# Evaluating Web Site Structure based on Navigation Profiles and Site Topology

Alberto Simões[1], Anália Lourenço[2], and José João Almeida[3]

[1] Centro de Estudos Humanísticos, Universidade do Minho
ambs@ilch.uminho.pt
[2] Departamento de Eng. Biológica, Universidade do Minho
analia@deb.uminho.pt
[3] Departamento de Informática, Universidade do Minho
jj@di.uminho.pt

**Abstract.** This work aims at pointing out the benefits of a topology-oriented wide scope, but differentiated, profile analysis. The goal was to conciliate advanced common website usage profiling techniques with the analysis of the website's topology information, outputting valuable knowledge in an intuitive and comprehensible way. Server load balancing, crawler activity evaluation and Web site restructuring are the primary analysis concerns and, in this regard, experiments over six month data of a real-world Web site were considered successful.

## 1 Introduction

Usually, Web usage profiling is performed by specialised analysts that are capable of mining clickstreams and interpreting different kinds of outputs. Yet, even do it is not always possible to make Webmasters participate in Web analysis activities, they should be primary end-users and thus, results should be made more comprehensible.

Profile reporting is somewhat challenging due to the complexity and detail of the information. Site topology provides a valuable help by tagging the sequence of requests to site structure, i.e., mapping profiles into the site's oriented graph representation [5].

Graphs are a natural representation of topologies. Nevertheless, site topologies tend to embrace a high number of Web pages and hyperlinks, leading to high dimensional graphs that are not easily rendered or analysed. While providing similar navigation and analysis capabilities, layered graph visualisation is considered a better approach.

In this scenario, we propose an integrated approach to topology-based differentiated profiling. Data visualisation was the primary design concern and addressed both data comprehensibility and readability and graph navigability.

## 2 Topology-Oriented Profiling

The development of the proposed approach involved three main activities: topology automatic retrieval and processing, differentiated profiling analysis and graph layered visualisation. Topology information is issued as the most adequate support to provide profiling insights. Differentiated profiling ensures focused processing and analysis of the two foremost groups of users, i.e., regular users and Web crawlers. Finally, the visualisation of the topology, i.e., the site graph, takes into consideration site's dimension and structuring, preserving data comprehensibility and readability.

### 2.1 Case Study

Natura Web site, property of the Language Specification and Processing Group[4] of the Computer Science Department of the University of Minho, complied with all these requirements and thus, was considered suitable for this experiment. Natura is a Natural Language Processing (NLP) research project focused on Portuguese language and its Web site supports the project research activities, the group's academic Web pages related to Natura and general NLP and the homepages of some of the project's members. Although this is a non-commercial site, the diversity of its contents is quite appealing in terms of differentiated usage profiling. Scientific publications, academic events, software and other NLP resources are mainly visited by students and researchers. Yet, the music repository embracing poems, lyrics, accords, music scores and karaoke files, attracts not only regular users and general searchers, but also focused music-related retrievers.

These experiments used six month server log data (Table 1).

| **Metric** | **Value** |
|---|---|
| Unique user agents | 30 324 |
| Unique hosts | 61 0127 |
| Unique requests | 44 673 |
| Unique referrers | 2 855 |
| Total number of requests | 7 198 999 |
| Volume of traffic | 1 795.09 GB |

**Table 1.** General information about server log data.

### 2.2 Topology Extraction and Processing

Topology extraction is supported by breadth-first crawler harvesting whereas the user may specify multiple seeds and harvesting stop points. These parameters specialise broad crawling in order to prevent crawling into irrelevant or problematic areas. For example, Concurrent Version System (CVS) access points and mailing list archives. The crawler outputs two classes of graph entities: nodes

[4] Natura Project Web site `http://natura.di.uminho.pt`

and edges. Nodes represent the URLs of site documents while edges stand for the hyperlinks relating those URLs.

Topology processing aims the simplification of the crawled topology and includes three main processes: node clustering, graph simplification and node depth calculation.

