei); Xiao, BJ (Xiao, Bingjia)

ABSTRACT: The quasi-steady state operation is one of the main purposes of EAST tokamak, and more than 400 s discharge pulse has been achieved in the past campaigns. The acquired data amount increases continuously with the discharge length. At the same time to meet the requirement of the upgrade and improvement of the diagnostic systems, more and more data acquisition channels have come into service. Some new diagnostic systems require high sampling rate data acquisition more than 10MSPS. In the last campaign 2014, the data streaming is about 2000MB/s and the total data amount is more than 100TB. How to store the huge data continuously becomes a big problem. A new data storage system based on Lustre has been designed to solve the problem. All the storage nodes and servers are connected to InfiniBand FDR 56Gbps network. The maximum parallel throughput of the total storage system is about 10GB/s. It is easy to expand the storage system by adding I/O nodes when more capacity and performance are required in the future. The new data storage system will be applied in the next campaign of EAST. The system details are given in the paper. (C) 2016 Elsevier B.V. All rights reserved.

Author(s) affiliation: [Wang, Feng; Chen, Ying; Li, Shi; Yang, Fei; Xiao, Bingjia] Chinese Acad Sci, Inst Plasma Phys, Hefei, Anhui, Peoples R China.
[Xiao, Bingjia] Univ Sci & Technol China, Sch Nucl Sci & Technol, Hefei, Anhui, Peoples R China.
[Yang, Fei] Anhui Med Univ, Dept Comp Sci, Hefei, Anhui, Peoples R China.
Reprint Address: Wang, F (reprint author), Chinese Acad Sci, Inst Plasma Phys, Hefei, Anhui, Peoples R China.
E-mail Author(s) affiliation: wangfeng@ipp.ac.cn
Times Cited: 0
Number of references: 6
Tags: Big data - Supporting technologies, Big data

A distributed frequent itemset mining algorithm using Spark for Big Data analytics (China) 2015

Author(s): Zhang, F (Zhang, Feng); Liu, M (Liu, Min); Gui, F (Gui, Feng); Shen, WM (Shen, Weiming); Shami, A (Shami, Abdallah); Ma, YL (Ma, Yunlong)

ABSTRACT: This paper proposes an efficient distributed frequent itemset mining algorithm (DFIMA) which can significantly reduce the amount of candidate itemsets by applying a matrix-based pruning approach. The proposed algorithm has been implemented using Spark to further improve the efficiency of iterative computation. Numeric experiment results using standard benchmark datasets by comparing the proposed algorithm with the existing algorithm, parallel FP-growth, show that DFIMA has better efficiency and scalability. In addition, a case study has been carried out to validate the feasibility of DFIMA.

Author(s) affiliation: [Zhang, Feng; Liu, Min; Gui, Feng; Ma, Yunlong] Tongji Univ, Sch Elect & Informat Engn, Shanghai 201804, Peoples R China.
[Shen, Weiming] Tongji Univ, Key Lab Embedded Syst & Serv Comp, Shanghai 201804, Peoples R China.
[Zhang, Feng; Shami, Abdallah] Univ Western Ontario, Dept Elect & Comp Engn, London, ON N6A 5B9, Canada.
Reprint Address: Ma, YL (reprint author), Tongji Univ, Sch Elect & Informat Engn, Shanghai 201804, Peoples R China.
E-mail Author(s) affiliation: zhangfeng0726@gmail.com; lmin@tongji.edu.cn; guifengleaf@gmail.com; wshen@ieee.org; ashami2@ uwo.ca; evanma@tongji.edu.cn
Times Cited: 3
Number of references: 42
Tags: Big data - Supporting technologies, Big data

The Gamma Matrix to Summarize Dense and Sparse Data Sets for Big Data Analytics (USA) 2016

Author(s): Ordonez, C (Ordonez, Carlos); Zhang, YQ (Zhang, Yiqun); Cabrera, W (Cabrera, Wellington)

ABSTRACT: We propose a parallel array operator, based on a specific form of matrix multiplication, that computes a comprehensive data summarization matrix. By deriving equivalent equations based on the summarization matrix, statistical methods are adapted to work in two phases: (1) Parallel summarization of the data set in one pass; (2) Iteration exploiting the summarization matrix

in many intermediate computations. We prove our summarization matrix captures essential statistical properties of the data set and it allows iterative algorithms to work faster in main memory, by decreasing the number of times the data set is scanned, and by reducing the number of CPU operations. Specifically, we show our summarization matrix benefits statistical models, including PCA, linear regression, and variable selection. From a systems perspective, we carefully study the efficient computation of the summarization matrix on the SciDB parallel array DBMS and how to exploit it in the R language statistical system. To achieve best performance, we introduce two specialized array operators for dense and sparse data sets, respectively. We present an experimental evaluation comparing SciDB, R, a columnar DBMS (a fast SQL engine), and Spark (a popular Hadoop system). Our experiments show R working together with SciDB eliminates main memory and performance limitations from R. More importantly, our R+SciDB prototype is significantly faster and more scalable than Spark and the columnar DBMS.

