International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Intrusion Detection Using Navie Bayes by Analyzing Big Data

B. Geetha Kumari, Jageti Padmavthi
Assistant Professor, Department of CSE, G. Narayanamma Institute of Technology and Science, Hyderabad, Telangana, India

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.

Keywords : Big Data Processing, Data-Driven Model , HACE, WBG Big Data, LHC, IDNB

I. INTRODUCTION

Every day, we create 2.5 quintillion bytes of data - so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few such colossal amount of data that is being produced continuously is what can be coined as Big Data. Big Data decodes previously untouched data to derive new insight that gets integrated into business operations. However, as the amounts of data increases exponential, the current techniques are becoming obsolete. Dealing with Big Data requires comprehensive coding skills, domain knowledge and statistics.

Despite being Herculean in nature, Big Data applications are almost ubiquitous- from marketing to scientific research to customer interests and so on. We can witness Big Data in action almost everywhere today. From Facebook which handles over 40 billion photos from its user base to CERN‘s Large Hydron Collider (LHC) which generates 15PB a year to Walmart which handles more than 1 billion customer transactions in an hour. Over a year ago, the World Bank organized the first WBG Big Data Innovation Challenge which brought forward several unique ideas applying Big Data such as big data to predict poverty and for climate smart agriculture and fore user-focused Identification of Road Infrastructure Condition and safety and so on.

Big Data:

In this digital era, analysts have enormous amounts of data available on hand. Big Data is the term for a collection of unstructured, semi-structured and structured datasets whose volume, complexity and rate of growth make them difficult to be captured, managed, processed or analyzed by using the typical database software tools and technologies. Different varieties are in the form of text, video, image, audio, webpage log files, blogs, tweets, location information, sensor data etc. Discovering useful insight from such huge datasets requires smart and scalable analytics services, programming tools and applications. In the name only we can define that Big Data is having the Large amount of data such as peta bytes of data. Big Data is not a solution for the all types of problems. Big Data Analytics is the solution for all types of problems in data mining. Every sorts of data are not a Big Data. Big data is having the characteristics such as Velocity,

Variety, Veracity and Volume and it is also defined as 4V's, the 4V's are Velocity, Variety, Veracity and Volume.

Volume: This essentially concerns the large quantities of data that is generated continuously. Initially storing such data was problematic because of high storage costs. However with decreasing storage costs, this problem has been kept somewhat at bay as of now. However this is only a temporary solution and better technology needs to be developed. Smartphones, E-Commerce and social networking websites are examples where massive amounts of data are being generated. This data can be easily distinguishes between structured data, unstructured data and semi-structured data.

Velocity: In what now seems like the pre-historic times, data was processed in batches. However this technique is only feasible when the incoming data rate is slower than the batch processing rate and the delay is much of a hindrance. At present times, the speed at which such colossal amounts of data are being generated is unbelievably high. Take Facebook for example it generates 2.7 billion like actions per day and 300 million photos amongst others roughly amounting to 2.5 million pieces of content in each day while Google Now processes over 1.2 trillion searches per year worldwide.

Variety: Veracity is one of the best most important technology trend in the Big Data. In relational database all data is well formatted. It is in form of structred. Structured data means well defined data with using group of rules such as name in the format of text, date in date format, amount should be in the form of numerical and having two decimals. Unstructured data is the main important concept in the Big data. Unstructured data means it is not in the well defined format such as images, audio files, video files, a tweet these all are different but people can express their ideas and thoughts. One of the important goal of the big data is to make sense of unstructured data by analyzing it.

Veracity: Trustworthy of the data is called as veracity. It is also refer as abnormality, noise and biases in data. Veracity is the biggest challenge compare with velocity and volume in Big Data. We should check the process of data is done perfectly and mine the meaningful data to achive the veracity in big data analysis.

Big Data processing, as the information comes from multiple, heterogeneous, autonomous sources with complex and evolving relationships, and keeps growing. Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages. The proposed shows 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 modelling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution. We proposed a new intrusion classification scheme which can effectively improve the classification performance in the situation that only few training data are available.

The proposed scheme is able to incorporate flow correlation information in to the classification process. Compared to contemporary approaches, IDNB (Intrusion Detection using Naive Bayes) demonstrates higher malicious behaviour detection rates in certain circumstances while does not greatly affect the network performances. NB is one of the earliest classification methods applied in intrusion detection system which is an effective probabilistic classifier employing the Bayes' theorem with naive feature independence assumptions. Furthermore, for each communication process, both the source and the destination are not malicious. NB classifier is that it only requires a small amount of training data to estimate the parameters of a classification model.

