

Data Mining with Big Data using Spectral Clustering

[1] RAVINDRA NAYAK BHUKYA
Mtech(Cse)
Assistant Professor
rabeendranayak@yahoo.com

AU College of Engineering, Andhra University, Visakhapatnam, AP.

[2] Dr. S. VIZIANANDA ROW
MTech, Ph.D
ASSOCIATE PROFESSOR
vizianandarow_s@yahoo.com

AU College of Engineering, Andhra University, Visakhapatnam, AP.

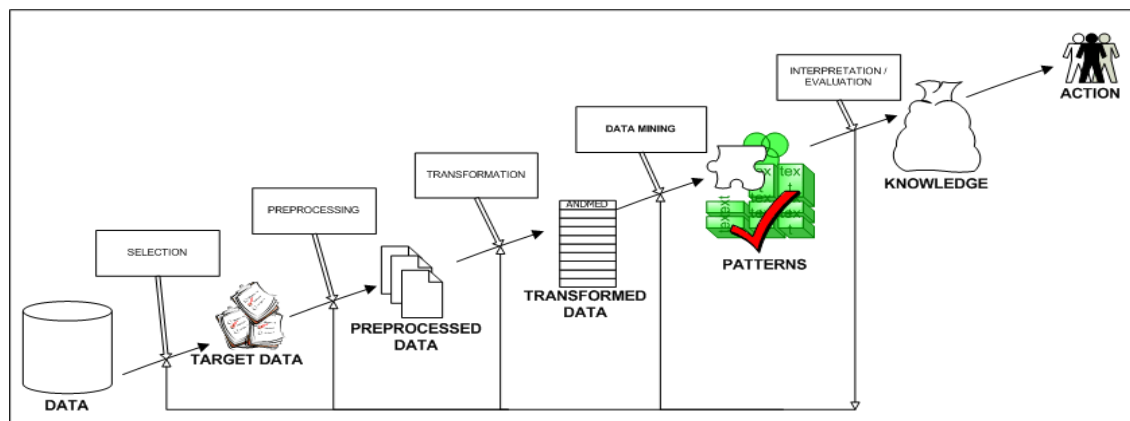
Abstract :

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that

characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

1. INTRODUCTION :

What is Data Mining?



Structure of Data Mining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

How Data Mining Works?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. **Generally, any of four types of relationships are sought:**

- **Classes:** Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations:** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns:** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- 1) Extract, transform, and load transaction data onto the data warehouse system.
- 2) Store and manage the data in a multidimensional database system.
- 3) Provide data access to business analysts and information technology professionals.
- 4) Analyze the data by application software.
- 5) Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

□ **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

□ **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where $k=1$). Sometimes called the k -nearest neighbor technique.

□ **Rule induction:** The extraction of useful if-then rules from data based on statistical significance.

□ **Data visualization:** The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Characteristics of Data Mining:

□ **Large quantities of data:** The volume of data so great it has to be analyzed by automated techniques e.g. satellite

information, credit card transactions etc.

□ **Noisy, incomplete data:** Imprecise data is the characteristic of all data collection.

□ **Complex data structure:** conventional statistical analysis not possible

□ **Heterogeneous data stored in legacy systems**

Benefits of Data Mining:

1) It's one of the most effective services that are available today. With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to them

2) An analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customers

3) An endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns)

4) Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors

5) Fast paced and prompt access to data along with economic processing techniques have made data mining

one of the most suitable services that a company seek

Advantages of Data Mining:

1. Marketing / Retail:

Data mining helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign...etc. Through the results, marketers will have appropriate approach to sell profitable products to targeted customers.

Data mining brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

2. Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from historical customer's data, the bank and financial institution can determine good and bad loans. In addition, data mining helps banks detect fraudulent credit card transactions to protect credit card's owner.

3. Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal

control parameters. For example semiconductor manufacturers has a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even has defects. Data mining has been applying to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

4. Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activities.

5. Law enforcement:

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

6. Researchers:

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

2. SYSTEM ANALYSIS

EXISTING SYSTEM:

- The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process

within a “tolerable elapsed time.” The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.

- The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

DISADVANTAGES OF EXISTING SYSTEM:

- The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.
- The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III).
- At Tier III, the data mining challenges concentrate on algorithm designs in tackling the difficulties raised by the Big Data volumes, distributed data distributions, and by

complex and dynamic data characteristics.

