Mining Big Data: Current Status, and Forecast to the Future

Wei Fan
Huawei Noah's Ark Lab
Hong Kong Science Park
Shatin, Hong Kong
david.fanwei@huawei.com

Albert Bifet
Yahoo! Research Barcelona
Av. Diagonal 177
Barcelona, Catalonia, Spain
abifet@yahoo-inc.com

ABSTRACT

Big Data is a new term used to identify the datasets that due to their large size and complexity, we can not manage them with our current methodologies or data mining software tools. *Big Data mining* is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. The Big Data challenge is becoming one of the most exciting opportunities for the next years. We present in this issue, a broad overview of the topic, its current status, controversy, and forecast to the future. We introduce four articles, written by influential scientists in the field, covering the most interesting and state-of-the-art topics on Big Data mining.

1. INTRODUCTION

Recent years have witnessed a dramatic increase in our ability to collect data from various sensors, devices, in different formats, from independent or connected applications. This data flood has outpaced our capability to process, analyze, store and understand these datasets. Consider the Internet data. The web pages indexed by Google were around one million in 1998, but quickly reached 1 billion in 2000 and have already exceeded 1 trillion in 2008. This rapid expansion is accelerated by the dramatic increase in acceptance of social networking applications, such as Facebook, Twitter, Weibo, etc., that allow users to create contents freely and amplify the already huge Web volume. Furthermore, with mobile phones becoming the sensory gateway to get real-time data on people from different aspects, the vast amount of data that mobile carrier can potentially process to improve our daily life has significantly outpaced our past CDR (call data record)-based processing for billing purposes only. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, to buses, railway stations and airports) are all loosely connected. Trillions of such connected components will generate a huge data ocean, and valuable information must be discovered from the data to help improve quality of life and make our world a better place. For example, after we get up every morning, in order to optimize our commute time to work and complete the optimization before we arrive at office, the system needs to process information from traffic, weather, construction, police activities to our calendar schedules, and perform deep optimization under the tight time constraints. In all these applications, we are facing significant challenges in leveraging the vast amount of data, including challenges in (1) system capabilities (2) algorithmic design (3) business models.

As an example of the interest that Big Data is having in the data mining community, the grand theme of this year's KDD conference was *'Mining the Big Data'*. Also there was a specific workshop *BigMine'12* in that topic: *1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*[1]. Both events successfully brought together people from both academia and industry to present their most recent work related to these Big Data issues, and exchange ideas and thoughts. These events are important in order to advance this Big Data challenge, which is being considered as one of the most exciting opportunities in the years to come.

We introduce Big Data mining and its applications in Section 2. We summarize the papers presented in this issue in Section 3, and discuss about Big Data controversy in Section 4. We point the importance of open-source software tools in Section 5 and give some challenges and forecast to the future in Section 6. Finally, we give some conclusions in Section 7.

2. BIG DATA MINING

The term 'Big Data' appeared for first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title of "Big Data and the Next Wave of InfraStress" [9]. Big Data mining was very relevant from the beginning, as the first book mentioning 'Big Data' is a data mining book that appeared also in 1998 by Weiss and Indrukya [34] . However, the first academic paper with the words 'Big Data' in the title appeared a bit later in 2000 in a paper by Diebold [8]. The origin of the term 'Big Data' is due to the fact that we are creating a huge amount of data every day. Usama Fayyad [11] in his invited talk at the KDD BigMine'12 Workshop presented amazing data numbers about internet usage, among them the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 milion tweets per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of zettabytes, and it is growing around 40% every year. A

[1] `http://big-data-mining.org/`

new large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, Twitter are starting to look carefully to this data to find useful patterns to improve user experience. Alex 'Sandy' Pentland in his 'Human Dynamics Laboratory' at MIT, is doing research in finding patterns in mobile data about what users do, and not in what people says they do [28].
We need new algorithms, and new tools to deal with all of this data. Doug Laney[19] was the first one in talking about 3 V's in Big Data management:

- Volume: there is more data than ever before, its size continues increasing, but not the percent of data that our tools can process
- Variety: there are many different types of data, as text, sensor data, audio, video, graph, and more
- Velocity: data is arriving continuously as streams of data, and we are interested in obtaining useful information from it in real time

Nowadays, there are two more V's:

- Variability: there are changes in the structure of the data and how users want to interpret that data
- Value: business value that gives organization a compelling advantage, due to the ability of making decisions based in answering questions that were previously considered beyond reach

Gartner[15] summarizes this in their definition of Big Data in 2012 as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.
There are many applications of Big Data, for example the following [17; 2]:

- Business: costumer personalization, churn detection
- Technology: reducing process time from hours to seconds
- Health: mining DNA of each person, to discover, monitor and improve health aspects of every one
- Smart cities: cities focused on sustainable economic development and high quality of life, with wise management of natural resources

These applications will allow people to have better services, better costumer experiences, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [17].

