ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of
DEVELOPMENT RESEARCH

International Journal of Development Research
Vol. 06, Issue, 08, pp.8842-8846, August, 2016

Full Length Research Article

INCONSISTENSIES IN BIG DATA

***,[1]Suriya Begum, [1]Swapnil Konar and [2]Ashhar**

[1]Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India
[2]Department of Computer Science and Engineering, Bangalore, India

ARTICLE INFO

Article History:
Received 30th May, 2016
Received in revised form
21st June, 2016
Accepted 19th July, 2016
Published online 24th August, 2016

Key Words:
Big Data,
Inconsistency,
Big Data Analysis,
Temporal,
Spatial.

ABSTRACT

In today's world, Big Data (BD) plays a vital role. It has many applications in various domains. There are a lot of issues and challenges in BD. A lot of research has been carried out in this domain. In this paper, we have discussed in depth an important issue, namely, inconsistency in BD. Different types of inconsistencies has been discussed.

INTRODUCTION

Today, the advancement of sciences, engineering and technologies, the human endeavours, and the social and economic activities have collectively created a torrent of data in digital form. This Big Data (BD) phenomenon will only get intensified and diversified in the years to come. To turn this BD phenomenon into a positive force for good has drawn tremendous and intensified interest from an ever-increasing set of BD stakeholders. As BD becomes an increasingly popular buzzword, we must not lose sight of the fact that research issues behind BD and Big Data Analysis (BDA) are embedded in multi-dimensional scientific and technological spaces (zhangd@ecs.csus.edu,.http://gaia.ecs.csus.edu/~zhangd.).There are many issues related to BD. One such issue is inconsistencies in BD In this paper, we have highlighted this important issue of inconsistencies in BD. And their impact on the outcome of BDA. Inconsistencies are ubiquitous in the real world, manifesting themselves in a plethora of human behaviours and decision-making processes for which BD are acquired, integrated, analyzed, and utilized in an attempt to create values and accelerate scientific discoveries. Once captured in BD, inconsistencies can occur at various granularities of knowledge content, from data, information, knowledge, meta-knowledge, to expertise. If not handled properly, inconsistencies can have adverse impact on the quality of the outcomes in BDA process. Inconsistencies can also exhibit in reasoning methods, heuristics, or problem-solving approaches of various analysis tasks, creating challenges for BDA. In the paper, we describe classifications for four types of inconsistencies in BD. It turns out that BD inconsistencies can be utilized as important heuristics for improving the performance in various analysis tasks and the quality of outcomes in BDA. The recently proposed inconsistency-induced learning, or learning, offers a promising approach towards proper handling of BD inconsistencies.

****Corresponding author: Suriya Begum***
Professor, Department of Computer Science and Engineering, New Horizon College of Engineering, Bangalore, India.

Inconsistencies in big Data Analysis

Inconsistencies are common place in human behaviours and decision-making processes for which BD are acquired, fused, and represented. Once captured in big data, inconsistent or

conflicting phenomena can occur at various granularities of knowledge content, from data, information, knowledge, meta-knowledge, to expertise, and can adversely affect the quality of the outcomes of BDA process. Inconsistencies can also manifest themselves in reasoning methods, heuristics, or problem-solving approaches of various analysis tasks, creating challenges for BDA.

Example: Let X and Y be a set of data instances and a set of labels for data instances, Respectively. Given a dataset S and two data elements *did* ∈ S and *do* ∈ S, *did* = (*x*, *y*) and *Did* = (*x* ′, *y* ′), where *x*, *x* ′□∈ X, and *y*, *y* ′□∈ Y. *did* and *do* are data instances with *inconsistent labels* when the following holds:

$(x = x') \wedge \square (y \neq y') \wedge \square (y \supset \neg y') \wedge \square (y' \square \supset \neg y)$.