PROPOSED SYSTEM:

- We propose a HACE theorem to model Big Data characteristics. The characteristics of HACH make it an extreme challenge for discovering useful knowledge from the Big Data.
- The HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations.
- To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

ADVANTAGES OF PROPOSED SYSTEM:

- Provide most relevant and most accurate social sensing feedback to better understand our society at realtime.

MODULES:

1. Integrating and mining biodata
2. Big Data Fast Response
3. Pattern matching and mining
4. Key technologies for integration and mining
5. Group influence and interactions

3. MODULES DESCRIPTION:

Integrating and mining biodata:

We have integrated and mined biodata from multiple sources to decipher and utilize the structure of biological networks to shed new insights on the functions of biological systems. We address the theoretical underpinnings and current and future enabling technologies for integrating and mining biological networks. We have expanded and integrated the techniques and methods in information acquisition, transmission, and processing for information networks. We have developed methods for semantic-based data integration, automated hypothesis generation from mined data, and automated scalable analytical tools to evaluate simulation results and refine models.

Big Data Fast Response:

We propose to build a stream-based Big Data analytic framework for fast response and real-time decision making.

- Designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing
- Building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data, as well as accurately predict the trend of the data in the future; and
- A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

Pattern matching and mining:

We perform a systematic investigation on pattern matching, pattern mining with wildcards, and application problems as follows:

- ◆ Exploration of the NP-hard complexity of the matching and mining problems,
- ◆ Multiple patterns matching with wildcards,
- ◆ Approximate pattern matching and mining, and
- ◆ Application of our research onto ubiquitous personalized information processing and bioinformatics

Key technologies for integration and mining:

We have performed an investigation on the availability and statistical regularities of multisource, massive and dynamic information, including cross-media search based on information extraction, sampling, uncertain information querying, and cross-domain and cross-platform information polymerization. To break through the limitations of traditional data mining methods, we have studied heterogeneous information discovery and mining in complex inline data, mining in data streams, multigranularity knowledge discovery from massive multisource data, distribution regularities of massive knowledge, quality fusion of massive knowledge.

Group influence and interactions:

- ◆ Employing group influence and information diffusion models, and deliberating group interaction rules

- in social networks using dynamic game theory
- ◆ Studying interactive individual selection and effect evaluations under social networks affected by group emotion, and analyzing emotional interactions and influence among individuals and groups, and

- ◆ Establishing an interactive influence model and its computing methods for social network groups, to reveal the interactive influence effects and evolution of social networks.

4. OUTPUT SCREENS :



Contents of directory /result

Goto : /result go

[Go to parent directory](#)

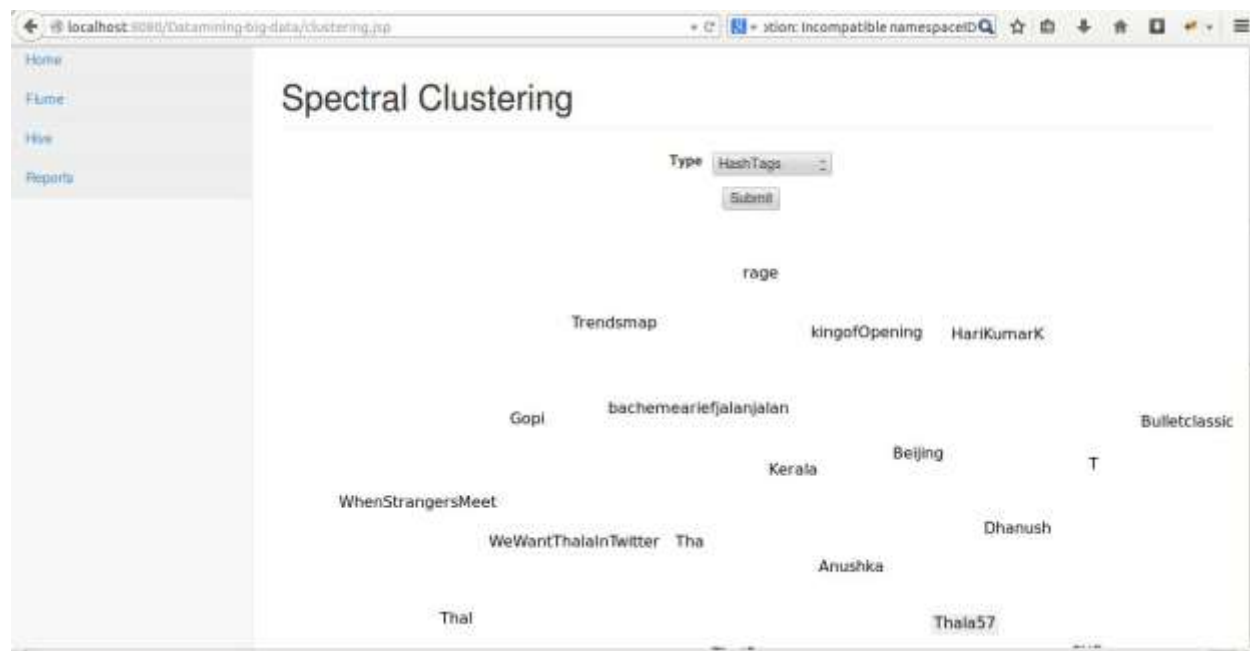
Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1415493253241	file	294.85 KB	1	64 MB	2014-11-09 01:34	rw-r--r--	eacefgo	supergroup

