**Using Support Vector Machines to Classify Student Attentiveness for the Development of Personalized Learning Systems**

There have been many studies in which researchers have attempted to classify student attentiveness. Many of these approaches depended on a qualitative analysis and lacked any quantitative analysis. Therefore, this work is focused on bridging the gap between qualitative and quantitative approaches to classify student attentiveness. Thus, this research applies machine learning algorithms (K-means and SVM) to automatically classify students as attentive or inattentive using data from a consumer RGB-D sensor. Results of this research can be used to improve teaching strategies for instructors at all levels and can aid instructors in implementing personalized learning systems, which is a National Academy of Engineering Grand Challenge. This research applies machine learning algorithms to an educational setting. Data from these algorithms can be used by instructors to provide valuable feedback on the effectiveness of their instructional strategies and pedagogies. Instructors can use this feedback to improve their instructional strategies, and students will benefit by achieving improved learning and subject mastery. Ultimately, this will result in the students' increased ability to do work in their respective areas. Broadly, this work can help advance efforts in many areas of education and instruction. It is expected that improving instructional strategies and implementing personalized learning will help create more competent, capable, and prepared persons available for the future workforce.

Clustering algorithms, Classification algorithms, Support vector machines, Education, Databases, Machine learning, algorithms, Learning systems

# The application of machine learning algorithm in underwriting process

This paper firstly analyses the actual underwriting methods of Chinese life insurance companies, and points out the merits and shortcomings of these methods. Then the incomplete database of insurance company is mined by the data mining's association rule algorithm. Thirdly the support vector machine (SVM) is applied to the underwriting process to classify the applicants. Finally the directions for improving this algorithm are pointed out. The algorithm proposed in this paper has promising future in underwriting process.

Machine learning algorithms , Hazards , Support vector machines , Association rules , Data mining , Insurance , Support vector machine classification , Databases , Machine learning , Kernel

# A new heuristic of the decision tree induction

Decision tree induction is one of the useful approaches for extracting classification knowledge from a set of feature-based instances. The most popular heuristic information used in the decision tree generation is the minimum entropy. This heuristic information has a serious disadvantage-the poor generalization capability [3]. Support vector machine (SVM) is a classification technique of machine learning based on statistical learning theory. It has good generalization. Considering the relationship between the classification margin of support vector machine(SVM) and the generalization capability, the large margin of SVM can be used as the heuristic information of decision tree, in order to improve its generalization capability. This paper proposes a decision tree induction algorithm based on large margin heuristic. Comparing with the binary decision tree using the minimum entropy as the heuristic information, the experiments show that the generalization capability has been improved by using the new heuristic.

Decision trees , Support vector machines , Entropy , Support vector machine classification , Machine learning , Classification tree analysis , Induction generators , Inverse problems , Machine learning algorithms , Cybernetics

# An incremental learning approach for restricted boltzmann machines

Determination of model complexity is a challenging issue to solve computer vision problems using restricted boltzmann machines (RBMs). Many algorithms for feature learning depend on cross-validation or empirical methods to optimize the number of features. In this work, we propose an learning algorithm to find the optimal model complexity for the RBMs by incrementing the hidden layer. The proposed algorithm is composed of two processes: 1) determining incrementation necessity of neurons and 2) computing the number of additional features for the increment. Specifically, the proposed algorithm uses a normalized reconstruction error in order to determine incrementation necessity and prevent unnecessary increment for the number of features during training. Our experimental results demonstrated that the proposed algorithm converges to the optimal number of features in a single layer RBMs. In the classification results, our model could outperform the non-incremental RBM.

Neurons , Complexity theory , Computational modeling , Training , Error analysis , Training data , Standards

# Chinese Text Categorization study based on feature weight learning

Text categorization (TC) is an important component in many information organization and information management tasks. Two key issues in TC are feature coding and classifier design. The Euclidean distance is usually chosen as the similarity measure in K-nearest neighbor classification algorithm. All the features of each vector have different functions in describing samples. So we can decide different function of every feature by using feature weight learning. In this paper text categorization via K-nearest neighbor algorithm based on feature weight learning is described. The numerical experiments prove the validity of this learning algorithm.

Text categorization , Machine learning , Nearest neighbor searches , Testing , Euclidean distance , Machine learning algorithms , Classification tree analysis , Decision trees , Support vector machines , Support vector machine classification

# Application of machine learning techniques to Web-based intelligent learning diagnosis system

This work proposes an intelligent learning diagnosis system that supports a Web-based thematic learning model, which aims to cultivate learners' ability of knowledge integration by giving the learners the opportunities to select the learning topics that they are interested, and gain knowledge on the specific topics by surfing on the Internet to search related learning courseware and discussing what they have learned with their colleagues. Based on the log files that record the learners' past online learning behavior, an intelligent diagnosis system is used to give appropriate learning guidance to assist the learners in improving their study behaviors and grade online class participation for the instructor. The achievement of the learners' final reports can also be predicted by the diagnosis system accurately. Our experimental results reveal that the proposed learning diagnosis system can efficiently help learners to expand their knowledge while surfing in cyberspace Web-based "theme-based learning" model.

