International Conference on "Recent Trends in Technology and its Impact on Economy of India"
Guru Nanak College for Girls, Sri Mukstar Sahib, Punjab (India) (ICRTTIEI-17)
24th October 2017, www.conferenceworld.in ISBN: 978-93-86171-74-0

A Review on Big Data and Its Various Constraints (Phishing Attack)

Er. Himanshi[1], Er. Dhruv Kumar Sharma[2]

ABSTRACT

'Big Data' is the application of specialized techniques and technologies to process very large sets of data. These data sets are often so large and complex that it becomes difficult to process using on-hand database management tools. The tendency of large data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separated smaller sets with the same total amount of data, allowing associations to be found to spot business trends, prevent diseases battle crime and so on. The challenges in Big Data can be broadly divided in the two categories: Engineering and Semantic. Engineering challenges include data management activities such as query, and storage efficiently. Semantics include what should be the structure of the data. Big Data is slowly becoming ubiquitous. Every area of business, health or general living standards now can implement big data analytics. Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one's advantage.

Keyword*: *Big Data, Malicious, Ripper

I. INTRODUCTION

'Big Data' is the application of specialized techniques and technologies to process very large sets of data. These data sets are often so large and complex that it becomes difficult to process using on-hand database management tools. Examples include web logs, call records, medical records, military surveillance, photography archives, video archives and large-scale e-commerce.

The amount of data generated every day in the world is massive. The increasing volume of digital and social media and internet of things, is fuelling it even further. The rate of data growth is surprising and this data comes at a speed, with variety (not necessarily structured) and contains wealth of information that can be a key for gaining an edge in competing businesses. Ability to analyze this enormous amount of data is bringing a new era of productivity growth, innovation and consumer surplus. "Big data is the term for a collection of data sets so large and complex that it becomes difficult to process it using traditional database management tools (i.e. RDBMS, it challenging to handle such huge data volume, it can't categorized unstructured data, it lacks in high velocity because it's designed for fixed data holding rather than fast growth, even if RDBMS is used handle and store "Big data", it will turn out to be very expensive) or data processing application. The challenges include the areas of capture, curation, storage, search, sharing, transfer, analysis, and visualization of this data.

With the above-mentioned attributes of big data, data is massive, comes at a speed and highly unstructured that it doesn't fit conventional relational database structures. With so much insight hidden in this data, an alternative way to process this enormous data is necessary. Big corporations could be well resourced to handle this task but

ISBN: 978-93-86171-74-0

the amount of data being generated every day easily outgrows this capacity. Cheaper hardware, cloud computing and open source technologies have enabled processing big data at a much cheaper cost. Lot of data means lot of hidden insights. The ability to quickly analyze big data means the possibility to learn about customers, market trends, marketing and advertising drives, equipment monitoring and performance analysis and much more. And this is an important reason that many big enterprises are in a need of robust big data analytics tools and technologies.

II. WHY LARGE DATA?

The tendency of large data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separated smaller sets with the same total amount of data, allowing associations to be found to spot business trends, prevent diseases battle crime and so on.

a. Data Sets Grow In Size

Big data sets are increasingly being gathered by abundant information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identifications (RFID) readers, and wireless sensor networks. The challenge for large enterprises is determining who should own big data initiatives that overlap the entire organization.

b. **How to Handle**

Big data is difficult to work with using most relational database management system and desktop statistics and visualization packages, requiring instead the hugely parallel software running on tens, hundreds, or even thousands of servers. Big data differs to depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyse the data set in its domain. Some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to re-evaluate data management options. For others, it may take tens or hundreds of terabytes before data size becomes an important thought.