The presence of *did* and *do* in **S** is referred to as *Data Inconsistency*. When subjecting a machine learning algorithm to a dataset **S** that contains data inconsistency, the model thus learned will have a reduced predictive accuracy. We need to recognize types of inconsistencies for different types of BD. For instance, for location-based or time series Data, temporal or spatial inconsistencies will dominate, whereas for unstructured text data, inconsistencies pertaining to antonym, negation, mismatched value, structural or lexical contrasts or world knowledge will occupy a commanding position. In addition, it is necessary to differentiate categories of inconsistent phenomena at different levels of data, information, knowledge, meta-knowledge.

Inconsistencies at Data Level instances (symbolic, numeric, categorical, waveform, etc.) and different types of labels. Inconsistencies at Information Level manifest in terms of functional dependencies or associations. Inconsistencies at Knowledge Level. Inconsistencies display in declarative or procedural beliefs; Meta-knowledge inconsistencies are demonstrated through control strategies or learning decisions. There are different BDA tasks or objectives, such as prediction, classification, regression, association analysis, clustering, and outlier analysis. Which type of inconsistencies has what impact on which analytic objective is yet another issue to be investigated? The goal is to utilize inconsistencies as valuable heuristics in guiding the development of inconsistency-specific tools to help assist tasks in BDA. One Example is Inconsistency-Induced Learning, or i2Learning in, that allows inconsistencies to be utilized as stimuli to initiate learning episodes that lead to the resolution of data or knowledge inconsistencies, or refined/augmented knowledge, which in turn improves the performance of a system (http://gaia.ecs. csus.edu/~zhangd).

What Actually it is ?

Today, the advancement of sciences, engineering and technologies, the human endeavours, and the social and economic activities have collectively created a torrent of data in digital form. This BD phenomenon will only get intensified and diversified in the years to come. Many domains and economic sectors can benefit from the BD push: life and physical sciences, medicine, education, healthcare, location-based services, manufacturing, retail, communication and media, government, transportation, banking, insurance, financial services, utilities, environment, and energy industry. In this paper, we first take a close look on an important issue: inconsistencies in BD. Inconsistencies are ubiquitous in the real world, manifesting themselves in a plethora of human behaviours and decision-making processes for which big data are acquired, integrated, analyzed, and utilized in an attempt to create values and accelerate scientific discoveries.

Once captured in BD, inconsistencies can occur at various granularities of knowledge content, from data, information, knowledge, meta-knowledge, to expertise. If not handled properly, inconsistencies can have adverse impact on the quality of the outcomes in BDA process. Inconsistencies can also exhibit in reasoning methods, heuristics, or problem-solving approaches of various analysis tasks, creating challenges for BDA.

Temporal Inconsistencies

When datasets contain a temporal attribute, data items with conflicting circumstances may coincide or overlap in time. The time interval relationships between conflicting data items can result in partial temporal inconsistency or complete temporal inconsistency. Temporal inconsistencies have been utilized as problem-solving heuristics in IBM Watson open-domain QA system where temporal reasoning is deployed to "detect inconsistencies between dates in the clue and those associated with a candidate answer" (http://gaia.ecs.csus.edu/~zhangd; Arul Murugan *et al*., 2014). In a temperature time-series data, a temperature recording of 35°F in July in New Orleans would be inconsistent with the context. In human electrocardiogram time-series data, a prolonged period of low value output in the ECG is inconsistent with the normal heart rhythm pattern, an indication for atria premature contraction (Maximilian). Table 1 gives a list of temporal inconsistencies.

Table 1.

