The WebDruid Roadmap

Fabien Chevalier

< fabien@juliana-multimedia.com >

v. 0.6

Table of Contents

I)Abstract:

This document is the repository of features i plan to add to 'The WebDruid'.

This document is supposed to reflect what i have in mind for The WebDruid future,
and involves changes that i decided to incorporated in the WebDruid which come for two sources:

- My own audit of 'The Webalizer', which i've achieved at the end of august 2004.
- Suggestions from numerous Webalizer users, who would like additional features on their favourite statistics package.

If you think of something great that is not present here, fill free to get in touch in me and say what you think!
I can't promise i'll do all what people want me to do, as i have only a limited timeframe for this projet, but if many people ask for it, it will be included in this roadmap.

If you think one of these features in worth you spending time on it, then drop me a note so that i can help you in your work!

II)Definitions:

These words will be used with the following meanings:

<u>The user</u>:
The end user of The WebDruid, the one who browses the result HTML pages.

<u>The administrator</u>:
The one who will compile, make The WebDruid run on the target system, and who will write the text configuration file 'webdruid.conf'.

III) New functionalities to be added:

A) How it will be done:

This deels with two kinds of needs:

- While keeping its core functionalities, make it also a useful tool for basic users. The information should be easier to understand.
- Have more statistical information

The detail of wished modifications is listed in the two sections below.

Each functionality is explained with the following scheme:

- The Name of the functionality
- Its description
- The estimated time required to code and do first level testing.

Each functionality can be coded separately. There is no preferred order.

B) Modifications towards a more easy to understand WebdDruid:

Name:

Title for URLs.

Functionalities:

The WebDruid displays a lot of URLs in its statistics table.
The URLs are not always useful for owner of web sites which didn't made it themselves.
Thus, titles should be displayed most of the time ... but maybe not for the actual 'View All Referrers' and 'View all URLs', as it would generate a lot of requests.
Titles should be retrieved by sending a “GET” request to the machine hosting the web site, and seeking for the <TITLE> </TITLE> couple.
Caution should be taken while coding, as it is likely that The WebDruid will send http GET requests to the server hosting the target web site. This should not disturb the statistics, nor load the web server too much.
Caching of titles will have to be studied
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 5

Name:

Online Glossary.

Functionalities:

The WebDruid has a lot of user documentation, but it is located in the source tree, which means you must have the source if you want to have explanations on words like Hits, Visits, Pages and so on...
The goal here is to have the end user terms explained in a glossary accessible through links within the tables of the WebDruid.
Please note that technical documentation on 'how to run, compile...' and everything else should remain in the README files.
Technical Note: README files will be updated to explain this functionality.

Scheduled days of work: 1

Name:

Online Help.

Functionalities:

The WebDruid displays all his tables, charts without any explanation.
A command-line and/or config-file parameter will be added to display a small paragraph explaining the information before the tables/charts of each section.
The default will be to turn it on. The parameter will be global for all tables/charts.
Technical Note: README files and man pages will be updated to explain this functionnality.

Scheduled days of work: 2

Name:

Better charts.

Functionalities:

The goal here is to add scales to the charts, what would make them more useful when there is no table to complete them.
Should be added the possibility to change the background color of the charts, or better, to make it to be transparent.
The 'usage' chart should be split into 3 distinct charts. The current layout with one big chart and two small sticked to it could be recreate by using an HTML table.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 5

Name:

Trivial Information.

Functionalities:

Most users are a bit afraid of the number of columns. Even for computer engineer, it can be difficult to understand without reading the manual.
Thus, it could be wise to be able to disable Hits, Files, Pages, Sites, Kbytes, and to keep only Visits.

A command-line and/or config-file parameter will be added to enable/disable this functionality. The default will be to turn it off.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 1

Name:

Root Page.

Functionalities:

In the top urls table, the root is displayed as '/'. It would be nice to have it a labeled as 'Home Page'.

Scheduled days of work: < 1

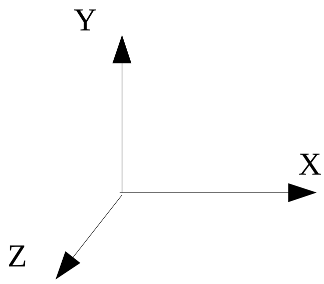
C) Modifications to produce more statistics:

Name:

3D chart.

Functionalities:

The goal here is to provide informations for more than a year, and to provide information on evolution for the same month through the years.
The x axis will be the months, form january to december.
The y axis will be the number of visits/pages/files. If it makes the image too loaded, the chart will have either to be split in three charts, or some items in {visits, pages, files} will have to be removed.
The z axis will be the years, with the newest years at the back.



A command-line and/or config-file parameter will be added to enable/disable this functionality. The default will be to turn it on.
Technical Note: README files and man pages will be updated to explain this functionality.