Machine learning , Intelligent systems , Learning systems , Education , Internet , Educational institutions , Courseware , Hybrid intelligent systems , Bayesian methods , Support vector machines

# A new semi-supervised support vector machine learning algorithm based on active learning

Semi-supervised support vector machine is an extension of standard support vector machine in machine learning problem in real life. However, the existing semi-supervised support vector machine algorithm has some drawbacks such as slower training speed, lower accuracy, etc. This paper presents a semi-supervised support vector machine learning algorithm based on active learning, which trains early learner by a spot of labeled-data, selects the best training samples for training and learning by active learning and reduces learning cost by deleting non- support vector. Simulative experiments have shown that the algorithm may get good learning effect at less learning cost.

Machine learning , Support vector machines , Machine learning algorithms , Support vector machine classification , Semisupervised learning , Predictive models , Educational institutions , Petroleum , Costs , Pattern recognition

# Deep Transfer Learning via Restricted Boltzmann Machine for Document Classification

Transfer learning aims to improve a targeted learning task using other related auxiliary learning tasks and data. Most current transfer-learning methods focus on scenarios where the auxiliary and the target learning tasks are very similar: either (some of) the auxiliary data can be directly used as training examples for the target task or the auxiliary and the target data share the same representation. However, in many cases the connection between the auxiliary and the target tasks can be remote. Only a few features derived from the auxiliary data may be helpful for the target learning. We call such scenario the deep transfer-learning scenario and we introduce a novel transfer-learning method for deep transfer. Our method uses restricted Boltzmann machine to discover a set of hierarchical features from the auxiliary data. We then select from these features a subset that are helpful for the target learning, using a selection criterion based on the concept of kernel-target alignment. Finally, the target data are augmented with the selected features before training. Our experiment results show that this transfer method is effective. It can improve classification accuracy by up to more than 10%, even when the connection between the auxiliary and the target tasks is not apparent.

Training , Machine learning , Vectors , Accuracy , Kernel , Data models , Random variables

**Apply Support Vector Machine for CRM Problem**

Data mining in the CRM aiming at learning available knowledge from the customer relationship by machine learning or statistical method to instruct the strategic behavior so that obtain the most profit. In recent years, Support vector machine (SVMs) has been proposed as a power tool in machine leaning and data mining. This paper applies the SVMs to resolve the practical CRM problem in a company. The final results report the good general performance of SVMs for CRM problem.

Support vector machines , Machine learning , Data mining , Support vector machine classification , Machine learning algorithms , Risk management , Statistical analysis , Customer relationship management , Cybernetics , Educational institutions

# A user-centric machine learning framework for cyber security operations center

To assure cyber security of an enterprise, typically SIEM (Security Information and Event Management) system is in place to normalize security events from different preventive technologies and flag alerts. Analysts in the security operation center (SOC) investigate the alerts to decide if it is truly malicious or not. However, generally the number of alerts is overwhelming with majority of them being false positive and exceeding the SOC's capacity to handle all alerts. Because of this, potential malicious attacks and compromised hosts may be missed. Machine learning is a viable approach to reduce the false positive rate and improve the productivity of SOC analysts. In this paper, we develop a user-centric machine learning framework for the cyber security operation center in real enterprise environment. We discuss the typical data sources in SOC, their work flow, and how to leverage and process these data sets to build an effective machine learning system. The paper is targeted towards two groups of readers. The first group is data scientists or machine learning researchers who do not have cyber security domain knowledge but want to build machine learning systems for security operations center. The second group of audiences are those cyber security practitioners who have deep knowledge and expertise in cyber security, but do not have machine learning experiences and wish to build one by themselves. Throughout the paper, we use the system we built in the Symantec SOC production environment as an example to demonstrate the complete steps from data collection, label creation, feature engineering, machine learning algorithm selection, model performance evaluations, to risk score generation.

Learning systems , Data models , Predictive models , Computer security , Machine learning algorithms , Mathematical model

# Comparison of Classification Techniques used in Machine Learning as Applied on Vocational Guidance Data

Recent developments in information systems as well as computerization of business processes by organizations have led to a faster, easier and more accurate data analysis. Data mining and machine learning techniques have been used increasingly in the analysis of data in various fields ranging from medicine to finance, education and energy applications. Machine learning techniques make it possible to deduct meaningful further information from those data processed by data mining. Such meaningful and significant information helps organizations to establish their future policies on a sounder basis, and to gain major advantages in terms of time and cost. This study applies classification algorithms used in data mining and machine learning techniques on those data obtained from individuals during the vocational guidance process, and tries to determine the most appropriate algorithm.

