Automatic Extraction of Structure, Content and Usage Data Statistics of Web Sites

Ioannis Paparrizos
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
(EPFL) CH-1015 Lausanne, Switzerland

ioannis.paparrizos@epfl.ch

Vassiliki Koutsonikola, Lefteris Angelis, Athena Vakali
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece

{vkoutson, lef, avakali}@csd.auth.gr

ABSTRACT
In this paper we present a web mining tool which automatically extracts the structure, content and usage data statistics of web sites. This work inspired by the fact that web mining consists of three axes: web structure mining, web content mining and web usage mining. Each one of those axes is using the structure, content and usage data respectively. The scope is to use the developed multi-thread web crawler as a tool to automatically extract from web pages data that are associated with each one of those three axes in order afterwards to compute several useful descriptive statistics and apply advanced mathematical and statistical methods. A description of our system is provided as well as some experimentation results.

Categories and Subject Descriptors
H.4.m [**Information Systems**] Miscellaneous. I.1.2 [**Computing Methodologies**]: Algorithms – *Nonalgebraic algorithms*.

General Terms
Algorithms, Design, Experimentation.

Keywords
Crawling, Classification Algorithm, Web Mining. Structure Content and Usage data.

1. INTRODUCTION
With the explosive growth of information sources available on World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find the desired information resources. Web mining refers to the extraction of interesting and potentially useful patterns and implicit information derived from the content and structure of web resources or usage activity. Thus, depending on data that are being used, web mining consists of three axes known as web structure mining, web content mining and web usage mining [2].

Inspired by that categorization and as according to authors' knowledge there is no tool to automatically extract from web pages data that are associated with each one of these categories, we developed a multi-thread web crawler for this purpose. By collecting data in those three categories it is easy afterwards to compute several useful statistics, apply advanced mathematical and statistical methods and visualize them to identify important trends and patterns. The ability to choose among in-depth crawl (following only inner links – same domain) or web crawl (following only outer links – different domains) or both is provided. This tool uses the breadth first search (BFS) algorithm to visit the new hyperlinks, it can efficiently crawl thousands of web pages and it collects statistics in the above mentioned categories, exporting them in formats that can be easily used in famous statistical packages.

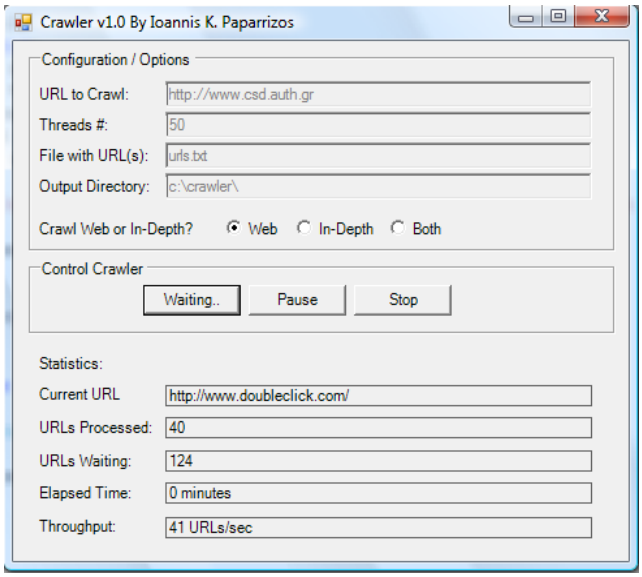

Figure 1. Crawler interface

2. SYSTEM DESCRIPTION
In the base of our system there is a main worker which is responsible to schedule new jobs to the slave workers. There is a queue in which the main worker keeps all web pages that haven't been visited yet. It also keeps track of all web pages in order to not visit them again. When the main worker is ready, it starts scheduling jobs to the slave workers. Each one of the slave workers is responsible to visit a specific web page, collect data according to an html tags and attributes classification algorithm and compute statistics based on these data. Furthermore, it collects all hyperlinks of this web page and together with the statistics it sends them to the main worker. Main worker checks if among the new hyperlinks there are some that already have been visited and adds the rest to the queue. When a slave worker completes a job, the main worker allocates a new one until queue is empty. The collected statistics are stored in a format that is

HT'10, June 13–16, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0041-4/10/06.

