zapthink
zapnote

Doc. ID: ZTZN-0128-1
Released: Dec. 27, 2001

ZAPTHINK ZAPNOTE™

LOGML
REPRESENTING WEB STATISTICS AND USAGE INFORMATION

September, 2001

Analyst: Ronald Schmelzer

Abstract

The key concept behind the idea of data mining using web server log file information is that web site users are inherently "telling" the web server where the relevant information they are seeking is located, and what important information is needed so web masters can appropriately position the information for maximum efficiency. Rather than simply crawling through every page on a website, not always an efficient mechanism for obtaining this information, LOGML can be used as an XML repository for actual usage patterns.

Analyzing Web Site Traffic

The increasing need to target information to the appropriate audience has spurred the development of a host of technologies and techniques for analyzing customer and web site visitor behavior. The key to many of these technologies is the use of web site log file information. Even though all web servers store information about all the visitors who hit a particular site, the formats they use are often different, proprietary, and mostly not XML-based. A recent research effort out of the Rensselaer Polytechnic Institute (RPI) hopes to solve this problem and to mine the rich fields of data present in these log files by proposing an XML-based log file format called LOGML.

The key concept behind the idea of data mining using log file information is that web site users are inherently "telling" the web server where the relevant information they are seeking is located, and what important information is needed so web masters can appropriately position the information for maximum efficiency. Rather than simply crawling through every page on a website, not always an efficient mechanism for obtaining this information, LOGML can be used as an XML repository for actual usage patterns.

LOGML: A Research Project in XML-based Web Site Analysis

Led by researcher John Punin, the LOGML format has its roots in Punin's research on the organization of web sites. From his point of view, the structure and interaction with web sites is well-organized and controlled. Since the structure of the website resembles a "graph" structure where pages can be identified as nodes in a graph and the various links between pages as edges, a great deal of information can be gleaned both from the web site structure as well as usage of the site. An effort called Graph Modeling Language (GML) was created that represented the graph structure of web sites, and Punin followed up with development of XGMML, a GML successor. Everything was started based on John's research on organization of web sites.

Mainly an academic effort under the mentorship of Mukkai Krishnamoorthy, Punin's advisor and professor at RPI, LOGML is an XML-based application designed to describe log reports and usage patterns of web servers. Just as relational and object-oriented databases can be "mined" to collect useful information about key areas of user interest, mining web usage logs can be used to identify customer choices as well as usage patterns and proper site organization. Log files indicate which web pages are most frequently accessed by web visitors, and this usage can be represented as a web graph in XGMML. This graph structure can then be annotated using the log information to guide web site developers and managers in proper site development and construction. The resultant information can be delivered to

users by means of summary reports that comprise useful web site information such as client sites, types of browsers and the usage time statistics. User activity on the web site can be further detailed by extracting a subgraph from the derived web site graph. This subgraph can then be analyzed to investigate general user activity on the web site.

LOGML can also be used as a compressed format for collection of key user information. Since usage information follows set patterns, the structure of XML can be used to aggregate repetitive log information. Experimentally, Punin, Krishnamoorthy ,and Zaki have shown that typical log file compression over currently used formats is about half the size for the same information.

Continued Development of LOGML

Further development of LOGML is along mainly academic, rather than commercial, lines. Once the structure of web sites is fully explored, attention will be focused on various data mining techniques, such as clustering. These techniques can be used to identify the location and frequency of which data is accessed. With an XML- based log file format such as LOGML, users can apply several well-known algorithms for locating and defining clusters using web site log file information.

Further directions for Punin's research includes integration with concepts of the Semantic Web to illustrate how semantically related concepts can be derived from actual web site usage. The log file information allows a developer to automatically define semantic information about a website without having to create from scratch or crawl a web site. In essence, the user is creating the semantics automatically for the user, which further allows web site designers to create systems that are more targeted, efficient, and relevant.

Other XML Formats

Since the LOGML effort is mainly academic, rather than commercial, it is not intended to compete with or even pose a practical alternative to commercially available log file formats. However, it is hoped that commercial vendors can take advantage of the research and create more intelligent, XML-based formats for use in data mining and other information collection. Even though the LOGML format itself is not a commercial product, it is being used in a variety of different applications. WWWPal makes use of LOGML as part of a larger web site analysis, visualization, and structure analysis application. Logreport.org is looking to use the format as part of their open source log file analysis effort, and OpenJGraph utilizes LOGML and XGMML for its web site graphic functionality.

Key Conclusions & Recommendations

- LOGML is mostly an academic effort and not one that is being commercially supported. However, vendors and users interested in representation of web log information will be interested in the work developed here.

Profile: LOGML	(September, 2001)
Date Founded:	
Funding: Research effort	
Specifications:	
• LOGML	
URL: http://www.cs.rpi.edu/~puninj/	
Main Phone:	
Contacts:	
John Punin puninj@cs.rpi.edu	

Related Research

- *Web Services Technologies and Trends* Report (ZT-WEBSERV)

About ZapThink, LLC

ZapThink is an IT market intelligence firm that provides trusted advice and critical insight into XML, Web Services, and Service Orientation. We provide our target audience of IT vendors, service providers and end-users a clear roadmap for standards-based, loosely coupled distributed computing – a vision of IT meeting the needs of the agile business.

ZapThink's role is to help companies understand these IT products and services in the context of SOAs and the vision of Service Orientation. ZapThink provides market intelligence to IT vendors who offer XML and Web Services-based products to help them understand their competitive landscape and how to communicate their value proposition to their customers within the context of Service Orientation, and lay out their product roadmaps for the coming wave of Service Orientation. ZapThink also provides implementation intelligence to IT users who are seeking guidance and clarity into how to assemble the available products and services into a coherent roadmap to Service Orientation. Finally, ZapThink provides demand intelligence to IT vendors and service providers who must understand the needs of IT users as they follow the roadmap to Service Orientation.

ZapThink's senior analysts are widely regarded as the "go to analysts" for XML, Web Services, and SOAs by vendors, end-users, and the press. They are in great demand as speakers, and have presented at conferences and industry events around the world. They are among the most quoted industry analysts in the IT industry.

ZapThink was founded in October 2000 and is headquartered in Waltham, Massachusetts. Its customers include Global 1000 firms, public sector organizations around the world, and many emerging businesses. ZapThink Analysts have years of experience in IT as well as research and analysis. Its analysts have previously been with such firms as IDC and ChannelWave, and have sat on the working group committees for standards bodies such as RosettaNet, UDDI, CPExchange, ebXML, EIDX, and CompTIA.

Call, email, or visit the ZapThink Web site to learn more about how ZapThink can help you to better understand how XML and Web Services impact your business or organization.

ZAPTHINK CONTACT:
ZapThink, LLC
11 Willow Street
Suite 200
Waltham, MA 02453
Phone: +1 (781) 207 0203
Fax: +1 (786) 524 3186
info@zapthink.com