Data Mining with Big Data

Greeshma G S
PG scholar in CSE
Younus College of Engineering and Technology,
Vadakkevila, Kollam, Kerala, India, pin: 691010

Mr. Arun Raj S
Asst. Professor, IT
Younus College of Engineering and Technology,
Vadakkevila, Kollam, Kerala, India, pin: 691010

***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.**

I.INTRODUCTION

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.

Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. As another example, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real-time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: “a picture is worth a thousand words ,” the billions of pictures on Flicker are a treasure tank for us to explore the human society, social events , public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data.

Due to 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. For example, the square kilometer array (SKA) in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

II.VECTORS OF BIG DATA

We can often describe Big Data Using three V’s including Volume, Velocity and Variety respectively.

1 Volume

Volume refers to the vast amounts of data generated every second. We are not talking about Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between 2008, the same amount of data will soon be generated every minute. New big data tools use distributed systems so that we can store and analyze data across databases that are dotes around anywhere in the world.

2 Velocity

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyze the data while is being generated without ever putting it into databases.

3 Variety

Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured such as text, images, video, voice etc. With big data technology we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, and video or voice recordings.

These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data. In a naï̈ve sense, we can imagine that a number of blind men are trying to size up a giant elephant (see Figure 2.1), which will be the Big Data in this context. The goal of each blind Man is to draw a picture (or conclusion) of the elephant according to the part of information he collects during the process. Because each person's view is limited to his local region, it is not surprising that the blind men will each conclude independently that the elephant "feels" like a rope, a hose, or a wall, depending on the region each of them is limited to. To make the problem even more complicated, let us assume that the elephant is growing rapidly and its pose changes constantly, and each blind man may have his own Information sources that tell him about biased knowledge about the elephant.

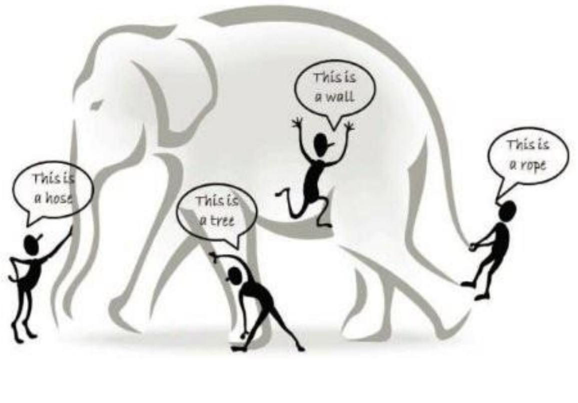

Figure 2.1: The blind men and the giant elephant

Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources to help draw a best possible picture to reveal the genuine gesture of the elephant in a real- time fashion. Indeed, this task is not as simple as asking each blind man to describe his feelings about the elephant and then getting an expert to draw one single picture with a combined view, concerning that each individual may speak a different language (heterogeneous and diverse information sources) and they may even have privacy concerns about the messages they deliberate in the information exchange process.

III.DATA MINING

Data mining refers to extracting knowledge from large amounts of data stored in databases, data warehouses or other information repositories. Some people treat data mining same as Knowledge discovery while some people view data mining essential step in process of knowledge discovery. Here is the list of steps involved in knowledge discovery process.

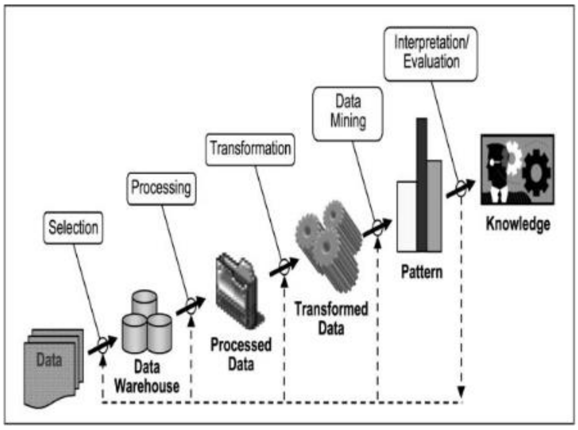

Figure 3.1: Knowledge discovery process

- Data Cleaning - In this step the noise and inconsistent data is removed.
- Data Integration - In this step multiple data sources are combined.
- Data Selection - In this step relevant to the analysis task are retrieved from the database.
- Data Transformation - In this step data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining - In this step intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation - In this step, data patterns are evaluated.
- Knowledge Presentation - In this step, knowledge is represented.

IV. DATA MINING CHALLENGES WITH BIG DATA

For an intelligent learn ing database system to handle Big Data, the essential key is to scale up to the e xceptionally large volu me of data and provide treatments for the characteris tics of big data. Fig. 2 s hows a conceptual vie w of the Big Data processing fra me work, which includes three tiers fro m inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorith ms (Tier III).