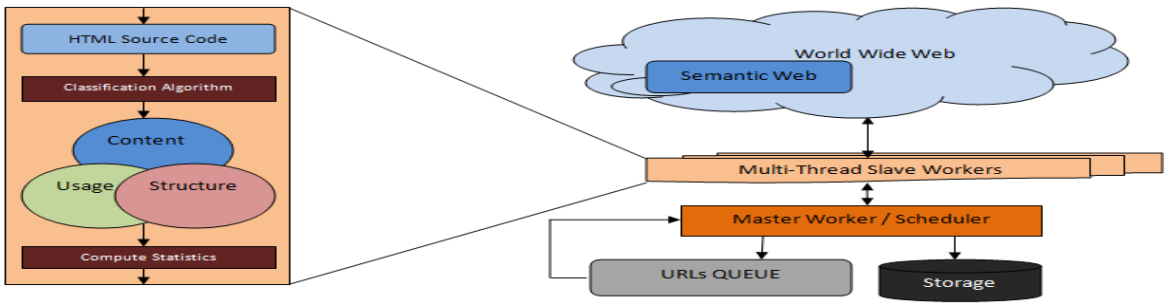

Figure 2. System description

easy to be parsed (comma separated values – CSV) and be used in famous statistical packages. Last, but not least, in the development of this tool several known problems of crawling procedure have been taken into consideration such as performance and throughput, network-based problems, http protocol issues and content-based problems.

3. CLASSIFICATION ALGORITHM

The algorithm works in levels and has three main categories with several sub-categories. The main categories are the Structure (S), Content (C), and Usage (U). Each one of them has several sub-categories too. For example in Structure (S) there are sub-categories such as S1. Documents Format, S2. Web Application Formats, S3. Style etc. Each one of those sub-categories has several sub-categories too (ex. S11, S12, S13 etc.). The scope is in each level the summary percentage of all sub-categories to be 100%. Thus, computing the lowest level, we are able to find out the percentages of each one of the sub-categories in a higher level and finally in the first level which is the most important.

This algorithm takes into consideration the source code of the web page and the messages being exchanged between the client/crawler and the web server. In html code there are tags, attributes, MIME types and file extensions that can be associated with each one of the three main categories. For example, the <a href="…"> tag, the text/css MIME type and the .css file extension corresponds to Structure category, the <meta name="…"> tag, the application/rdf+xml MIME type and the .rdf file extension corresponds to Content category. In Usage category several messages that are being exchanged between the client and the web server have been taken into consideration, such as message about encoding (UTF-8, UTF-16 etc.), language (US English, UK English, Greek etc.), web server (Apache, IIS etc.) and protocols (http, https etc.) that are being used. Having created a list with all mime types, file extensions, html tags and attributes, and messages being exchanged, the algorithm can map which one of those belong to which one of the three categories (structure, content, usage).

In Figure 2, each of the slave workers applies the above classification algorithm to each web page that it visits. Having those data collected, the slave worker is able to compute statistics. Having percentages of structure, content and usage data we are able to apply now more advanced methods which will help us identify problems and patterns of structure, content and usage data in the crawled web pages. A visualization example - based on Compositional Data Analysis [1] - of data collected from in-depth crawl of myspace.com using the CodaPack software [3] is presented in Figure 3. The Compositional Data Analysis is a technique widely used in sciences such as biology, chemistry and geology in order to identify the sub-components of a component. In the ternary diagram each web page of myspace.com is presented as a point in an equilateral triangle with distances from the sides depending on its Structure, Content and Usage percentage. Obviously, further clustering and classification techniques can be applied.

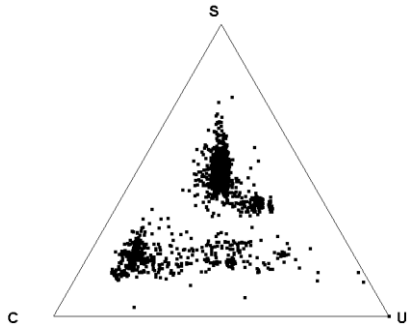

Figure 3. Ternary Plot for myspace.com (S=Structure, C=Content, U=Usage)

4. CONCLUSION AND FUTURE WORK

In this paper we presented a tool which can be used to automatically extract data that are associated with each one of the three axes of web mining. The collected statistics can be used for further analysis which may give us useful results about problems that can present in the structure, content or usage data of a web site. For example, in our analysis there were web pages with 100% usage, and 0% structure and content data. That implies that this web page is either a blank page or a redirect page which is important to know as a site may lose users or advertised sites may lose hits and views. On the other side, web pages with similar statistics can be clustered together in order to identify outliers and patterns that are common in the web pages of a cluster.

5. REFERENCES

[1] Aitchison J. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd, 1986.

[2] Osmar Rachid Za. Resource And Knowledge Discovery From The Internet And Multimedia Repositories. Technical report, Phd dissertation, Simon Fraser University, March 1999.

[3] Thió-Henestrosa S., Gómez O., Cepero R., CODAPACK 3D. A new version of Compositional Data Package. 3rd Compositional Data Analysis Workshop. Girona, 2008.