IRINA NEAGA & YUQIUGE HAO

TOWARDS BIG DATA MINING AND DISCOVERY

Introduction

We live in an era of big data that has embedded a huge potential and increased information complexity and risks such as insecurity as well as information overload and irrelevance. Also business intelligence and analytics are important in dealing with the magnitude and impact of data driven problems and solutions in the contemporary society and economy. Analysts, computer scientists, economists, mathematicians, political scientists, sociologists, and other scholars are clamouring for access to the massive quantities of data in order to extract meaningful information and knowledge. Very large data sets are generated by and about organisations, people, and their collaboration and interactions in the digital ecosystems and physical spaces. Diverse groups argue about the potential benefits, limitations, and risks of accessing and analysing huge amounts of data such as financial data, genetic sequences, social media interactions, medical records, phone/email logs, government records, and other digital traces generated by people and organisations.

With the development of internet communication and collaboration, data is playing a central and crucial role. Lots of data intensive applications occur in recent years such as the Google+, Twitter, LinkedIn and Facebook etc. All these data intensive driven applications generate and process massive data usually stored in the cloud.

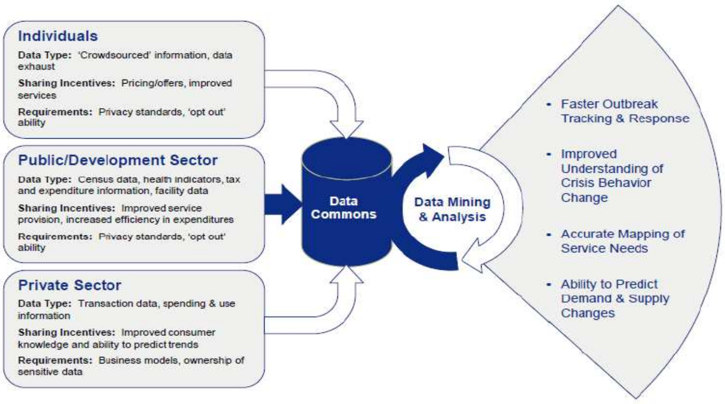
Individuals
Data Type: 'Crowdsourced' information, data exhaust
Sharing Incentives: Pricing/offers, improved services
Requirements: Privacy standards, 'opt out' ability
Public/Development Sector
Data Type: Census data, health indicators, tax and expenditure information, facility data
Sharing Incentives: Improved service provision, increased efficiency in expenditures
Requirements: Privacy standards, 'opt out' ability
Private Sector
Data Type: Transaction data, spending & use information
Sharing Incentives: Improved consumer knowledge and ability to predict trends
Requirements: Business models, ownership of sensitive data
Data Commons
Data Mining & Analysis
Faster Outbreak Tracking & Response
Improved Understanding of Crisis Behavior Change
Accurate Mapping of Service Needs
Ability to Predict Demand & Supply Changes

Figure 1 – Complex Data Infrastructure/Ecosystem (source: World Economic Forum, 2012)

Big data could be very beneficial to resolve critical issues providing the potential of new insights for the advancements of medical especially cancer research, global security, discovering and predicting terrorism activities, and dealing with socio-economic and environmental issues.

Big data could be interpreted as a complex data infrastructure and new powerful data technologies and management solutions are needed and will be directed to improve the decision making processes and forecasting through application of advanced data exploratory studies, data mining, predictive analytics and knowledge discovery as presented in figure 1.

The main key characteristics that define big data are volume, velocity, variety and value. Veracity could be also considered an additional characteristic. The related big data models are presented in figure 2.

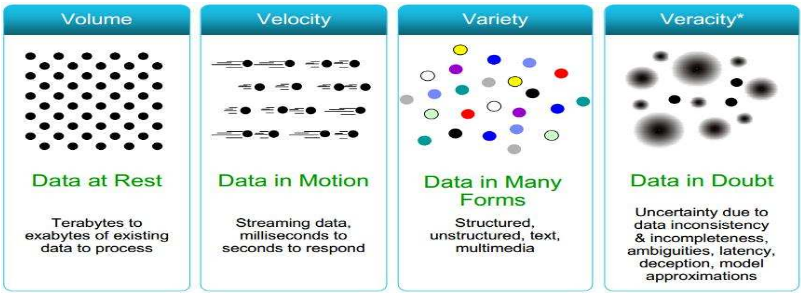
Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia
Veracity*
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Figure 2 – Big Data Characteristics based on the source: McKinsey Global Institute, 2011

On the other hand because of the characteristics of the cloud, it is an enabler of big data acquisition, and software processing strategies. Based on Gartner's estimation, 50% of data will be stored on the cloud by 2016 (Schouten, 2012). However in the reality, cloud has not been widely used for data analytics especially in practical applications.

