ISSN 2250-3153

A Data Mining: Overview to Distributed Systems

Ms. Rupali Chikhale

G. H. Raisoni Institute of Information Technology, Nagpur

Abstract- Distribution of data and computation allows for solving larger problems and execute applications that are distributed in nature. Data mining technology has emerged as a means for identifying patterns and trends from large quantities of data. The Data Mining technology normally adopts data integration method to generate Data warehouse, on which to gather all data into a central site, and then run an algorithm against that data to extract the useful Module Prediction and knowledge evaluation. Applications from various domains have adopted this technique to perform data analysis efficiently. Several issues need to be addressed when such techniques apply on data these are bulk at size and geographically distributed at various sites. The system contains modules for secure distributed communication, database connectivity, organized data management and efficient data analysis for generating a global mining model. Performance evaluation of the system is also carried out and presented. New technologies are emerging to make big data analytics possible and cost-effective. This paper describe system architecture and distributed data mining, also known as multi agent based distributed data mining, in terms of significance, system overview, existing systems, and research trends.

Index Terms- Data mining; distributed systems; reliability; performance.

I. INTRODUCTION

The widespread use of computers and the advance in database technology have provided huge amounts of data. The explosive growth of data in databases has generated an urgent need for efficient data mining techniques to discover useful information and knowledge.

Distributed data mining (DDM) is a fast growing area which deals with the problem of finding data patterns in an environment with distributed data and computation. In current era most of the data analysis systems require centralized storage of data, the increasing merger of computation with communication is more demand data mining environments that can utilize the full advantage of distributed computation. Reliability of distributed no longer can be assured by static design because distributed systems are increasingly large, heterogeneous and dynamic. On the one side, large-scale computing grids, clouds and clusters provide computing and data resources with hundreds or even up to thousands of nodes.

Distributed Systems today grow dynamically in terms of new applications, hardware and network components, users and workload changes. Complex interactions between the different layers of a distributed system make systems and effects of faults hard to understand such that faulty behavior and poor performance cannot always be distinguished. Our approach to reliable operation of distributed systems is based on building a dynamic model for the distributed systems from monitored system data.

A fundamental challenge for DDM is to develop mining techniques without having to communicate data unnecessarily. Such functionality is required for reasons of efficiency, accuracy and privacy. In addition, appropriate protocols, languages, and network services are required for mining distributed data to handle the required metadata and mapping.

