ISSN 2348 – 9928

Doi:01.0401/ijaict.2014.04.02 Published on 05 (11) 2015

DATA DRIVEN DYNAMIC CORRELATION TECHNIQUE FOR COMPLEXITY REDUCTION IN BIG DATA

Prof. L. Hariharan
Assistant Professor,
Department of Computer Science and Engineering,
M. Kumarasamy College of Engineering,
Karur, Tamilnadu, India

Prof. G. Gurumoorthy
Assistant Professor,
Department of Information Technology,
Shree Venkateshvara Hi-Tech Engineering College,
Gobi, Tamilnadu, India

Abstract— Big Data is huge data sets with sizes above the ability of commonly used software tools to manage, and process the data within a tolerable minimum time. We present common correlation dynamics which reduces difficulty and characterizes the features of the Big Data explosion. This DDDC model involves demand-driven aggregation of information sources, drawing out and analysis, user interest modeling, and security and seclusion considerations.

Keywords— Big Data, Data Mining, Heterogeneity, Autonomous Sources, Complex and Evolving Associations.

I. INTRODUCTION

The era of Big Data has arrived astonishingly in the past few years. Numerous data are produced in the form of documents, chatting messages, audio, video, and applications and they are spread in the web. It will be harder to analyze these enormous data and we need more intricate algorithms and applications for mining these heterogeneous data. Also Big data has property of autonomous sources with complex and evolving relationships. In the internet every day quintillion bytes of data are created and Our capability for data generation has never been so powerful and enormous ever since last few centuries.

The remainder of the paper is structured as follows: In Section 2, we summarize the key challenges for Big Data mining.. in Section 3 we propose a DDDCT (Dynamic Correlation Technique) to process mining with Big Data, Related work is discussed in Section 4, and we conclude the paper in Section 5.

II. DATA MINING CHALLENGES WITH BIG DATA

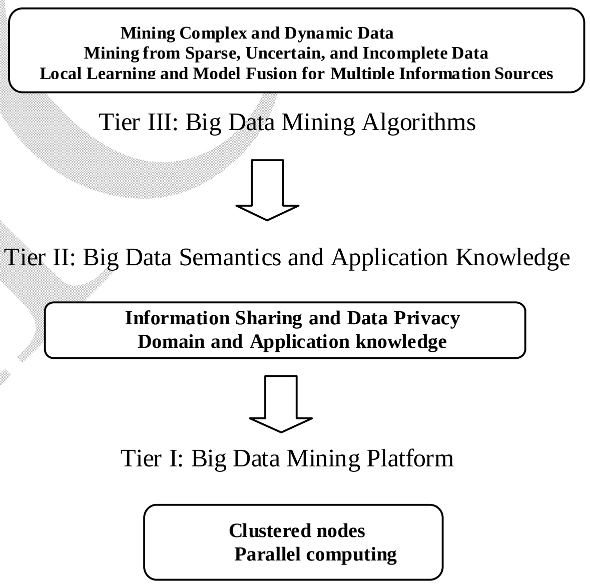
Mining Complex and Dynamic Data
Mining from Sparse, Uncertain, and Incomplete Data
Local Learning and Model Fusion for Multiple Information Sources
Tier III: Big Data Mining Algorithms
Tier II: Big Data Semantics and Application Knowledge
Information Sharing and Data Privacy
Domain and Application knowledge
Tier I: Big Data Mining Platform
Clustered nodes
Parallel computing

Fig 1: A Big Data processing framework: The research challenges form a three tier structure and center around the "Big Data mining platform" (Tier I), which focuses on low-level data accessing and computing. Challenges on information sharing and privacy, and Big Data application domains and knowledge form Tier II, which concentrates on high-level semantics, application domain knowledge, and user privacy issues. The outmost circle shows Tier III challenges on actual mining algorithms an effectual computing platform will have to take dispersed large-scale data storage.