**Node Clustering** There are some specific website areas or kind of web applications that have hundred of similar pages, all based on a similar template. These pages share the same set of inbounds and outbounds, as shown in figure 1.

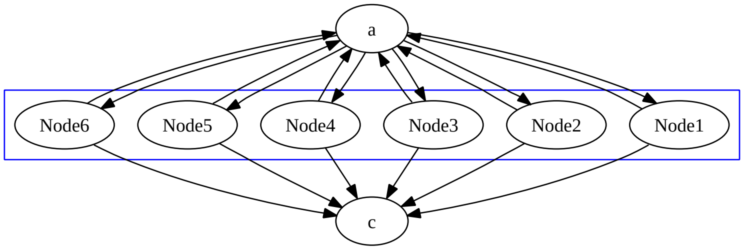


**Fig. 1.** An example of a cluster candidate.

This means that all these nodes can be collapsed in a single node, representing them all, without losing any kind of information, but reducing the size of the graph, making it easier to analyse and to visualize. Therefore, the graph is inspected in order to find areas that can be simplified.

**Graph Simplification** There are some pages that include a big set of inbounds and outbounds from and to the same web page. Also, some pages include a big set of hyperlinks to themselves, like the usual index presented in the top of webpages, that link to different sections in that same page. Figure 2 exemplifies how simplification is performed regarding multiple, or self-reference, hyperlinks.

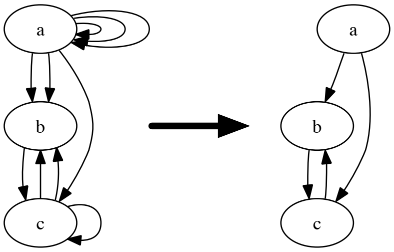


**Fig. 2.** Example of graph simplifications.

**Depth Computation** It is not plausible to assume that visits always start at the root or the first level nodes nor to expect that their pattern will follow a straightforward depth-first or breadth-first pattern. Depending on previous visits and Web site indexing, visits can be initiated at any level and may look into different (related or non-related) levels of contents. Therefore, the calculation of node depth (Figure 3) was considered an important profiling asset.

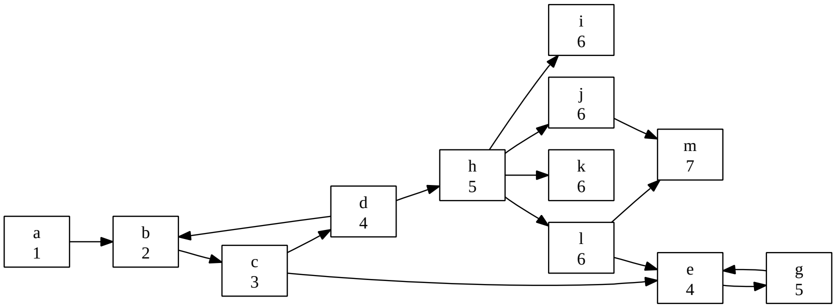


**Fig. 3.** Graph depth calculation based on SPPP Dijkstra's algorithm.

### 2.3 Profile Analysis

Conventionally, Web analysis focuses on regular users' most common traverse patterns. Yet, regular users are not the only users whose attention Webmasters wish to capture. Nowadays, Web site visibility is greatly dependent on general and focused purpose search indexes. Also, there are a considerable number of other crawlers traversing the Web sites gathering all sorts of information. Even though they are not primary users, their presence indicates that someone is interested in that kind of contents and such interest may return some profit if Webmasters take into consideration those profiles as well.

Due to the ever growing similarities between crawler and regular user patterns, usage differentiation is a challenging task. Web crawlers are widespread and standard detection heuristics are unable to cope with the continuous evolving of the technology. Navigation pattern mining seems to be the most reasonable approach to the problem as it naturally encompasses changes in navigation patterns and does not imply the maintenance of any catalog.

