Visual Data Mining of Log Files

Guillermo Francia, III Monica Trifas Dorothy Brown Rahjima Francia Chrissy Scott
Department of Mathematical, Computing and Information Sciences
Jacksonville State University
700 Pelham Road North
Jacksonville, AL 36265

Abstract Data mining is based on a simple analogy. The growth of data warehousing has created mountains of data. The mountains represent a valuable resource to the enterprise. But to extract value from these mountains, we must "mine" for the gold in data warehouses and data marts. Everywhere that there are data warehouses, data mines are also constructed.

Data visualization has the ability to present a great deal of information in a user friendly format. It is well known that humans comprehend visual information much quicker and more efficiently than verbal information. "A picture is worth a thousand words." Successful visualizations can reduce the time it takes to get the information, make sense out of it, and enhance creative thinking. Great strides have been made in the area of computer generated data visualizations in recent years.

This paper discusses visual data mining techniques for analyzing real forensic data.

INTRODUCTION

A. Data Mining

Data is the basic form of information that needs to be collected, managed, mined and interpreted to create knowledge. Discovering the patterns, trends, and anomalies in massive data represents one of the grand challenges of the information age. The more data someone has to handle, the more difficult it is to effectively analyze and draw meaningful conclusions from it. Data mining uses analytic technologies to quickly explore mountains of data and to provide the usable information the user needs. Data mining is a multidisciplinary field drawing upon works from statistics, database technology, artificial intelligence, pattern recognition, machine learning, information theory, control theory, information retrieval, high-performance computing, and data visualization. The aim of data mining is to extract implicit, previously unknown and potentially useful patterns and models from data [8].

Data mining derives its name from the similarities between searching for valuable business information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides.

Data mining applications have been shown to be highly effective in addressing important business problems, research activities, and engineering solutions. We expect to see a continuing trend in the building and deployment of data mining and knowledge discovery applications for crucial business and scientific decision support systems.

B. Data Visualization

Data visualization is a powerful tool in the field of Information Science. Information Science, along with the field of Computer Science, saw phenomenal growth after World War II. At this time there was an "information explosion" - an exponential growth of scientific publications and literature due to the many advances in science and technology since the beginning of the twentieth century [12]. Discovering information became the new "Gold Rush". Researchers, professionals, and businesses raced each other to find this precious commodity [12]. Dictionaries state that information science deals with the collection, storage and retrieval of information. Those in the field of Computer Science, however, claim it as "a field of professional practice and scientific inquiry addressing the problem of effective communication of knowledge records [12]".

Visual representation and interaction technologies provide a mechanism allowing the user to see and understand a large amount of information at once. The human mind can indisputably understand complex information received through visual channels. Based on this ability, visual analytics facilitates the methodical reasoning process.

Creating effective visual representations is a labor-intensive process that requires a solid understanding of the visualization pipeline, characteristics of the data to be displayed, and the tasks to be performed. An efficient technique for visual representations must employ cognitive and perceptual principles that can be deployed through engineered, reusable components. Visual representation principles must address all types of data, scale and information complexity, enable knowledge discovery, and facilitate analytical reasoning.

Visual analytics software uses visual representations and interactions to accelerate rapid insight into complex data. Visual representations translate data into a visible form that highlights important features including commonalities and anomalies. These visual representations allow the users to perceive salient aspects of their data quickly. The cognitive reasoning process can be augmented through perceptual reasoning, by visual representations. Thus, the analytical reasoning process becomes faster and more focused.

It is a challenge to create well-constructed visual representations. In the field of scientific visualization, data often corresponds to real-world objects and phenomena. In scientific visualization, the goal is to reproduce these real-world representations as faithfully as computationally feasible. However, most visual analytic problems manipulate

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 531–535.

abstract information; therefore the researcher has to select the best representation for the information.

Visual representations invite the user to explore the data. This exploration requires that the user be able to interact with the data to understand trends and anomalies, isolate and reorganize information as appropriate, and engage in the analytical reasoning process. The analyst gains insight through these interactions.

The design of visual representations of information has been ongoing for centuries. Over the past 20 years, the increasing speed and availability of computers allowed information visualization researchers to create dynamic and interactive computer-mediated visual metaphors for depicting abstract information.

