Le graphe du Web

Jean-Loup Guillaume et Matthieu Latapy

1 Introduction

Depuis quelques années, Internet, et le Web en particulier, révolutionnent les modes de communication des êtres humains et leur rapport à l'information. Non seulement le nombre de personnes *navigant* sur le Web suit une croissance exponentielle, mais de plus en plus d'individus participent eux-même à la richesse du Web en créant des pages sur des sujets extrêmement diversifiés. Ainsi, plusieurs millions de nouvelles pages sont créées quotidiennement dans le monde, et au moins autant sont modifiées, le nombre total de pages s'élevant à plusieurs milliards.

Toutes ces pages et les liens hypertexte qui les relient les unes aux autres forment un graphe : chaque page Web est associée à un sommet et les hyperliens entre pages Web sont les arcs du graphe. Cet énorme graphe joue un rôle important dans un grand nombre d'applications : création de moteurs de recherche efficaces, détection de communautés, etc. Il est cependant extrêmement difficile de l´étudier, à cause de sa taille et de son évolution rapide.

La méthode générale utilisée pour explorer le graphe est la suivante : on commence avec une première page dans laquelle on cherche tous les liens hypertexte. Chacun de ces liens pointe sur une page. On recommence alors l'opération à partir de chacune de ces pages en allant chercher tous les liens qu'elle contient, et ainsi de suite. Le processus s'arrête quand on ne trouve plus de nouvelles pages. Avec cette méthode, si l'on pouvait visiter mille pages par seconde (dans la réalité c'est plutôt cent), il faudrait plus d'un mois pour connaître toutes les pages accessibles depuis la page initiale et ce, quelle que soit la puissance des machines utilisées. Pendant ce temps, plusieurs centaines de millions de pages auront vu le jour, et de nombreuses autres auront été supprimées ou déplacées. Par conséquent l'image qu'on obtient est non seulement partielle mais aussi faussée par l'évolution du graphe. On peut voir cette image du graphe du Web comme une image radar : seule une petite partie du graphe est connue à un instant donné et le reste évolue pendant qu'on observe cette petite partie. Il y a un grand nombre d'autres problèmes qui rendent la récupération du graphe du Web impossible en pratique. Par exemple, si une page n'est pointée par aucune autre, elle ne sera jamais découverte avec cette méthode. De façon plus générale, si une page n'est pas atteignable à partir de la page initiale on ne la trouvera jamais.

Par conséquent, l'étude du graphe du Web passe pas la récupération de *parties* du graphe, les plus grandes et les plus représentatives possibles. Les propriétés de ces parties sont alors étudiées comme s'il s'agissait du graphe entier, mais il faut garder à l'esprit que ce n'est pas le cas.

2 À quoi ressemble le graphe du Web

Depuis quelques années, les chercheurs tentent d'avoir une idée de la structure globale du Web. Pour atteindre cet objectif, trois approches principales ont été utilisées : un point de vue macroscopique où on regarde le graphe "de loin", et on s'intéresse à sa structure à grande échelle, un point de vue microscopique où on recherche dans le graphe des petites structures qui se répètent souvent, et enfin une approche statistique où on calcule certaines statistiques pertinentes sur le graphe.

Approche macroscopique : Une étude récente [2] a proposé la description suivante du graphe du Web. Il serait composé de quatre zones distinctes : un gros noyau dans lequel on peut aller de n'importe quelle page à n'importe quelle autre en suivant des liens, un ensemble de pages depuis lesquelles on peut aller dans le noyau mais qui n'est pas accessible depuis ce dernier, un troisième ensemble qui est accessible depuis le noyau mais à partir duquel on ne peut pas atteindre le noyau, et finalement un grand nombre de pages à partir desquelles on ne peut pas atteindre le noyau et qui ne sont pas accessibles à partir de ce noyau. De plus, ces quatre parties ont approximativement la même taille. Cette structure a reçu le nom de nœud papillon (voir Figure 1).