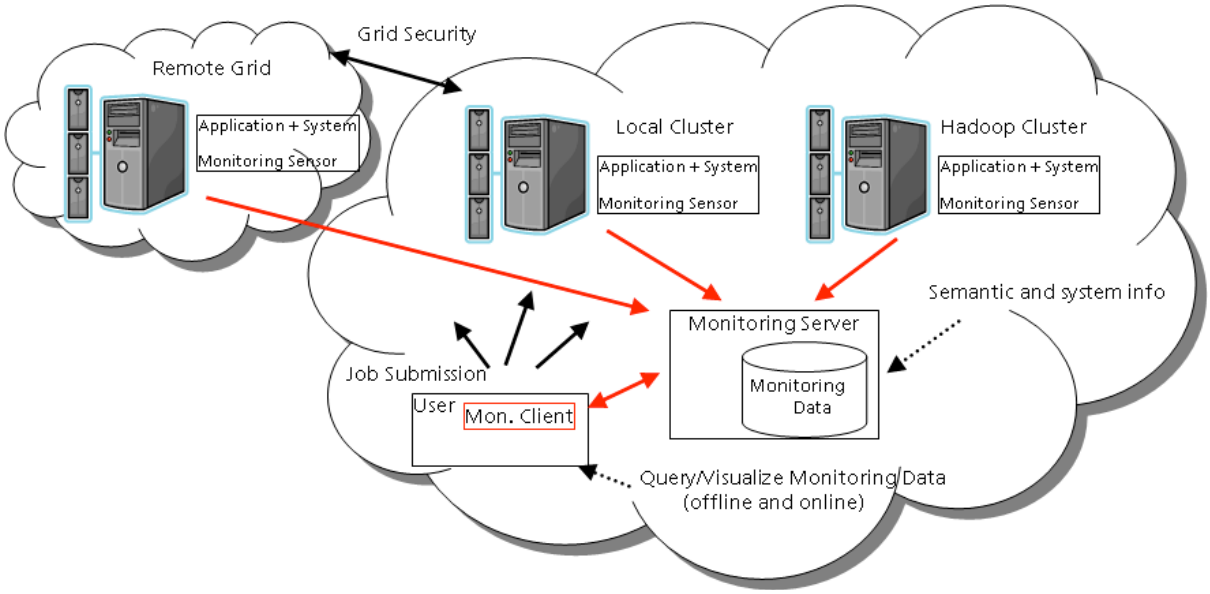
Distributed Data Mining Architecture

Our proposed mining architecture is a client/server-based system developed for performing knowledge discovery from large distributed sources of data. Due to the diversity of mining algorithms and the diversity of data sources, it is difficult to generate a mining model by combining mining rules on different sites. Our proposed system works independently to combine result from different sites. This section describes the abstract architecture model of the Distributed Data Mining and the interaction between its various subsystems. The architecture has the following subsystems: communication, mining, analyzing, and database.

We want to evaluate collected data of the system with the help of data mining techniques in order to build a model of the system. This model can be used for online prediction of the system's behavior and thus gives the opportunity to react even before faults or errors occur.

System Architecture

In the following, an initial architecture is presented that enables data miners to build a system model for execution time prediction and performance fault detection for distributed data mining algorithms. Fig. 4 depicts the architecture, which is currently being implemented



The following depicts a 5-Step approach on predictive maintenance for distributed systems:

• Step 1 - **Data Gathering (logging, online monitoring)**: This step provides the basis for building the model of the system. Observational data from the operational system is gathered either in an online or an off-line setting. Using data from the real system for model building allows for much more realistic models than a-priori static models or simulations.

• Step 2 - **Model Building**: This step is based on the data mining analysis of the monitoring data. E.g., feature-selection and classification algorithms can be used to determine which parameters of the system are most correlated to failures or performance problems

• Step 3 - **Online-Monitoring and Prediction**: Given that a model of the system gained in Step2 exists, selected parameters can be monitored and short term prediction on potential problems can be made on the basis of this model. Also, predictions regarding parameter settings of the distributed system become possible.

• Step 4 - **Preventive Measures**: In case that the model predicts reliability or performance problems based on the actual observations, measures might be taken to prevent problems, instance system parameters might be changed or load might be reduced

• (Step 5 - **Adaptation of the model**): As the actual system might emerge over the time, adaptations of the model itself might become necessary. Basically, the model building process taking place in Step 2 on the data from Step 1 will have to be checked against long term changes in the actual monitored data. The model adaptation might require human interaction, but also could be foreseen automatically, for instance when re-adjusting thresholds or expected delays.

Distributed Data Mining on Grids

- The Grid extends the distributed and parallel computing paradigms allowing resource negotiation, dynamical allocation, heterogeneity, open protocols and services.
- As Grids and Clouds became well accepted computing infrastructures it is necessary to provide data mining services, algorithms, and applications.
- Those may help users to leverage Grid/Cloud/… capability in supporting high-performance distributed computing for solving their data mining problems in a distributed way.

Grid services for distributed data mining

- Exploiting the SOA model and the Web Services Resource Framework (WSRF) it is possible to define basic services for supporting distributed data mining tasks in Grids
- Those services can address all the aspects that must be considered in data mining and in knowledge discovery processes
- data selection and transport services,
- data analysis services,
- knowledge models representation services, and
- Visualization services.

It is possible to define services corresponding to:

Single Steps - that compose a KDD process such as preprocessing, filtering, and visualization.

Single Data Mining Tasks - such as classification, clustering, and association rules discovery.

Distributed Data Mining Patterns - such as collective learning, parallel classification and meta-learning models.

Data Mining Applications or KDD processes - including all or some of the previous tasks expressed through a multi-step workflow.

Data mining Grid services

- This collection of data mining services can constitute an Open Service Framework for Grid-based Data Mining Open Service Framework for Grid-based Data Mining
- Allowing developers to program distributed KDD processes as a composition of single and/or aggregated services available over a Grid.