	Conflicting case
Partial	Time intervals of two inconsistent events are partially overlapping.
Complete	Time intervals of two inconsistent events coincide or satisfy containment.
Anomalous value	A time-series data has an anomalous value.
Contextual	A time-series data has an anomalous instance in a given context.
Motif	Time-series data has a segment of data values that reoccurs and is anomalous

Explanation

Given the types of temporal constraints, there are many types of no-goods that could theoretically arise; illustrates six abstract examples. In these examples, the TAP is trying to insert activity B into the plan, which causes the no-good; we refer to B as the goal activity. Activity B is constrained to be placed "between" A and C; more precisely, start (B) is constrained by end (A) and end (B) is constrained by start(C). Both A and C are activities, except in case 4 where A is a sol event. Thick vertical lines indicate the temporal boundaries of the no-good. The arrows represent edges in the no-good; the

dotted arrows represent edge-pairs through the origin. The edges adjacent to B may have non-zero lower bounds; we refer to these as *buffers*. In case 4, the line on the arrow from end (B) to start(C) indicates a positive buffer; this means that C cannot start earlier than m time units *after* the end of B, where the buffer is +m. In case 6, there is a negative buffer on the edge from end(A) to start(B); this means that B cannot start earlier than m time units *before* the end of A, where the buffer is –m. In case 5, the middle unlabelled block indicates a *backwards* chain of activities; in general, there could be several of these in a no-good, causing it to zigzag in direction.

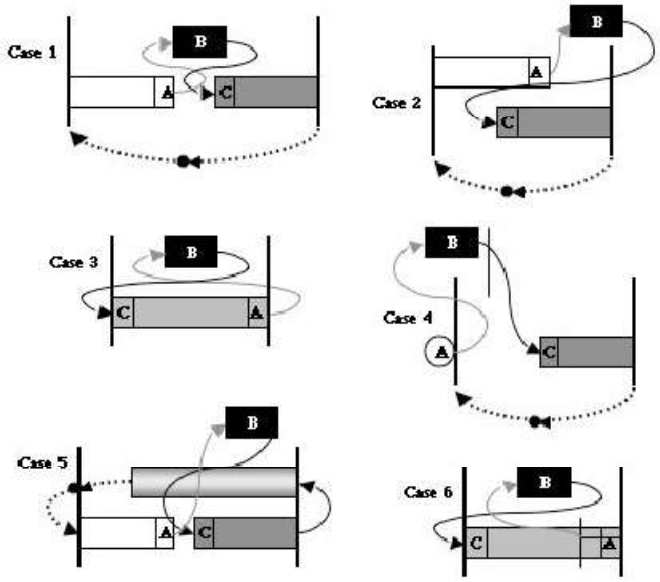

Figure 1.

Spatial Incosistencies

When datasets include geometric or spatial dimension, data items are often about objects in space that have geometric properties (location, shape) and that observe spatial relations (topological, directional and distance relations) as shown in Table 2. Spatial inconsistencies can arise from the geometric representation of objects (a spatial object having multiple conflicting geometric locations), spatial relations between objects (violations of spatial constraints with regard to some spatial relation), or aggregation of composite objects (different representations of the same object from different sources resulting in violation of the constraint that objects must have unique geometric representation).

Text Inconsistencies

The Text inconsistencies usually originate from various sources like emails, blogs, social media etc. These inconsistencies can highly alter the integrity of the data. Whenever two texts are referring to the same event or entity, then they are said to be co-reference. The co-reference is a mandatory condition for text inconsistencies (Arul Murugan *et al*., 2014).

Table 2.

	Conflicting case
Geometric location	A spatial object has conflicting geometric locations.
Geometric shape	A spatial object has conflicting geometric shapes.
Topological	Violation of topological constraints.
Directional	Violation of directional properties.
Distance	Violation of distance properties.
Scaling induced	Different geometric representations of a spatial object at different scales.
Semantic constraint	Violation of semantic integrity constraints.
Structural constraint	Violation of structural integrity constraints of geometric primitives.
Integration induced	Different representations of the same spatial object from different sources resulting in violation of the constraint that objects must have unique geometric representation.

Table 3.

