Journal of Network and Computer Applications 37 (2014) 194–205

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

journal homepage: www.elsevier.com/locate/jnca

Data summarization for network traffic monitoring

Demetris Hoplaros*, Zahir Tari, Ibrahim Khalil

School of Computer Science and Information Technology, RMIT University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 7 May 2012
Received in revised form
27 December 2012
Accepted 18 February 2013
Available online 28 April 2013

Keywords:
Network monitoring
Clustering
Data mining
Frequent itemsets
Summarization

ABSTRACT

Network traffic monitoring is a very difficult task, given the amount of network traffic generated even in small networks. One approach to facilitate this task is network traffic summarization. Data summarization is a key concept in data mining. However, no current measures exist in order to facilitate the evaluation of summaries. This paper presents four metrics which can be used to characterize data summarization results. *Conciseness* and *Information Loss* have already been defined, but we modified *Information Loss*, due to the fact that it was biased towards recurring attributes across individual summaries. We also propose two additional metrics, *Interestingness* and *Intelligibility*. Using the proposed metrics, we evaluated existing summarization techniques on well known network traffic datasets. We also proposed a summarization technique, based on an existing one but incorporating the proposed metrics as objective function. In order to further demonstrate the usability of the metrics, we performed classification on summarized datasets, showing that the metrics can be used to facilitate the selection of summaries for performing data mining. Using the summarized datasets with a reasonable conciseness, we were able to achieve similar results in terms of accuracy, but at a fraction of the running time, proportional to the conciseness of the summarized dataset.

1. Introduction

Data is essential to all functions in all domains. It is envisaged that each year Internet traffic doubles (Coffman and Odlyzko, 2002) and according Parkinson's Law, as long as there is storage, data will keep expanding (Berglas, 2008). One of the challenges apart from storage is that in order to obtain knowledge in any domain one needs to analyze the data (Zins, 2007). A summary of the input data could be easier and faster to analyze and still obtain similar knowledge. For example, in computer networks, a network administrator needs to monitor the activity of the network (Qin et al., 2011). Even for a small company network, the amount of data generated from different network applications, e.g. Email, FTP, HTTP and P2P applications is huge, which cannot be analyzed easily (Keys et al., 2005). Furthermore, network traffic is generated at very fast rates, making it impossible for administrators to monitor a network in real-time (Wang et al., 2005). Consequently, a summary of the network traffic is very helpful for network managers to quickly assess what is happening in the network.

A summary of a document is essentially a concise version of the original (Karras, 2009). The aim of data summarization is, given an input dataset, to provide a more concise view of the provided dataset (Chandola and Kumar, 2005; Saint-Paul et al., 2005). Summarization has been widely explored in many domains, including transactional databases (Chandola and Kumar, 2006), network data streams (Chandola and Kumar, 2005; Hu et al., 2009), Intrusion Detection Systems (IDS) (Chandola et al., 2007; Portnoy et al., 2001; Zhu, 2011), Point of Sales data (POS) (Agrawal et al., 1993) and natural text (Yu and Ren, 2009). Although this is not an exhaustive list, it can give an idea that summarization techniques are applied in many domains and they have proven to be effective in obtaining data out of large datasets, that is easier to interpret.

Knowledge discovery from data is a process, which can sometimes be very time consuming. In the case of intrusion detection (Zhu, 2011), research has shown that it can be performed on summarized datasets, and still produce a very high approximation of the results that would be produced on the original dataset, but with higher efficiency (Chandola et al., 2007). Summarization then can be a step before performing data mining (as seen in Fig. 1) which can speed up the process of knowledge discovery (Su, 2011).

In Fig. 2 we observe that network traffic monitoring can be added to the knowledge discovery process presented in Fig. 1. In this scenario, the data which is to be analyzed is Network Traffic. Traffic Analysis and Anomaly Detection can be compared to the Data Mining part of the knowledge discovery process. We can identify three applications of data summarization in network traffic monitoring and intrusion detection:

- Summarizing network traffic, which can give an overview of what is going on in the network to the administrator.

* Corresponding author. Tel.: +61 401189232.
E-mail addresses: demetris.hoplaros@rmit.edu.au, dhoplaros@gmail.com (D. Hoplaros).

1084-8045/$ - see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jnca.2013.02.021

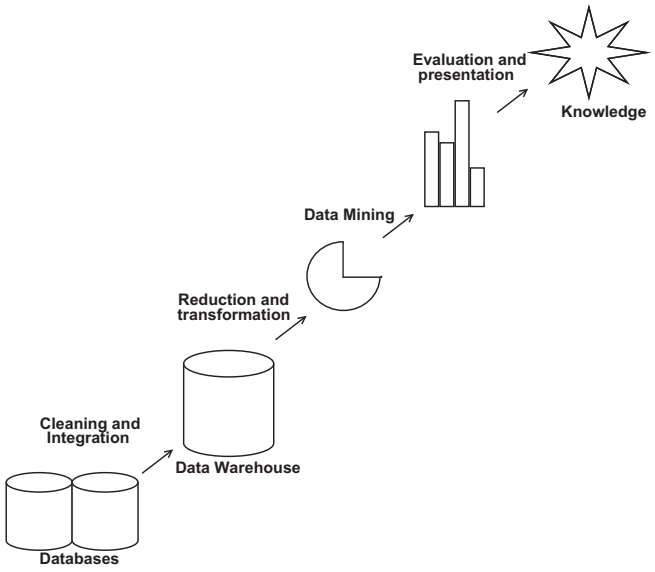

Fig. 1. Architecture of the knowledge discovery process (Singhal and Jajodia, 2006).

