WUMprep
Logfile Preparation for Data Mining with WUM

Carsten Pohle

May 20, 2003

Abstract

This document describes the process of preparing Web server logfiles for dataming using the miner `WUM`. It is also the main documentation for the `WUMprep` suite of Perl scripts used for this purpose.

This document is slighly out of date. Use it with care ;-)

1 Introduction

`WUMprep` is a collection of Perl scripts supporting data preparation for data mining Web server logfiles. It's primary purpose is to be used in conjunction with the Web usage miner `WUM`, but `WUMprep` might also be used standalone or in conjunction with other tools for Web log analysis. This document is intended to give an overview of data preparation using the `WUMprep` tools.

Prototypically, preparing Web server log files for mining with `WUM` requires the following steps:

1. Conversion of the logfile into the "extended cookie" format
2. Removal of irrelevant requests
3. Removal of duplicate requests
4. *Optional:* Try to resolve host IP addresses into hostnames
5. Definition of sessions
6. Removal of robot requests
7. Application specific data preparation

Each of these steps is supported by certain Perl scripts, each of them having its own inline-documentation, explaining the usage and the underlying algorithms in greater detail. It can be accessed by invoking the command `perldoc script.pl` on the command line, where script.pl is replaced with the Perl scrip's filename. (Please note that you have to specify the script's complete path if the script directory is not contained in the `PATH` environment variable.)

All options and parameters for the `WUMprep` scripts are stored in a file called `wumprep.conf`. A template of this file is included in the directory containing the `WUMprep` Perl scripts. This template is well documented and should be self-explaining. The configuration file is expected to reside in the directory containing the logfiles to be processed.

1.1 Logfile conversion

DEPRECATED! REWRITE THIS SECTION!

Virtually every Web server writes logfiles of the received requests and answers. Depending on the used server software, the records of these logfiles may contain different kinds and numbers of fields.

To keep the `WUMprep` scripts simple, they have been designed to support only one of the several logfile formats, referred to as the "extended cookie format". A sample log line is presented in Figure 1.

The "extended cookie format" serves as a generic format most logfiles can be converted into. In the `WUMprep` suite, the script `logConv.pl` does the logfile conversion. See the script documentation for details about the supported source log formats.

```
picasso.wiwi.hu-berlin.de - - [10/Dec/1999:23:06:31 +0200]
"GET /index.html HTTP/1.0" 200 3540 "http://www.berlin.de/"
"Mozilla/3.01 (Win95; I)" "VisitorID=10001; SessionID=20001"
```

Figure 1: Sample "extended cookie format" log line

1.2 Removing irrelevant requests

The idea behind the `WUM` mining model is to analyze usage patterns. For this purpose, we are interested in information about the paths visitors take when traversing a Web site, as is included in Web server logfiles. These logfiles not only contain requests to the pages comprising the Web site, but also requests of images, scripts etc. embedded in these pages. These "secondary" requests are not needed for the analysis and thus irrelevant – they must be removed from the logs before mining.

The script `logFilter.pl` is designed to perform this task of data cleaning.

1.3 Removing duplicate requests

If a network connection is slow or a server's respond time is low, a visitor might issue several successive clicks on the same link before the requested page is finally showed in his browser. Those duplicate requestes are noise in the date and should be removed.

This is the script's `logFilter.pl` second job. It detects such duplicates in the log and drops all but the first occurences.

1.4 Resolving host IP addresses

Depending on the Web server configuration, either a host's IP address or its hostname is logged. For data preparation purposes, knowing the hostnames has some advantages about working with IP addresses. For example, many proxy servers of major internet service providers identify themselfes as proxies in their hostnames. Those log entries could be removed to improve the accuracy of the data mining results when user identification relies on hostnames.

Most IP addresses can be resolved to hostnames with appropriate DNS queries. This job is done by the script `reverseLookup.pl`.

1.5 Definition of sessions

For further data preparation and data mining tasks, it is neccessary to divide logfiles into user sessions. A session is a contiguous series of requests from a single host. Multiple sessions of the same host can be divided by measuring a maximal page view time for a single page, using a user/session identifcation cookie or defining one or more pages as "session-starters".

In the `WUMprep` suite, `sessionize.pl` is the script that supports this task. It prefixes each host field in the log with a session identifyer. For details about the criteria used for session identification, please resort to the script's inline documentation.

1.6 Removing robot requests

On many Websites, a significant fraction of the requests stem from robots, indexers, spiders or agents. Since these requests are generated automatically, their traces in the logfile do not represent human browsing behaviour and thus adulterate mining results.

To distinguish between human users and hosts that are robots, there exist several heuristics. They are implemented in the script `removeRobots.pl` and desribed in the script's inline documentation.

1.7 Further data preparation steps

The data preparation steps described so far can be viewed as "generic" ones, applying to most Web usage mining tasks. Now, any irrelevant or disturbing data have been removed and the logs are divided into single user sessions.

What follows now is application specific data preparation, for which no generic algorithms are provided by `WUMprep`.