The availability of cloud based solutions has dramatically lowered the cost of storage, amplified by the use of commodity hardware even on a "pay as-you-go" basis that is directed to effectively and timely processing large data sets. The big data could be delivered in form of "as-a-service". Google BigQuery [https://cloud.google.com/products/big-query] is only one example of applying big data solutions in a cloud based platform.

In cloud computing, data and software applications are defined, developed and implemented as services. These services have defined a multi-layered infrastructure and are described as follows (Grace, 2010; Mell and Grance, 2009):

1. *Software as a Service* (SaaS), where applications are hosted and delivered online via a web browser offering traditional desktop functionality
2. *Platform as a Service* (PaaS), where the cloud provides the software platform for systems (as opposed to just software)
3. *Infrastructure as a Service* (IaaS), where a set of virtual computing resources, such as storage and computing capacity, are hosted in the cloud; customers deploy and run only their own applications for obtaining the needed services.

On the other hand it is also recognised the tension between big data approaches, and solutions versus information security and data privacy requirements. The big data might enable the violation of the privacy and information security breaches and by consequence decreasing the trust in data defined as a service in the cloud. Big data stored and processed in the cloud could lack a centralized control and ownership.

According to McKinsey & Co (2011) Big Data is seen as "the next frontier for innovation, competition and productivity" and as such the related applications will contribute to economic growth. The positive impacts of big data provide a huge potential for for organisations. In order to achieve these aspirations several issues should be analysed and discussed in the context of complex systems and using systems approaches such as holistic thinking and system dynamics.

Therefore major issues are emerging and this work-in-progress attempts to discuss a few key aspects directed to the development and adopting data mining techniques and strategies for big data.

Background and Research Approach

Demirkan and Delen (2013) have defined some research directions including dealing with affordable analytics for big data. This means using open-source, free-of-charge data/text mining algorithms and associated commercial tools (e.g. R, RapidMiner, Weka, Gate, etc.) New approaches need to provide solutions for moving these tools to the cloud and produce efficient and affordable applications for discovering knowledge and patterns from very large/big data sets directed to support business intelligence and decision support systems applications.

The principles of data/information-as-a-service, data/information-security-as-a-service, and analytics-as-a-service are explained in the context of using service oriented architecture.

However the cloud platforms are not completely following service oriented thinking and even more there is a debate that cloud computing is different of service oriented architectures, and grid computing.

The main motivation of adopting cloud computing for analytics applied for large (big) data sets could be because cloud solutions are accessible outside the a web based organisation communication secured with firewalls. Cloud based business analytics are also cost effective, easy to set up and test. The results are easy to be shared outside the organisations. Greg Sheldon, CIO of Elite Brands said "The biggest benefit, is to be able to access a large amounts of information from anywhere you have web access, specifically on an iPad. This is beneficial to our field sales team when information is needed on the fly." (Fields, 2013:2)

The main research questions are related but not limited to the following aspects:

1. In the context of big data and cloud computing how analytics (e.g. data mining), information and knowledge management disciplines and approaches will evolve?
2. What should be the techniques, strategies and practices to increase the benefits and minimise the big data risks?
3. The potential to reduce the growing number of security breaches and cyber-security risks and increase organisational awareness, business agility and resilience.
4. The existing legislation such as data protection law, regulations and standards how should evolve. Moreover, the ethics issues will be considered.