[Go back to DFS home](#)

Local logs

[Log directory](#)

This is [Apache Hadoop](#) release 1.2.1



Spectral Clustering

Type: HashTags

rage

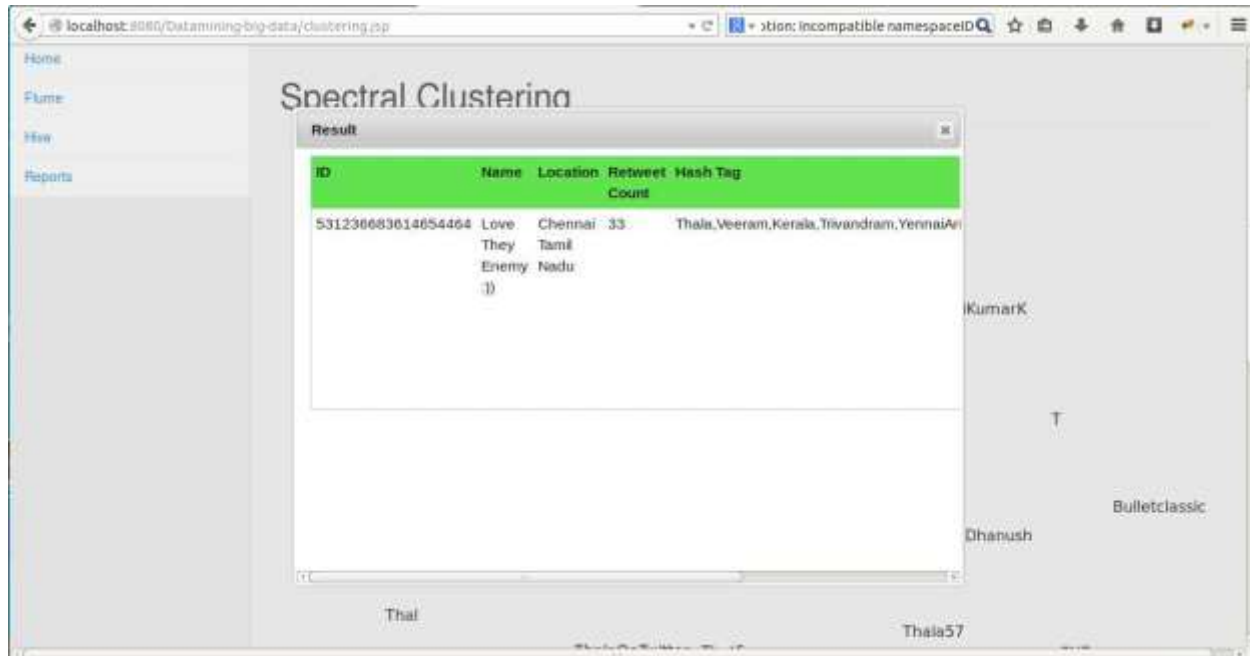
Trendsmap kingofOpening HariKumarK

Gopl bachemeariefjalanjalan Bulletclassic

Kerala Beijing T

WhenStrangersMeet WeWantThalainTwitter Tha Anushka Dhanush

Thal Thala57



5. CONCLUSION

Driven by real-world applications and key industrial stakeholders and initialized by national funding agencies, managing and mining Big Data have shown to be a challenging yet very compelling task. While the term Big Data literally concerns about data volumes, our HACE theorem suggests that the key characteristics of the Big Data are 1) huge with heterogeneous and diverse data sources, 2) autonomous with distributed and decentralized control, and 3) complex and evolving in data and knowledge associations. Such combined characteristics suggest that Big Data require a “big mind” to consolidate data for maximum values [27].

