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 applications and services based on Internet, data acquisition and storage technologies, petabyte-sized network data collections, are becoming more and more common, 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 on these large volumes of network data can be dramatically scaled up to gain interesting insights.**

This paper proposes a comprensive system, named NEMICO, designed to digging deep into data network data for targeted traffic flow analysis (e.g., characterization of traffic flows, identification of interesting correlations at a different level of abstraction, detect anomalies). NEMICO include new approaches that contribute to a paradigm-shift in distributed data mining, by addressing issues raised by the characteristics of Big Data such as data sparsity, horizontal scaling, and parallel computation.

I. INTRODUCTION

An important issue in network traffic analysis is to profile communications, detect anomalies or security threats, and identify recurrent patterns. To these aims, the analysis could be performed on: (i) Packet payloads, (ii) traffic metrics, and (iii) 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 include studying correlations among data [1], [2], extracting information for prediction [3], grouping network data with similar properties [4]. However, due to the continuous growth in network speed, peta-bytes of data may be transferred through a network every day. These "big data" collections stress the limits of existing data mining techniques and sets new horizons for the design of innovative techniques to address data analysis.

The challenge of our work is to design and develop a comprensive system able to offer a variety of network analytics services to cloud users. More specifically, this paper presents the NEMICO system to efficiently discover different kinds of interesting knowledge by means of a distributed computing model. NEMICO consists of a series of distributed MapReduce jobs run in the cloud. Each job performs a different step in the knowledge discovery process ranging from the network data acquisition to knowledge exploitation.

II. THE NEMICO ARCHITECTURE

The building blocks of the NEMICO architecture are shown in Figure **??**. To effectively support analysts in discovering different and interesting kinds of knowledge, a broad variety of data mining algorithms will be integrated in the system such as exploratory techiques (e.g., association rules, clustering) and prediction ones (e.g., classification and regression algorithms).

The NEMICO implementation consists of a series of distributed jobs run in the cloud to cover the overall process of discovering useful knowledge from network data. Each job receives as input the result of one or more preceding jobs and performs one of the steps required for mining network data. Currently, each job is performed by one or more MapReduce tasks run 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 [6], a tool that allows users to collect network and transport layer measurements. Tstat rebuilds TCP connections by matching sequence numbers on data segments with the corresponding acknowledgement (ACK) numbers. 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. Among others, the supported measurements comprise (i) the Round-Trip-Time (RTT) observed on a TCP flow, which indicates the minimum time lag between the observation of a TCP segment and the observation of the corresponding ACK, (ii) the TCP port of the server, (iii) the number of packets, and (iv) the class of application layer service (e.g., HTTP,VIDEO) assigned by Tstat. Note that Tstat measurements comprise both continuous values (e.g., the RTT) and discrete values (e.g., the class of service).

B. Data preparation

To suit the raw data to the subsequent data mining step some established data preprocessing steps are applied to the

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

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. In some cases (e.g., the association rule mining algorithms) discretization is not mandatory but strongly recommended because considering continuously valued attributes could bias the mining result (e.g., the association rules extracted from continuously valued attributes are unlikely to occur frequently in the source data). The discretization step can be performed either automatically (by using established techniques [**?**]) or semi-automatically (by partitioning continuous value ranges into appropriate bins based on the prior knowledge about the measurement domains). Due to the nature of the analyzed network data, manual data discretization is often preferable.

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 [**?**]. Hence, to successfully apply association rule mining algorithms the acquired data have to be tailored to the transactional data format.

Labelling. Supervised data mining techniques (e.g., classification) require the labelling of one or more data attributes as classes to drive the prediction process. Hence, if the subsequent data mining process comprises supervised analyses experts have to label one or more network measurements as reference classes.

Taxonomy generation. The mining process for extracting more abstract and interesting correlations among data (e.g., generalized association rules) is driven by taxonomies. A taxonomy is a hierarchy of aggregations over values of one attribute (e.g., TCP port) and it is usually represented as a tree. NEMICO allows either to automatically infer interesting taxonomies directly from the data. To this aim different algorithms have been devised and implemented to automatically extract taxonomies for the considered attributes (i.e., port number, packet number) since some attributes (e.g., port number) are actually hierarchical attributes while others are numerical ones.In NEMICO taxonomies could also be provided directly by the user.