- Those services should exploit other basic Grid services for data transfer and management for data transfer, replica management, data integration and querying.

By exploiting the Grid services features it is possible to develop data mining services accessible every time and everywhere.

- This approach may result in
- Service-based distributed data mining applications
- Data mining services for virtual organizations.
- Distributed data analysis services on demand.
- A sort of knowledge discovery eco-system formed of a large numbers of decentralized data analysis services.

II. SUMMARY

- New HPC infrastructures allow us to attack new problems, BUT require to solve more challenging problems.
- New programming models and environments are required
- Data is becoming a BIG player, programming data analysis applications and services is a must.
- New ways to efficiently compose different models and paradigms are needed.
- Relationships between different programming levels must be addressed.
- In a long-term vision, pervasive collections of data analysis services and applications must be accessed and used as public utilities.
- We must be ready for managing with this scenario.

III. CONCLUSION

In this paper, we highlight the problem of the increase in complexity, diversity and scale of data. We introduce a separation of concerns between data mining and integration (DMI) process development and the mapping, optimization and enactment of these processes. We postulate this separation of concerns will allow handling separately the user and application diversity and the system diversity and complexity issues simultaneously.

We introduced an initial architecture for observing the distributed systems and algorithm executions that allows for model building and on-line monitoring based on predicting upcoming reliability and performance problems with previously generated models. Users are enabled to take preventive measures for increased reliability or performance. We presented a concrete scenario of applying this approach in the field of distributed data mining.

REFERENCES

[1] S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta. Distributed data mining in peer-to-peer networks. Internet Computing,IEEE, 10(4):18–26, 2006.

[2] V. Gorodetsky, O. Karsaev, and V. Samoilov. Infrastructural Issues for Agent-Based Distributed Learning. In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, pages 3–6. IEEE Computer SocietyWashington, DC, USA, 2006.

[3] http://www.ijric.org/volumes/Vol3/11Vol3.pdf

[4] Banks, T. Web Services Resource Framework (WSRF) | Primer v1.2. Tech. rep., OASIS, May 2006.

[5] Bettina Berendt, Bamshad Mobasher, Myra Spiliopoulou, and Jim Wiltshire, Measuring the Accuracy of Sessionizers for Web Usage Analysis, Workshop on Web Mining at the First SIAM International Conference on Data Mining, 2001.

[6] J. Pitkow, In search of reliable usage data on the WWW, Sixth International World Wide Web Conference, 1997.

[7] http://si.deis.unical.it/~talia/dpa08_talia.pdf

[8] fhttp://www.intel.in/content/dam/www/public/us/en/documents/white-papers/distributed-data-mining-paper.pdf

[9] Hoffmann, G.; Malek, M., "Call Availability Prediction in a Telecommunication System: A Data Driven Empirical Approach," Reliable Distributed Systems, 2006. SRDS '06. 25th IEEE Symposium on , vol., no., pp.83-95, 2-4 Oct. 2006

[10] Khayat, N.: Semantic Instrumentation and Measurement of Data Mining Algorithms, Technical Report on R&D 2, Hochschule Bonn-Rhein-Sieg, 2009.

[11] Wirth, P,: Monitoring von Data Mining Algorithm in verteilten Umgebungen, Master Thesis Hochschule Bonn-Rhein-Sieg (to be submitted).

[12] Duan, R.; Prodan, R.; Fahringer, T., "Short Paper: Data Mining-based Fault Prediction and Detection on the Grid," High Performance Distributed Computing, 2006 15th IEEE International Symposium on , vol., no., pp.305-308

[13] Parthasarathy,S., and Subramonian,R., (1999), “Facilitating Data Mining on a network of workstations”, to appear in Advances in Distributed Data Mining, (eds) Hillol Kargupta and Philip Chan, AAAI Press.

AUTHORS

First Author – Ms. Rupali Chikhale, M.C.A., B.Sc. (CS), G. H. Raisoni Institute of Information Technology, Nagpur. Email: rupali.chikhale@raisoni.net, rsg2409@gmail.com, M - 9423102992