Efforts and Challenges of Big Data Mining and Discovery

Considering big data a collection of complex and large data sets that are difficult to process and mine for patterns and knowledge using traditional database management tools or data processing and mining systems a briefing of the existing efforts and challenges is provided in this paragraph. While presently the term big data literally concerns about data volumes, Wu et al. (2013) have introduce HACE theorem that described the key characteristics of the big data as (1) huge with **h**eterogeneous and diverse data sources, (2) **a**utonomous with distributed and decentralized control, and (3) **c**omplex and **e**volving in data and knowledge associations. Generally, business intelligence applications are using data analytics that are grounded mostly in data mining and statistical methods and techniques. These strategies are usually based on the mature commercial software systems of RDBMS, data warehousing, OLAP, and BPM. Since the late 1980s, various data mining algorithms have been developed mainly within the artificial intelligence, and database communities. In the IEEE 2006 International Conference on Data Mining (ICDM), the 10 most influential data mining algorithms were identified based on expert nominations, citation counts, and a community survey (Chen et al, 2012). In ranked order, these techniques are as follows C4.5, k-means, SVM (support vector machine), Apriori, EM (expectation maximization), PageRank, AdaBoost, kNN (k-nearest neighbors), Naïve Bayes, and CART (Wu et al, 2007). These algorithms are for classification, clustering, regression, association rules, and network analysis. Most of these well known data mining algorithms have been implemented and deployed in commercial and open source data mining systems (Witten et al. 2011).

Chen at al. (2012) have compared data base management systems and analytics as well as ETL with using MapReduce and Hadoop. Hadoop was originally a (distributed) file system approach applying the MapReduce framework that is a software approach introduced by Google in 2004 to support distributed computing on large/big data sets. Recently, Hadoop has been developed and used as a complex ecosystem that includes a wider range of software systems, such as HBase (a distributed table store), Zookeeper (a reliable coordination service), and the Pig and Hive high-level languages that compile down MapReduce components (Rabkin and Katz, 2013). Therefore in the recent conceptual approaches Hadoop is primarly considered an ecosystem or an infrastructure or a framework and not just the file system alongside MapReduce components.

The big data and cloud computing frameworks include the Google MapReduce, Hadoop Reduce, Twister, Hadoop++, Haloop, and Spark etc. which are used to process big data and run computational tasks. The cloud databases are used to store massive structured and semi-structured data generated from different types of applications. The most important cloud databases include the BigTable, Hbase, and HadoopDB. In order to implement an efficient big data mining and analysis framework, the data warehouse processing is also important. The most important data warehouse processing technologies include the Pig, Hive etc.

Strambei (2012) suggests a different conceptual interpretation of the OLAP technology considering the emergence of web services, cloud computing and big data. One of the most important consequences could be widely open access to web analytical technologies. The related approach has evaluated the OLAP Web Services viability in the context of the cloud based architectures.

There are also a few reported practical applications of big data mining in the cloud. Patel et al. (2012) have explored a practical solution to big data problem using the Hadoop data cluster, Hadoop Distributed File System alongside Map Reduce framework, and a big data prototype application scenario. The results obtained from various experiments indicate promising results to address big data problem.

The challenges for moving beyond existing data mining and knowledge discovery techniques (NESSI, 2012, Witten et al, 2011) are defined as follows:

1. a solid scientific foundation to be able to select an adequate analytical method and a software design solution
2. new algorithms (and demonstrate the efficiency and scalability, etc.) and machine learning techniques
3. the motivation of using cloud architecture for big data solutions and how to achieve the best performance of implementing data analytics using cloud platform (e.g. big data as a service)
4. dealing with data protection and privacy in the context of exploratory or predictive analysis of big data
5. software platforms and architectures alongside adequate knowledge and development skills to be able to implement them
6. a genuine ability to understand not only the data structures (and the usability for a given processing method), but also the information and business value that is extracted from big data.

Concluding Remarks

The big data movement has energized the data mining, knowledge discovery in data bases and associated software development communities, and it has introduced complex, interesting questions for researchers and practitioners. As organizations continue to increase the amount and values of collected data formalizing the process of big data analysis and analytics becomes overwhelming. In this paper, we discuss some existing approaches and have analysed the main research issues of big data mining, knowledge, and patterns discovery in a data intensive cloud computing environment. This research will be progressed providing theoretical and practical approaches that will be tested through the development of a case study for the application of big data.

