Structure intrinsèque du Web

Fabien Mathieu, Laurent Viennot

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Structure intrinsèque du Web

Fabien Mathieu et Laurent Viennot

N° 4663

Décembre 2002

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4663--FR+ENG

INRIA
ROCQUENCOURT

Structure intrinsèque du Web

Fabien Mathieu et Laurent Viennot

Thème 1 — Réseaux et systèmes
Projet Gyroweb

Rapport de recherche n° 4663 — Décembre 2002 — 7 pages

Abstract: The web graph has been widely adopted as the core describing the web structure [2]. 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 coding of the web graph [5, 4].

Locality in the web can be further modelled by a *clustered graph* defined by adding the prefix tree of URLs to the web graph. Its internal nodes are the common prefixes of URLs and its leaves are the URLs themselves. As shown by Figure 1, a prefix ordering of URLs according to this tree allows us 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 sparser than D. The blocks of D that can be observed on the diagonal are sets of pages strongly related together.

Key-words: Web, Graph, Tree, Blocks, Clusters

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Local Structure in the Web

Résumé : Le graphe du web a largement été adopté pour représenter la structure du web [2]. En revanche, le lien entre le graphe du web et la localisation des pages web est rarement utilisé. Pourtant, il a déjà été remarqué que la plupart des hyperliens étaient de nature locale (i.e. reliant deux pages d'un même serveur) et que cela permettait de réaliser un encodage efficace du graphe du web [5, 4].

La localité dans le graphe du web se formalise en introduisant la notion de *graphe en clusters*, défini comme l'adjonction de l'arbre des préfixes des URLs au graphe. Les nœuds internes sont les préfixes communs des URLs tandis que les feuilles sont les URLs elles-mêmes. Comme le montre la figure 1, un tri des URLs dans l'ordre lexicographique associé à cet arbre nous permet d'observer la localité des hyperliens directement à partir de la matrice d'adjacence M du graphe du web. Ainsi triée, M se décompose visuellement en deux termes : $M = D + S$, où D est diagonale par blocs et S une matrice creuse D. Les blocs de D correspondent à des pages fortement reliées entre elles.

Mots-clés : Web, Graphes, Arbres, Blocs, Clusters

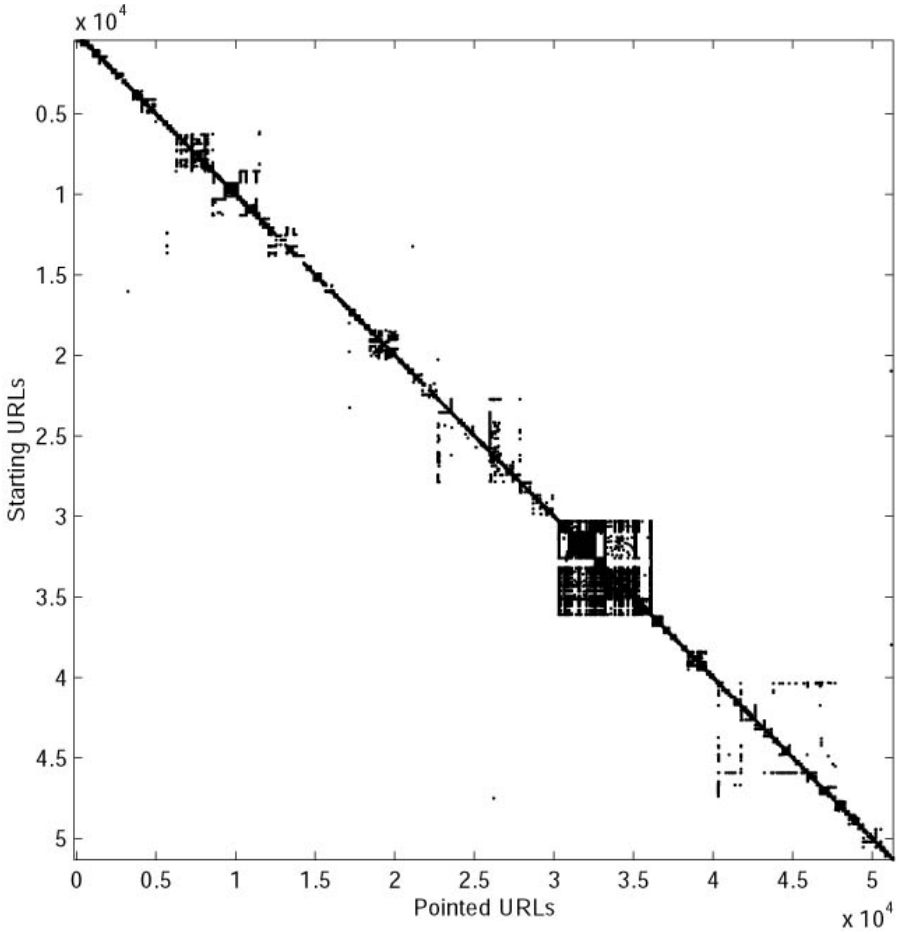

Figure 1: Zoom sur la diagonale de la matrice d'adjacence

La représentation classique du web comme un graphe n'utilise pas l'intégralité des données structurelles du web. L'écriture standardisée des URLs, que l'on peut représenter par un arbre, n'est pas prise en compte. Nous allons donc définir une double structure, avec le graphe des hyperliens d'un côté, la localisation dans l'arbre des URLs de l'autre : le découpage en clusters d'un graphe.