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are

required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along

with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future.

We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

6. REFERENCES

- [1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 603-630, Dec. 2012.
- [2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 707-734, Dec. 2012.
- [3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," *Science*, vol. 337, pp. 337-341, 2012.
- [4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," *ACM Crossroads*, vol. 19, no. 1, pp. 20-23, 2012.
- [5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 523-547, Dec. 2012.
- [6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," *Nature*, vol. 489, pp. 49-51, 2012.
- [7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *J. Computational Science*, vol. 2, no. 1, pp. 1-8, 2011.
- [8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," *Science*, vol. 323, pp. 892-895, 2009.
- [9] J. Bughin, M. Chui, and J. Manyika, *Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch*. McKinsey Quarterly, 2010.
- [10] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, vol. 329, pp. 1194-1197, 2010.
- [11] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," *Proc. 17th ACM Int'l Conf. Multimedia, (MM '09)*, pp. 917-918, 2009.
- [12] R. Chen, K. Sivakumar, and H. Kargupta, "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data," *Knowledge and Information Systems*, vol. 6, no. 2, pp. 164-187, 2004.
- [13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks,"

Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

[14] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06), pp. 281-288, 2006.

[15] G. Cormode and D. Srivastava, "Anonymized Data: Generation, Models, Usage," Proc. ACM SIGMOD Int'l Conf. Management Data, pp. 1015-1018, 2009.

[16] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J. McPherson, "Ricardo: Integrating R and Hadoop," Proc. ACM SIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998, 2010.

[17] P. Dewdney, P. Hall, R. Schilizzi, and J. Lazio, "The Square Kilometre Array," Proc. IEEE, vol. 97, no. 8, pp. 1482-1496, Aug. 2009.

[18] P. Domingos and G. Hulten, "Mining High-Speed Data Streams," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '00), pp. 71-80, 2000.

[19] G. Duncan, "Privacy by Design," Science, vol. 317, pp. 1178-1179, 2007.

[20] B. Efron, "Missing Data, Imputation, and the Bootstrap," J. Am. Statistical Assoc., vol. 89, no. 426, pp. 463-475, 1994.

[21] A. Ghoting and E. Pednault, "Hadoop-ML: An Infrastructure for the Rapid Implementation of Parallel Reusable Analytics," Proc. Large-Scale Machine Learning: Parallelism and Massive Data Sets Workshop (NIPS '09), 2009.

[22] D. Gillick, A. Faria, and J. DeNero, MapReduce: Distributed Computing for Machine Learning, Berkley, Dec. 2006.

[23] M. Helft, "Google Uses Searches to Track Flu's Spread," The New York Times, <http://www.nytimes.com/2008/11/12/technology/internet/12flu.html>. 2008.

[24] D. Howe et al., "Big Data: The Future of Biocuration," Nature, vol. 455, pp. 47-50, Sept. 2008.

[25] B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience," Nature, vol. 482, p. 308, 2012.

[26] "IBM What Is Big Data: Bring Big Data to the Enterprise," <http://www-01.ibm.com/software/data/bigdata/>, IBM, 2012.

[27] A. Jacobs, "The Pathologies of Big Data," Comm. ACM, vol. 52, no. 8, pp. 36-44, 2009.

[28] I. Kopanas, N. Avouris, and S. Daskalaki, "The Role of Domain Knowledge in a Large Scale Data Mining Project," Proc. Second Hellenic Conf. AI: Methods and Applications of Artificial Intelligence, I.P. Vlahavas, C.D. Spyropoulos, eds., pp. 288-299, 2002.

[29] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data," Proc. VLDB Endowment, vol. 5, no. 12, 2032-2033, 2012.

[30] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," J. Cryptology, vol. 15, no. 3, pp. 177-206, 2002.

[31] W. Liu and T. Wang, "Online Active Multi-Field Learning for Efficient Email Spam Filtering," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 117-136, Oct. 2012.

[32] J. Lorch, B. Parno, J. Mickens, M. Raykova, and J. Schiffman, "Shoroud: Ensuring Private Access to Large-Scale Data in the Data Center," *Proc. 11th USENIX Conf. File and Storage Technologies (FAST '13)*, 2013.

[33] D. Luo, C. Ding, and H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining," *Proc. IEEE 12th Int'l Conf. Data Mining*, pp. 489-498, 2012.

[34] J. Mervis, "U.S. Science Policy: Agencies Rally to Tackle Big Data," *Science*, vol. 336, no. 6077, p. 22, 2012.