2.1 Global Pulse: "Big Data for development"

To show the usefulness of Big Data mining, we would like to mention the work that Global Pulse is doing [33] using Big Data to improve life in developing countries. Global Pulse is a United Nations initiative, launched in 2009, that functions as an innovative lab, and that is based in mining Big Data for developing countries. They pursue a strategy that consists of 1) researching innovative methods and techniques for analyzing real-time digital data to detect early emerging vulnerabilities; 2) assembling free and open source technology toolkit for analyzing real-time data and sharing hypotheses; and 3) establishing an integrated, global network of Pulse Labs, to pilot the approach at country level.
Global Pulse describe the main opportunities Big Data offers to developing countries in their White paper "Big Data for Development: Challenges & Opportunities"[22]:

- Early warning: develop fast response in time of crisis, detecting anomalies in the usage of digital media
- Real-time awareness: design programs and policies with a more fine-grained representation of reality
- Real-time feedback: check what policies and programs fails, monitoring it in real time, and using this feedback make the needed changes

The Big Data mining revolution is not restricted to the industrialized world, as mobiles are spreading in developing countries as well. It is estimated than there are over five billion mobile phones, and that 80% are located in developing countries.

3. CONTRIBUTED ARTICLES

We selected four contributions that together shows very significant state-of-the-art research in Big Data Mining, and that provides a broad overview of the field and its forecast to the future. Other significant work in Big Data Mining can be found in the main conferences as KDD, ICDM, ECML-PKDD, or journals as "Data Mining and Knowledge Discovery" or "Machine Learning".

- **Scaling Big Data Mining Infrastructure: The Twitter Experience** by *Jimmy Lin and Dmitriy Ryaboy (Twitter,Inc.)*. This paper presents insights about Big Data mining infrastructures, and the experience of doing analytics at Twitter. It shows that due to the current state of the data mining tools, it is not straightforward to perform analytics. Most of the time is consumed in preparatory work to the application of data mining methods, and turning preliminary models into robust solutions.

- **Mining Heterogeneous Information Networks: A Structural Analysis Approach** by *Yizhou Sun (Northeastern University) and Jiawei Han (University of Illinois at Urbana-Champaign)*. This paper shows that mining heterogeneous information networks is a new and promising research frontier in Big Data mining research. It considers interconnected, multi-typed data, including the typical relational database data, as heterogeneous information networks. These semi-structured heterogeneous information network models leverage the rich semantics of typed nodes and links in a network and can uncover surprisingly rich knowledge from interconnected data.

- **Big Graph Mining: Algorithms and discoveries** by *U Kang and Christos Faloutsos(Carnegie Mellon Universitiy)*. This paper presents an overview of mining big graphs, focusing in the use of the PEGASUS tool, showing some findings in the Web Graph and Twitter social network. The paper gives inspirational future research directions for big graph mining.

- **Mining Large Streams of User Data for Personalized Recommendations** by *Xavier Amatriain (Netflix)*.

This paper presents some lessons learned with the Netflix Prize, and discuss the recommender and personalization techniques used in Netflix. It discusses recent important problems and future research directions. Section 4 contains an interesting discussion about if we need more data or better models to improve our learning methodology.

4. CONTROVERSY ABOUT BIG DATA

As Big Data is a new hot topic, there have been a lot of controversy about it, for example see [7]. We try to summarize it as follows:

- There is no need to distinguish Big Data analytics from data analytics, as data will continue growing, and it will never be small again.
- Big Data may be a hype to sell Hadoop based computing systems. Hadoop is not always the best tool [23]. It seems that data management system sellers try to sell systems based in Hadoop, and MapReduce may be not always the best programming platform, for example for medium-size companies.
- In real time analytics, data may be changing. In that case, what it is important is not the size of the data, it is its recency.
- Claims to accuracy are misleading. As Taleb explains in his new book [32], when the number of variables grow, the number of fake correlations also grow. For example, Leinweber [21] showed that the S&P 500 stock index was correlated with butter production in Bangladesh, and other funny correlations.
- Bigger data are not always better data. It depends if the data is noisy or not, and if it is representative of what we are looking for. For example, some times Twitter users are assumed to be a representative of the global population, when this is not always the case.
- Ethical concerns about accessibility. The main issue is if it is ethical that people can be analyzed without knowing it.
- Limited access to Big Data creates new digital divides. There may be a digital divide between people or organizations being able to analyze Big Data or not. Also organizations with access to Big Data will be able to extract knowledge that without this Big Data is not possible to get. We may create a division between Big Data rich and poor organizations.