Machine learning , Data mining , Classification algorithms , Machine learning algorithms , Algorithm design and analysis , Learning systems , Computers

# Fitting and Prediction for Crack Propagation Rate Based on Machine Learning Optimal Algorithm

Establishing fatigue crack propagation rate is the key to forecasting structure fatigue lifetime, nine parameters fatigue crack propagation rate model and McEvily model are widely applied at present, but it is very complex to realize these models, partial derivative must be calculated and there is large deviation between fitted static parameter and actual value and physical conception isn't clear. In accordance with the disadvantage above methods, Based on optimum parameter selection with grid search and cross validation, we presented optimal common machine learning algorithm (least squares support vector machine-LSSVM) method for fatigue crack propagation rate forecast. Complicated and strong nonlinear fatigue crack propagation rate curve was simulated by network design and conformation of LSSVM learning algorithm and the optimized SVM parameters were selected by the method of network searching and cross validation. Compared the errors with output value of the optimized model and output value from nine parameters fatigue crack propagation rate fitting model, LSSVM whose parameter was optimized with cross validation had excellent ability of nonlinear modeling and generalization. It provided a simple and feasible intelligent approach for material fatigue analysis.

Machine learning , Machine learning algorithms , Fatigue , Predictive models , Algorithm design and analysis , Support vector machines , Least squares methods , Design optimization , Life estimation , Lifetime estimation

# Financial Data Mining Based on Support Vector Machines and Ensemble Learning

With the rapid development of e-commerce, financial data mining has been one of the most important research topics in the data mining community. Support vector machines (SVMs) and ensemble learning are two popular techniques in the machine learning field. In this paper, support vector machines and ensemble learning are used to classify financial data respectively. The experiments conducted on the public dataset show that compared with SVMs, ensemble learning achieves obvious improvement of performance.

Data mining , Support vector machines , Machine learning , Machine learning algorithms , Boosting , Support vector machine classification , Risk management , Pattern recognition , Learning systems , Automation

# A heuristic algorithm to incremental support vector machine learning

Incremental learning techniques are possible solutions to handle vast data as information from Internet updating gets faster. Support vector machine works well for incremental learning model with impressive performance for its outstanding power to summarize the data space in a concise way. This paper proposes a heuristic algorithm to incremental learning with SVM taking the possible impact of new training data to history data into account. The idea of this heuristic algorithm is that the partition difference set has less elements, and existing hyperplane is much closer to the optimal one. New support vectors in this algorithm consist of existing support vectors and partition difference set of new training data and history data by separating hyperplane. The algorithm improves classification precision by adding partition difference set, and decreases the computation complexity by constructing new classification hyperplane on support vector set. The experimental results show that this heuristic algorithm is efficient and effective to improve the classification precision.

Heuristic algorithms , Machine learning algorithms , Support vector machines , Machine learning , Support vector machine classification , Training data , Partitioning algorithms , Data engineering , History , Educational institutions

# Statistical Machine Learning Used in Integrated Anti-Spam System

IASS is the integrated anti-spam system, which adopts machine learning to filter spam in a intelligent, flexible, precise, and self-adaptive way. The methods of linear classification based on optimal separating hyperplane and K-means clustering are used in action recognition layer. The method of improved naive Bayes is used in content analysis layer. The application of machine learning helps improve the performance of IASS.

Machine learning , Electronic mail , Frequency , Learning systems , Cybernetics , Intelligent networks , Filters , Unsolicited electronic mail , Protection , Postal services

# Dynamic single machine scheduling using Q-learning agent

Single machine scheduling methods have attracted a lot of attentions in recent years. Most dynamic single machine scheduling problems in practice have been addressed using dispatching rules. However, no single dispatching rule has been found to perform well for all important criteria, and no rule takes into account the status or the other resources of system's environment. In this research, an intelligent agent-based single machine scheduling system is proposed, where the agent is trained by a new improved Q-learning algorithm. In such scheduling system, agent selects one of appropriate dispatching rules for machine based on available information. The agent was trained by a new simulated annealing-based Q-learning algorithm. The simulation results show that the simulated annealing-based Q-learning agent is able to learn to select the best dispatching rule for different system objectives. The results also indicate that simulated annealing-based Q-learning agent could perform well for all criteria, which is impossible when using only one dispatching rule independently.

Single machine scheduling , Dispatching , Intelligent agent , Job shop scheduling , Simulated annealing , Machine intelligence , Scheduling algorithm , Manufacturing , Routing , Job production systems

# Sensitivity based Growing and Pruning method for RBF network in online learning environments

How to define the architecture of classifiers dynamically is one of the major research topics in online learning. This paper presents a new online learning algorithm for Radial Basis Function Network named Sensitivity Based Neurons Growing and Pruning Method for RBF network (SBGAP). The performance of SBGAP is evaluated experimentally by comparing accuracy and the number of neurons with the existing methods. The experimental results show that SBGAP achieve litter higher accuracy with fewer hidden units in most situations.