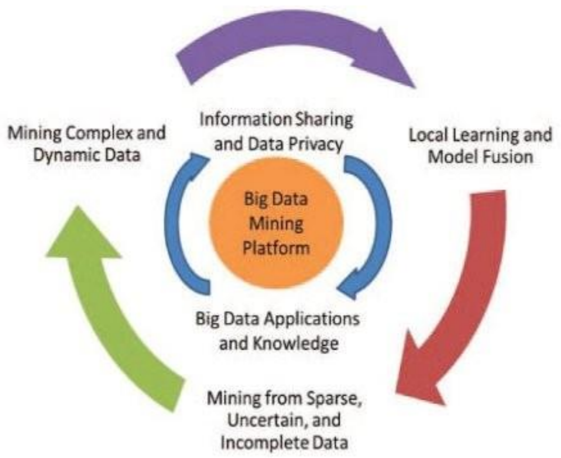

Figure 4.1: A Big Data processing framework

4.1 Tier I: Big Data Mining Platfor m

The challenges at Tier I focus on data accessing and arith metic co mputing procedures. Because Big Data are often stored at different locations and data volumes may continuous ly grow, an effect ive computing platform will have to take distributed large-scale data storage into cons ideration for computing. For e xa mp le, typical data min ing algorith ms require all data to be loaded into the main me mo ry, this , however, is becoming a clear technical barrier for Big Data because moving data across diffe rent locations is expensive, even if we do have a super large ma in me mo ry to hold all data for computing.

In typical data min ing systems, the min ing procedures require computational intensive computing units for data analysis and comparis ons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors. For s ma ll scale data mining tasks, a single desktop computer, wh ich contains hard disk and CPU processors, is sufficient to fulfill the data min ing goals. Indeed, many data mining algorith m are designed for this type of problem settings. For mediu m scale data min ing tasks, data are typically large (and possibly distributed) and cannot be fit into the main me mory. Co mmon solutions are to rely on para llel co mputing or collective min ing to sample and aggregate data fro m different sources and then use parallel computing progra mming to carry out the min ing process.

For Big Data mining, because data scale is far beyond the capacity that a single personal computer (PC) can handle, a typical Big Data processing fra me work will re ly on c luster computers with a high-perfor mance computing platform, with a data mining tas k being deployed by running some parallel programming tools, such as MapReduce or Enterpris e Control Language (ECL), on a la rge number of computing nodes (i.e., c lusters). The role of the s oftware co mponent is to ma ke sure that a single data mining task, such as finding the best match of a query fro m a database with billions of records, is split into many sma ll tas ks each of which is running on one or mu ltip le co mputing nodes.

4.2 Tier II: Big Data Se mantics and Applic ati on Knowle dge

The challenges at Tier II center on semantics and domain knowledge fo r d ifferent Big Data applications . Se mantics and applicat ion knowledge in Big Data re fer to nu merous aspects related to the regulations , policies, us er knowledge, and do main informat ion. 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 (T ier III). For e xa mp le, depending on different doma in applications , the data privacy and informat ion sharing mechanis ms between data producers and data consumers can be significantly diffe rent. The two most important issues at this tier include data sharing and privacy; and doma in and application knowledge.

4.2.1 In formation Sha ring and Data Privacy Information s haring is an u ltimate goal for all s ys tems involving multiple part ies. While the motivation for sharing is clea r, a rea l-world concern is that Big Data applicat ions are related to sens itive informat ion, such as banking transactions and med ical records . Simple data e xchanges or transmis sions do not resolve privacy concerns . For e xa mple, knowing people's locations and their preferences , one can enable a variety of useful location-bas ed services , but public dis clos ure of an individual's locations /movements over time can have serious consequences for privacy. To protect privacy, two co mmon approaches are to res trict acces s to the data, s uch as adding certification or access control to the data entries, so sensitive information is accessible by a limited group of users only, and anonymize data fields such that sensitive information cannot be pinpointed to an individual record. For the first approach, common challenges are to design secured certification or access control mechanisms, such that no sensitive information can be misconducted by unauthorized ind ividuals. For data anonymization, the main objective is to inject randomness into the data to ensure a number of privacy goals. For example, the most common k-anonymity privacy measure is to ensure that each individual in the database must be indis tinguishable from k _ 1 others. Common anonymizat ion approaches are to use suppression, generalization, perturbation, and permutation to generate an altered version of the data, which is, in fact, some uncertain data.

4.2.2 Domain and Application Knowledge

Domain and application knowledge provides ess ential information for designing Big Data mining algorithms and systems. In a simple case, domain knowledge can help identify right features for modeling the underlying data e.g., blood glucose level is clearly a better feature than body mass in diagnosing Type II diabetes. The domain and application knowledge can also help design achievable business objectives by using Big Data analytical techniques.