This paper demonstrates the viability of using Log Parser and Mineset in visual data mining. In our endeavor, we have taken real forensic data from two web servers and a desktop computer to generate visual images for data mining and interpretation. This work is part of an on-going project in the area of visualization and management of digital forensic data.

MS LOG PARSER TOOLKIT 2.2

A log is a record composed of log entries containing information about the events occurring within an organization's systems and networks. Previously, logs were used primarily for troubleshooting problems, but logs now serve many functions within most organizations, such as optimizing system and network performance, recording the actions of users, and providing data useful for investigating malicious activity [16].

Log management is vital for organizations. The information contained within these log entries are useful for performing auditing and forensic analysis, supporting an organization's internal investigations, establishing baselines, and identifying operational trends and long-term problems [16].

Filtering through a large number of log entries can be almost impossible. Once log entries are gathered, one needs to sort, aggregate, normalize, and correlate them in order to produce functional data with significant patterns. One such tool that is capable of doing the aforementioned is the MS Log Parser Toolkit 2.2.

The Log Parser tool first appeared in 2000 as a utility to test the logging mechanism of Microsoft's Internet Information Services (IIS). This allowed users to retrieve and display all the fields from a single log file in any of the three text-logging formulas supported by IIS. Since then, as tests became more complex, specifically the filtering through log entries, Microsoft saw an immediate need for a log management tool. Version 2.0 was the first available version outside Microsoft. MS Log Parser Version 2.2, which shipped in January 2005, was designed and engineered with the vision of helping users achieve their data-processing goals in a simple, fast, and powerful way [4].

Log Parser gives you a way to create a data processing channel by mixing and matching input formats and output formats as needed, using a query written in Structured Query Language (SQL). Input formats can be thought of as SQL tables containing data to be processed and output formats as SQL tables that receive the results of the data processing. Thus, the Log Parser contains an SQL-like engine core that is proficient enough in performing input and output processing of web log files [4].

Once logs are gathered, there is a need for further processing in order to make it more perceptible to users. Log data are displayed in a human-readable format for functional reporting or monitoring for anomalies. One of the most exciting features of Log Parser version 2.2 is its ability to automatically generate graphical charts based on the queried information. One can generate dozens of chart types, including bar charts, pie charts, line charts, and more. An example of a 3D bar chart that was generated by Log Parser is shown in Fig.1 [4].

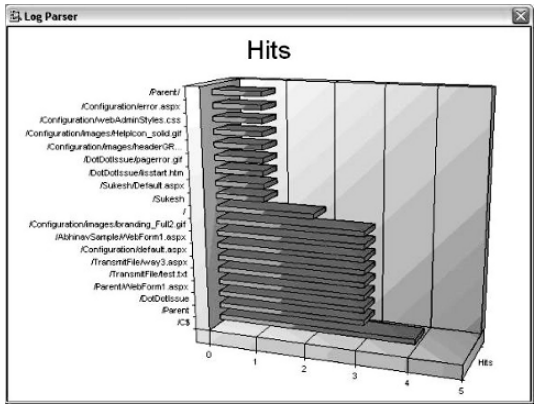

Fig. 1. Using the Log Parser

MINESET

Mineset is a commercial visual data mining product from Silicon Graphics, which was founded in 1982. Their initial focus was to introduce the market to new technologies that would allow users to interact with their data in 3D [13]. Mineset was first released by Silicon Graphics in 1996 primarily as a data mining and visualization product [1]. On October 23, 2003, Silicon Graphics announced an agreement with Purple Insight for the distribution of Mineset Data Mining and Real-Time 3D Visualization Software [15]. Purple Insight is a software and services company. It is the world's premier provider of Visual Data Mining solutions [11]. Silicon Graphics and Purple insight have the same vision in regards to the power of visualization [15].

Mineset makes use of a three tier architecture and is depicted in Fig.2. The first tier includes the Tool Manager and visualization tools. It is called the client tier. The visualization tools use mining algorithms to generate and display data and visual models. The Tool Manager is a graphical user interface that allows users to interface with Mineset. The second tier is