Author(s) affiliation: [Ordonez, Carlos; Zhang, Yiqun; Cabrera, Wellington] Univ Houston, Dept Comp Sci, Houston, TX 77204 USA.
Reprint Address: Ordonez, C (reprint author), Univ Houston, Dept Comp Sci, Houston, TX 77204 USA.
E-mail Author(s) affiliation: ordonez@cs.uh.edu; yzhang@cs.uh.edu; wcabrera@cs.uh.edu
Times Cited: 0
Number of references: 28
Tags: Big data - Supporting technologies, Big data

Handling big data: research challenges and future directions (Greece) 2016

Author(s): Anagnostopoulos, I (Anagnostopoulos, I.); Zeadally, S (Zeadally, S.); Exposito, E (Exposito, E.)

ABSTRACT: Today, an enormous amount of data is being continuously generated in all walks of life by all kinds of devices and systems every day. A significant portion of such data is being captured, stored, aggregated and analyzed in a systematic way without losing its "4V" (i.e., volume, velocity, variety, and veracity) characteristics. We review major drivers of big data today as well the recent trends and established platforms that offer valuable perspectives on the information stored in large and heterogeneous data sets. Then, we present a classification of some of the most important challenges when handling big data. Based on this classification, we recommend solutions that could address the identified challenges, and in addition we highlight cross-disciplinary research directions that need further investigation in the future.

Author(s) affiliation: [Anagnostopoulos, I.] Univ Thessaly, Nea Ionia, Greece.
[Zeadally, S.] Univ Kentucky, Lexington, KY 40506 USA.
[Exposito, E.] Univ Toulouse, Toulouse, France.
Reprint Address: Zeadally, S (reprint author), Univ Kentucky, Lexington, KY 40506 USA.
E-mail Author(s) affiliation: szeadally@uky.edu
Times Cited: 0
Number of references: 68
Tags: Big data - Supporting technologies, Big data

Knowledge discovery from social media using big data-provided sentiment analysis (SoMABiT) (Germany) 2015

Author(s): Bohlouli, M (Bohlouli, Mahdi); Dalter, J (Dalter, Jens); Dornhofer, M (Dornhoefer, Mareike); Zenkert, J (Zenkert, Johannes); Fathi, M (Fathi, Madjid)

ABSTRACT: This paper presents and discusses the technological and scientific focus of SoMABiT as a social media analysis platform using big data technology. Sentiment analysis has been employed in order to discover knowledge from social media. The use of MapReduce and the development of a distributed algorithm towards an integrated platform that can scale for any data volume and provide social media-driven knowledge is the main novelty of the proposed concept in comparison to the state-of-the-art technologies.

Author(s) affiliation: [Bohlouli, Mahdi; Dalter, Jens; Dornhoefer, Mareike; Zenkert, Johannes; Fathi, Madjid] Univ Siegen, D-57076 Siegen, Germany.
Reprint Address: Bohlouli, M (reprint author), Univ Siegen, Dept Elect Engn & Comp Sci, Inst Knowledge Based Syst, Hoelderlinstr 3, D-57076 Siegen, Germany.
E-mail Author(s) affiliation: mahdi.bohlouli@uni-siegen.de
Times Cited: 1
Number of references: 63
Tags: Big data - Supporting technologies, Big data

Optical storage arrays: a perspective for future big data storage (Australia) 2014

Author(s): Gu, M (Gu, Min); Li, XP (Li, Xiangping); Cao, YY (Cao, Yaoyu)

ABSTRACT: The advance of nanophotonics has provided a variety of avenues for light-matter interaction at the nanometer scale through the enriched mechanisms for physical and chemical reactions induced by nanometer-confined optical probes in nanocomposite materials. These emerging nanophotonic devices and materials have enabled researchers to develop disruptive methods of tremendously increasing the storage capacity of current optical memory. In this paper, we present a review of the recent advancements in nanophotonics-enabled optical storage techniques. Particularly, we offer our perspective of using them as optical storage arrays for next-generation exabyte data centers.