	Conflicting case
Complementary	- Miami Heat was in the 2012 NBA final. - Miami Heat was not in the 2012 NBA final.
Mutual exclusive	- Sea cucumber is animal. - Sea cucumber is vegetable.
Inheritance	- Penguin cannot fly. - Penguin is bird and bird can fly, hence penguin can fly.
Synonym	- The system has a fast response time. - The system's response time is not rapid.
Antonym	- The system has a fast response time. - The system has a slow response time.
Anti-subsumption	- John is a surgeon. - John is not a doctor.
Anti-supertype	- BigDog is not a robot. - BigDog is a legged squad support system.
Asymmetric	- John is married to Jane. - Jane is not married to John.
Anti-inverse	- John is parent of Mike. - Mike is not child of John.
Mismatching	- M_5 is a mobile agent that runs in both Android and iOS environments. - M_5 does not run in Android environment.
Disagreeing	- M_5 has a memory of 10 GB. - M_5 has a memory of 5000 MB.
Contradictory	- M_5 was developed in March 2013. - M_5 was deployed in December 2012.

FUNCTIONAL DEPENDENCY INCONSISTENCIES

Many big datasets are stored, aggregated, and cleaned through the help of relational database systems where functional dependencies (FD) or conditional functional dependencies play a critically important role in enforcing the integrity constraints for the database. Violations of such functional dependencies or conditional functional dependencies will result in inconsistencies in data and information.

Table 4.

	Conflicting case
Single FD	Violation of single functional dependency
Multiple FD	Violation of multiple functional dependencies.
Conditional FD	Violation of conditional functional dependencies

Inconsistence Induced Learning

A framework for inconsistency-induced learning, (i2Learning), has been proposed in. i2 Learning accommodates perpetual or lifelong learning by allowing successive learning episodes to be triggered through inconsistencies an agent encounters during its problem solving episodes. Learning in the framework is accomplished through the continuous knowledge refinement and/or augmentation in order to overcome encountered can be incrementally improved with each learning episode i2Learning offers an overarching structure that facilitates the growth and expansion of various inconsistency-specific learning strategies. The essential idea behind i2Learning is to identify the cause of inconsistency and then apply cause-specific heuristics to resolve inconsistencies. For instance, if an inconsistent phenomenon stems from irrelevant features, then we can deploy a search algorithm that discerns relevant features from irrelevant ones.

We can then overcome inconsistencies by excluding irrelevant features from participating in the analysis process. If an inconsistent case arises as a result of property inheritance, then the heuristic rules of subclass-specificity superseding super class generality can be utilized to resolve inheritance inconsistencies. In the context of big data and big data analysis, i2Learning can also play an active role in improving the data quality by reconciling the inconsistencies found in the datasets, in refining or augmenting knowledge for analysis, modelling or interpretation of big data, and in helping enhance big data applications. For instance, in crowd sourced. Customer ranking application for goods or services, comments made by customers invariably contain inconsistencies (text, temporal, or spatial). Treating customers' comments as a knowledge base that contains pockets of incompatible opinions, we can apply i2Learning algorithms to resolve or overcome the inconsistencies customers' comments, which in turn refines or augments this "knowledge base" to render a more coherent and accurate ratings of the goods or services. This process is continuous and perpetual, with each new inconsistent.

Conclusion

In this paper, We focus our attention on one of the challenges, inconsistencies in Big Data and their impact on BDA. We examine four types of inconsistencies in BD, namely, temporal inconsistencies, spatial inconsistencies, text inconsistencies, and functional dependency inconsistencies. The contribution of this work lies in the fact that articulating explicitly the types of inconsistent phenomena in BD can help pave the way to improve the quality of BDA

Future Work

Our future work can be carried out in the following Directions. Details of other frequently encountered types of inconsistencies in BD and their impact on BDA still need to be fleshed out. Empirical study is planned on utilizing i2Learning algorithms with some real world dataset to improve the analysis results.

REFERENCES

Arul Murugan R1, Angora R2,Boopathi R3 ," BIG DATA: PRIVACY AND INCONSISTENCY ISSUES ",IJRET: International Journal of Research in Engineering and Technology eosin : 2319-1163 | pass: 2321-7308 .Volume: 03 Special Issue: 07|May-2014.

