Available online at www.sciencedirect.com

ScienceDirect

Information Sciences 178 (2008) 694–713

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

Privacy preserving data mining of sequential patterns for network traffic data ☆

Seung-Woo Kim [a], Sanghyun Park [a,*], Jung-Im Won [b], Sang-Wook Kim [b]

[a] *Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, 120-749 Seoul, Republic of Korea*
[b] *College of Information and Communications, Hanyang University, Republic of Korea*

Received 2 November 2006; received in revised form 3 August 2007; accepted 25 August 2007

Abstract

As the total amount of traffic data in networks has been growing at an alarming rate, there is currently a substantial body of research that attempts to mine traffic data with the purpose of obtaining useful information. For instance, there are some investigations into the detection of Internet worms and intrusions by discovering abnormal traffic patterns. However, since network traffic data contain information about the Internet usage patterns of users, network users' privacy may be compromised during the mining process. In this paper, we propose an efficient and practical method that preserves privacy during sequential pattern mining on network traffic data. In order to discover frequent sequential patterns without violating privacy, our method uses the *N*-repository server model, which operates as a single mining server and the *retention replacement* technique, which changes the answer to a query probabilistically. In addition, our method accelerates the overall mining process by maintaining the *meta tables* in each site so as to determine quickly whether candidate patterns have ever occurred in the site or not. Extensive experiments with real-world network traffic data revealed the correctness and the efficiency of the proposed method.

Keywords: Data mining; Sequential pattern; Network traffic; Privacy

1. Introduction

The number of computers connected to the Internet and exchanging data via the Internet have dramatically increased, owing to the rapid advance of network technology. Recently, a new kind of data mining has appeared in which researchers extract useful knowledge from network traffic data that are automatically gathered by a remote server [6,12,15,19,27]. Identifying patterns of network intrusions and differentiating anomalous network activity from normal network traffic data are typical examples.

☆ This work was partially supported by the Korea Research Foundation Grant funded by the Korean Government (KRF-2005-041-D00651) and the ITRC support program supervised by the IITA (IITA-2005-C1090-0502-0009).

* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.
E-mail addresses: kimsw@cs.yonsei.ac.kr (S.-W. Kim), sanghyun@cs.yonsei.ac.kr (S. Park), jiwon@hanyang.ac.kr (J.-I. Won), wook@hanyang.ac.kr (S.-W. Kim).

0020-0255/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2007.08.022