Local Structure in the Web

Fabien Mathieu, Laurent Viennot

Local Structure in the Web

Fabien Mathieu
Gyroweb - LIAFA, LIRMM
34392 Montpellier Cedex 5 France
fmathieu@clipper.ens.fr

Laurent Viennot
Gyroweb - LIAFA, INRIA Rocquencourt
F-78153 Le Chesnay (France)
Laurent.Viennot@inria.fr

ABSTRACT

The web graph has been widely adopted as the core describing the web structure [4]. However, little attention has been paid to the relationship between the web graph and the location of the pages. It has already been noticed that links are often local (i.e. from a page to another page of the same server) and this can be used for efficient encoding of the web graph [9,7].

Locality in the web can be further modelled by the *clustered graph* induced by the prefix tree of URLs. The web tree's internal nodes are the common prefixes of URLs and its leaves are the URLs themselves. A prefix ordering of URLs according to this tree allows to observe local structure in the web directly on the adjacency matrix M of the web graph. M splits in two terms : $M = D + S$, where D is diagonal by blocks and S is a very sparse matrix. The blocks of D that can be observed along the diagonal are sets of pages strongly related together.

Keywords

URL-clustered, web graph, adjacency matrix, local structure

1. INTRODUCTION

Classically, the web structure is modeled as a graph. As we show, this representation doesn't capture all the structural data of the web. If we assume the tree of the URls is worth it, we can define a double structure, with the hyperlinks graph on one side and the URLs' tree on the other : the clustered graph of the web.

The *clustered graph* structure was first introduced by Feng in [5] for giant graphs representation. It appears in many domains where structured diagrams can be found. It is mainly used for graph drawing. The problem is then to cluster a given graph to allow better drawing[2]. In the case of the web graph, an intrinsic clusterization is given by the URLs.

2. DEFINITION

A clustered partition of a graph $G(V, E)$ is given by a tree T whose leaves are V. Each internal node n defines a cluster $V_T(n)$. This cluster is the subset of V given by the leaves of the subtree rooted at n. For example, the cluster of the root of the tree is V.

3. URL-CLUSTERED WEB GRAPH

The web graph can naturally be clustered by the tree structure of the URLs (Uniform Resource Locators [1]). The web can be viewed as a collection of file hierarchies (one per web server) which can be naturally gathered in a single tree using the URLs of these files. To further refine the tree, we can gather together all servers within a given domain by adding a node for each domain :

For example, the URL `http://smith.mysite.org/fishing/index.html` splits into *http*, *org*, *mysite*, *smith*, *fishing* and *index.htm*, yielding the path from the root to the corresponding leave in the tree as shown in Figure 1.

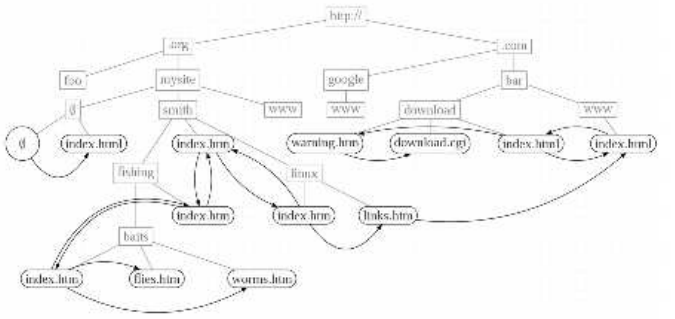


Figure 1 : Example of clustered web graph.

4. STRUCTURAL WEB SITE DEFINITION

Considering most of the webmasters try to organize their sites, we expect that the clustered web graph will be intimately linked to the notion of web site. This is confirmed by observing the adjacency matrix M of a 8 million URLs crawl from *.fr* where URLs are sorted according to a prefix ordering of the tree of the URLs (see Figures 2 and 3). First, M obviously splits in two terms $M = D + S$, where D is diagonal by blocs and S a ver sparse matrix. The site and sub-site structure visually appears along the diagonal as squares on the tree sorted adjacency matrix. The observation of the tree sorted adjacency matrix allows to identify the web sites as some of the clusters of the URL-clustered web graph.

