Expert Systems With Applications 64 (2016) 536–548

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Discovering users with similar internet access performance through cluster analysis

Tania Cerquitelli*, Antonio Servetti, Enrico Masala

Control and Computer Engineering Department, Politecnico di Torino, Corso Duca degli Abruzzi, 24 – 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 29 March 2016
Revised 1 July 2016
Accepted 3 August 2016
Available online 4 August 2016

Keywords:
Cluster analysis
Internet access performance
Anomaly detection
Network monitoring

ABSTRACT

Users typically subscribe to an Internet access service on the basis of a specific download speed, but the actual service may differ. Several projects are active collecting internet access performance measurements on a large scale at the end user location. However, less attention has been devoted to analyzing such data and to inform users on the received services. This paper presents MiND, a cluster-based methodology to analyze the characteristics of periodic Internet measurements collected at the end user location. MiND allows to discover (i) *groups of users* with a similar Internet access behavior and (ii) *the (few) users* with somehow anomalous service. User measurements over time have been modeled through histograms and then analyzed through a new two-level clustering strategy. MiND has been evaluated on real data collected by Neubot, an open source tool, voluntary installed by users, that periodically collects Internet measurements. Experimental results show that the majority of users can be grouped into homogeneous and cohesive clusters according to the Internet access service that they receive in practice, while a few users receiving anomalous services are correctly identified as outliers. Both users and ISPs can benefit from such information: users can constantly monitor the ISP offered service, whereas ISPs can quickly identify anomalous behaviors in their offered services and act accordingly.

1. Introduction

Currently the vast majority of people use the Internet service for a wide range of everyday activities. Internet access is obtained by signing a contract between the subscriber (i.e., the final user) and an Internet Service Provider (ISP). Each subscription is linked to a maximum theoretical download speed, which sometimes cannot be achieved due to many factors (e.g., technical issues, service delivery optimization, business rules). Thus, the received service, in particular the download speed experienced in practice, may differ from the advertised value, and neither the users nor the ISP might easily detect such fact.

Different projects have been developed to monitor the Internet access performance on a large scale by frequently measuring the download speed at the end user location. Open source tools, such as NDT (2016) and Neubot (Nexa Center, 2016), are voluntarily installed on user computers and they can provide basic information, e.g., the received download speed in the last few minutes, to the users. Furthermore, the collected data (partially anonymized) are also stored in publicly-available repositories for further inspection. An interesting but relatively unexplored research issue is how to analyze the large volume of collected measurements over time to verify whether the service received by the users is coherent with the one of other users with the same subscription or if there are anomalies. The latter information is, in general, useful for both users and ISPs. Users might be informed of the disservice which might be otherwise unnoticed or difficult to detect, and ISPs might be alerted so that they can discover potentially unexpected network behavior.

In this paper we propose a novel data analytics methodology, named MiND (Mining Neubot Data), aiming at analyzing the statistical distribution of active measurements of Internet access download speed to address two research questions: (i) Statistical behaviors of the Internet access performance received at user locations are sufficiently similar to be clustered in groups? (ii) It is possible to detect some anomalous patterns in the Internet access performance that deserve to be investigated in-depth to understand their root causes?

To address the previous questions, we employed an exploratory analytics technique, i.e., cluster analysis. This analysis method identifies groups of objects that share similar properties. Since it does not require previous knowledge of data (i.e., class labels, which in our case are anomalous services and services coherent with the one of other users with the same subscription), it has

* Corresponding author. Fax: + 39 011 0907099.
E-mail addresses: tania.cerquitelli@polito.it (T. Cerquitelli), antonio.servetti@polito.it (A. Servetti), enrico.masala@polito.it (E. Masala).

http://dx.doi.org/10.1016/j.eswa.2016.08.025
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

been widely exploited in many application domains, such as web page content (Chehreghani, Abolhassani, & Chehreghani, 2009), social networks (van Dam & van de Velden, 2015), medical data (Cerquitelli, Chiusano, and Xiao, 2016; Combes & Azema, 2013), network data (Baralis, Bianco, Cerquitelli, Chiaraviglio, & Mellia, 2013).

In our context, MiND analyzes the statistical distribution of the download speed measurements over time (through a frequency histogram) collected at the user locations to group Internet users into homogeneous and cohesive groups according to the broadband access service that they really experience. In case of users with a regular access service, most of the download speed measurements are close to their maximum download speed and there are few or no occurrences of speed values below that threshold. Moreover, it is normal that the measured speed occasionally vary (i.e., few measurements are much lower than that the maximum download speed). However, when the distribution of the download speed measurements is anomalous over time, it may be a symptom of the fact that the ISP might not be able to provide the expected service with good reliability. From the point of view of the single user, if the user experiences a download speed similar to the one of a group of other users in a given considered collection we may assume that users receive a service coherent with the subscribed one. Otherwise, we assume that an anomalous behavior has been detected. In the latter case, both the user and the ISP should be informed: users might be interested to know that in practice they receive a service different from the subscribed one, whereas ISP might have the opportunity to investigate further the unexpected network behavior and eventually fix it.