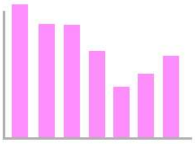
Scheduled days of work: 5

Name:

Time for top Urls.

Functionalities:

The column for average time and its standard deviation should be added to the top urls table.
A bar chart should be added to provide visual feedback. The x axis will be for time and the y axis will be for the numer of connections.
The bar number n will display the number of users who stayed between (n-1) and n minutes on this page.



The mean and standard deviation will always be displayed.
A command-line and/or config-file parameter will be added to enable/disable the bar charts. The default will be to turn it on.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 2

Name:

List of downloaded medias.

Functionalities:

Nowadays web sites tend to contain more and more medias. It is current to have images, sounds, videos, documents, which are not part of the web site design, but which are needed by their own.

The goal here is to have a monthly table for each kind of medias with the number of downloads.

A command-line and/or config-file parameter will be added to enable/disable the functionality. The default will be to turn it on.

A global command-line and/or config-file parameter will be added to limit the maximum number of entries in the top tables. The default will be 50.

Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 3

Name:

Users countries.

Functionalities:

'The WebDruid' uses reverse DNS lookups to try to guess the country of the site user. This has a few bottlenecks: Some users remain 'unresolved', some are seen as .net, .com... which is not very useful. And attempt to solve this problem has been made by the use of the GeoIP library, but brings other kind of issues (How to easily maintain the database?).
The suggestion here is to use the 'Accept-Language' HTTP header field, as described by the RFC 2616, as a country approximation.
This will require that interested user add this information to their log format. The precise 'how-to-do' will have to be detailed in the README for the Apache web server.

Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 2

Name:

Flop URLs.

Functionalities:

This is supposed to be the dual of the current 'Top URLs'.

A command-line and/or config-file parameter will be added to enable/disable the functionality. The default will be to turn it off.

A command-line and/or config-file parameter will be added to limit the maximum number of entries in the flop table. The default will be 10.

Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 2

Name:

Dynamic Pages support.

Functionalities:

Most of the following is from a webdruid user. I think it is pretty clear...

read_blog.php?blog_id=1

is no the same than:

read_blog.php?blog_id=2

We need « something » to be able to differentiate these two pages.

I guess you could add a new config option named
"DynamicPage" with the
following arguments:

DynamicPage script_url important_param1= i_param2=
i_param3= ...

script_url will be, for example /read-blog.php (to match read-blog on the document root) or just index.php (to match every request to index.php)

the parameters are the name of the "get" params as in engines.list.

Now, when you are processing the logs, you will have to check after accepting the request as an interesting one (it's not an img, etc) if the current request matches one of the script_urls. Then, you break the query string by "&", get only the important parameters, abc
sort them (see later) and use the script url plus the important parameters as the requested url

* You have to sort the parameters because most of the web scripting languages treat urls with the get parameters in different order as if it were the same.
So:
index.php?a=1&b=1&c=1&d=1
and
index.php?d=1&c=1&b=1&a=1

will call refer to the same file, but if you have "a=" and "c=" as important parameters, you will end up using 2 different urls after creating the basic url (url+imp params) for the same content:

index.php?a=1&c=1
in the 2nd,
index.php?c=1&a=1

If you order the parameters, you will get the same url even if the user asks for the same url with the params in different order.

Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 5

D) Non visible modifications:

Name:

XML Output.

Functionalities:

Current output is directly generated as HTML formats.
Its has some issues:
Output module is a mix of data computation and user interface definition.

- It is not easily possible to fine tune the reports.
- It is impossible to change totally the report without rewriting all the output module and the computation algorithm.
- UI is written in C, which may not be the best langage to write HTML.

So the suggestion here is to first write a master XML document.
Then people may easy generate GUIs based on this XML document.

Scheduled days of work: 5

Name:

Documentation rewrite.

Functionalities:

I want to get rid of this 90K long README file.
The new one will be an OpenOffice.org document, which will allow me to generate easily PDF and XHTML help document.

Scheduled days of work: 2

Name:

Custom home page.

Functionalities:

```
Here is the comments of a user:

Isn't there an option to specify which page is the
home page? We have websites like "griho.udl.es/ipo" ,
and the graph will have no home page, probably because
it expects to find the homepage in "griho.udl.es/"
Unfortunately, I strip all URLS not starting with
"griho.udl.es/ipo" so that the expected homepage will
never appear in that graph because it is not even seen
by webdruid. If there was an option in the conf file
where I could specify "griho.udl.es/ipo/pres.html" as
the home page, then the graph would be correct and way
more understandable. For example:


Homepage  /ipo/pres.html
```

Scheduled days of work: 1

D) Summary:

Patch Name	***# Days***
Title for URLs	5
Online Glossary	1
Online Help	2
Better charts	5
Trivial Information	1
Root Page	1
3D chart	5
Time for top Urls	2
List of downloaded medias	3
User countries	2
Flop URLs	2
Dynamic pages support	5
XML modularization	5
Documentation rewrite	2
Custom home page	1
TOTAL	**42**

IV) Obsoleted features:

This is a list of the features, which have been deprecated.
Reason is indicated along.