Correspondence

Dr Irina Neaga, School of Management, Plymouth Business School, Plymouth, Devon, United Kingdom. E-mail: irina.neaga@plymouth.ac.uk

Yuqiuge Hao, Industrial Management Unit, Department of Production, University of Vaasa, Vaasa, Finland. E-mail:yuqiuge.hao@uwasa.fi

Authors' brief bios

Dr. Irina Neaga is a lecturer researching on data, information, and knowledge based approaches, systems and strategies applied in international collaborative logistics, decisions and operations within the School of Management (Plymouth Business School) at Plymouth University, Plymouth, United Kingdom. Prior to this she has worked for several years as a researcher at Loughborough University, Leicestershire, UK where she has contributed to the European funded research and large interdisciplinary projects funded by RCUK, Technology Strategy Board and industry. She has substantial working experience in academia, and industry on large-scale, strategic partnerships of systems engineering and innovation through research, education, knowledge transfer and consultancy. She has worked in higher education, manufacturing industry, and research consortia in Finland, Canada, The Netherlands and Romania. Her research portfolio includes IT as a Utility, cloud computing adoption and information security management in global supply chains.

Yuqiuge Hao is a doctoral student within the Department of Production from Vaasa University, Vaasa, Finland, primarily researching on the application of big data in manufacturing, and cloud ERP systems. She contributes to the project **Ad**aptive **V**irtual **ENT**erprise Manufac**TUR**ing **E**nvironment (Adventure) which is a Small or Medium-Scale Focused Research Project (STREP) funded by the European Seventh Framework Programme in Virtual Factories and Enterprises. She received a Master degree in Computer Science from the University of Stockholm, Stockholm, Sweden.

References

Chen, H., Chaing, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly, 36, 4, pp. 1165-1188

Chong, R.F. (2012) Changing the World: Big Data and the Cloud, The Atlantic. Available at: http://www.theatlantic.com/sponsored/ibm-cloud-rescue/archive/2012/09/changing-the-world-big-data-and-the-cloud/262065/ [Accessed by: 20 May, 2013].

Demirkan, H. and Delen, D. (2013) Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud, Decision Support Systems, 55, 1, pp. 412–421.

European Commission (2010) The Future of Cloud Computing - Opportunities for European Cloud Beyond 2010, European Commission Public Report.

Fields, E. (2013) Why Business Analytics in the Cloud?,Tableau Software White Paper.

Grace, L. (2012) Basics about Cloud Computing, Software Engineering Institute, Carnegie Mellon University, USA at: http://www.sei.cmu.edu/library/assets/whitepapers/ Cloudcomputingbasics.pdf.

Mell, P. and Grance, T. (2009) The NIST definition of cloud computing v15. Version 15 available at http://csrc.nist.gov/groups/SNS/cloud-computing/cloud-def -v15.doc

McKinsey Global Institute (2011) Big Data: The next frontier for innovation, competition and productivity.

NESSI (2012) Big Data – A New World of Opportunities, White Paper.

Rabkin, A. and Katz, R.H. (2013) How Hadoop Clusters Break, IEEE Software, Feature: Big Data, pp. 88-94

Patel, A.B., Birla, M. and Nair, U. (2012) Addressing Big Data Problem Using Hadoop and Map Reduce, NIRMA University Conference on Engineering, NUi-CONE, pp. 1-5.

Rittinghouse, J.W. and Ransome, J.F. (2010) Cloud Computing Implementation, Management and Security, CRC Press Taylor and Francis.

Schouten, E., (2012) Big Data 'as a Service'. The Atlantic. Available at:
http://www.theatlantic.com/sponsored/ibm-cloud-rescue/archive/2012/09/
big-data-as-a-service/262461/ [Accessed by: 20 May, 2013]

Strambei, C. (2012) OLAP Services on Cloud Architecture, Journal of Software & Systems Development, IBIMA Publishing.

Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: Practical Machine Learning Tools and Techniques. 3rdedtion. Morgan Kaufamann series in data management systems.

World Economic Forum (2012). Big Data, Big Impact: New Possibilities for International Development. World Economic Forum Report.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H.,McLachlan, G. J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 Algorithms in Data Mining, Knowledge and Information Systems, 14,1, pp. 1-37.

Wu, X. , Zhu, X., Wu, G., Ding, W. (2013) Data Mining with Big Data, Knowledge and Data Engineering, IEEE Transactions, in press World Economic Forum (2012). Big Data, Big Impact: New Possibilities for International Development. World Economic Forum Report.