ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Mining of Big Data Sets on HADOOP

S.Manasa
M.Tech,
Dept of Computer Science and Engineering,
G.Pulla Reddy Engineering College (Autonomous),
Kurnool - 518 007, AP, India.

Y.Rama Mohan
Assitant Professor,
Dept of Computer Science and Engineering,
G.Pulla Reddy Engineering College (Autonomous),
Kurnool - 518 007, AP, India.

ABSTRACT:

Detection of the things that we are emerging from the growth of big data, social networks have been proposed. The traditional time-frequency-based techniques may not be appropriate in this case. Referring to the emergence of the things that we focus on the social aspects of these networks. In particular, we replies, Dynamic (intentionally or unintentionally) with reference to the user focus on the production of consumer mentions and retweets. The user of a social network, we propose mentioning the behavior of the probability model, and the model proposed by the measured anomalies to detect the emergence of a new thing. Overall, the social network is not only an outstanding scores from hundreds of users to post / view reference contacts to identify emerging issues. We have many real data sets collected from Twitter to show our method. Reference-anomaly-based techniques for the proposed experiments, at least in the early text-anomaly-based techniques to detect things like the new show, and the topic of the text is the most advance in some cases poorly posts.

I.INTRODUCTION:

In general, (sometimes called data or knowledge discovery) data mining to extract useful information from different perspectives, and the process of analyzing the data - information that can be used to increase revenues, expenses, or both cuts. Data mining software is one of a number of analytical tools to analyze the data. It's the users, to analyze information from many angles or pointed to summarize the classification and recognition of relationships. Technically, data mining has been evolving in a separate transaction and analytical systems for large-scale information technology fields in large relational databases.While correlations or patterns among dozens process of finding, data mining provides the link between the two.

Data mining software, based on open-ended questions that the user can analyze the stored transaction data relationships and patterns. There are many types of analytical software: statistics, machine learning, and neural networks. The data stored in the data used to detect pre-determined groups. For example, users can visit a restaurant chain could mine customer purchase data in order to identify and usually they do. These information items daily specials.Data logical relationships or grouped according to user preferences used by the traffic increase. For example, the data can be mined to identify the market or can be mined to identify the user affinities.Data groups. For example, behavior patterns and trends to anticipate diaper beer dug an example of associative mining.Data.

For example, the probability of an outdoor equipment retailer, backpack, sleeping bags and hiking shoes are bought on the basis of a consumer to buy predicted. Software.Present professionals.Analyze extract useful data format data by application, business analysts and information technology system.Provide data in a multidimensional database, data access, transform, and load data warehouse transaction system.Store and manage on the basis of such concepts as the natural evolution of a model The combination of genetic mutation and natural selection process structure.Optimization methods that learn through training and resemble biological neural networks, linear predictive models table.Non as a graph or tree-shaped structures that represent sets of decisions .

The decisions to produce an additional classification rules. Classification and Regression Trees techniques specific decision tree (CART) Chi Square Automatic Interaction Detection (CHAID) there. CART and CHAID decision tree methods used for the classification of a dataset. You will be given the records are expected to result in a new (unclassified) dataset provides the applicable regulations.

CHAID segments to create a multi-way splits CART departments while using the chi-square test dataset by creating 2-way splits. It is generally very similar to a historical dataset k CART record (s) to a dataset based on a combination of classes CHAID.A technique that classifies each record in the data preparation is less than the required (where k = 1). Sometimes the statistical significance of the data based on the extraction of useful if-then rules that technique.The k-nearest neighbor. sThe visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.It's one of the most effective services that are available today.

With the help of data mining, one can discover precious information about the customers and their behavior for a specific set of products and evaluate and analyze, store, mine and load data related to themAn analytical CRM model and strategic business related decisions can be made with the help of data mining as it helps in providing a complete synopsis of customersAn endless number of organizations have installed data mining projects and it has helped them see their own companies make an unprecedented improvement in their marketing strategies (Campaigns).Data mining is generally used by organizations with a solid customer focus. For its flexible nature as far as applicability is concerned is being used vehemently in applications to foresee crucial data including industry analysis and consumer buying behaviors.Fast paced and prompt access to data along with economic processing techniques have made data mining one of the most suitable services that a company seek.

II.RELATED WORK:

Dynamic networks have recently being recognized as a powerful abstraction to model and represent the temporal changes and dynamic aspects of the data underlying many complex systems. Significant insights regarding the stable relational patterns among the entities can be gained by analyzing temporal evolution of the complex entity relations. This can help identify the transitions from one conserved state to the next and may provide evidence to the existence of external factors that are responsible for changing the stable relational patterns in these networks. This paper presents a new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real-world applications show that the method is efficient and scalable. Web crawlers are essential to many Web applications, such as Web search engines, Web archives, and Web directories, which maintain Web pages in their local repositories. In this paper, we study the problem of crawl scheduling that biases crawl ordering toward important pages. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known PageRank as the importance metric. In order to score URLs, the proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We conduct a large-scale experiment using publicly available data sets to examine the effect of each feature on crawl ordering and evaluate the performance of many algorithms. The experimental results verify the efficacy of our schemes. In particular, compared with the representative RankMass crawler, the FPR-title-host algorithm reduces computational overhead by a factor as great as three in running time while improving effectiveness by 5 % in cumulative PageRank.Identifying social influence in networks is critical to understanding how behaviors spread. We present a method that uses in vivo randomized experimentation to identify influence and susceptibility in networks while avoiding the biases inherent in traditional estimates of social contagion.

