



Key Word Based Mining on Twitter

Mr. Sunil Bandari

Department of Computer science and Engineering Aurora's Technological and Research Institute

Email id: sunil_chinni@yahoo.co.in

Mrs. K. Kavitha

(Associate Professor) Department of Computer science and Engineering Aurora's Technological and Research Institute

Email id: kavithakbjr@gmail.com

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 characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

I INTRODUCTION

In the era of Big Data, every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century. As another example, on 4 October 2012, the first presidential debate between President Barack Obama and Governor Mitt Romney triggered more than 10 million tweets within 2 hours. Among all these tweets, the specific moments that generated the most discussions actually revealed the public interests, such as the discussions about medicare and vouchers. Such online discussions provide a new means to sense the public interests and generate feedback in real time, and are mostly appealing compared to generic media, such as radio or TV broadcasting. Another example is

Flickr, a public picture sharing site, which received 1.8 million photos per day, on average, from February to March 2012 [35]. Assuming the size of each photo is 2 megabytes (MB), this requires 3.6 terabytes (TB) storage every single day. Indeed, as an old saying states: "a picture is worth a thousand words," the billions of pictures on Flickr are a treasure tank for us to explore the human society, social events, public affairs, disasters, and so on, only if we have the power to harness the enormous amount of data.

The above examples demonstrate the rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a "tolerable elapsed time." The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions [40]. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible. For example, the square kilometer array (SKA) [17] in radio astronomy consists of 1,000 to 1,500 15-meter dishes in a central 5-km area. It provides 100 times more sensitive vision than any existing radio telescopes, answering fundamental questions about the Universe. However, with a 40 gigabytes (GB)/second data volume, the data generated from the SKA are exceptionally large. Although researchers have confirmed that interesting patterns, such as transient radio anomalies [41] can be discovered from the SKA data, existing methods can only work in an offline fashion and are incapable of handling this Big Data scenario in real time. As a result, the unprecedented data volumes require an effective data analysis and

prediction platform to achieve fast response and real-time classification for such Big Data.

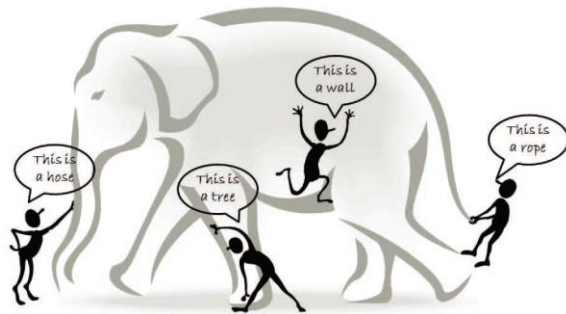


Fig. 1. The blind men and the giant elephant: the localized (limited) view of each blind man leads to a biased conclusion.

II SYSTEM ANALYSIS

Existing System :

- The rise of Big Data applications where data collection has grown tremendously and is beyond the ability of commonly used software tools to capture, manage, and process within a “tolerable elapsed time.” The most fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly infeasible.
- The unprecedented data volumes require an effective data analysis and prediction platform to achieve fast response and real-time classification for such Large amount of Data.

Disadvantages Of Existing System :

The challenges at Tier I focus on data accessing and arithmetic computing procedures. Because Big Data are often stored at different locations and data volumes may continuously grow, an effective computing platform will have to take distributed large-scale data storage into consideration for computing.

- The challenges at Tier II center around semantics and domain knowledge for different Big Data applications. Such information can provide additional benefits to the mining process, as well as add technical barriers to the Big Data access (Tier I) and mining algorithms (Tier III).

- 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.

Proposed System :

- We propose a HACE theorem to model Big Data characteristics. The characteristics of HACH make it an extreme challenge for discovering useful knowledge from the Big Data.
- The 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.
- To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data.

Advantages Of Proposed System :

Provide most relevant and most accurate social sensing feedback to better understand our society at realtime.

III LITERATURE SURVEY

1) Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,

For modeling and presenting the temporal changes and dynamic aspects of data underlying multiple complex systems, the dynamic networks are found as a powerful abstraction. We can obtain Significant insights related to the constant relational patterns among the entities by analyzing temporal evolution of the complex entity relations. This can help identify the transitions from one conserved state to the next and may give proof to the existence of external factors that are responsible for changing the stable relational patterns in these networks.

A new data mining techniques are represented in this article and 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. The results of experiments(based on multiple datasets from real-world applications) prove that the method is efficient and scalable.



2) Novel Approaches to Crawling Important Pages

Web crawlers are important for many Web applications, such as Web archives and Web search engines and Web directories, that contain Web pages in their local repositories. Here, in our project, we observe the problem of crawl scheduling that biases crawl ordering toward important pages. A set of crawling algorithms are proposed for effective and efficient crawl ordering by giving priority for important pages with the help of PageRanks as the importance metric. Our proposed algorithms utilize various features for scoring URLs. The features that our proposed algorithms utilize may include inter-host links, partial link structure, topic relevance, and page titles.

For examining the each crawling feature's effect on crawling order and also for evaluating the performance of our algorithms, we conduct a big experiment with the help of the data sets that are available publicly. The obtained results for our experiments verify the efficacy of our system. 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

3) Identifying Influential and Susceptible Members of Social Networks

For understanding the behavior spread of social networks, It is critical to identify the social influence in networks. So, we here, in this system present a method that does use in vivo randomized experimentation for identifying the 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

In social networks (twitter), the younger users are more susceptible to influence than older users, women are not more influential than that of men, 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.

4) Big Privacy: Protecting Confidentiality in Big Data

A big amount of information about individuals, such as energy usage, demographic information, internet activity, social interactions, and communication patterns are gathered and analyzed by many statistical agencies, survey organizations, medical centers, and Web and social networking companies. Wide dissemination of micro data (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, the information producers cannot release micro data as collected, so that the sensitive attributes' information such as personal data of user can not be miss used (for privacy reason). If the we fail to secure the confidential data that we agreed to do, then it can cause harm to data providers and also for data subjects. It is also illegal to do so, especially in government and research settings. For example, if one reveals confidential data covered by the Confidential Information Protection and Statistical Efficiency Act of a country like defence system's confidential data, one is subject to a maximum of \$250,000 in fines and a five year prison term.

5) Analyzing Collective Behavior from Blogs Using Swarm Intelligence

With a high 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 individuals' collective functional study has implications to predictive analytics, business intelligence, customer relationship management, and observing online collective action as manifested by various flash mobs, the Arab Spring (2011) and other such events.

In our scheme, we present a nature-inspired theory for modeling collective functions from the observed data on blogs with the help of swarm intelligence, where the aim is to model accurately and predict the future functionality of a big population after studying their actions in 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. Results were obtained in trend prediction using ant colony based pheromone classifier and



CHI statistical measure. We provide empirical guidelines for choosing perfect parameters for the our scheme, conclude with interesting observations, and envision future research directions.

IV IMPLEMENTATION

Modules :

1. Integrating and mining Large data
2. Fast Response
3. Pattern matching and mining
4. Key technologies for integration and mining
5. Group influence and interactions

Integrating and mining bio data :

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 integration, 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 information 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 regularities 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.

V 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 Project suggests that the key characteristics of the Big Data are 1) huge with multiple properties of User Data, 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.

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.