Remarque : La structure de *clustered graph* n'est pas qu'un objet *ad-hoc* introduit pour étudier le web. Introduite à l'origine par Feng dans [3] pour aider à la visualisation de grands graphes, elle peut apparaître dans de très nombreux contextes : par exemple, le graphe des relations humaines (les sommets sont des personnes et A pointe vers B si A connaît B) peut se compléter de manière pertinente en un *clustered graph*, où l'arbre vient de la localisation (pays, (état), ville, quartier, rue, ...).

1 Définition

Un découpage en clusters d'un graphe $G(V, E)$ est la donnée d'un couple (T, φ), où T est un arbre à $|V|$ feuilles et φ une bijection entre les feuilles de T et les sommets de G. À chaque nœud n de l'arbre correspond une partie $V_{(T,\varphi)}(n)$ de V définie par :

$$V_{(T,\varphi)}(n) := \begin{cases} \{\varphi(n)\}, \text{ si } n \in \text{ feuilles}(T) \\ \bigcup_{m \text{ fils de } n} V_{(T,\varphi)}(m), \text{ si } n \notin \text{ feuilles}(T) \end{cases}$$

2 Clusterisation canonique du graphe web

Le graphe du web $\mathcal{G}_{web}$ possède une clusterisation (T, φ) naturelle induite par l'écriture des URLs :

- La racine de T est étiquetée par «http».
- toute URL est de la forme (certaines chaînes pouvant être vides) :

  `http://sous-serv.serveur.domaine.zone:port/rep1/.../repn/fichier.ext`

  Elle se place comme feuille de l'arbre en suivant le chemin (les nœuds internes sont créés à la volée s'ils n'existent pas)

  http→zone→domaine→serveurs.sous-serveurs.port→rép$_1$ →... →rép$_n$ →fichier.ext
- Les différents fils d'un même nœud sont rangés (arbitrairement) par ordre lexicographique.

Par construction, l'arbre que nous construisons ainsi est bien un découpage en clusters du graphe du web. La figure 2 est un exemple de la structure typique d'un graphe clusterisé, construit à partir de quelques URLs arbitraires.

3 Application

Si l'on considère que la plupart des concepteurs de site essaient d'avoir une certaine organisation, aussi bien au niveau du parc de machines qu'ils ont éventuellement à disposition qu'au niveau des répertoires et sous-répertoires, on comprend que le graphe en clusters du web va être intimement lié à la notion de site web, comme tendent à le confirmer les différentes observations faites sur la matrice d'adjacence M d'un crawl effectué sur plus de huit millions de sommets en juin 2001 (voir figure 1). L'organisation des sites se retrouve sur la matrice, et une décomposition $M = D + S$, où D est diagonale par blocs et S une matrice creuse D, apparaît. Les blocs coïncident, après vérification, remarquablement bien avec les nœuds internes du découpage canonique en clusters ; en y regardant de plus près, ce découpage autorise une vision récursive (et plus fine) du problème de la définition d'un site web. Par exemple :

- Une société peut être hébergée sur `www.big-compagny.com/*`,
- son département de recherche situé sur `www.big-compagny.com/research/*`
- et les pages de John Smith[1] dans `www.big-compagny.com/research/~smith/*`.

Nous pouvons dès lors entrevoir les moyens de définir structurellement ce qu'est un site web. Plutôt que d'adopter la définition intuitive proposée par [1], nous appellerons *site web* un ensemble organisé de pages proches dans l'arbre du web et fortement connectées entre elles dans le graphe du web. Et plutôt que de travailler sur l'ensemble des partitions des sommets du graphe, ce qui n'est guère réalisable en pratique, il est possible de se restreindre (au moins pour commencer) aux partitions induites par les nœuds internes de l'arbre des clusters.

Les applications d'une partition du graphe du web en *sites web* sont nombreuses. On pourrait par exemple faire la distinction entre les liens internes de navigation à l'intérieur d'un site et les liens externes, information utile si l'on fait l'hypothèse que les seconds sont plus importants que les premiers. Le graphe-quotient issu d'une telle partition serait beaucoup plus petit, mais néanmoins pertinent. Des algorithmes comme le PageRank utilisé par Google pourrait être calculé de manière distribué sur plusieurs niveaux, par exemple intra-site et inter-site.

4 Conclusion - futurs travaux

Pour ces raisons, nous croyons que la structure du web est mieux modélisée par le concept de graphe clusterisé. Le découpage en clusters d'un graphe peut formellement être défini comme un arbre et un graphe reliant les feuilles de l'arbre. La représentation du web en termes de clusters d'un graphe peut aider à identifier structurellement les sites web, par exemple en sélectionnant les nœuds internes de l'arbre comme ensemble de candidats initiaux. D'autres applications sont possibles :

[1] Un chercheur de la société...