Neurons , Accuracy , Sensitivity , Training , Radial basis function networks , Heart , Machine learning

# A New Fuzzy Multicategory Support Vector Machines Classifier

This paper proposes a new fuzzy multicategory support vector machines (FMSVM) classifier. The main idea is that the proposed FMSVM uses knowledge of the ambiguity associated with the membership of samples for a given class and the relative location of samples to the origin. Compared with the existing SVMs, the new proposed FMSVM that uses the L2-norm in the objective function has the improvement in aspects of classification accuracy and reducing the effects of noises and outliers.

Support vector machines , Support vector machine classification , Fuzzy sets , Machine learning , Noise reduction , Cybernetics , Mathematics , Computer science , Research and development , Electronic mail

# Improved Fuzzy Multicategory Support Vector Machines Classifier

This paper investigates an improved fuzzy multicategory support vector machines classifier (IFMSVM). It uses knowledge of the ambiguity associated with the membership of data samples of a given class and relative location to the origin, to improve classification performance with high generalization capability. In some aspects, classifying accuracy of the new algorithm is better than that of the classical support vector classification algorithms. Numerical simulations show the feasibility and effectiveness of this algorithm

Support vector machines , Support vector machine classification , Testing , Machine learning , Mathematics , Computer science , Electronic mail , Classification algorithms , Numerical simulation , Quadratic programming

# A survey on active learning strategy

Active learning is a hot topic in machine learning field. The main task of active learning is to automatically select the representative instances for efficiently reducing the sample complexity. This paper presents a brief survey of active learning regarding selection methods, query strategies, applications and other related works.

Machine learning , Training , Classification algorithms , Uncertainty , Learning systems , Complexity theory , Support vector machines

# A new nu-support vector machine for training sets with duplicate samples

Analyzed theoretically, /spl nu/-SVM was found to be over-dependent on each training sample, even if the samples have same value. This dependence would result in more time for training, more support vectors and more decision time. In order to overcome this problem, we propose a new /spl nu/-SVM. This new /spl nu/-SVM multiplies each slack variable in the objective function by a weight factor, and automatically computes each weight factor by the number of corresponding samples with same value before training. Theoretical analysis and the results of experiments show that the new /spl nu/-SVM has the same classification precision rate as the standard /spl nu/-SVM and the new /spl nu/-SVM is faster than the /spl nu/-SVM in training and decision if the training sets have same value samples.

Support vector machines , Support vector machine classification , Machine learning , Chemical analysis , Information technology , Petroleum , Chemical technology , Information analysis , Character recognition , Speech recognition

# Restricted Boltzmann Machine for Nonlinear System Modeling

In this paper, we use a deep learning method, restricted Boltzmann machine, for nonlinear system identification. The neural model has deep architecture and is generated by a random search method. The initial weights of this deep neural model are obtained from the restricted Boltzmann machines. To identify nonlinear systems, we propose special unsupervised learning methods with input data. The normal supervised learning is used to train the weights with the output data. The modified algorithm is validated by modeling two benchmark systems.

Nonlinear systems , Machine learning , Probability distribution , Training , Search methods , Unsupervised learning , Benchmark testing

# Application research of support vector machine in E-Learning for personality

In order to accurately build the learner's learning style in E-Learning, according to the needs and preferences to provide personalized learning materials and harmonious human-computer interaction environment. This paper combines Felder-Silverman learning style with support vector machine technology, and use machine learning technologies for learners to build dynamic learning style. Through the analysis of the Emotion and recognition interaction of the personalized E-Learning based on statistical learning theory and support vector machine technology, it demonstrates the correctness and feasibility using support vector machine to build learning styles. The combination of support vector machine, emotion and recognition interaction in the personalized E-Learning makes great contribution to build human-computer interaction environment.

Machine learning , Support vector machines , Electronic learning , Training , Educational institutions , Materials , Brain modeling

# From bits to information with learning machines: theory and applications

Summary form only given. Learning is becoming the central problem in trying to understand intelligence and in trying to develop intelligent machines. The paper outlines some previous efforts in developing machines that learn. It sketches the authors's work on statistical learning theory and theoretical results on the problem of classification and function approximation that connect regularization theory and support vector machines. The main application focus is classification (and regression) in various domains-such as sound, text, video and bioinformatics. In particular, the paper describe the evolution of a trainable object detection system for classifying objects-such as faces and people and cars-in complex cluttered images. Finally, it speculates on the implications of this research for how the brain works and review some data which provide a glimpse of how 3D objects are represented in the visual cortex.