The present work used the pattern mining approach introduced in [4,3], involving the semi-automatic labeling of a training set of Web sessions and tree model induction. Besides crawler and regular user sessions there were identified browser-related application sessions (Table 2).

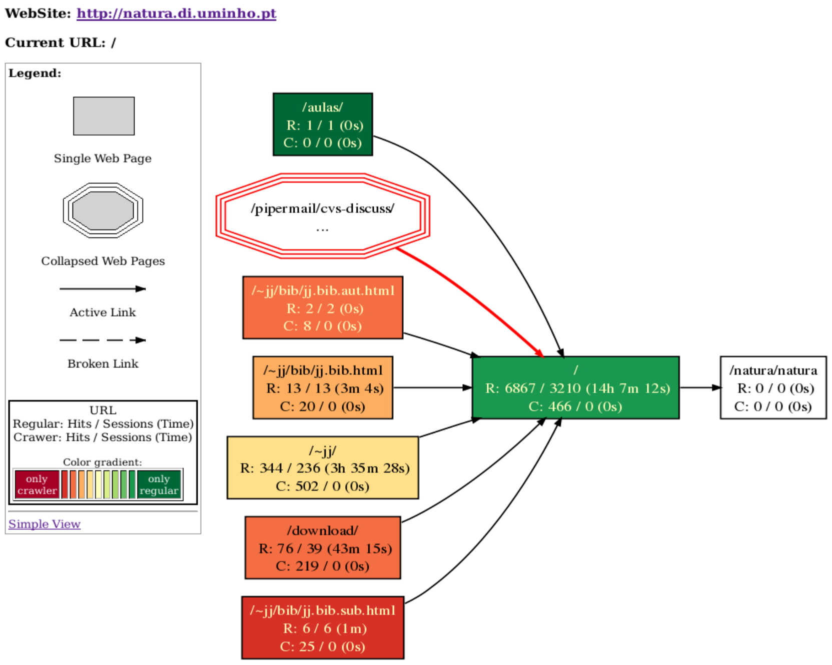
| **Month** | Total | Crawler | Regular | Application | Unknown |
|---|---|---|---|---|---|
| january | 166 490 | 59 879 | 98 875 | 6 592 | 1 144 |
| february | 175 192 | 66 091 | 103 163 | 4 670 | 1 268 |
| march | 256 649 | 120 829 | 130 126 | 4 187 | 1 507 |
| april | 222 203 | 102 445 | 115 041 | 3 376 | 1 341 |
| may | 339 413 | 135 151 | 196 535 | 5 621 | 2 106 |
| june | 318 937 | 151 324 | 161 511 | 4 067 | 2 035 |

**Table 2.** Statistics about differentiated Web sessions.

### 2.4 Topology-based Profile Evaluation

A web application was conceived in order to support topology-based profile visualisation. This tool is written in Perl and uses GraphViz software [2] to perform on-the-fly graph rendering while profiles are retrieved from a MySQL database.

As it is very hard to visualise any site graph as a whole [1] and profile analysis would not be comprehensible, the tool uses a layered visualisation approach. Each layer includes the current node and all nodes that are directly connected to it (Figure 4).



**WebSite:** http://natura.di.uminho.pt

**Current URL:** /

**Fig. 4.** Layered perspective of Natura's Web site root.

Node representation includes the corresponding URL and profile statistics. There are two profile views: the general view and the expanded view. The general view presents the number of hits for crawler or regular users, while the expanded view also includes session counts and estimated time spent. Also, a colour gradient provides an intuitive view of node's crawler vs user load balance. Node colour is based on a gradient from green to red, where greener nodes are mostly accessed by regular users and redder node are mostly accessed by Web crawlers. Node clusters are represented by an octagon shape with a sample non-clickable URL. On the other hand, edge representation is twofold: plain arrows for active links, and dashed arrows for broken links (Figure 5).