Author(s) affiliation: [Gu, Min; Li, Xiangping; Cao, Yaoyu] Swinburne Univ Technol, Fac Sci Engn & Technol, Ctr Microphoton, Hawthorn, Vic 3122, Australia.
Reprint Address: Gu, M (reprint author), Swinburne Univ Technol, Fac Sci Engn & Technol, Ctr Microphoton, Hawthorn, Vic 3122, Australia.
E-mail Author(s) affiliation: mgu@swin.edu.au

Times Cited: 29
Number of references: 64
Tags: Big data - Supporting technologies, Big data

Parallel Processing Systems for Big Data: A Survey (China) 2016

Author(s): Zhang, YQ (Zhang, Yunquan); Cao, T (Cao, Ting); Li, SG (Li, Shigang); Tian, XH (Tian, Xinhui); Yuan, L (Yuan, Liang); Jia, HP (Jia, Haipeng); Vasilakos, AV (Vasilakos, Athanasios V.)

ABSTRACT: To explore new research opportunities and assist users in selecting suitable processing systems for specific applications, this survey paper will give a high-level overview of the existing parallel data processing systems categorized by the data input as batch processing, stream processing, graph processing, and machine learning processing and introduce representative projects in each category. As the pioneer, the original MapReduce system, as well as its active variants and extensions on dataflow, data access, parameter tuning, communication, and energy optimizations will be discussed at first. System benchmarks and open issues for big data processing will also be studied in this survey.

Author(s) affiliation: [Zhang, Yunquan; Cao, Ting; Li, Shigang; Yuan, Liang; Jia, Haipeng] Chinese Acad Sci, Inst Comp Technol, State Key Lab Comp Architecture, Beijing 100190, Peoples R China.
[Tian, Xinhui] Chinese Acad Sci, Inst Comp Technol, Adv Comp Syst Res Ctr, Beijing 100190, Peoples R China.
[Vasilakos, Athanasios V.] Lulea Univ Technol, Dept Comp Sci Elect & Space Engn, S-97187 Lulea, Sweden.
Reprint Address: Zhang, YQ (reprint author), Chinese Acad Sci, Inst Comp Technol, State Key Lab Comp Architecture, Beijing 100190, Peoples R China.
E-mail Author(s) affiliation: zyq@ict.ac.cn; caoting@ict.ac.cn; lishigang@ict.ac.cn; tianxinhui@ict.ac.cn; yuanliang@ict.ac.cn; jiahaipeng@ict.ac.cn; vasilako@ath.forthnet.gr

Times Cited: 0
Number of references: 74
Tags: Big data - Supporting technologies, Big data

The rise of "big data" on cloud computing: Review and open research issues (Malaysia) 2015

Author(s): Hashem, IAT (Hashem, Ibrahim Abaker Targio); Yaqoob, I (Yaqoob, Ibrar); Anuar, NB (Anuar, Nor Badrul); Mokhtar, S (Mokhtar, Salimah); Gani, A (Gani, Abdullah); Khan, SU (Khan, Samee Ullah)

ABSTRACT: The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and Cloud computing, big data storage systems, and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized. (C) 2014 Elsevier Ltd. All rights reserved.

Author(s) affiliation: [Hashem, Ibrahim Abaker Targio; Yaqoob, Ibrar; Anuar, Nor Badrul; Mokhtar, Salimah; Gani, Abdullah] Univ Malaya, Fac Comp Sci & Informat Technol, Kuala Lumpur 50603, Malaysia.
[Khan, Samee Ullah] N Dakota State Univ, NDSU CIIT Green Comp & Commun Lab, Fargo, ND 58108 USA.

Reprint Address: Hashem, IAT (reprint author), Univ Malaya, Fac Comp Sci & Informat Technol, Kuala Lumpur 50603, Malaysia.
E-mail Author(s) affiliation: targio@siswa.um.edu.my; ibraryaqoob@siswa.um.edu.my; badrul@um.edu.my; salimah@um.edu.my;
abdullah@um.edu.my; samee.khan@ndsu.edu
Times Cited: 84
Number of references: 126
Tags: Big data - Supporting technologies, Big data

Social big data: Recent achievements and new challenges (Spain) 2016

Author(s): Bello-Orgaz, G (Bello-Orgaz, Gema); Jung, JJ (Jung, Jason J.); Camacho, D (Camacho, David)
Source: INFORMATION FUSION Volume: 28 Pages: 45-59 DOI: 10.1016/j.inffus.2015.08.005 Published: MAR 2016

ABSTRACT: A number of libraries such as Mahout and SparkMLib have been designed to develop new efficient applications based on machine learning algorithms. The combination of big data technologies and traditional machine learning algorithms has generated new and interesting challenges in other areas as social media and social networks. These new challenges are focused mainly on problems such as data processing, data storage, data representation, and how data can be used for pattern mining, analysing user behaviours, and visualizing and tracking data, among others. In this paper, we present a revision of the new methodologies that is designed to allow for efficient data mining and information fusion from social media and of the new applications and frameworks that are currently appearing under the "umbrella" of the social networks, social media and big data paradigms. (C) 2015 Elsevier B.V. All rights reserved.