The intelligent learning database system [1] to handle Big Data, the crucial key is to scale up to the extremely large volume of data and provide treatments for the characteristics featured by the aforementioned HACE theorem. Fig. 2 shows a abstract view of the Big Data processing framework, which includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III). The confronts at Tier I focus on data admittance and arithmetic computing techniques. Big Data are often stored at dissimilar localities and data quantities may continuously grow into consideration for computing. For example, archetypal data mining algorithms require all data to be laden into the main memory, this, however, is becoming a clear technical fence for Big Data because moving data across different locations is posh, even if We do have a large main memory to hold all data for computing. The confronts at Tier II center around semantics and area knowledge for different Big Data

Corresponding Author: Prof. L. Hariharan, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.04.02 Published on 05 (11) 2015

utilities. Such information can offer additional benefits to the mining process, as well as add technical blockades to the Big Data access (Tier I) and mining algorithms (Tier III). For example, depending on different field applications, the data seclusion and information sharing mechanisms among data producers and data consumers can be considerably different. Sharing feeler network data for applications like water quality monitoring may not be discouraged, whereas releasing and sharing movable users' location information is clearly not acceptable for majority, if not all, applications.

In addition to the above privacy issues, the application domains can also offer additional information to profit or guide Big Data mining algorithm designs. For example, in market bin transactions data, each transaction is considered sovereign and the discovered knowledge is typically represented by finding highly connected items, possibly with respect to different sequential and/or spatial restrictions. In a public network, on the other hand, users are linked and share dependency structures. The information is then represented by user communities, fortunate in each group, and social influence modeling, and so on. Therefore, understanding semantics and application information is important for both low-level data access and for eminent mining algorithm designs.

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, sparse, heterogeneous, uncertain, incomplete, and multisource 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 fusion is tested and relevant information is fed back 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 promise of smooth development of each stage, but also a purpose of Big Data processing. In the following, we elaborate challenges with respect to the three tier framework in Fig. 1.

2.1 Tier I: Big Data Mining Platform

In typical data mining systems, the mining procedures require computational rigorous computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient right of entry to, at least, two types of resources: data and computing processors. For small scale data mining tasks, a lone desktop computer, which contains hard disk and CPU processors, is sufficient to complete the data mining goals. Indeed, many data mining algorithm are intended for this type of problem settings. For medium scale data mining tasks, data are characteristically large (and possibly distributed) and cannot be fit into the main memory. General solutions are to rely on parallel computing [2], [3] or collective mining [12] to sample and cumulative data from different sources and then use parallel computing programming (such as the Message Passing Interface) to carry out the mining process.

For Big Data mining, because data scale is far beyond the ability that a single personal computer (PC) can handle, a archetypal Big Data processing structure will rely on group computers with a high-performance computing platform, with a data mining task being deployed by running some parallel programming tools, such as MapReduce or Enterprise Control Language (ECL), on a large number of computing nodes (i.e., clusters). The role of the software component is to create sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or numerous computing nodes. For example, as of this writing, the world most influential super computer Titan, which is deployed at Oak Ridge National Laboratory in Tennessee, contains 18,688 nodes each with a 16-core CPU.

Such a Big Data system, which blends both hardware and software components, is barely available without key engineering stockholders' support. In fact, for decades, companies have been making trade decisions based on transactional data stored in relational databases. Big Data mining offers opportunities to go ahead of traditional relational databases to rely on less structured data: weblogs, social media, e-mail, sensors, and photographs that can be mined for helpful information. Major commerce intelligence companies, such IBM, Oracle, Tera data, and so on, have all marked their own products to help customers obtain and organize these diverse data sources and coordinate with customers' obtainable data to find new insights and capitalize on hidden relationships.

2.2 Tier II: Big Data Semantics and Application Knowledge

Semantics and function knowledge in Big Data refer to many aspects related to the regulations, policies, user knowledge, and area information. The two most significant issues at this tier include 1) data sharing and privacy; and 2) field and application knowledge. The former provides answers to determine concerns on how data are maintained, accessed, and shared; whereas the latter focuses on answering questions like "what are the beneath- lying applications and "what are the knowledge or patterns users mean to discover from the data "

2.2.1 Information Sharing and Data Privacy

Information sharing is an ultimate goal for all systems involving manifold parties. While the motivation for sharing is clear, a real-world apprehension is that Big Data applications are related to responsive information, such as banking transactions and medical records. Simple data exchanges or transmissions do not determine privacy concerns [19], [4], [5]. For example, knowing people's locations and their preferences, one can allow a variety of helpful location-based services, but public disclosure of an individual's locations/movements over time can have grave consequences for privacy. To protect privacy, two common approaches are to 1) limit access to the data, such as adding certification or admission control to the data entries, so sensitive information is accessible by a limited collection of users only, and 2) anonymize data fields such that sensitive information cannot be located to an individual record [15]. For the first approach, common challenges are to intend secured certification or access manage mechanisms, such that no responsive information can be misconduct by unauthorized individuals. For data anonymization, the major objective is to inject randomness into the data to make sure a number of solitude goals. For example, the most common k-anonymity privacy gauge is to ensure that each individual in the database must be indistinguishable from k — 1 others. Common anonymization approaches are to use suppression, generalization, perturbation, and variation to generate an altered account of the data, which is, in fact, some uncertain data.