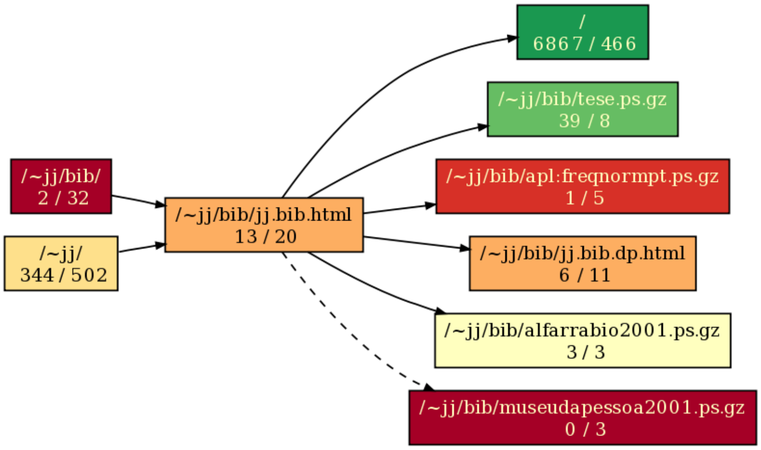


**Fig. 5.** Example of hyperlink representations.

Looking at the generated topology Webmasters can easily identify hot and dead site areas and broken links. Besides, they may evaluate differentiated profiling metrics tracking down overlapping traverse areas, i.e., site areas that attract both users and crawlers. Large overlapping areas are interpreted as good indexing indicators, while distant usage areas indicate poor or inadequate indexing. Based upon the list of crawlers that were profiled, it is also possible to argue about indexing flaws, both in terms of crawling purposes and site structuring.

As the period of analysis is specified by the user, it is possible to assess server load over distinct periods of time as well as compare the impact of restructuring actions.

## 3 Final Remarks

Profile analysis provides relevant insights about user interests and current navigation patterns. Yet, regular users and crawlers are not alike and should be analysed separately aiming at their particular traverse purposes.

The proposed approach extended profiling with site topology. Layered graph viewing supported profile visualisation and graph features represented site's most common semantics. Intuitively, node and edge shapes represent different site resources and their associations while node colour balances differentiated usage.

Data visualisation was the primary design concern and addressed both data comprehensibility and readability and graph navigability. At first, if not aware of the real structure of the Web site, a naive user may consider topology navigation confusing, but the Webmaster will quickly find out his way.

In terms of future work, both profiling and data visualisation may be enriched. Node information can be more thorough. Clusters should identify the list of collapsed URLs along with the associated regular/crawler profiling metrics. Also, it would be relevant to provide information about the most common navigation patterns highlighting the preferred hyperlinks. Regarding the web application, the usage of GraphViz is an overhead to the server running the application, and therefore we should opt for an HTML canvas-based rendering tool.

## Acknowledgments

This work is funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-010049

## References

1. J. Cugini and J. Scholtz. VISVIP: 3D visualization of paths through web sites. In *Tenth International Workshop on Database and Expert Systems Applications*, pages 259–263, Florence, Italy, 1999.
2. E. R. Gansner and S. C. North. An open graph visualization system and its applications to software engineering. *Journal of Software — Practice and Experience*, 30(11):1203–1233, 2000.
3. Anália Lourenço and Orlando Belo. Applying clickstream data mining to real-time web crawler detection and containment. In H.-J. Lenz and R. Decker, editors, *Advances in Data Analysis — Proceedings of the 30th Annual Conference of The Gesellschaft für Klassifikation, XVI*, pages 351–358, 2006.
4. Anália Lourenço and Orlando Belo. Catching web crawlers in the act. In *Sixth International Conference on Web Engineering (ICWE'06)*, 2006.
5. D. Oikonomopoulou, M. Rigou, S. Sirmakessis, and A. Tsakalidis. Full-coverage web prediction based on web usage mining and site topology. In *Web Intelligence, IEEE/WIC/ACM International Conference on (WI'04)*, pages 716–719, 2004.