4.3 Tier III: Big Data Mining Algorithms

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. The circle at Tier III contains three stages. First, s parse, heterogeneous, uncertain, incomplete, and mu ltisource data are preprocessed by data fusion techniques. Second, complex and dynamic data are mined after preprocessing. Third, the global knowledge obtained by local learning and model fus ion is tested and relevant information is fedback to the preprocessing stage. Then, the model and parameters are adjusted according to the feedback. In the whole process, information sharing is not only a promis e of smooth development of each stage, but also a purpose of Big Data processing.

V. RELATED WORKS

5.1 Big Data Mining Platforms (Tier I)

Due to the multisource, massive, heterogeneous, and dynamic characteristics of application data involved in a distributed environment, one of the most important characteristics of Big Data is to carry out computing on the petabyte (PB), even the exabyte (EB)-level data with a complex computing process. Therefore, utilizing a parallel computing infrastructure, its corresponding programming language support, and software models to efficiently analyze and mine the distributed data are the critical goals for Big Data processing to change from “quantity” to “quality.”

Currently, Big Data processing mainly depends on parallel programming models like MapReduce, as well as providing a cloud computing platform of Big Data services for the public. MapReduce is a batch-oriented parallel computing model. There is still a certain gap in performance with relational databases. Improving the performance of MapReduce and enhancing the real-time nature of large-scale data processing have received a significant amount of attention, with MapReduce parallel programming being applied to many machine learning and data mining algorithms. Data mining algorithms usually need to scan through the training data for obtaining the statistics to solve or optimize model parameters. It calls for intensive computing to access the large-scale data frequently. To improve the efficiency of algorithms, Chu propos ed a general-purpose parallel programming method, which is applicable to a large number of machine learning algorithms based on the simple MapReduce programming model on multicore processors. Ten class ical data mining algorithms are realized in the frame work, including locally weighted linear regression, k-Means, logistic regression, naive Bayes, linear s upport vector machines, the independent variable analysis, Gaussian discriminate analysis, expectation maximization, and back- propagation neural networks. With the analysis of these classical machine learn ing algorithms, we argue that the computational operations in the algorithm learning process could be transformed into a summation operation on a number of training data sets. Summation operations could be performed on different subsets independently and achieve penalization executed easily on the MapReduce programming platform. Therefore, a large-scale data set could be divided into several subsets and assigned to multiple Mapper nodes. Then, various s ummation operations could be performed on the Mapper nodes to collect intermediate results. Finally, learn ing algorithms are executed in parallel through merg ing summation on Reduce nodes. Ranger proposed a MapReduce-based application programming interface Phoenix, which supports parallel programming in the environment of multicore and multiprocessors ystems, and realized three data mining algorithms including k-Means, principal component analysis, and linear regression. Gillick improved the MapReduce’s implementation mechanism in Hadoop, evaluated the algorithms' performance of single-pass learning, iterative learning, and query-based learning in the MapReduce framework, studied data sharing between computing nodes involved in parallel learning algorithms, distributed data storage, and then showed that the MapReduce mechanisms suitable for large-scale data mining by testing series of standard data mining tasks on medium-size clusters. Papadimitriou and Sun proposed a distributed collaborative aggregation (DisCo) framework using practical distributed data preprocessing and collaborative aggregation techniques. The implementation on Hadoop in an open source MapReduce project showed that DisCo has perfect scalability and can process and analyze massive data sets (with hundreds of GB). To improve the weak scalability of traditional analysis software and poor analysis capabilities of Hadoop systems, Das conducted a study of the integration of R (open source statistical analysis software) and Hadoop. The in-depth integration pushes data computation to parallel processing, which enables powerful deep analysis capabilities for Hadoop. Wegener achieved the integration of Weka (an open-source machine learning and data mining software tool) and MapReduce. Standard Weka tools can only run on a single machine, with a limitation of 1-GB memory. After algorithm parallelization, Weka breaks through the limitations and improves performance by taking the advantage of parallel computing to handle more than 100-GB data on Map Reduce clusters. Ghoting proposed Hadoop-ML, on which developers can easily build task-parallel or data-parallel machine learning and data mining algorithms on program blocks under the language runtime environment.

5.2 Big Data Semantics and Application Knowledge (Tier II)

In privacy protection of massive data, Ye proposed a multilayer rough set model, which can accurately describe the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criteria in the anonymization process, and designed a dynamic mechanism for balancing privacy and data utility, to solve the optimal generalization/refinement order for classification. A recent paper on confidentiality protection in Big Data summarizes a number of methods for protecting public release data, including aggregation, suppression, data swapping, adding random noise, or simply replacing the whole original data values at a high risk of disclosure with values synthetically generated from simulated distributions. For applications involving Big Data and tremendous data volumes, it is often the case that data are physically distributed at different locations, which means that users no longer physically possess the storage of their data. To carry out Big Data mining, having an efficient and effective data access mechanism is vital, especially for users who intend to hire a third party to process their data. Under such a circumstance, users' privacy restrictions may include no local data copies or downloading and all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many others. In Wang a privacy-preserving public auditing mechanism for large scale data storage (such as cloud computing systems) has been proposed. The public key-based mechanism is used to enable third-party auditing (TPA), so users can safely allow a third party to analyze their data without breaching the security settings or compromising the data privacy.