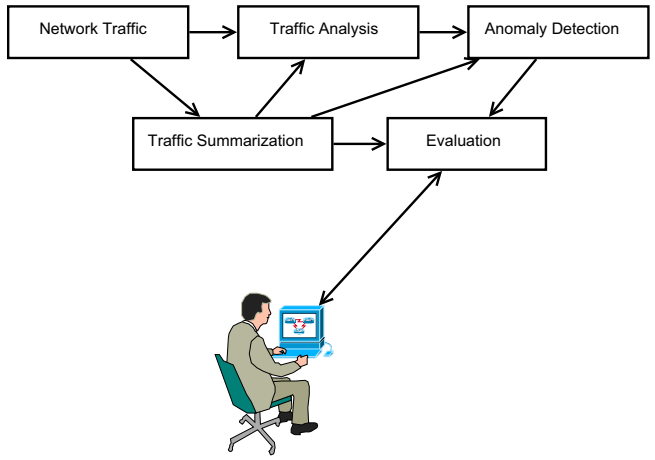

Fig. 2. Network traffic monitoring architecture and summarization applicability.

As stated above, network traffic can be huge even for small networks. A concise representation of the network traffic can facilitate network traffic monitoring.

- Summarizing network traffic, to be used as input to an Anomaly Detection algorithm.

  Intrusion Detection algorithms can be very expensive. Anomaly Detection algorithms (Xie et al., 2011) can use summarized datasets as inputs and approximate the results of using the normal dataset.
- Summarizing the alarms generated by the Anomaly Detection algorithm.

  Intrusion Detection alarms can also be huge, even if there can be a lot of alarms for the same attack. A summary of these alarms could facilitate the administrator's duty.

In all three aforementioned scenarios, the aim is to succinctly represent the data, in a way that will be efficient for both the administrator and the analysis algorithm to analyze and yield results.

There are many existing data summarization techniques, employing different concepts such as frequent itemset mining (Chandola and Kumar, 2005; Vreeken et al., 2010) and clustering (Mampaey and Vreeken, 2010; Aggarwal et al., 2005; Ha-Thuc et al., 2008), and are aimed towards different applications, such as data streams summarization (Cormode and Muthukrishnan, 2005; Hu et al., 2009), relational databases (Jin et al., 2005; Jagadish et al., 2004; Saint-Paul et al., 2005) and transactional data (Chandola and Kumar, 2005; De Bie et al., 2011). However, the evaluation for most of these techniques is based solely on their performance. Furthermore, of the few that introduce measures to characterize their results, only *Summarization ratio* (or *Compaction Gain*) and *Information Loss* (or *Distortion*) are used, and Summarization ratio is actually a parameter. More appropriate metrics could be used by both experts and data analysis algorithms as heuristics to select or disregard a particular summary for further analysis.

In this paper we present the following contributions:

- Definition of two new metrics to be used for Data Summarization (*Interestingness* and *Intelligibility*) and modification of an existing one (*Information Loss*).

  We modify one existing metric (*Information Loss*), which was biased towards recurring attributes and introduce two additional metrics to be used in the data level of summarization, *Intelligibility* and *Interestingness*. These measures can be used at the data level, therefore provide the means to the user or data mining algorithm to select the most appropriate summary to evaluate. We then evaluated three existing techniques using the proposed metrics, using known datasets. So far, summarization techniques could only be compared based on the output size, and more recently with Information Loss. Using more metrics, users can get a better idea on which summary and/or summarization technique is more appropriate given an input dataset and application.
- Proposal of a new Summarization Technique, incorporating the proposed metrics.

  Metrics can also be used as an objective function. We incorporated our metrics to an existing technique, in a manner that can easily be parameterized, therefore allowing easy modification for specific applications. This way results can be more specific, if a user requires specific upper bound on any of the measures. We also demonstrate that summarization can be helpful in data preprocessing, when dealing with large datasets, by performing classification on summarized datasets. The network traffic datasets we used are annotated, and are better tailored for classification usage. We demonstrated that a summary of the dataset can also be used for classification, and still yield similar results in terms of accuracy, given a reasonable conciseness.

In the next section we present related work, discussing the approach used to evaluate their results.

2. Related work

As presented in the Introduction, Summarization has been widely explored in many domains and applications, using a variety of techniques. In general, the aim is to succinctly represent an input dataset. In this section we will present related work on Data Summarization, along with the proposed summarization metrics in each approach.

In Xiang et al. (2010) the authors agree that there are no universal standards regarding what is a good summary or a good summarization technique. The aim of their technique is to succinctly represent a transactional database, employing the notion of *hyperrectangles*, which is the Cartesian product of a set of transactions and a set of items. In order to describe the effectiveness of