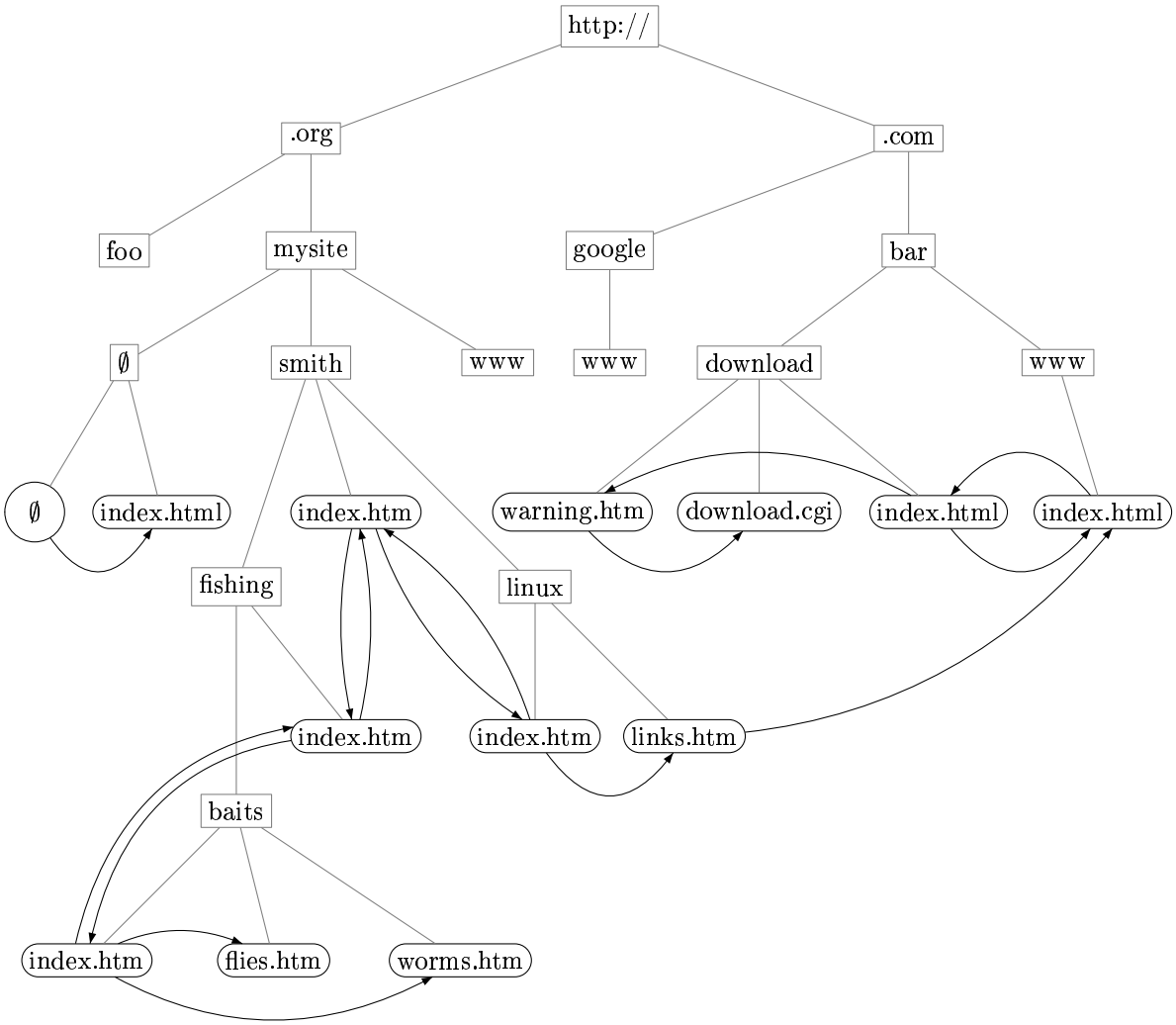

Figure 2: Exemple de *clustered graph*

- Modèles probabilistes du graphe du web basés sur la structure d'arbre-graphe.
- Calcul des corrélations entre les paramètres de l'arbre du web et ceux du graphe du web. Par exemple, des corrélations sur les distances permettrait d'avoir un meilleur contrôle sur la qualité des crawls réels.

References

[1] Nick Craswell, David Hawking, and Stephen Robertson. Effective site finding using link anchor information. In *Research and Development in Information Retrieval*, pages 250–257, 2001.

[2] Andrei Broder et al. Graph structure in the web. In *Proc. 9th International World Wide Web Conference*, pages 309–320, 2000.

[3] Qingwen Feng, Robert F. Cohen, and Peter Eades. How to draw a planar clustered graph. *Journal of the ACM*, 959:21–31, 1995.

[4] J. Guillaume, M. Latapy, and L. Viennot. Efficient and simple encodings for the web graph, 2002.

[5] K. Randall, R. Stata, R. Wickremesinghe, and J. Wiener. The link database: Fast access to graphs of the web, 2001.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399