2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing

NEMICO: Mining network data through cloud-based data mining techniques

Elena Baralis, Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, Luigi Grimaudo, Fabio Pulvirenti
Dipartimento di Automatica e Informatica
Politecnico di Torino - Torino, Italy
Email: {name.surname}@polito.it

***Abstract*—Thanks to the rapid advances in Internet-based applications, data acquisition and storage technologies, petabyte-sized network data collections are becoming more and more common, thus prompting the need for scalable data analysis solutions. By leveraging today's ubiquitous many-core computer architectures and the increasingly popular cloud computing paradigm, the applicability of data mining algorithms to these large volumes of network data can be scaled up to gain interesting insights. This paper proposes NEMICO, a comprehensive Big Data mining system targeted to network traffic flow analyses (e.g., traffic flow characterization, anomaly detection, multiple-level pattern mining). NEMICO comprises new approaches that contribute to a paradigm-shift in distributed data mining by addressing most challenging issues related to Big Data, such as data sparsity, horizontal scaling, and parallel computation.**

I. INTRODUCTION

Important issues in network traffic analysis are communication profiling, anomaly or security threat detection, and recurrent pattern discovery. Traffic analyses are commonly performed on: (i) packet payloads, (ii) traffic metrics, or (iii) some statistical features computed on traffic flows. A significant research effort has been devoted to the application of data mining techniques to network traffic analysis. The proposed approaches address the discovery of significant correlations among data [1], [2], the extraction of kowledge useful for prediction [3], and the clustering of network data with similar properties [4]. However, due to the continuous growth in network speed, petabytes of data may be transferred through a network every day. These "big data" collections stress the limits of existing data mining techniques and thus they set new horizons for the design of innovative data mining approaches.

The goal of this work is to design and develop a comprehensive system which provides cloud users with a variety of network analytics services. More specifically, this paper presents NEMICO (NEtwork MIning in the ClOud), a data mining system focused on efficiently discovering interesting knowledge from Big network datasets by means of distributed approaches. NEMICO consists of a series of distributed MapReduce jobs running in the cloud. Each job performs a different step of the knowledge discovery process, ranging from network data acquisition to knowledge exploitation.

The research leading to these results has received funding from the European Union under the FP7 Grant Agreement n. 619633 (Integrated Project "ONTIC")

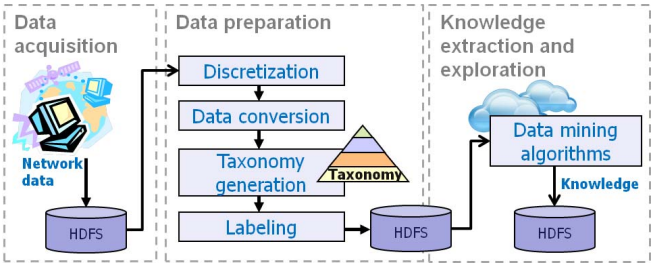

Fig. 1. Architecture of NEMICO

II. THE NEMICO ARCHITECTURE

The building blocks of the NEMICO architecture are shown in Figure 1. The system has been thought to support the integration of a variety of data mining algorithms, including supervised approaches (e.g., classification and regression algorithms) and unsupervised ones (e.g., association rule mining, clustering).

The NEMICO architecture consists of a series of distributed jobs running in the cloud, which cover all the steps of the knowledge discovery process from network data. Each job takes as input the result of one or more preceding jobs and it performs a specific step of the network data mining process. Currently, each job is performed by one or more MapReduce tasks running on a Hadoop cluster.

A. Data acquisition

NEMICO acquires massive amounts of network traffic measurements through passive traffic sniffing. A passive probe located on the Internet access link of an edge network is exploited to acquire incoming and outgoing packets flowing on the link. Traffic monitoring is performed by Tstat [5], a tool that allows users to collect network and transport layer measurements. The acquired datasets consist of a set of records, each one corresponding to a different TCP flow. To ensure measurement reliability, only the TCP flows that last more than ten packets (i.e., the long-lived flows) are considered.

NEMICO stores the collected network traffic data in an HDFS distributed file system. It can successfully acquire and handle all the network measurements provided by Tstat (i.e., the Round-Trip-Time (RTT) observed on a TCP flow, the class of application layer service (e.g., HTTP,VIDEO) assigned by Tstat).

978-1-4799-7881-6/14 $31.00 © 2014 IEEE
DOI

B. Data preparation

To suit the raw data to the subsequent data mining step some established data preprocessing steps are applied to the raw network traffic measurements. A brief description of the main data preparation steps is given below.

Discretization. Discretization concerns the transformation of continuous values into discrete ones. Since some data mining algorithms are unable to cope with continuously valued data, measurement values are discretized prior to running the algorithms. The discretization step can be performed either automatically by using established techniques [6] or semi-automatically by partitioning continuous value ranges into appropriate bins based on the prior knowledge about the measurement domains.