III. CHARACTERISTICS OF BIG DATA

Big Data is important because it enables organizations to gather, store, manage, and manipulate vast amounts data at the right speed, at the right time, to gain the right insights. Big data generators must create scalable data (Volume) of different types (Variety) under controllable generation rates (Velocity), while maintaining the important characteristics of the raw data (Veracity), the collected of data can bring to the intended process, activity or predictive analysis/hypothesis. Therefore, these four characteristics have been used to define Big Data, also known as 4 V's (volume, velocity, variety and veracity). The following sub clauses go into further depth on these characteristics :

ISBN: 978-93-86171-74-0

Figure 1: 4 V's in big data

- **Volume:**The name 'Big Data' itself is related to a size which is vast. Size of data plays very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon volume of data. Hence, 'Volume' is one characteristic which needs to be considered while dealing with 'Big Data'.
- **Variety:**Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Now days, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. is also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.
- **Velocity:**The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks and social media sites, sensors, mobile devices, etc. The flow of data is massive and continuous.
- **Variability**: This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

IV. WHY BIG DATA SHOULD MATTER TO YOU

The real issue is not that you are obtaining large amounts of data. It is what you do with the data that counts. The hopeful vision is that establishments will be able to take data from any source, harness relevant data and analyze it to find answers that enable to

- Cost reductions
- Time reductions.
- New product development and optimized offerings.
- Smarter business decision making.

The combining of big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.

ISBN: 978-93-86171-74-0

- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use click stream analysis and data mining to detect fraudulent behavior.

V. CHALLENGES OF BIG DATA

The challenges in Big Data can be broadly divided in the two categories: Engineering and Semantic. Engineering challenges include data management activities such as query, and storage efficiently. The semantic challenge is determining the meaning of information from large volumes of unstructured data. The analysis of Big Data involves multiple distinct phases which includes data acquisition and recording, information exaction and cleaning, data integration, aggregation and representation, query processing, data modeling and Interpretation. Each of these phases introduces challenges. Heterogeneity, scale, timeliness, complexity and privacy are certain challenges of big data mining.

• **Heterogeneity and Incompleteness:** The difficulties of big data analysis derive from its large scale as well as the presence of mixed data based on different patterns or rules (heterogeneous mixture data) in the collected and stored data. In the case of complicated heterogeneous mixture data, the data has several patterns and rules and the properties of the patterns vary greatly. Data can be both structured and unstructured. It may exists in the form of email attachments, images, pdf documents, medical records, X rays, voice mails, graphics, video, audio etc. and they cannot be stored in row/ column format as structured data.Incomplete data creates uncertainties during data analysis and it must be managed during data analysis. Doing this correctly is also a challenge. Incomplete data refers to the missing of data field values for some samples. The missing values can be caused by different realities, such as the malfunction of a sensor node, or some systematic policies to intentionally skip some values.

• **Scale and Complexity:** Managing large and rapidly increasing volumes of data is a challenging issue. Traditional software tools are not enough for managing the increasing volumes of data. Data analysis, organization, retrieval and modeling are also challenges due to scalability and complexity of data that needs to be analyzed.

• **Timeliness:** As the size of the data sets to be processed increases, it will take more time to analyze. In some situations results of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed by preventing the transaction from taking place at all. A large data set, it is often necessary to find elements in it that meet a specified criterion. Scanning the entire data set to find suitable elements is obviously impractical. In such cases Index structures are created in advance to permit finding qualifying elements quickly. The problem is that each index structure is designed to support only some classes of criteria.

ISBN: 978-93-86171-74-0

VI. APPLICATIONS OF BIG DATA

Big Data is slowly becoming ubiquitous. Every area of business, health or general living standards now can implement big data analytics. Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one's advantage. The major applications of Big Data have been listed below:

Figure 2: Application of Big Data

- **In Banking:** The use of customer data invariably raises privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information. Research indicates that 62% of bankers are cautious in their use of big data due to privacy issues. Further, outsourcing of data analysis activities or distribution of customer data across departments for the generation of richer insights also amplifies security risks. Such as customers' earnings, savings, mortgages, and insurance policies ended up in the wrong hands. Such incidents reinforce concerns about data privacy and discourage customers from sharing personal information in exchange for customized offers.
- **In Telecom:** Now a day's big data is used in different fields. In telecom, service providers are trying to compete in the cutthroat world of telecom services. Where more and more subscribers rely on over-the-top (OTT) players as providers of value-added services are focused on increasing revenue and enhancing the customer experience as key business objectives.
- **In Health:** Big data analytics has helped healthcare improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solutions. Some areas of improvement are more inspirational than actually implemented.Many health care stakeholders have under invested in information technology because of uncertain returns.The nature of health care industry itself also creates challenges: while there are many players, there is no way to easily share data among different providers or facilities, partly because of privacy concerns. Even within a single hospital, or pharmaceutical company, important information often remains soloed within one group or department because organizations lack procedures for integrating data and communicating findings.

ISBN: 978-93-86171-74-0

- **In Utilities industries:** Big data is a critical element to solving key business problems for utility companies. It can turn the information from smart meter and smart grid projects into meaningful operational insights and understandings about their customer's behavior. As smart grid and smart meters become crucial to the industry, they will likely start generating hundreds of terabytes of data every year and unstructured text data compiled from maintenance records and Twitter feeds. The accuracy, breadth and depth of these new data points present new opportunities for the utility companies that are prepared to take advantage of them.
- **In Automotive industries:** The automotive industry continues to face a growing number of challenges and pressures. Cost pressure, competition, globalization, market shifts, and volatility are all increasing. At the same time, big data and analytics today offer previously unthinkable possibilities for tackling these and many other challenges automakers face. Our global team of automotive specialists has authored this collection of articles with the hopes of sharing what possibilities analytics offers your company and what you should consider when considering an analytics-driven initiative.
- **In Retail:** From traditional brick and mortar retailers and wholesalers to current day e-commerce traders, the industry has gathered a lot of data over time. This data, derived from customer loyalty cards, POS scanners, RFID etc. is not being used enough to improve customer experiences on the whole. Any changes and improvements made have been quite slow.Applications of big data in the Retail and Wholesale industry.Big data from customer loyalty data, POS, store inventory, local demographics data continues to be gathered by retail and wholesale stores.In New York's Big Show retail trade conference in 2014, companies like Microsoft, Cisco and IBMpitched the need for the retail industry to utilize big data for analytics and for other uses including:
  - Optimized staffing through data from shopping patterns, local events, and so on
  - Reduced fraud
  - Timely analysis of inventory

  Social media use also has a lot of potential use and continues to be slowly but surely adopted especially by brick and mortar stores. Social media is used for customer prospecting, customer retention, promotion of products, and more.

VII. MALICIOUS URL

URLs have become a common channel to facilitate Internet criminal activities such as drive-by download, spamming and phishing. Many attackers try to use these URLs for spreading malicious programs or stealing identities. **Kaspersky La. b** reported that browser- based attacks in 2014 increased from 956,393,693 to 1,595,587,670 and 82.36% of these used malicious URLs. The Anti- malicious Working Group (APWG) also reported that malicious attacks using malicious URLs increased from 93,462 to 123,486 in the second half of 2014. There is an urgent need to develop a mechanism to detect malicious URLs from the high volume, high velocity, and high variety data, high veracity data. Many information security companies and organizations offer

ISBN: 978-93-86171-74-0

malicious URL detection including Google's safe browsing and Trend micro's web reputation services. The blacklist is a universal solution for protecting users from the malicious URLs; examples include Phish Tank, SORBS, and URIBL. These services provide a list of malicious URLs reported by volunteers or collected by web crawlers and verified by some reliable back-end system. The content-based analysis service BLADE is also a well-known solution for detecting malicious URLs. These services download the web page content and analyze it for malicious content. The content for analyzing is time-consuming and consumes bandwidth. Therefore, content based analysis services will be integrated with a blacklist or a cache mechanism to improve the performance and avoid re-analyzing the same URL. However, content-based analysis methods are not a practical solution for the large volume of URLs and the speed at which new URLs can be created. The filtering mechanism to be used before the content based analysis to remove the bulk of benign URLs, reducing the volume of URLs on which content-based analysis needs to be performed.