Du Zhang , "GRANULARITIES AND INCONSISTENCIES IN BIG DATA ANALYSIS", International Journal of Software Engineering and Knowledge Engineering , Department of Computer Science, California Sate University Sacramento, 95819-6021, USA,zhangd@ecs.csus.edu,.http://gaia.ecs.csus.edu/~zhangd.

Du Zhang, "INCONSISTENCIES OF BIG DATA", Department of Computer Science, California Sate University Sacramento, 95819-6021, USA, zhangd@ecs.csus.edu,.http://gaia.ecs.csus.edu/~zhangd.

http://ictpost.com/big-data-will-help-unearth-inconsistencies-in-indian-healthcare/

http://people.mpi-inf.mpg.de/alumni /d5/2014/ mdylla// publications/ BTW11.pdf

http://searchbusinessanalytics.techtarget.com/feature/Eliminate-inconsistent-data-by-ousting-its-evil-twin-sisters

http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0515_1931.pdfwww.arpnjournals.com

http://www.computerworld.com/article/2878080/test-shows-big-data-text-analysis-inconsistent-inaccurate.html

http://www.rroij.com/open-access/investigations-on-methods-developed-foreffective-discovery-of-functionaldependencies.pdf

http://www.slideshare.net/minujoseph/inconsistencies-in-big-data

http://www.stsci.edu/institute/conference/iwpss/plenary-d1-Bresina.pdf

http://www.worldscientific.com/doi/abs/10.1142/S0218194013500241?journalCode=ijseke

https://books.google.co.in/books?id=kLW8BAAAQBAJ&pg=PA199&lpg=PA199&dq=big+data+inconsistency&source=bl&ots=WO9OWZS- =big% 20data % 20 inconsistency&f=false

https://books.google.co.in/books?id=p7d1BwAAQBAJ&pg=PA9&lpg=PA9&dq=big+data+inconsistency&source=bl&ots=70e7rQbVUs&sig=_ce4g- 20inconsistency&f=false

IEEE Xplore Abstract (Abstract) - Inconsistencies in big data Ieeexplore.ieee.org

John, L. Berezina and Paul H. Morris, "EXPLANATION AND RECOMMENDATION FOR TEMPORAL INCONSISTENCIES ",

Maria del Pilar Angeles ,Lachlan M. Mackinnon ,"Detection and Resolution of Data Inconsistencies, and Data Integration using Information Quality criteria",School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh,EH14 4AS pilar@macs.hw.ac.uk ,School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh,EH14 4AS lachlan@macs.hw.ac.uk.

Maximilian Della_ Mauro Suzie Martin Theo bald ,"Resolving Temporal Conflicts in Inconsistent RDF Knowledge Bases", {mdylla, msozio,mtb}@mpi-inf.mpg.de Max-Planck Institute for Informatics (MPI-INF) .Saarbrucken, Germany.

NASA Ames Research CentreMoffett Field, CA 94043,John.L.Bresina@nasa.gov
Paul.H.Morris@nasa.gov.
people.mpi-inf.mpg.de
www.rroij.com
www.stsci.edu

