# LOGML - XML Language for Web Usage Mining

John Punin
Department of Computer Science, Rensselaer Polytechnic Institute
puninj@cs.rpi.edu

Mukkai Krishnamoorthy
Department of Computer Science, Rensselaer Polytechnic Institute
moorthy@cs.rpi.edu

Mohammed Zaki
Department of Computer Science, Rensselaer Polytechnic Institute
zaki@cs.rpi.edu

## ABSTRACT

We propose a new XML language, LOGML. LOGML is a web-log report description language. We generate web-log reports in LOGML format for a web site from web log files and the web graph. In this paper, we further illustrate the usefulness of this XML application with a web data mining example. We provide sample results, namely frequent patterns of users in a web site, with our web usage mining algorithm.

## Keywords

XML, LOGML, XGMML, Web Usage Mining, Web Characterization, Web Graph, WWWPal System, Frequent Pattern Mining

## 1. INTRODUCTION

Recently XML has gained wider acceptance in both commercial and research establishments. In this paper, we suggest a new XML language and a web data mining application which utilizes it to extract complex structural information.

Log Markup Language (LOGML) [8] is an XML 1.0 application designed to describe log reports of web servers. Web data mining is one of the current hot topics in computer science. Mining data that has been collected from web server logfiles, is not only useful for studying customer choices, but also helps to better organize web pages. This is accomplished by knowing which web pages are most frequently accessed by the web surfers. Further we produce summary reports, comprising of information such as client sites, types of browsers and the usage time statistics. We also gather the client activity in a web site as a subgraph of the web site graph. This subgraph can be used to get better understanding of general user activity in the web site. In LOGML, we create a new XML vocabulary to structurally express the contents of the logfile information.

Recently web data mining has been gaining a lot of attention because of its potential commercial benefits. For example, consider a web log database at a popular site, where an object is a web user and an attribute is a web page. The mined patterns could be the sets or sequences of most frequently accessed pages at that site. This kind of information can be used to restructure the web-site, or to dynamically insert relevant links in web pages based on user access patterns. Furthermore, click-stream mining can help E-commerce vendors to target potential online customers in a more effective way, at the same time enabling personalized service to the customers.

Web Mining is an umbrella term that refers to mainly two distinct tasks. One is Web Content Mining [9], which deals with problems of automatic information filtering and categorization, intelligent search agents, and personalize web agents. Web Usage Mining [9] on the other hand relies on the structure of the site, and concerns itself with discovering interesting information from user navigational behavior as stored in web access logs.

The focus of this paper is on using web usage mining. While extracting simple information from web logs is easy, mining complex structural information is very challenging. Data cleaning and preparation constitute a very significant effort before mining can even be applied. The relevant data challenges include: elimination of irrelevant information such as image files and cgi scripts, user identification, user session formation, and incorporating temporal windows in the user modeling. After all this pre-processing, one is ready to mine the resulting database.

The proposed LOGML language has been designed to facilitate this web mining process in addition to storing additional detailed summary information extracted from web logs. Using the LOGML generated documents the pre-processing steps of mining are considerably simplified. We also propose a new mining paradigm, called Frequent Pattern Mining, to extract increasingly informative patterns from the LOGML database.

## 2. LOG MARKUP LANGUAGE (LOGML)

Web-log reports are the compressed version of logfiles. Web masters in general save web server logs in several files. Usually each logfile contains a single day of information. Due to disk space limitation, old log data gets deleted to make room for new log information. Generally, web masters generate HTML reports of the logfiles and do not have problems keeping them for a long period of time as the HTML reports are an insignificant size. If a web master likes to generate reports for a large period of time, he has to combine several HTML reports to produce a final report. LOGML is conceived to make this task easier. Web masters can generate LOGML reports of logfiles and combine them on a regular basis without much effort. LOGML files can be combined with XSLT [2] to produce HTML reports. LOGML offers the flexibility to combine them with other XML applications, such as SVG [3], to produce graphics of the statistics of the reports. LOGML can also be combined with RDF [4] to provide some metadata information about the web server that is being analyzed. LOGML is based on Extensible Markup and Modeling Language (XGMML) [7]. LOGML document can be seen as a snapshot of the web site as the

user visits web pages and traverses hyperlinks. It also provides a succinct way to save the user sessions. In the W3C Working Draft "Web Characterization Terminology & Definitions Sheet", the user session is defined as "a delimited set of user clicks across one or more Web servers" [5].