Author(s) affiliation: [Bello-Orgaz, Gema; Camacho, David] Univ Autonoma Madrid, Dept Comp Sci, E-28049 Madrid, Spain.
[Jung, Jason J.] Chung Ang Univ, Dept Comp Engn, Seoul 156756, South Korea.
Reprint Address: Jung, JJ (reprint author), Chung Ang Univ, Dept Comp Engn, Seoul 156756, South Korea.
E-mail Author(s) affiliation: gema.bello@uam.es; j2jung@gmail.com; david.camacho@uam.es
Times Cited: 15
Number of references: 152
Tags: Big data - Supporting technologies, Big data

A Software Framework for Efficient IoT Contexts Acquisition and Big Data Analytics (South Korea) 2014

Author(s): Kim, MK (Kim, Moon Kwon); La, HJ (La, Hyun Jung); Kim, SD (Kim, Soo Dong)
Source: JOURNAL OF INTERNET TECHNOLOGY Volume: 15 Issue: 6 Pages: 939-947 DOI: 10.6138/JIT.2014.15.6.06 Published: NOV 2014

ABSTRACT: In this paper, we present a framework for efficiently acquiring and analyzing IoT contexts by utilizing big data approaches. We first discuss characteristics of IoT contexts, and define the requirement of a framework for managing IoT contexts. And, we present the design model and an implementation of the framework. Finally, we show the result of applying the framework in developing Safety Surveillance System for secure buildings.

Author(s) affiliation: [Kim, Moon Kwon; Kim, Soo Dong] Soongsil Univ, Dept Comp Sci, Seoul, South Korea.
Reprint Address: La, HJ (reprint author), Soongsil Univ, Dept Comp Sci, Seoul, South Korea.
E-mail Author(s) affiliation: mkdmkk@gmail.com; hjla80@gmail.com; sdkim777@gmail.com
Times Cited: 0
Number of references: 26
Tags: Big data - Supporting technologies, Big data - Acquisition, Big data

A View on Fuzzy Systems for Big Data: Progress and Opportunities (Spain) 2016

Author(s): Fernandez, A (Fernandez, Alberto); Carmona, CJ (Jose Carmona, Cristobal); del Jesus, MJ (Jose del Jesus, Maria); Herrera, F (Herrera, Francisco)
Source: INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE SYSTEMS Volume: 9 Special Issue: SI Pages: 69-80 DOI: 10.1080/18756891.2016.1180820 Supplement: 1 Published: 2016

ABSTRACT: In this paper, we will carry out an overview of the main existing proposals on the topic, analyzing the design of these models. Additionally, we will discuss those problems related to the data distribution and parallelization of the current algorithms, and also its relationship with the fuzzy representation of the information. Finally, we will provide our view on the expectations for the future in this framework according to the design of those methods based on fuzzy sets, as well as the open challenges on the topic.

Author(s) affiliation: [Fernandez, Alberto; Jose del Jesus, Maria] Univ Jaen, Dept Comp Sci, Jaen 23071, Spain.
[Jose Carmona, Cristobal] Univ Burgos, Dept Civil Engn, Burgos 09006, Spain.
[Herrera, Francisco] Univ Granada, Dept Comp Sci & Artificial Intelligence, CITIC UGR, Res Ctr Informat & Commun Technol, E-18071 Granada, Spain.
[Herrera, Francisco] King Abdulaziz Univ, Fac Comp & Informat Technol, Jeddah 21413, Saudi Arabia.

Reprint Address: Fernandez, A; del Jesus, MJ (reprint author), Univ Jaen, Dept Comp Sci, Jaen 23071, Spain.
Carmona, CJ (reprint author), Univ Burgos, Dept Civil Engn, Burgos 09006, Spain.
Herrera, F (reprint author), Univ Granada, Dept Comp Sci & Artificial Intelligence, CITIC UGR, Res Ctr Informat & Commun Technol, E-18071 Granada, Spain.
Herrera, F (reprint author), King Abdulaziz Univ, Fac Comp & Informat Technol, Jeddah 21413, Saudi Arabia.
E-mail Author(s) affiliation: alberto.fernandez@ujaen.es; cjcarmona@ubu.es; mjjesus@ujaen.es; herrera@decsai.ugr.es

Times Cited: 1
Number of references: 75
Tags: Big data - Supportig technologies, Big data ■

ABOUT THIS PUBLICATION