One of the main benefits of the data anomization-based information sharing approaches is that, once anonymized, data can be freely shared crossways different[20]bash without involving preventive access controls. This naturally leads to one more explore area namely privacy protecting data mining [18], where manifold parties, each holding some amenable data, are trying to attain a ordinary data mining object without sharing any responsive information within the data. This solitude preserving mining goal, in practice, can be solved through two types of approaches including 1) using particular message protocols, such as Yao's protocol , to request the distributions of the entire data set, rather than requesting the real values of each record, or 2) designing particular data mining methods to increase knowledge from anonymized data (this is inherently similar to the unsure data mining methods).

2.2.2 Domain and Application Acquaintance

Sphere and claim knowledge provides essential information for scheming Big Data mining algorithms and systems. In a easy case, domain knowledge can help identify right features for copy the underlying data (e.g., blood glucose level is clearly a better feature than body accumulation in diagnosing Type II diabetes). The area and application knowledge can also help plan attainable business objectives by using Big Data analytical techniques.

2.3 Tier III: Big Data Mining Algorithms

2.3.1 Limited Learning and Model Fusion for Multiple Information Sources

As Big Data applications are featured with autonomous springs and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically exorbitant due to the potential transmission charge and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each place often leads to biased decisions or models, just like the elephant and blind men case [21], [22].

Under such a circumstance, a Big Data mining system has to allow an information exchange and fusion mechanism to ensure that all distributed sites (or information sources) can work jointly to achieve a global optimization goal. Model mining and correlations are the key steps to ensure that models or patterns discovered from manifold information sources can be consolidated to meet the global mining objective. By exchanging prototypes between multiple sources, new global patterns can be synthetized by aggregating patterns crosswise all sites. At the knowledge level, model association analysis investigates the significance between models generated from different data sources to decide how relevant the data sources are correlated with each other, and how to form precise decisions based on models[23], [24].

2.3.2 Mining from Sparse, Uncertain, and Incomplete Data.

Spare, uncertain, and incomplete data are decisive features for Big Data applications. Being sparse, the number of data points is too few for drawing steadfast conclusions. This is normally a complication of the data dimensionality issues, where data in a high-dimensional space do not show clear trends or distributions. For most machine learning and data mining algorithms, high-dimensional spare data significantly decline the reliability of the models derived from the data. Common approaches are to use dimension reduction or feature selection [6] to reduce the data dimensions or to carefully include additional samples to lessen the data scarcity, such as broad unsupervised learning methods in data mining. Uncertain data are a special type of data realism where each data field is no longer deterministic but is theme to some random/error distributions. This is mainly linked to area specific applications with inaccurate data readings and collections. For example, data produced from GPS tools are inherently uncertain, mainly because the knowledge barrier of the device limits the precision of the data to convinced levels.

Each recording location is represented by a mean value plus a inconsistency to indicate expected errors. For data solitude related applications, users may intentionally insert randomness/errors into the data to stay anonymous. For uncertain data, the major confront is that each data item is represented as model distributions but not as a single value, so most existing data mining algorithms cannot be straight applied. Common solutions are to take the data distributions into deliberation to estimate model parameters. For example, error alert data mining [9] utilizes the mean and the discrepancy values with respect to each single data item to construct a Naïve Bayes model for classification. Similar advances have also been applied for decision trees or database queries.

Incomplete data refer to the missing of data field values for some samples. The missing values can be caused by different realities, such as the breakdown of a sensor node, or some systematic policies to intentionally hop some values While most modern data mining algorithms have inherent solutions to handle missing values (such as ignoring data fields with missing values), data accusation is an established research field that seeks to impute missing values to produce improved models (compared to the ones built from the original data). Many imputation methods [20] exist for this purpose, and the main approaches are to fill most frequently observed values or to construct learning models to predict possible values for each data field, based on the observed values of a given instance.