5. TOOLS: OPEN SOURCE REVOLUTION

The Big Data phenomenon is intrinsically related to the open source software revolution. Large companies as Facebook, Yahoo!, Twitter, LinkedIn benefit and contribute working on open source projects. Big Data infrastructure deals with Hadoop, and other related software as:

- Apache Hadoop [3]: software for data-intensive distributed applications, based in the MapReduce programming model and a distributed file system called Hadoop Distributed Filesystem (HDFS). Hadoop allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reduce tasks use the output of the maps to obtain the final result of the job.
- Apache Hadoop related projects [35]: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others.
- Apache S4 [26]: platform for processing continuous data streams. S4 is designed specifically for managing data streams. S4 apps are designed combining streams and processing elements in real time.
- Storm [31]: software for streaming data-intensive distributed applications, similar to S4, and developed by Nathan Marz at Twitter.

In Big Data Mining, there are many open source initiatives. The most popular are the following:

- Apache Mahout [4]: Scalable machine learning and data mining open source software based mainly in Hadoop. It has implementations of a wide range of machine learning and data mining algorithms: clustering, classification, collaborative filtering and frequent pattern mining.
- R [29]: open source programming language and software environment designed for statistical computing and visualization. R was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets.
- MOA [5]: Stream data mining open source software to perform data mining in real time. It has implementations of classification, regression, clustering and frequent item set mining and frequent graph mining. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software. The `streams` framework [6] provides an environment for defining and running stream processes using simple XML based definitions and is able to use MOA, Android and Storm. SAMOA [1] is a new upcoming software project for distributed stream mining that will combine S4 and Storm with MOA.
- Vowpal Wabbit [20]: open source project started at Yahoo! Research and continuing at Microsoft Research to design a fast, scalable, useful learning algorithm. VW is able to learn from terafeature datasets. It can exceed the throughput of any single machine network interface when doing linear learning, via parallel learning.

More specific to Big Graph mining we found the following open source tools:

- PEGASUS [18]: big graph mining system built on top of MAPREDUCE. It allows to find patterns and anomalies in massive real-world graphs. See the paper by U. Kang and Christos Faloutsos in this issue.

- GraphLab [24]: high-level graph-parallel system built without using MapReduce. GraphLab computes over dependent records which are stored as vertices in a large distributed data-graph. Algorithms in GraphLab are expressed as vertex-programs which are executed in parallel on each vertex and can interact with neighboring vertices.

6. FORECAST TO THE FUTURE

There are many future important challenges in Big Data management and analytics, that arise from the nature of data: large, diverse, and evolving [27; 16]. These are some of the challenges that researchers and practitioners will have to deal during the next years:

- Analytics Architecture. It is not clear yet how an optimal architecture of an analytics systems should be to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture of Nathan Marz [25]. The Lambda Architecture solves the problem of computing arbitrary functions on arbitrary data in realtime by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer. It combines in the same system Hadoop for the batch layer, and Storm for the speed layer. The properties of the system are: robust and fault tolerant, scalable, general, extensible, allows ad hoc queries, minimal maintenance, and debuggable.
- Statistical significance. It is important to achieve significant statistical results, and not be fooled by randomness. As Efron explains in his book about Large Scale Inference [10], it is easy to go wrong with huge data sets and thousands of questions to answer at once.
- Distributed mining. Many data mining techniques are not trivial to paralyze. To have distributed versions of some methods, a lot of research is needed with practical and theoretical analysis to provide new methods.
- Time evolving data. Data may be evolving over time, so it is important that the Big Data mining techniques should be able to adapt and in some cases to detect change first. For example, the data stream mining field has very powerful techniques for this task [13].
- Compression: Dealing with Big Data, the quantity of space needed to store it is very relevant. There are two main approaches: compression where we don't loose anything, or sampling where we choose what is the data that is more representative. Using compression, we may take more time and less space, so we can consider it as a transformation from time to space. Using sampling, we are loosing information, but the gains in space may be in orders of magnitude. For example Feldman et al. [12] use coresets to reduce the complexity of Big Data problems. Coresets are small sets that provably approximate the original data for a given problem. Using merge-reduce the small sets can then be used for solving hard machine learning problems in parallel.
- Visualization. A main task of Big Data analysis is how to visualize the results. As the data is so big, it is very difficult to find user-friendly visualizations. New techniques, and frameworks to tell and show stories will be needed, as for example the photographs, infographics and essays in the beautiful book "The Human Face of Big Data" [30].
- Hidden Big Data. Large quantities of useful data are getting lost since new data is largely untagged file-based and unstructured data. The 2012 IDC study on Big Data [14] explains that in 2012, 23% (643 exabytes) of the digital universe would be useful for Big Data if tagged and analyzed. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed.