The current implementation of NEMICO includes the implementation of both data discretization and conversion which are performed by a single map only job. Each record is processed by the map function and, if the number of packets is above the threshold (e.g., 10?? packets), the corresponding discretized version is generated as output of the mapping step. This task entails an inherently parallel elaboration, considering that can be applied independently to each record.

C. Knowledge extraction

The knowledge extraction block of NEMICO focuses on data mining algorithms to find implicit, previously unknown, and potentially useful information from large volume of network data. NEMICO includes novel data mining algorithms that contribute to a paradigm-shift in distributed data mining. The analytics algorithms support studying correlations among data (e.g., association rules at different levels of abstraction), grouping data with similar properties (e.g., clustering), and extracting interesting knowledge for prediction (e.g., classification, regression). The design of these new approaches addresses the following open issues.

Sparse data distribution. Since Big data collections have large cardinality and/or high-dimensional data, they are usually characterized by an inherent sparseness and variable distribution. Aimed at addressing these issues, 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 to re-analyze the entire dataset.

Algorithm optimization. Available large-scale data mining algorithms are poorly optimized for cloud computing environments. Hence, our algorithm have been designed to improve the scalability of the existing techniques and their performance in massively parallel computing environments.

Horizontal scalability. To apply the mining techniques to petabyte-scale datasets, such as network traffic, we integrated into NEMICO 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 includes Hadoop-based data mining algorithms that enable the extraction of interesting correlations among network data at a different level of abstractions (i.e, association rule mining [5] and misleading multiple-level patterns [7]).

D. Knowledge categorization and selection

Explorative data mining approaches, such as association rule mining and clustering algorithms, may identify interesting knowledge, which may be both huge in size and complex in structure. To ease exploitation of the mined knowledge, different interestingness measures (e.g., chi square, support expectation, collective strength) to reduce and evaluate the amount of extracted knowledge are needed. For a detailed review on measures for interesting rules and quality indexes for clustering see [8] and [9] respectively. The current implementation of NEMICO includes COSA METTIAMO??

Furthermore, only for the discovery of correlations when coping with relatively large or complex transactional network datasets the number of mined rules could be so large that a manual inspection becomes unfeasible. To overcome this issue, NEMICO propose a classification of the rules into groups according to their semantics in the network domain.

III. PRELIMINARY EXPERIMENTS

The current implementation of NEMICO has been validated on real network traffic datasets obtained by performing different capture stages on a backbone link of a nation-wide ISP in Italy that offers us three different vantage points. The dataset has size 192.56 GB and it consists of 413,012,989 records, i.e., one record for each bi-directional TCP flow).

The MapReduce jobs of the NEMICO were developed in Java using the new Hadoop Java APIs. The experiments were performed on a cluster of 5 nodes running Cloudera's Distribution of Apache Hadoop (CDH4.5). Each cluster node is a 2.67 GHz six-core Intel(R) Xeon(R) X5650 machine with 32 Gbyte of main memory running Ubuntu 12.04 server with the 3.5.0-23-generic kernel. All the reported execution times are real times obtained from the Cloudera Manager web control panel.

We evaluated the scalability of the proposed architecture by measuring the speedup achieved increasing the number of Hadoop cluster nodes. Specifically, we considered three configurations: 1 node, 3 nodes, and 5 nodes.

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] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

[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*. New York, NY, USA: ACM Press, 2006, pp. 281–286.

[5] 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.

[6] 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.

[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] Hilderman R. and Hamilton H. J., *Knowledge Discovery and Measures of Interest*. The Springer International Series in Engineering and Computer Science, Vol. 638, 2001.

[9] Pang-Ning T. and Steinbach M. and Kumar V., *Introduction to Data Mining*. Addison-Wesley, 2006.