Machine learning , Learning systems , Statistical learning , Function approximation , Support vector machines , Support vector machine classification , Focusing , Bioinformatics , Object detection , Face detection

# Machine learning for automatic acquisition of Chinese linguistic ontology knowledge

Due to the complexity and flexibility of natural language, automatic linguistic knowledge acquisition and its application research becomes difficult. In this paper, we present a machine learning method to automatically acquire Chinese linguistic ontology knowledge from typical corpus. This study, first, defined the description frame of Chinese linguistic ontology knowledge, and then, automatically acquired the usage of a Chinese word with its co-occurrence of context in using semantic, pragmatics, syntactic, etc from the corpus, final, the above information and their representation act as Chinese linguistic ontology knowledge bank. We completed two groups of experiments, i.e. documents similarity computing, text reordering for information retrieval. Compared with previous works, the proposed method solves the inferior precision of nature language processing.

Machine learning , Ontologies , Natural languages , Natural language processing , Knowledge acquisition , Tagging , Learning systems , Information retrieval , Computer science , Knowledge engineering

# Integrated access to big data polystores through a knowledge-driven framework

The recent successes of commercial cognitive and AI applications have cast a spotlight on knowledge graphs and the benefits of consuming structured semantic data. Today, knowledge graphs are ubiquitous to the extent that organizations often view them as a “single source of truth” for all of their data and other digital artifacts. In most organizations, however, Big Data comes in many different forms including time series, images, and unstructured text, which often are not suitable for efficient storage within a knowledge graph. This paper presents the Semantics Toolkit (SemTK), a framework that enables access to polyglot-persistent Big Data stores while giving the appearance that all data is fully captured within a knowledge graph. SemTK allows data to be stored across multiple storage platforms (e.g., Big Data stores such as Hadoop, graph databases, and semantic triple stores) - with the best-suited platform adopted for each data type - while maintaining a single logical interface and point of access, thereby giving users a knowledge-driven veneer across their data. We describe the ease of use and benefits of constructing and querying polystore knowledge graphs with SemTK via four industrial use cases at GE.

Semantics , Big Data , Data models , Ontologies , Time series analysis , Databases , Triples (Data structure)

# Web-based collaborative big data analytics on big data as a service platform

As data has been increasing explosively due to development of social networks and cloud computing, there has been a new challenge for storing, processing, and analyzing a large volume of data. The traditional technologies do not become a proper solution to process big data so that a big data platform has begun to emerge. It is certain that big data platform helps users develop analysis service effectively. However, it still takes a long time to collect data, develop algorithms and analytics services. We present a collaborative big data analytics platform for big data as a service. Developers can collaborate with each other on the platform by sharing data, algorithms, and services. Therefore, this paper describes big data analytics platform that effectively supports to manage big data and develop analytics algorithms and services, collaborating with data owners, data scientists, and service developers on the Web. Finally, we introduce a CCTV metadata analytics service developed on the platform.

Big data , Portals , Collaboration , Web services , Monitoring , Streaming media

# A timeline visualization system for road traffic big data

The rapid converging of big data and IoT (Internet of Things) technologies provides more opportunities in the area of road traffic applications. In this paper, we discuss a timeline visualization tool which enables us to better understand of traffic behaviors from road traffic big data.

Roads , Big data , Data visualization , Data mining , Conferences , Cloud computing , Image color analysis

# Towards a big data requirements engineering artefact model in the context of big data software development projects: Poster extended abstract

In this paper, we describe our ongoing research aimed at defining a Requirements Engineering Artefact Model (REAM) in the context of Big Data software applications. This model aims to provide a “big picture” of the Requirements Engineering work products created and used in Big Data software development projects. REAM are important tools that can be used as references for the definition of domain-specific RE models, system life-cycle processes and artefact-centered processes, currently bereft in the Big Data Software Engineering research.

Big Data , Software , Requirements engineering , Data models , Computational modeling , Context modeling , Software engineering

# Sharing best practices for the implementation of Big Data applications in government and science communities

The Federal Big Data Working Group supports the Federal Big Data Initiative but is not endorsed by the Federal Government or its agencies. This working group uses meetups with onsite and virtual participation to share best practices for the implementation of Big Data applications in government and science communities. Decision-makers and the scientific community interact with data science in order to take advantage of the Big Data transformation of how information is used in science, decision support, data discovery and data publishing. The working group federates use cases, data publications, solutions and technologies. The range of topics is illustrated in a keynote and panel discussion at a recent Big Data conference and in a summary of recent working group meetups.

Big data , Government , Semantics , Communities , Conferences , Best practices , Data privacy

# Design of a scalable data stream channel for big data processing

This paper outlines big data infrastructure for processing data streams. Our project is distributed stream computing platform that provides cost-effective and large-scale big data services by developing data stream management system. This research contributes to advancing feasibility of big data processing for distributed, real-time computation even when they are overloaded.

Big data , Ports (Computers) , Data models , Twitter , Sparks , Synchronization

# Gradient-based signatures for big multimedia data

With the continuous increase of heterogeneous multimedia data, the question of how to access big multimedia data efficiently has become of crucial importance. In order to provide fast access to complex multimedia data, we propose to approximate content-based features of multimedia objects by means of generative models. The proposed gradient-based signatures epitomize a high quality content-based approximation of multimedia objects and facilitate efficient indexing and query processing at large scale.