Estimation in a representative sample of 1.3 million Facebook users showed that younger users are more susceptible to influence than older users, men are more influential than women, women influence men more than they influence other women, and married individuals are the least susceptible to influence in the decision to adopt the product offered. Analysis of influence and susceptibility together with network structure revealed that influential individuals are less susceptible to influence than noninfluential individuals and that they cluster in the network while susceptible individuals do not, which suggests that influential people with influential friends may be instrumental in the spread of this product in the network. A tremendous amount of data about individuals – e.g., demographic information, internet activity, energy usage, communication patterns and social interactions – are being collected and analyzed by many national statistical agencies, survey organizations, medical centers, and Web and social networking companies.

Wide dissemination of microdata (data at the granularity of individuals) facilitates advances in science and public policy, helps citizens to learn about their societies, and enables students to develop skills at data analysis. Often, however, data producers cannot release microdata as collected, because doing so could reveal data subjects' identities or values of sensitive attributes.

Failing to protect confidentiality (when promised) is unethical and can cause harm to data subjects and the data provider. It even may be illegal, especially in government and research settings. For example, if one reveals confidential data covered by the U. S. Confidential Information Protection and Statistical Efficiency Act, one is subject to a maximum of $250,000 in fines and a five year prison term.

With the rapid growth of the availability and popularity of interpersonal and behavior-rich resources such as blogs and other social media avenues, emerging opportunities and challenges arise as people now can, and do, actively use computational intelligence to seek out and understand the opinions of others.

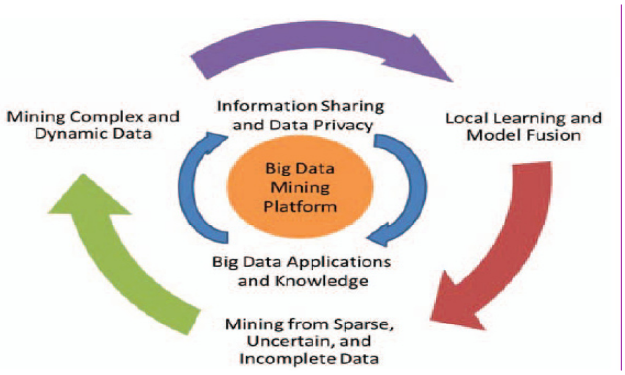
The study of collective behavior of individuals has implications to business intelligence, predictive analytics, customer relationship management, and examining online collective action as manifested by various flash mobs, the Arab Spring (2011) and other such events.

In this article, we introduce a nature-inspired theory to model collective behavior from the observed data on blogs using swarm intelligence, where the goal is to accurately model and predict the future behavior of a large population after observing their interactions during a training phase.

Specifically, an ant colony optimization model is trained with behavioral trend from the blog data and is tested over real-world blogs. Promising results were obtained in trend prediction using ant colony based pheromone classier and CHI statistical measure.

We provide empirical guidelines for selecting suitable parameters for the model, conclude with interesting observations, and envision future research directions.

SYSTEM ARCHITECTURE:


Mining Complex and Dynamic Data
Information Sharing and Data Privacy
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and Knowledge
Mining from Sparse, Uncertain, and Incomplete Data

III.SYSTEM PREMELIES

A.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.

B.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.

C.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

D.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.

E.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.

The rise of Big Data applications where data collection has grown tremen dously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions.

In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Big Data.

IV.CONCLUSION:

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

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies.

Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data.

At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources, along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains.

With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

REFERENCES:

[1] R. Ahmed and G. Karypis, “Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 603-630, Dec. 2012.

[2] M.H. Alam, J.W. Ha, and S.K. Lee, “Novel Approaches to Crawling Important Pages Early,” Knowledge and Information Systems, vol. 33, no. 3, pp 707-734, Dec. 2012.

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

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

[5]S. Banerjee and N. Agarwal, “Analyzing Collective Behavior from Blogs Using Swarm Intelligence,” Knowledge and Information Systems, vol. 33, no. 3, pp. 523-547, Dec. 2012.

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

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

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

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

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

[11] E.Y. Chang, H. Bai, and K. Zhu, “Parallel Algorithms for Mining Large-Scale Rich-Media Data,” Proc. 17th ACM Int’l Conf. Multimedia, (MM ’09,) pp. 917-918, 2009.

[12] R. Chen, K. Sivakumar, and H. Kargupta, “Collective Mining of Bayesian Networks from Distributed Heterogeneous Data,” Knowledge and Information Systems, vol. 6, no. 2, pp. 164-187, 2004.

[13] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, “Efficient Algorithms for Influence Maximization in Social Networks,” Knowledge and Information Systems, vol. 33, no. 3, pp. 577-601, Dec. 2012.

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