## 2.1 Structure of LOGML Documents

A typical LOGML document has three sections under `logml` element (the root element). The first section is a graph that describes the log graph of the visits of users to web pages and hyperlinks. This section uses XGMML [7] to describe the graph and its root element is the `graph` element. The second section is the additional information of log reports such as top visiting hosts, top user agents, and top keywords. The third section is the report of the user sessions. Each user session is a subgraph of the log graph. The subgraphs are reported as a list of edges that refer to the nodes of the log graph. Each edge of the user sessions also has a timestamp for when the edge was traversed. This timestamp helps to compute the total time of the user session. The complete LOGML specification, DTD, LOGML Schema (XML Schema) and examples can be found at [8].

# 3. USING LOGML FOR WEB DATA MINING

In this section, we propose solving a wide class of mining problems that arise in web data mining, using a novel, generic framework, which we term Frequent Pattern Mining (FPM). FPM not only encompasses important data mining techniques like discovering associations and frequent sequences, but at the same time generalizes the problem to include more complex patterns like tree mining and graph mining. These patterns arise in complex domains like the web. Association mining, and frequent subsequence mining are some of the specific instances of FPM that have been studied in the past [1, 11, 10, 6, 12]. In general, however, we can discover increasingly complex structures from the same database. Such complex patterns include frequent subtrees, frequent DAGs and frequent directed or undirected subgraphs. As one increases the complexity of the structures to be discovered, one extracts more informative patterns.

The same underlying LOGML document that stores the web graph, as well as the user sessions, which are subgraphs of the web graph, can be used to extract increasingly complex and more informative patterns. Given a LOGML document extracted from the database of web access logs at a popular site, one can perform several mining tasks. The simplest is to ignore all link information from the user sessions, and to mine only the frequent sets of pages accessed by users. The next step can be to form for each user the sequence of links they followed, and to mine the most frequent user access paths. It is also possible to look at only the forward accesses of a user, and to mine the most frequently accessed subtrees at that site. Generalizing even further, a web site can be modeled as a directed graph, since in addition to the forward hyperlinks, it can have back references, creating cycles. Given a database of user accesses (with full information about their traversals, including forward and backward links) one can discover the frequently occurring subgraphs.

We applied the Eclat association mining algorithm [11] to a real LOGML document from the RPI web site (one day's logs). There were 200 user sessions with an average of 56 distinct nodes in each session. It took us 0.03s to do the mining with 10% minimum support. An example frequent set found is shown below:

```
FREQUENCY = 22 , NODE IDS =  25854 5938 25649 25650 25310 16511
   http://www.cs.rpi.edu/~sibel/poetry/poems/nazim_hikmet/turkce.html
   http://www.cs.rpi.edu/~sibel/poetry/sair_listesi.html
   http://www.cs.rpi.edu/~sibel/poetry/frames/nazim_hikmet_1.html
   http://www.cs.rpi.edu/~sibel/poetry/frames/nazim_hikmet_2.html
   http://www.cs.rpi.edu/~sibel/poetry/links.html
   http://www.cs.rpi.edu/~sibel/poetry/nazim_hikmet.html
```

# 4. REFERENCES


[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In U. Fayyad and et al, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, Menlo Park, CA, 1996.

[2] J. Clark and S. Deach. Extensible Stylesheet Language (XSL), Version 1.0. `http://www.w3.org/TR/WD-xsl`, 2000.

[3] J. Ferraiolo. Scalable Vector Graphics (SVG) 1.0 Specification. `http://www.w3.org/TR/SVG/`, 2000.

[4] O. Lassila and R. Swick. Resource description framework (RDF) model and syntax specification. `http://www.w3.org/TR/REC-rdf-syntax/`, 1999.

[5] B. Lavoie and H. F. Nielsen. Web Characterization Terminology & Definitions Sheet. `http://www.w3.org/1999/05/WCA-terms/`, 1999.

[6] H. Mannila, H. Toivonen, and I. Verkamo. Discovering frequent episodes in sequences. In *1st Intl. Conf. Knowledge Discovery and Data Mining*, 1995.

[7] J. Punin and M. Krishnamoorthy. Extensible Graph Markup and Modeling Language (XGMML) Specification. `http://www.cs.rpi.edu/~puninj/XGMML/draft-xgmml.html`, 1999. Status: INFORMATIONAL.

[8] J. Punin and M. Krishnamoorthy. Log Markup Language (LOGML) Specification. `http://www.cs.rpi.edu/~puninj/LOGML/draft-logml.html`, 2000. Status: INFORMATIONAL.

[9] R. Cooley, B. Mobasher, and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. In *8th IEEE Intl. Conf. on Tools with AI*, 1997.

[10] R. Srikant and R. Agrawal. Mining generalized association rules. In *21st VLDB Conf.*, 1995.

[11] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372-390, May-June 2000.

[12] M. J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, 42(1/2), Jan/Feb 2001. Special issue on Unsupervised Learning.