For most Big Data applications, privacy concerns focus on excluding the third party (such as data miners) from directly accessing the original data. Common solutions are to rely on some privacy-preserving approaches or encryption mechanisms to protect the data. A recent effort by Lorch indicates that users' "data access patterns" can also have severe data privacy issues and lead to disclosures of geographically co-located users or users with common interests. In their system, namely Shround, users' data access patterns from the servers are hidden by using virtual disks. As a result, it can support a variety of Big Data applications, such as micro blog search and social network queries, without compromising the user privacy.

5.3 Big Data Mining Algorithms (Tier III)

To adapt to the multisource, massive, dynamic Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods, designing a data mining mechanism from a multisource perspective, as well as the study of dynamic data mining methods and the analysis of stream data. The main motivation for discovering knowledge from massive data is improving the efficiency of single-source mining methods.

Wu proposed and established the theory of local pattern analysis, which has laid a foundation for global knowledge discovery in multisource data mining. This theory provides a solution not only for the problem of full search, but also for finding global models that traditional mining methods cannot find. Local pattern analysis of data processing can avoid putting different data sources together to carry out centralized computing.

Data streams are widely used in financial analysis, online trading, medical, testing, and so on. Static knowledge discovery methods cannot adapt to the characteristics of dynamic data streams, such as continuity, variability, rapidity, and infinity, and can easily lead to the loss of useful information. Therefore, effective theoretical and technical frameworks are needed to support data stream mining. Knowledge evolution is a common phenomenon in real world systems. For example, the clinician's treatment programs will constantly adjust with the conditions of the patient, such as family economic status, health insurance, the course of treatment, treatment effects, and distribution of cardiovascular and other chronic epidemiological changes with the passage of time. In the knowledge discovery process, concept drifting aims to analyze the phenomenon of implicit target concept changes or even fundamental changes triggered by dynamics and context in data streams. According to different types of concept drifts, knowledge evolution can take forms of mutation drift, progressive drift, and data distribution drift, based on single features, multiple features, and streaming features.

VI.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, the key characteristics of the Big Data are huge with heterogeneous and diverse data sources, autonomous with distributed and decentralized control, and 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.

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.

REFERENCE

[1] "*IBM What Is Big Data: Bring Big Data to the Enterprise*, http://www01.ibm.com/software/data/bigdata/, IBM, 2012.

[2] F. Michel, "*How Many Photos Are Uploaded to Flickr Every Day and Month?*" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[3] A. Rajaraman and J. Ullman, *Mining of Massive Data Sets*. Cambridge Univ. Press, 2011.

[4] C. Reed, D. Thompson, W. Majid, and K. Wagstaff, "*Real Time Machine Learning to Find Fast Transient Radio Anomalies: A Semi-Supervised Approach Combining Detection and RFI Excision*," Proc. Int'l Astronomical Union Symp. Time Domain Astronomy, Sept. 2011.

[5] X. Wu, "*Building Intelligent Learning Database Systems*," AI Magazine, vol. 21, no. 3, pp. 61-67, 2000.

[6] 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.

[7] J. Shafer, R. Agrawal, and M. Mehta, "*SPRINT: A Scalable Parallel Classifier for Data Mining*," Proc. 22nd VLDB Conf., 1996.

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

[9] 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.

[10] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "*Evaluating MapReduce for Multi-Core and Multiprocessor Systems*," Proc. IEEE 13th Int'l Symp. High Performance Computer Architecture (HPCA '07), pp. 13-24, 2007.

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

[12] S. Das, Y. Sismanis, K.S. Beyer, R. Gemulla, P.J. Haas, and J.McPherson, "*Ricardo: Integrating R and Hadoop*," Proc. ACMSIGMOD Int'l Conf. Management Data (SIGMOD '10), pp. 987-998.2010.

[13] D. Wegener, M. Mock, D. Adranale, and S. Wrobel, "*Toolkit-Based High-Performance Data Mining of Large Data on MapReduce Clusters*," Proc. Int'l Conf. Data Mining Workshops (ICDMW '09),pp. 296-301, 2009.

[14] X. Wu, C. Zhang, and S. Zhang, "*Database Classification for Multi-Database Mining*," Information Systems, vol. 30, no. 1, pp. 71-88, 2005.

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