Curing Computer Logorrhea (excessive wordiness) with Sisyphus and LogHound

SAND2008-2702C

Jon Stearley
jrstear@sandia.gov
Sandia National Laboratories (US)*

Risto Vaarandi
risto.vaarandi@mil.ee
Cooperative Cyber Defence Centre of Excellence (EU)

Modern supercomputers and networks easily involve tens or hundreds of thousands of interrelated devices, each producing streams of time-stamped text (logs) as their primary record of normal – and abnormal – events. Malfunction and misuse must be localized and remedied quickly in order to minimize costly downtime or exposure of sensitive information. Efficiently finding the few lines of critical text among literally millions of lines of computer-generated text requires automated anomaly detection.

The Sisyphus log data mining toolkit provides powerful but easy-to-use tools for quickly finding important events in logs. It automatically ranks log files and colorizes their content based on information theory, quantitatively and intuitively answering the questions, "what is the strangest log file, and why?" It incorporates LogHound to automatically infer word patterns, helping to sift the subtle-but-significant from the blatant-but-boring. LogHound is developed by Risto Vaarandi of Estonia, and is thus the primary topic of this European Project Showcase submission.

During the past decade, a number of methods have been proposed for event log data mining (e.g., see [1]), with many of them being based on the breadth-first Apriori algorithm for mining frequent itemsets [2]. The development of LogHound was motivated by the shortcomings of existing methods. First, the methods are not efficient for mining longer patterns which are common in event logs [3]. Second, they have been mainly designed for mining event type patterns, ignoring patterns of other sorts (e.g., line patterns which can be used for writing signatures for event log monitoring tools). Third, existing depth-first log mining approaches assume that the event log fits into the main memory -- however, this is not the case for larger logs [3].

LogHound is a fast log mining solution that addresses these shortcomings. It employs breadth-first approach like Apriori, but uses several techniques for speeding up its work and reducing its memory consumption. It uses a hashing technique for detecting frequent words (frequent items) in a memory-efficient way, and caches frequently occurring log fragments in the main memory. It uses the itemset trie data structure like Apriori, but applies a special technique for building a reduced version of the trie (this technique is based on correlations between frequent words that are common in event logs) [3].

For the reasons of speed, LogHound has been written in C. It has several options for flexible preprocessing of the logs on-the-fly -- a user can specify a regular expression filter for processing certain log file lines only, matching lines can be converted with templates, etc. The user can also tune the memory usage of LogHound (e.g., set the cache size for log fragments), and mine closed frequent itemsets only [3]. LogHound has been employed for a variety of purposes, mainly for syslog log file analysis (about 20 institutions have reported the use of LogHound). A recently published paper describes its application for near-real-time mining of network traffic patterns from Cisco Netflow data [4].

Sisyphus provides a mature and intuitive web interface to LogHound, through which log analysts can surf emerging log patterns via their web browser. It is open-source and has been downloaded over 350 times since its initial release in early 2006. Case studies of positive production impact on supercomputers, computer centers, and networks have been published [4, 5], and work is ongoing. See http://www.cs.sandia.gov/sisyphus and http://www.estpak.ee/~risto/loghound/ for more information.

1. H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery 1(3), 1997, pp. 259-289.

2. R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. Proceedings of the 20th Conference on Very Large Data Bases, 1994, pp. 478-499.

3. R. Vaarandi. A Breadth-First Algorithm for Mining Frequent Patterns from Event Logs. Proceedings of the 2004 IFIP International Conference on Intelligence in Communication Systems, LNCS Vol. 3283, 2004, pp. 293-308.

4. Risto Vaarandi. Mining Event Logs with SLCT and LogHound. Proceedings of the 2008 IEEE/IFIP Network Operations and Management Symposium, 2008, pp. 1071-1074.

5. Jon Stearley, Adam Oliner. Bad Words: Finding Faults in Spirit's Syslogs. Workshop on Resiliency in High-Performance Computing, Lyon, France, 2008.

** Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.*