III. DCT – DYNAMIC CORRELATION TECHNIQUE

New active models for realized covariance matrices are proposed. The expected value of the recognized covariance matrix is specified in two steps: a model for each realized variance, and a model for the realized correlation matrix. The realized correlation model is a dynamic provisional correlation model. Assessment can be done in two steps as well, and a QML understanding is provided to each step, by assuming a Wishart restricted distribution. Moreover, the model is appropriate to large matrices since judgment can be done by the compound likelihood method.

3.1 Common Correlation Dynamics: Scalar Models

The most economical specification that we propose imposes a scalar dynamic equation on the conditional correlation matrix. A possible dynamic modernized equation for Rt, inspired by that of Tse and Tsui (2002) for multivariate GARCH models, is set by

$Rt = (1 - \alpha - \beta)R^{-} + \alpha Pt-1 + \beta Rt-1$, where

$Pt = \{diag(Ct)\}-1/2\ Ct\{diag(Ct)\}-1/2$

is the observed correlation matrix at time t. The factors α and β, and their sum, are controlled to lie between zero and one.

The matrix R^{-} is a factors that must satisfy the constraints of a correlation matrix, i.e. positive unambiguous symmetric with unit diagonal elements. Because Pt has unit diagonal elements, Rt is a fine defined correlation matrix for all t if initial matrix R0 is a correlation matrix.

IV. BIG DATA SEMANTICS AND APPLICATION KNOWLEDGE (TIER II)

In privacy protection of enormous data, Ye et al. [10] proposed a multilayer uneven set model, which can accurately explain the granularity change produced by different levels of generalization and provide a theoretical foundation for measuring the data effectiveness criterion in the anonymization process, and designed a active mechanism for balancing privacy and data utility, to explain the optimal generalization/refinement order for classification.

For applications involving Big Data and tremendous data volumes, it is often the case that data are actually distributed at different locations, which means that users no longer physically own the storage of their data. To carry out Big Data mining, having an competent and effective data access mechanism is vital, especially for users who mean to hire a third party (such as data miners

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.04.02 Published on 05 (11) 2015

or data auditors) to process their data. Under such a circumstance, users' solitude restrictions may comprise 1) no local data copies or downloading, 2) all analysis must be deployed based on the existing data storage systems without violating existing privacy settings, and many others. In Wang et al. [12], a privacy-preserving community auditing mechanism for large level data storage (such as cloud computing systems) has been proposed. The public key-based device is used to enable third-party auditing (TPA), so users can securely allow a third party to analyze their data without breaking the security settings or compromising the data privacy.

For most Big Data applications, privacy concerns center on excluding the third party (such as data miners) from straight accessing the original data. Common solutions are to rely on some privacy-preserving approaches or encryption mechanisms to defend the data.

4.1 Big Data Mining Algorithms (Tier III)

To adapt to the multisource, massive, active Big Data, researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source information discovery methods [13], designing a data mining method from a multisource perspective [14], [15], as well as the study of dynamic data mining methods and the examination of stream data [16], [17]. The main incentive for discovering knowledge from massive data is improving the efficiency of single-source mining methods. On the basis of plodding improvement of computer hardware functions, researchers continue to discover ways to improve the efficiency of knowledge detection algorithms to make them better for massive data. Because massive data are typically collected from different data sources, the knowledge detection of the massive data must be performed using a multisource mining mechanism. As real-world data often come as a data torrent or a characteristic flow, a well-established method is needed to discover knowledge and master the development of knowledge in the dynamic data source. Therefore, the massive, heterogeneous and real-time characteristics of multisource data provide necessary differences between single-source knowledge discovery and multisource data mining.

V. CONCLUSIONS

At the data level, the autonomous information sources and the variety of the data compilation 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 create altered data copies. Developing a safe and sound information sharing procedure is a major challenge. At the model level, the key confront is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to examine model correlations between distributed sites, and combine decisions from multiple sources to gain a best model out of the Big Data. At the organization 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 likely factors. A system needs to be carefully designed so that formless data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should assist form legitimate patterns to predict the trend and future.We regard Big Data as an emerging tendency 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 give most relevant and most precise social sensing feedback to better understand our civilization at real-time. We can further rouse the participation of the public audiences in the data production circle for societal and inexpensive events.

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. Mehta, 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. Multi- media, (MM '09,) pp. 917-918, 2009.

ISSN 2348 – 9928
Doi:01.0401/ijaict.2014.04.02 Published on 05 (11) 2015

[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 Kilometer 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] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding "Data Mining with Big Data", IEEE Trans. Knowledge And Data Engineering, vol. 26, no. 1, pp 97-107, JAN 2014

[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.