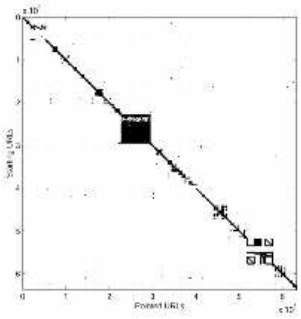
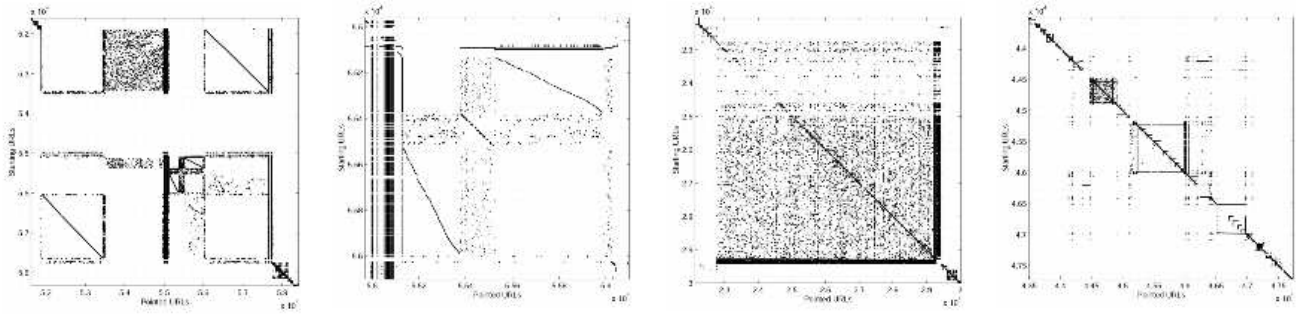


Figure 2 : The URL tree sorted adjacency matrix of 6.10^4 pages in an 8 million pages crawl of .fr.



[allmacintosh.easynet.fr, a site about computers software.] **[allmacintosh.(...).fr/osx, a sub-site about a specific OS.]** **[aemiaif.lip6.fr/jargon, an interactive dictionnary.]** **[algo.inria.fr, an INRIA team site.]**

Figure 3 : A visual approach of the local web structure : zoom on clusters of Figure 2

We can thus structurally define a web site as a cluster where most of the links are internal to the cluster. A simple algorithm for deciding if a cluster corresponds to a web site consists in computing the ratio of the number of internal links over the total number of links found in the pages of the cluster. If the ratio is greater than some threshold, we have identified a web site automatically. Indeed, more elaborate strategies would consider also links that are incoming from outside the cluster. A union of clusters could also be allowed as a web site. For example, several sons of an internal node may give clusters more thighed together than to the other sons. To our knowledge, the only related works about automatic site extraction are based on weakly disconnected components computation [3] or community detection [6] and do not use the tree structure of the web.

5. APPLICATIONS

The applications of a *web site* partition are numerous. It allows to distinguish ‘‘navigational’’ links and ‘‘real’’ links, assuming the latter are more important than the former. Search engines often restrain their responses to one (or a few) page per site, so that a single site can not monopolize a response. Partitionning the web graph could also allow to effectively distribute some algorithms.

6. PERSPECTIVES

Our model could be extended to cope with pages accessible through several URLs. This is linked to the problem of identifying uniquely web pages and requires more inquiry. Our model could integrate other protocols data (as ftp files). The most promising work consists in finding algorithms which automatically perform an efficient site-partition and distributed algorithms to take advantage of such partitions.

7. CONCLUSION

The web graph has been widely studied to model the web [4,8]. The URL-clustered web graph could be used to better model the web structure and its evolution.

8. REFERENCES

1. **T. Berners, L. Masinter, and M. M. Cahill.**
Rfc-1738 : Uniform ressource locators (url), 1994.
2. **M. Brinkmeier.**
Communities in graphs, 2002.
3. **C. H. Q. Ding, X. He, and H. Zha.**
A spectral method to separate disconnected and nearly-disconnected web graph components.
In *KDD 2001,* pages 275-280, 2001.
4. **A. B. et al.**
Graph structure in the web.
In *Proc. 9th International World Wide Web Conference*, pages 309-320, 2000.
5. **Q. Feng, R. F. Cohen, and P. Eades.**
How to draw a planar clustered graph.
***Journal of the ACM,* 959:21-31, 1995.**
6. **D. Gibson, J. M. Kleinberg, and P. Raghavan.**
Inferring web communities from link topology.
In *UK Conference on Hypertext,* pages 225-234, 1998.
7. **J. Guillaume, M. Latapy, and L. Viennot.**
Efficient and simple encodings for the web graph.
***In Proceedings of the 11-th international conference on the World Wide Web*, 2002.**
8. **L. Page, S. Brin, R. Motwani, and T. Winograd.**
The pagerank citation ranking : Bringing order to the web.
***Technical report, Computer Science Department, Stabdford University,* 1998.**
9. **K. Randall, R. Stata, R. Wickremesinghe, and J. Wiener.**
The link database: Fast access to graphs of the web.
***Research Report 175, Compaq Systems Research Center, Palo Alto, CA*, 2001.**