II. RESEARCH

1. An Automatic Extraction of Top-k Lists from the Web

Important source of structured information on the web is links. This paper is concerned with "top-k list" pages, which are web pages that specify a list of k instances of a particular query. Examples include "the 10 tallest building in the world" and "the top 20 best cricket players in India". We present an efficient algorithm that fetch the target lists with great accuracy even when the input pages contain other non-useful data of the same size or errors. The extraction of such lists can help develop existing knowledge bases about general consideration.

III. AUTOMATIC EXTRACTION OF DATA FROM DEEP WEB PAGE

There is large amount of information accessible to be mined from the World Wide Web. The information on the Web is in the form of structured and unstructured objects, which is known as data records. Such data records are necessary because essential information are available in these pages, e.g. lists of products and there detail information. It is important to extract such data records to provide proper information to user as per their concern. Manual approach, supervised learning, and automatic techniques are used to solve these problems. The manual method is not relevant for huge number of pages. It is a challenging work to retrieve appropriate and beneficial information from Web pages. Presently, numbers of web retrieval systems called web wrappers, web crawler have been invented. In this paper, some current techniques are inspected, then our work on web information extraction is presented. Experimental analysis on large number of real input web URL address selections indicates that our algorithm properly extracts data in most cases.

2. Survey on web mining techniques for Extraction of top k list

Today finding proper result within less time is important need but one more problem is that very poor percentage information available on web is useful and interpretable and which consumes lot of time to extract. The method for extracting information from top k web pages which contains top k instances of interested topic needed to deals with system. In contrast with other structured data like web tables Information in top-k lists contains valuable and exact information of rich, and interesting. Therefore top-k list are of higher quality as it can help to develop open domain knowledge bases to applications such as search for truth result.

3. Extracting general from web document:

In this paper, author proposed a new different technique for extraction of general lists from the web. Method uses basic premises on visual rendering of list and structural arrangement of items. The aim of system was to minimize the restrictions of existing work which deals with the principle of extracted lists. Several visual and structural features were combined for obtaining goal.

IV. EXISTING SYSTEM

- Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.
- Exploring the Big Data in this scenario is equivalent to aggregating heterogeneous information from different sources (blind men) to help draw a best possible picture to reveal the genuine gesture of the giant in a real-time fashion.
- Existing system shows a HACE [Heterogeneous autonomous Complex evolving] theorem that characterizes the features of the Big Data revolt, and implement a Big Data processing model.

LIMITATIONS / DISADVANTAGES

- Big Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns.
- Complex dependency structures underneath the data raise the difficulty for existing learning systems; they also offer exciting opportunities that simple data representations are incapable of achieving.
- Big Data complexity is represented in many aspects, including complex heterogeneous data types, complex intrinsic semantic associations in data, and complex relationship networks among data. That is to say, the value of Big Data is in its complexity.

PROPOSED SYSTEM

- In proposed system we implemented IDNB (intrusion detection by Naive bayes) for Big Data scheme. The main purpose of this implementation is to detect intrusion packets or data for increases the performance of Big Data Processing model.
- The proposed scheme is able to incorporate flow correlation information in to the classification process. IDNB (Intrusion Detection using Naive Bayes) demonstrates higher malicious behaviour detection rates in certain circumstances while does not greatly affect the network performances.
- NB is one of the earliest classification methods applied in intrusion detection system which is an effective probabilistic classifier employing the Bayes' theorem with naive feature independence assumptions. NB classifier is that it only requires a

small amount of training data to estimate the parameters of a classification model.

ADVANTAGES

- Detect intrusion packets data in client side when large complex data arrived.
- To minimize the effort of handling large complex data by using specialized tool used for securing network and checking available service.
- Provide security against hackers, malicious software, Denial of services.
- NB with feature discretization demonstrates not only significantly higher accuracy but also much faster classification speed.
- NB-based traffic classifier improves classification with a small set of training samples.

V. IMPLIMENTATION

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 integra-tion, 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 informa-tion 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 regula-rities 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.

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, 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.

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.

VII. 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. 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.

[14]. 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.

[15]. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," Science, vol. 337, pp. 337-341, 2012.

[16]. A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data," ACM Crossroads, vol. 19, no. 1, pp. 20-23, 2012.

[17]. 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.

[18]. E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," Nature, vol. 489, pp. 49-51, 2012.

[19]. J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," J. Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[20]. S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," Science, vol. 323, pp. 892-895, 2009.

[21]. J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. McKinSey Quarterly, 2010.

[22]. D. Centola, "The Spread of Behavior in an Online Social Network Experiment," Science, vol. 329, pp. 1194-1197, 2010.

[23]. 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.

[24]. 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.

[25]. 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.