Name:

Downloadable statistics.

Functionalities:

A link will be added somewhere to download a zip archive containing the whole site.
A command-line and/or config-file parameter will be added to enable/disable this functionality. The default will be to turn it off.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 1

Deprecation reason:

After more questionning, people don't see any interest for this feature *'Hey guy, all reports are avaible on the net! Why do you want me to store them?'*

Name:

User Interface Rewrite.

Functionalities:

The Webalizer should be runnable in two modes:

- an 'embedded' mode, looking similar to what actually exist.
- a 'web site' mode, which would use all new web design tools available: JavaScript, CSS. This should be viewable by any recent web browser, at least Mozilla 1.0 and Internet Explorer 5.0. If an administrator knows his users are using old browsers, he will instead use embedded mode, which MUST NOT require JavaScript or CSS.
  The Look and Feel of the new design could be like the sample provided with this document.
  In all cases, the cascading style sheets should be used to provide he ability for the users to customize the look of The Webalizer. All style sheets settings should be stored on a file named 'webalizer.css' and copied to the output root at the end of the process. A command-line and/or config-file parameter will be added to give the opportunity to replace the style sheet by a custom one.
  Most of the JavaScript code should be stored on a file named 'webalizer.js' and copied to the output root at the end of the process. This will prevent having to generate Javascript code from within 'The Webalizer' source code.

The 'web site' mode will be the default, and The Webalizer will switch to embedded mode if a command-line and/or config-file parameter is provided.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 7

Deprecation reason:

This one is to be split into smaller tasks.

V) Work completed:

This is a list of what has been done currently implemented. Notes included.

Name:

Search engines Page.

Functionalities:

The current table does not directly deal with the search engines. You have to configure your webalizer.conf with something like:

```
....
SearchEngine        google.    q=
GroupReferrer       google.    Google
SearchEngine        yahoo.     p=
GroupReferrer       yahoo.     Yahoo
....
```

The goal here is to have the list of search engine included in 'The Webalizer'.
The list will be stored in a text file, so that it remains editable by the end user, but will provide good default value.
A suggestion is to 'steal' it from the AWstats project, and then to join efforts to maintain it so that the work is not done twice.
This file will be updated with each release of the Webalizer.
The 'Search Engines' monthly table will contain the following columns:

- Search Engine: the name of the engine
- Key Expressions: will contain one line for each key expression which was used to reach the site.
- Number of requests: for each key expression.
- %: percentage relative to the whole list.
- Entry page: where the search engine has sent the user.

A second table will be built which will replace 'key expressions' with 'key words'.

Before these two tables, a table will be built, which will create a summary for each search engine, i.e. Without the key expressions/words ans the entry page.

A command-line and/or config-file parameter will be added to enable/disable the 'key expressions' table. The default will be to turn it on.
A command-line and/or config-file parameter will be added to enable/disable the 'key words' table. The default will be to turn it on.

Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 7

Notes:

Partially implemented. The global list has not been implemented yet. It will only be if some people do ask for it. The list with keywords hasn't been implemented too. It will only be if some people do ask for it.

Name:

Path chart.

Functionalities:

This is a new tool to provide visual feedback of what people are doing in the web site. The goal is to have the URLs as nodes, and links between urls will be labeled with the number users who went between these two nodes.
To achieve this goal the use of graphviz (http://www.research.att.com/sw/tools/graphviz/), the open source graph drawing software, will be of great help.
A command-line and/or config-file parameter will be added to enable/disable this functionality. The default will be to turn it on.
A command-line and/or config-file parameter will be added to limit the maximum number of arcs between nodes (which will imply a limitation of the node's number). It will have a default value that fulfills most needs.
Technical Note: README files and man pages will be updated to explain this functionality.

Scheduled days of work: 10

Notes:

Work is completely finished. There are a few issues which are still to be dealed with:

- Graph printing: du to it's sometimes heretic dimension, some users ave reported trouble in printing the graph.
- There is an issue that, when generating the graph, graphviz displays a message saying it cannot find fonts in directories he looks in.

Name:

W3C log format.

Functionalities:

Many people have asked for this.
The goal here is to test an integrate Webalizer's W3C patch.
This should allow us to support IIS 5.0 and 6.0 log files processing.

Scheduled days of work: 1

Notes:

Done – no more things to say!.

Name:

Multiple log files.

Functionalities:

Many people have asked for this.
The WebDruid must be able to read its log records from more than one file at a time. This will make life for a lot of people easier, especially those dealing with server farms.

Scheduled days of work: 3

Notes:

Done – no more things to say!.