Multimedia communication , Multimedia databases , Adaptation models , Data models , Indexing , Big data , Query processing

# Data confidentiality challenges in big data applications

In this paper, we address the problem of data confidentiality in big data analytics. In many fields, much useful patterns can be extracted by applying machine learning techniques to big data. However, data confidentiality must be protected. In many scenarios, data confidentiality could well be a prerequisite for data to be shared. We present a scheme to provide provable secure data confidentiality and discuss various techniques to optimize performance of such a system.

Protocols , Big data , Encryption , Parallel processing , Computational efficiency , Software

# Some key problems of data management in army data engineering based on big data

This paper analyzed the challenges of data management in army data engineering, such as big data volume, data heterogeneous, high rate of data generation and update, high time requirement of data processing, and widely separated data sources. We discussed the disadvantages of traditional data management technologies to deal with these problems. We also highlighted the key problems of data management in army data engineering including data integration, data analysis, representation of data analysis results, and evaluation of data quality.

Data engineering , Data analysis , Data integration , Big Data , Distributed databases , Data models , Uncertainty

# Linked Open Data mining for democratization of big data

Data is everywhere, and non-expert users must be able to exploit it in order to extract knowledge, get insights and make well-informed decisions. The value of the discovered knowledge from big data could be of greater value if it is available for later consumption and reusing. In this paper, we present an infrastructure that allows non-expert users to (i) apply user-friendly data mining techniques on big data sources, and (ii) share results as Linked Open Data (LOD). The main contribution of this paper is an approach for democratizing big data through reusing the knowledge gained from data mining processes after being semantically annotated as LOD, then obtaining Linked Open Knowledge. Our work is based on a model-driven viewpoint in order to easily deal with the wide diversity of open data formats.

Data mining , Big data , Resource description framework , Data models , Knowledge based systems , Data analysis , Proposals

# Pedigree-ing Your Big Data: Data-Driven Big Data Privacy in Distributed Environments

This paper introduces a general framework for supporting data-driven privacy-preserving big data management in distributed environments, such as emerging Cloud settings. The proposed framework can be viewed as an alternative to classical approaches where the privacy of big data is ensured via security-inspired protocols that check several (protocol) layers in order to achieve the desired privacy. Unfortunately, this injects considerable computational overheads in the overall process, thus introducing relevant challenges to be considered. Our approach instead tries to recognize the "pedigree" of suitable summary data representatives computed on top of the target big data repositories, hence avoiding computational overheads due to protocol checking. We also provide a relevant realization of the framework above, the so-called Data-dRIven aggregate-PROvenance privacypreserving big Multidimensional data (DRIPROM) framework, which specifically considers multidimensional data as the case of interest.

Big Data , Data privacy , Distributed databases , Proposals , Cloud computing , Protocols , Data models

# Data science vs big data @ UTM big data centre

Big data tsunami has hit Malaysia recently that has awakening the industry and academy communities to aggressively address the insight, hindsight and foresight challenges ensuring Malaysia to be among the top world players in big data information economy for the next decade. Rapid development of Information and Communication Technology (ICT) in this era is very significant due to increasing number of users accessing data keeps growing by the time. This phenomenon has been coined as big data. What is Big data??? We address big data as assets that needs unique platform to deal with bizarre behavior of datasets whose size is beyond the ability of typical data storage to manage, mine and analyze accordingly. This bizarre behavior requires three main personalities: volume, velocity, and variety that basically need new architecture, techniques, algorithms, and analytics to uncover the golden and hidden knowledge from information obesity. From these perspectives, we demonstrate our experiences in setting up our Data Science/Big Data platform, algorithms and tool to align with big data plug and play within the academic environment as well as our services to the community and industries.

Big data , Graphics processing units , Memory , Companies , Information technology

# Big Data Pre-Processing: Closing the Data Quality Enforcement Loop

In the Big Data Era, data is the core for any governmental, institutional, and private organization. Efforts were geared towards extracting highly valuable insights that cannot happen if data is of poor quality. Therefore, data quality (DQ) is considered as a key element in Big data processing phase. In this stage, low quality data is not penetrated to the Big Data value chain. This paper, addresses the data quality rules discovery (DQR) after the evaluation of quality and prior to Big Data pre-processing. We propose a DQR discovery model to enhance and accurately target the pre-processing activities based on quality requirements. We defined, a set of pre-processing activities associated with data quality dimensions (DQD's) to automatize the DQR generation process. Rules optimization are applied on validated rules to avoid multi-passes pre-processing activities and eliminates duplicate rules. Conducted experiments showed an increased quality scores after applying the discovered and optimized DQR's on data.