The main novelties of MiND are fourfold. (i) *Data transformation*. To highlight the relevance of Internet access in terms of bandwidth, collected measurements (download speed measurements repeated over time) have been represented through frequency histograms. Specifically, Internet bandwidths are divided into intervals (or bins) defined by a domain expert. Each histogram reports, for each bin, the total number of measurements performed by a single user. Thus, the histogram allows to compactly model all the measurements performed by the same user over time. (ii) *Two-level clustering strategy*. To correctly identify groups of users according to the download speed that they really experienced and to correctly identify anomalous patterns, a two-level clustering strategy has been proposed, based on the DBSCAN (Ester, Kriegel, Sander, & Xu, 1996b) and K-means (Hartigan & Wong, 1979) algorithms. The proposed strategy allows dealing with Internet access measurements including both noise and outlier data, as well as to group users into well-separated clusters. (iii) *A novel distance measure* has been proposed to drive the DBSCAN algorithm into correctly identifying noise and outliers. (iv) *Performance of all users are analyzed together.* Differently from previous works, MiND analyzes the statistical distribution of Internet access performance experienced by all users together to correctly model a comprehensive view of the network.

The proposed methodology has been thoroughly evaluated on real and heterogeneous datasets including data belonging to a single ISP in different geographical areas and data collected in different time intervals. Data have been collected by means of Neubot (Nexa Center, 2016), an open source software research project supported by the Nexa Center for Internet and Society of the Politecnico di Torino in Italy. The datasets used in this paper and the source code for the cluster analysis are published online in a public repository on Github (Servetti, 2016) together with a short description of the work. Experimental results demonstrate that MiND correctly identifies homogeneous and cohesive groups of users receiving a similar download speed. The MiND findings allow enhancing user awareness of the Internet access service that they really perceive and spotting anomalous network behavior that may require further analysis and investigation.

The paper is organized as follows. Section 2 summarizes the related work in the area concerning both Internet access measurement collection and their analysis. The proposed mining framework is described in Section 3 illustrating in details the algorithmic choices and how to optimally tune their parameters. A thoroughly experimental evaluation is presented in Section 4 showing the effectiveness and robustness of the proposed algorithms. Section 5 discusses the MiND findings and their possible exploitation from both the academic and managerial perspectives. Finally, Section 6 draws conclusions and discusses further developments.

2. Related work

Measurement of Internet access network speed is a popular field of investigation for multiple parties ranging from academia to governments (Marsan, 2013). On one hand, Internet regulators are actively supporting large scale network measurements to foster up to date and widespread monitoring of Internet access services in order to be able to compare broadband providers and to frame better policies to regulate them. On the other hand, users are becoming eager and eager to know how their Internet connection behaves both with respect to other ISPs and, inside the same ISP, compared to other users. For instance, in the case of Ookla Speed Wave (Ookla, 2016), group of users can compare results against each other and compete for achievements such as highest download speed and lowest latency badges.

Most of the available platforms for broadband measurements are targeted on collecting and analyzing aggregate information for interested organizations. Such platforms are based on *spot measurements* of the different access networks that ISPs offer as broadband connection to Internet users. Thus, a relatively small number of probe points on each provider are used by these platforms to make assumption on the ISP quality of service (e.g., average speed, percentage of satisfied users, etc.). These implementations are generally based on highly reliable measurements that are performed by dedicated hardware that must be delivered to the user and installed on his network. This class of platforms include: the RIPE Atlas project (RIPE, 2016), that was started in late 2010 and that now counts 6926 installed probes; the SamKnows project (SamKnows, 2016), that since 2008 is collaborating with governments and industries to benchmark broadband performance in several countries (e.g., the September 2013 campaign counted data from 6398 subscribers (Federal Communication Commission, 2014)); the Bismark project, that at the end of 2014 counts 420 devices deployed, largely in developing countries (BISmark, 2016).

Other platforms are oriented to informing users, as opposed to institutions and governments, about their specific Internet access service. Thus, to easily reach every potentially interested user, they are based on software applications that can be installed on different operating systems or used directly from the web browser. These implementations can characterize each single user connection with a very deep level of detail. In this scenario it is possible to distinguish between two schemes: user activated probes and periodic probes. The first scheme includes Ookla Speedtest.net and NDT where each test must be run directly by the user. Even if they are very popular (Ookla counts 5 million measurements each day and NDT 3 million measurements per month), both suffer from a relatively small number of *measurements per user* that clearly limits the ability to statistically characterize the behavior of the user's connection. For example, NDT completely lacks the concept of "user" because results are identified only by the client IP address which may be reused by several users over time. The second scheme includes Neubot, that provides a smaller number of mea-