7. CONCLUSIONS

Big Data is going to continue growing during the next years, and each data scientist will have to manage much more amount of data every year. This data is going to be more diverse, larger, and faster. We discussed in this paper some insights about the topic, and what we consider are the main concerns, and the main challenges for the future. Big Data is becoming the new Final Frontier for scientific data research and for business applications. We are at the beginning of a new era where Big Data mining will help us to discover knowledge that no one has discovered before. Everybody is warmly invited to participate in this intrepid journey.

8. ACKNOWLEDGEMENTS

We would like to thank Jimmy Lin, Dmitriy Ryaboy, Jiawei Han, Yizhou Sun, U Kang, Christos Faloutsos and Xavier Amatriain for contributing to this special section.

9. REFERENCES

[1] SAMOA, http://samoa-project.net, 2013.

[2] C. C. Aggarwal, editor. *Managing and Mining Sensor Data*. Advances in Database Systems. Springer, 2013.

[3] Apache Hadoop, http://hadoop.apache.org.

[4] Apache Mahout, http://mahout.apache.org.

[5] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa.cms.waikato.ac.nz/. *Journal of Machine Learning Research (JMLR)*, 2010.

[6] C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.

[7] d. boyd and K. Crawford. Critical Questions for Big Data. *Information, Communication and Society*, 15(5):662–679, 2012.

[8] F. Diebold. "Big Data" Dynamic Factor Models for Macroeconomic Measurement and Forecasting. *Discussion Read to the Eighth World Congress of the Econometric Society*, 2000.

[9] F. Diebold. On the Origin(s) and Development of the Term "Big Data". Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.

[10] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2010.

[11] U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modeling. `http://big-data-mining.org/keynotes/#fayyad`, 2012.

[12] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *SODA*, 2013.

[13] J. Gama. *Knowledge Discovery from Data Streams*. Chapman & Hall/Crc Data Mining and Knowledge Discovery. Taylor & Francis Group, 2010.

[14] J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.

[15] Gartner, http://www.gartner.com/it-glossary/big-data.

[16] V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza. Big data, big business: bridging the gap. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, BigMine '12, pages 7–11, New York, NY, USA, 2012. ACM.

[17] Intel. Big Thinkers on Big Data, http://www.intel.com/content/www/us/en/big-data/big-thinkers-on-big-data.html, 2012.

[18] U. Kang, D. H. Chau, and C. Faloutsos. PEGASUS: Mining Billion-Scale Graphs in the Cloud. 2012.

[19] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. *META Group Research Note, February 6*, 2001.

[20] J. Langford. Vowpal Wabbit, http://hunch.net/~vw/, 2011.

[21] D. J. Leinweber. Stupid Data Miner Tricks: Overfitting the S&P 500. *The Journal of Investing*, 16:15–22, 2007.

[22] E. Letouzé. Big Data for Development: Opportunities & Challenges. May 2011.

[23] J. Lin. MapReduce is Good Enough? If All You Have is a Hammer, Throw Away Everything That's Not a Nail! *CoRR*, abs/1209.2191, 2012.

[24] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, California, July 2010.

[25] N. Marz and J. Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, 2013.

[26] L. Neumeyer, B. Robbins, A. Nair, and A. Kesari. S4: Distributed Stream Computing Platform. In *ICDM Workshops*, pages 170–177, 2010.

[27] C. Parker. Unexpected challenges in large scale machine learning. In *Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, BigMine '12, pages 1–6, New York, NY, USA, 2012. ACM.

[28] A. Petland. Reinventing society in the wake of big data. *Edge.org, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data*, 2012.

[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[30] R. Smolan and J. Erwitt. *The Human Face of Big Data*. Sterling Publishing Company Incorporated, 2012.

[31] Storm, http://storm-project.net.

[32] N. Taleb. *Antifragile: How to Live in a World We Don't Understand*. Penguin Books, Limited, 2012.

[33] UN Global Pulse, http://www.unglobalpulse.org.

[34] S. M. Weiss and N. Indurkhya. *Predictive data mining: a practical guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[35] P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. *IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Companies,Incorporated, 2011.