Big Data , Optimization , Data models , Quality assessment

# A big data analytics framework for forecasting rare customer complaints: A use case of predicting MA members' complaints to CMS

Centers for Medicare & Medicaid Services (CMS) publishes Medicare Part C Star Ratings each year to measure the quality of care of Medicare Advantage (MA) contracts. One of the key measures is Complaints about the Health Plan, which is captured in Complaints Tracking Module (CTM). Complaints resulted in CTM are rare events: for MA contracts with 2-5 star ratings, number of complaints for every 1,000 members range from .10 to 1.84 over last 5 years. Reducing number of complaints is extremely important to MA plans as they impact CMS reimbursements to MA plans. Forecasting and reducing complaints is an extremely technically challenging task, and involves ethics considerations in patients' rights and privacy. In this research, we constructed a big data analytics framework for forecasting rare customer complaints. First, we built a big data ingestion pipelines on a Hadoop platform: a) Ingest MA plan's customer complaints data from CTM from past 3 years. b) Ingest health plan's call center data for MA members from past 3 years, including both structured data and unstructured text script for the calls. c) Ingest MA members' medical claims, including members' demographics and enrollment history. d) Ingest MA members' pharmacy claims. e) Integrate and unified data from above sources, and enrich the data with additional engineered features into a big wide table, one row per member for analysis and modeling. Second, we designed a unique decision tree based Large Ensemble with Over-Sampling (LEOS) algorithm, which mimics random forest but with extreme oversampling of target class to increase bias, and leverages the parallel computing of Hadoop clusters by generating thousands of fixed size training data sets, and for each such dataset training a decision trees with similar fixed tree structure, and ensemble them. Third, we validated our framework and LEOS learning algorithm with real data, and also discussed ethics issues we encountered in handling data and applying findings from research.

Decision trees , Training , Training data , Big Data , Prediction algorithms , Ethics , Contracts

# Big data gathering and mining pipelines for CRM using open-source

Customer Relationship Management (CRM) is currently the fastest growing sector of enterprise software, estimated to increase to $36.5B worldwide by 2017. CRM technologies increasingly use data mining primitives across multiple applications. At the same time, the growth of big data has led to the evolution of an open source big data software stack (primarily powered by Apache software) that rivals traditional enterprise database (RDBMS) stacks. New technologies such as Kafka, Storm, HBase have significantly enriched this open source stack, alongside more established technologies such as Hadoop MapReduce and Mahout. Today, enterprises have a choice to make regarding which stack they will choose to power their big data applications. However, there are no published studies in literature on enterprise big data pipelines built using open source components supporting CRM. Specific questions that enterprises have include: how is the data processed and analyzed in such pipelines? What are the building blocks of such pipelines? How long does each step of this processing take? In this work, we answer these questions for a large scale (serving over a 100M customers) industrial CRM pipeline that incorporates data mining, and serves several applications. Our pipeline has, broadly, two parts. The first is a data gathering part that uses Kafka, Storm, and HBase. The second is a data mining part that uses Mahout and Hadoop MapReduce. We also provide timings for common tasks in the second part such as data preprocessing for machine learning, clustering, reservoir sampling, and frequent itemset extraction.

Big data , Pipelines , Customer relationship management , Data mining , Benchmark testing , Storms , Software

# Digital construction of coal mine big data for different platforms based on life cycle

Big data has penetrated into various industries and business functions, and become important factors of production in the global economy. In the big data technology system, big data collection is the basis. The storage, analysis, integration and visualization of unstructured data and semi-structured data will become an important focus of the big data innovation. Traditional structured data will no longer be the core of big data. Based on the life cycle theory, using new digital technology, such as acquisition, processing, storage, organization and copyright protection, clusters of high concurrency retrieval and dynamic scheduling, intelligent digital display, coal mine industry information data can be collected and integrated, to realize centralized management, unified retrieval and joint exhibition of information resources, to provide technical means and reference for the digital construction of heterogeneous coal mine information data by means of big data thinking.

Coal , Information services , Metadata , Big Data , Distributed databases , Copyright protection , Coal mining

# A novel big-data processing framwork for healthcare applications: Big-data-healthcare-in-a-box

Herein we present a novel big-data framework for healthcare applications. Healthcare data is well suited for bigdata processing and analytics because of the variety, veracity and volume of these types of data. In recent times, many areas within healthcare have been identified that can directly benefit from such treatment. However, setting up these types of architecture is not trivial. We present a novel approach of building a big-data framework that can be adapted to various healthcare applications with relative use, making this a one-stop “Big-Data-Healthcare-in-a-Box”.

Medical services , Sparks , Big data , Java , Context , Computer architecture , Business

# Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory

Currently, on-line monitoring and measuring system of power quality has accumulated a huge amount of data. In the age of big data, those data integrated from various systems will face big data application problems. This paper proposes a data quality assessment system method for on-line monitoring and measuring system of power quality based on big data and data provenance to assess integrity, redundancy, accuracy, timeliness, intelligence and consistency of data set and single data. Specific assessment rule which conforms to the situation of on-line monitoring and measuring system of power quality will be devised to found data quality problems. Thus it will provide strong data support for big data application of power quality.