Data conversion. Since most algorithms are designed to handle only a subset of specific data formats, data conversion entails the transformation of the raw data into the expected data format. For example, most association rule mining algorithms are designed to cope with transactional data [6]. Hence, to apply association rule mining algorithms the acquired data have to be tailored to the transactional data format.

Taxonomy generation. The data mining process can be driven by semantics-based models (e.g., taxonomies or ontologies). These models are used to enrich the source data with multiple-level or multi-faceted information. For example, a taxonomy consists of set of is-a hierarchies built over the data attributes, which can be exploited to aggregate specific data values (e.g., the TCP ports) into meaningful higher-level categories. Since many data mining algorithms (e.g., [7]) allow pushing taxonomy-based information into the extraction process, NEMICO supports both the automatic taxonomy inference over a subset of specific network data attributes (e.g., port number, packet number) and the semi-automatic taxonomy construction.

Labeling. Supervised data mining techniques (e.g., classification) require the labeling of one data attribute as class label. Hence, if the data mining process comprises supervised analyses domain-experts have to specify the class attribute.

The current implementation of NEMICO includes a first implementation of all the described activities as map jobs.

C. Knowledge extraction and exploration

Knowledge extraction entails the application of data mining algorithms to find implicit, previously unknown, and potentially useful information from large volumes of network data. NEMICO comprises novel data mining algorithms that contribute to a paradigm-shift in distributed data mining. The analytics algorithms entail (i) discovering underlying correlations among traffic data (e.g., multiple-level associations among data equipped with taxonomies), (ii) grouping traffic flows with similar properties (e.g., clustering), and (iii) extracting models useful for prediction (e.g., classification, regression). The algorithms are designed to address the following issues.

Sparse data distribution. Since Big data collections have large cardinality and/or a high number of dimensions, they are usually characterized by an inherent sparseness. Aimed at addressing this issue, we propose incremental algorithms that are able to cope with result refinement over different incremental runs and that scale by adapting to new data without the need for re-analyzing the entire dataset.

Algorithm optimization. Most data mining algorithms are poorly optimized for cloud computing environments. Conversely, our algorithms have specifically been designed to scale in massively parallel computing environments.

Horizontal scalability. To apply the data mining techniques to petabyte-scale network traffic datasets, NEMICO integrates advanced analytics algorithms (e.g., association rule mining) with horizontally scalable approaches, such as those based on Map Reduce and shared columnar storage backends.

The current implementation of NEMICO comprises Hadoop-based data mining algorithms focused on the extraction of interesting and multiple-level correlations among network data (i.e., association rules [8] and misleading multiple-level patterns [7]). To make the results manageable for manual inspection the extracted patterns can be ranked according to established quality measures [6] to select top-k interesting ones or categorized into semantically related groups (e.g., all patterns related to a specific class of service). Both input data and generated knowledge are stored in an HDFS file system.

III. EXPERIMENTS AND CONCLUSIONS

The current implementation of NEMICO was developed in Java using the new Hadoop Java APIs. It was validated on real network traffic datasets with size 192.56 GB. The experiments were performed on a cluster of 5 nodes running the Cloudera's Distribution of Apache Hadoop (CDH4.5). More details of both cluster node and data can be found in [7]. For both data mining algorithms in NEMICO, we evaluated the speedup achieved increasing the number of Hadoop cluster nodes and the achieved results show that our approaches scale roughly linearly with the number of nodes and the speedup approximately corresponds to the number of cluster nodes. As future work we are planning to introduce visualization tools and new data mining algorithms to support a larger variety of traffic data analyses.

REFERENCES

[1] D. Apiletti, E. Baralis, T. Cerquitelli, and V. D'Elia, "Characterizing network traffic by means of the netmine framework," *Computer Networks*, vol. 53, no. 6, pp. 774–789, 2009.

[2] E. Baralis, L. Cagliero, T. Cerquitelli, V. D'Elia, and P. Garza, "Expressive generalized itemsets," *Information Sciences*, vol. 278, pp. 327–343, 2014.

[3] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Blinc: multilevel traffic classification in the dark." in *SIGCOMM*, 2005, pp. 229–240.

[4] J. Erman, M. Arlitt, and A. Mahanti, "Traffic classification using clustering algorithms," in *MineNet '06*, 2006, pp. 281–286.

[5] A. Finamore, M. Mellia, M. Meo, M. Munafò, and D. Rossi, "Experiences of internet traffic monitoring with tstat," *IEEE Network*, vol. 25, no. 3, pp. 8–14, 2011.

[6] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., 2005.

[7] E. Baralis, L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, L. Grimaudo, and F. Pulvirenti, "Misleading generalized itemset mining in the cloud," in *ISPA'14*, 2014.

[8] D. Apiletti, E. Baralis, T. Cerquitelli, S. Chiusano, and L. Grimaudo, "Searum: A cloud-based service for association rule mining," in *ISPA'13*, 2013, pp. 1283–1290.