The idea of incremental reduced-error pruning is due to Fürnkranz and Widmer (1994) and forms the basis for fast and effective rule induction. The RIPPER rule learner is due to Cohen (1995),RIPPER (Repeated Incremental Pruning to Produce Error Reduction) is one of the Classification rule algorithm. The RIPPER algorithm is a direct method, i.e. RIPPER extracts the rules directly from the data. The algorithm progresses through four phases: growth, pruning, optimization, selection. In the growth phase, one rule is generated by greedily adding attributes to the rule until the rule meets stopping criteria. In the following prune phase, each rule is incrementally pruned, allowing the pruning of any final sequence of the attributes, until a pruning metric is fulfilled. In the optimization stage each generated rule is further optimized by a) greedily adding attributes to the original rule and b) by independently growing a new rule undergoing a growth and pruning phase. Finally, in the selection phase, the best rules are kept and the other rules are deleted from the model (Pang-Tang, 2011). This algorithm scales almost linearly with the number of training examples and is particularly suited for building models from data sets with imbalanced class distribution. Ripper is a rule based learner that build a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules. RIPPER forms rules through a process of repeated growing and pruning, and the representation of the rules generated can be more powerful because it is not constrained by the arbores cent structure of the tree. During the growing phase, the rules are made more restrictive in order to fit the training data as closely as possible. During the pruning phase, the rules are made less restrictive in order to avoid over fitting, which can cause poor classification performance. The algorithm proceeds iteratively starting with an empty rule set, and in each iteration, the training data is split into a growing set and a pruning set, then a rule is grown from the growing set and immediately pruned or simplified based on the pruning set. For example, given a rule ABCD->y, RIPPER check whether D should be pruned first, followed by CD, BCD, etc. while the original rule covers only positive examples, the pruned rule may cover some of the negative example in the training set (Steinbach, 2011). If the error rate of the new rule on the pruning set does not exceed some threshold, the rule is added to the rule set representing the learned model and allexamples in the training data covered by this rule are removed before being split again for the next repetition. Otherwise, the iteration is stopped and the rule set is returned.

ISBN: 978-93-86171-74-0

The RIPPER algorithm builds a single rule in the following steps:

- Split currently uncovered examples into a growing and pruning set.
- On the growing set, it starts with an empty rule (a rule with no antecedent).
- Add a new condition into the rule antecedent as long as this addition maximize FOIL's information gain criterion
- Repeat Step 3 until no negative examples from the growing set are covered by this rule.
- Prune this immediately on the pruning set.

For two-class problems, RIPPER chooses the majority class as its default class and learns the rules for detecting the minority class. For multiclass problems, the classes are ordered according to their frequencies. In a multi-class situation, the rules generated from the RIPPER algorithm are ranked in ascending order based on the number of examples in the class. An unknown instance is tested against the rules in that order. The first rule that covers the test instance "fires" and the testing phase ends (Pan & Ding, 2006).

The RIPPER algorithm for multi-class classification is described in the following steps:

- RIPPER sort the classes in ascending order based on the class size.
- It chooses the smallest class as the positive class and the rest is considered as the negative class.
- A rule set for the positive class is learned.
- Repeat step 2 and 3 for the next smallest class

Problem in ripper algorithm

- The main reason is over-fitting. The algorithm learns too much detail about the attributes of the URLs data. In the pruning phase, the rules are made less restrictive in order to avoid overfitting, which can cause poor classification performance.
- The major disadvantages of RPPER, it is scaled poorly with training set size and had problems with noisy data.
- The major drawback of RIPPER is its greedy optimization algorithm and its tendency to over-fit the training data at times.