Data integrity , Power quality , Monitoring , Power measurement , Redundancy , Big Data , Business

# Research on analysis system of city price based on big data

This paper attempts to construct an analysis model of city price, by combining the big data system with the information of city price, so as to provide the reference for the government to implement the policy of the accurate price control. This paper considers that the analysis system of city price should include decision-making layer, supporting layer and showing layer on the conceptual level, and its transmission path should include the data collection, data management, data mining, decision making and safety protection. when it comes to the construction of the subsystem, technological tools such as data mining, cloud computing and visualization should be used, mainly in order to build data acquisition subsystem, data management subsystem, data analysis subsystem, data transmission subsystem and so on, and provide the graphic description for the corresponding technological path at the same time.

Urban areas , Big data , Government , Decision making , Data models , Data mining , Market research

# An OWL Ontology for Supporting Semantic Services in Big Data Platforms

In the last years, there was a growing interest in the use of Big Data models to support advanced data analysis functionalities. Many companies and organizations lack IT expertise and adequate budget to have benefits from them. In order to fill this gap, a model-based approach for Big Data Analytics-as-a-service (MBDAaaS) can be used. The proposed model, composed by declarative, procedural and deployment (sub) models, can be used to select a deployable set of services based on a set of user preferences shaping a Big Data Campaign (BDC). The deployment of a BDC requires that the selection of services has to be carried out on the basis of coherent and non conflictual user preferences. In this paper we propose an OWL ontology in order to solve this issue.

Ontologies , OWL , Data models , Big Data , Semantics , Analytical models

# Big data analysis of youth tobacco smoking trends in the United States

As large amounts of data are now generated in the healthcare industry, big data technologies are used to process these data. Tobacco smoking has been significant among the youth in the United States. In this research, we use big data techniques such as R and Tableau to explore the tobacco smoking trends among the youth in the United States. Results indicate that there is more number of youth male smokers than youth female smokers. The results also indicate that more than 51 percent of current youth smokers want to quit smoking.

Big Data , Market research , Data visualization , Surgery , Cardiovascular diseases

# Meta-analysis of big data security and privacy: Scholarly literature gaps

We collected 79,012 articles from 1916-2016 related to big data to determine which topics were being studied and how much of the literature was focused on privacy or security-related keywords. The analysis demonstrated that the big data paradigm commenced in late 2011 and the research production exponentially rose starting in 2012, which approximated a Weibull distribution that captured 82% of the variance (p<;.01). We found there were 13 dominant topics capturing 49% of the big data production in journals during 2011-2016 but privacy and security topics accounted for only 2% and this trend recently dropped to less than 1% Thus, we argued that we need to stimulate more big data privacy-security research.

Big data , Data privacy , Security , Production , Market research , Databases

# Where big data meets linked data: Applying standard data models to environmental data streams

In August 2015, a new seafloor observatory was deployed in Galway Bay, Ireland. The sensors on the observatory platform are connected by fibre-optic cable to a shore station, where a broadband connection allows data transfer to the Marine Institute's data centre. This setup involved the development of a new data acquisition system which takes advantage of open source streaming data solutions developed in response to the Big Data paradigm, in particular the Velocity aspect. This activity merges concepts from the arenas of both Big Data and Internet of Things where data standardisation is not normally considered. This paper considers the architecture implemented to stream marine data from instrument to end user and offers suggestions on how to standardise these data streams.

Instruments , Big data , Oceans , Observatories , Vocabulary , Standards , Sensors

# An iterative methodology for big data management, analysis and visualization

Big Data constitutes an opportunity for companies to empower their analysis. However, at the moment there is no standard way for approaching Big Data projects. This, coupled with the complex nature of Big Data, is the cause that many Big Data projects fail or rarely obtain the expected return of investment. In this paper, we present a methodology to tackle Big Data projects in a systematic way, avoiding the aforementioned problems. To this end, we review the state of the art, identifying the most prominent problems surrounding Big Data projects, best practices and methods. Then, we define a methodology describing step by step how these techniques could be applied and combined in order to tackle the problems identified and increase the success rate of Big Data projects.

Big data , Iterative methods , Data visualization , Data models , Systematics , Best practices , Data mining

# Big data analytics platform for flight safety monitoring

The conventional methods of data analytics for flight safety monitoring have met many bottlenecks. This paper analyzes the insufficiencies of the preliminary business process of an airline. For the purpose of meeting requirements of efficiency and accuracy and avoiding the drawbacks encountered before, the architectural framework of the flight safety monitoring platform utilizing big data technology is proposed and demonstrated by the function module structure and logical structure. The platform is implemented by dividing the system into five subsystems, namely data acquisition, data decoding, data storage, data analysis and visualization.

Safety , Business , Monitoring , Data visualization , Big Data , Data analysis , Decoding