Other Publications

1. Suriya Begum, Dr. Prashanth C.S.R, "Review of Load Balancing in cloud Computing", *International Journal of Computer Science Issues,* ISSN (Print): 1694-0784 | ISSN (Online): 1694-0814, Vol. 10, Issue 1, No. 2, January-2013 , pg. 343 – 352 .
2. Suriya Begum, Dr. Prashanth C.S.R, "Investigational Study of 7 Effective Schemes of Load Balancing in cloud Computing", *International Journal of Computer Science* Issues, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, Vol. 10, Issue 6, No. 1, November-2013, pg. 276 – 287.
3. Suriya Begum, Dr. Prashanth C.S.R, "Mathematical Modeling of Joint Routing and Scheduling for an Effective Load Balancing in cloud", *International Journal of Computer Application*, (0975 – 8887), Volume 104 – No.4, October-2014, pg. 32 – 38.
4. Suriya Begum, Dr. Prashanth C.S.R, "Stochastic Based Load Balancing Mechanism for Non-Iterative Optimization of Traffic in Cloud ", IEEE International Conference, SSN Colg of Engg, Chennai, 23-25 March-2016, WiSPNET 2016.
5. Suriya Begum, Vasanthi, Dr. Prashanth C.S.R, "Ensuring Data Security in Cloud Computing From Single To Multi Cloud For Multi Share Users", National Level Technical Paper Presentation, MVJ CE, Bangalore, 4th-5th October, FVCT-2012.
6. Suriya Begum, Archana, Dr. Prashanth C.S.R, "Performance Analysis of Cloud Computing Centres using M=G=m=m+r Queuing Systems", National Level Technical Paper Presentation, NHCE, Bangalore, 28th March, NCICT-2013.
7. Suriya Begum, Ayub Inamdar, "Priority Based Pre-Emptive Scheduling of Real Time Services Request with Task Migration for Cloud Computing", National Level Technical Paper Presentation, Akshaya IT Tumkur, 29th April Technika-2014
8. Suriya Begum, Asha C, Dr. Prashanth C.S.R, "A Novel Load Balancing Strategy for Effective Utilisation of Virtual Machines in Cloud", *International Journal of Computer Science and Mobile Computing*, ISSN : 2320-088X, Vol 4, Issue 6, June-2015, pg. 862-870.
9. Suriya Begum, Ananth Raju, Dr. Prashanth C.S.R, "Resource Management By Virtual Machines Migration In Cloud Computing", *International Journal of Computer Science and Mobile Computing*, ISSN : 2320-088X, Vol 4, Issue 12, December-2015, pg. 307-312.
10. Suriya Begum, Sachin, Ram Charan, Nikhit, Sai Prathap, "Analysis of Various Load Balancing Techniques in Cloud Environment", *International Journal of Computer Science and Mobile Computing*, ISSN: 2320-088X, Vol 5, Issue 2, Feb-2016, pg . 248 -254.
11. Suriya Begum, Venugopal, "Comparison Of Various Techniques In IoT For Healthcare System", *International Journal of Computer Science and Mobile Computing*, ISSN: 2320-088X, Vol 5, Issue 3, March-2016, pg. 59 – 66.
12. Suriya Begum, Kavya, "Analysis of Various Big Data Techniques For Security", *International Journal of Computer Science and Mobile Computing,* ISSN: 2320-088X, Vol 5, Issue 3, March-2016, pg. 54 – 58 .
13. Suriya Begum, Venugopal, "Analysis of Load Balancing Algorithms in Cloud Environment", *International Journal of Emerging Technology and Advanced Engineering*, ISSN: 2250-2459, Vol 6, Issue 4, April-2016, pg. 151 – 154.
14. Suriya Begum, Kavya, "A Study on Load Balancing techniques in Cloud Computing Environment", *International Journal of Emerging Technology and Advanced Engineering*, ISSN: 2250-2459, Vol 6, Issue 5, May-2016, pg. 72 – 74.
15. Suriya Begum, Kavya, Venugopal, "A Study on Load Balancing techniques in Cloud Computing Environment", 3rd International Conference On Convergent Innovative Technologies, Cambridge IT, Bangalore. ISSN (Online): 2319-6890, 20th May-2016,ICCIT-2016 .
16. Suriya Begum, Kavya, Venugopal, "A Study on Load Balancing techniques in Cloud Computing Environment", *International Journal of Engineering Research, ISSN(Online)* : 2319-6890, ISSN(Print) : 2347-5013, Vol 5, Issue Special 4, May-2016, pg.904-909 .
17. Suriya Begum, Rohit Mulay, Ashhar, "Near Field Communication: A Survey", *International Journal of Emerging Technology and Advanced Engineering*, ISSN: 2250-2459, Vol 6, Issue 6, June-2016, pg. 92 –9 7.

****