VII. CONCLUSION

Various techniques have been implemented in order to control the phishing attacks. Different tools and software are there to determine such sites. Most of the browsers are built with phishing alert functionality. Blacklist has been a promising approaches in past but dynamic nature of phishing sites demands more efficient methods. Different systems such as PhishZoo, PhishNet and LinkGuard are proposed in order to determine phish websites in real time. Heuristic based data mining techniques are also utilized in order to detect phishing websites. Methods using data mining algorithm first extract the features of the suspected site and check it with the classifier. Classifiers are the rules generated using data mining algorithms.

In this research work, C4.5, CART and REPTree data mining algorithms have been used to detect the phishing websites. To train and test the algorithms WEKA tool is used. A comparison has been made between these three algorithms based on success rate, error rate and accuracy. Analysis of the algorithms shows that C4.5 gives the

best accuracy and lowest error rate whereas REPTree gives lowest accuracy and highest error rate. CART shows results between these two algorithms.

REFERENCES

[1] Due, B., Kristiansen, M., Colomo-Palacios, R., &Hien, D. H. T. Introducing Web Data Topics: A Multicourse Experience Report from Norway. Proceedings of the 3rd International Conference on Technological Ecosystems for Enhancing Multiculturality, 2015, 8(2), 565–569.

[2] Shirudkar, K., &Motwani, D. Web-Data Security, 2015, 5(3), 1100–1109.

[3] J. Ma, L. Saul, S. Savage and G. Voelker, "Learning to Detect Malicious URLs", ACM Transactions on Intelligent Systems and Technology, (2011), 1(1), 30:1-30:24.

[4] H. S. Choi, B. B. Zhu and H. J. Lee, "Detecting Malicious Web Links and Identifying Their Attack Types", Proceedings of the 2nd USENIX Conference on Web application development (WebApps), USENIX Association Berkeley, (2011), 1(3), 1-12.

[5] B. Eshete, A. Villafiorita and K. Weldemariam, "BINSPECT: Holistic Analysis and Detection of Malicious Web Pages", Proceedings of the 8th International ICST Conference, SecureComm,(2012), 3(6), 1544-1562.

[6] W. Tao, S. Z. Yu and B. L. Xie, "A Novel Framework for Learning to Detect Malicious Web Pages", Proceedings of the International Forum on Information Technology and Applications (IFITA), (2010), 1(9), 212-220.

[7] W. Zhang, Y. X. Ding, Y. Tang and B. Zhao, "Malicious web page detection based on online Learning algorithm", Proceedings of the International Conference on Machine Learning, (2011), 17(6), 1914-1919.

[8] V. L. Le, I. Welch, X. Y. Gao and P. Komisarczuk, "Two-Stage Classification Model to Detect Malicious Web Pages", Proceedings of the International Conference on Advanced Information Networking and Application (AINA), (2011), 15(11), 113-120.

[9] M. Cova, C. Kruegel and G. Vigna, "Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code", Proceedings of the International World Wide Web Conference Committee (IW3C2), WWW, (2010), 44(1), 48-58.

[10] Y. H. Choi, T. G. Kim and S. J. Choi, "Automatic Detection for JavaScript Obfuscation Attacks in Web Pages through String Pattern Analysis", International Journal of Security and Its Applications, (2010), 22(7), 13-26.

[11] R. B. Basnet and A. H. Sung, "Classifying Phishing Emails Using Confidence-Weighted Linear

[12] Classifiers", Proceedings of the International Conference on Information Security and Artificial Intelligence (ISAI), (2010), 4(3), 108-112.

[13] R. B. Basnet and A. H. Sung, "Learning to Detect Phishing Webpages", Journal of Internet Services and Information Security (JISIS), (2014), 4(1), 21-39.

[14] K. Rieck, T. Krueger and A. Dewald, "Cujo: Efficient Detection and Prevention of Drive-by-

[15] DownloadAttacks", Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC), (2010), 3(5), 31-39.