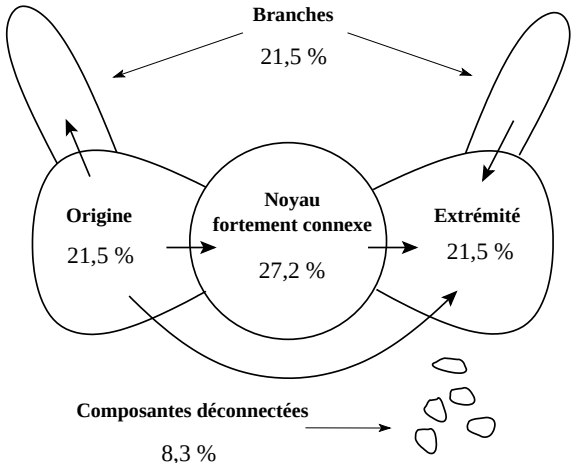

FIG. 1 – *Le graphe du Web vu comme un nœud papillon*

Avant cette étude, on pensait qu'on pouvait aller à peu près de n'importe quelle page Web à n'importe quelle autre en suivant les liens. Ce n'est pas du tout le cas : quand on part d'une page Web du noyau du nœud papillon, on ne peut atteindre que *la moitié* des pages du Web ! De plus, les suites de liens reliant les pages Web entre elles peuvent être très longues (plus de 900 clics de souris pour aller d'une page à l'autre), ce qui contredit à nouveau les suppositions antérieures. Une autre information importante, mise en évidence par la structure de nœud papillon, est la présence d'un cœur de pages Web très liées entre elles, qu'on imaginait beaucoup plus gros ; en fait, ce cœur ne constituerait qu'un quart du Web.

Approche microscopique : En regardant de plus près le graphe du Web, on remarque l'existence de petites structures qui apparaissent fréquemment. Parmi celles-ci on peut trouver des couples d'ensemble de pages Web tels que toutes les pages du premier ensemble

pointent vers toutes les pages du second, les pages du second ensemble ne pointant pas les unes sur les autres. Cette structure correspond à une communauté centrée autour d'un sujet de prédilection [3] : le premier ensemble contient les pages de "fans" qui mettent des liens vers leurs "stars" (voir Figure 2).

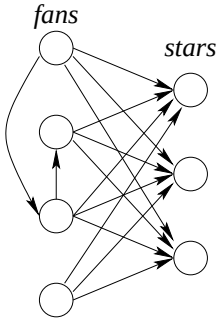

FIG. 2 – *Liens reliant les fans et les stars*

Par exemple, des passionnés de voitures pointeront tous vers les pages des grands constructeurs, qui eux ne pointent pas les uns sur les autres puisqu'ils sont concurrents. La recherche de ce type de micro-structures peut permettre de trouver automatiquement des ensembles de pages Web qui traitent d'un même sujet. D'autres types de structures existent qui correspondent chacune à un phénomène identifié. Citons par exemple les *clans* qui correspondent à des ensembles de pages pour lesquelles il suffit d'un très petit nombre de clics pour aller d'une page à l'autre.

Etudes statistiques : Il est aussi possible d'effectuer des études statistiques au niveau des pages Web [2]. Une page contient en moyenne 11 liens hypertexte, et on sait de plus que la probabilité qu'une page contienne i liens est proportionnelle à $i^{-2.1}$. Ceci signifie que la grande majorité des pages a peu de liens : il y a environ cinq fois moins de pages avec neuf liens que de pages avec un seul lien. Si, à l'inverse, on compte le nombre de pages qui pointent sur une page donnée, les résultats sont similaires, mais cette fois la proportion de pages peu pointées est encore plus importante (la probabilité est proportionnelle à $i^{-2.7}$).

Diverses autres études statistiques ont été menées pour déterminer la taille moyenne des pages Web, le nombre d'images qu'elles contiennent, etc. L'évolution du graphe commence également à être étudiée : plus de 7 millions de pages sont créées chaque jour, et on sait que certaines parties du graphe évoluent plus vite que d'autres [4].

3 Une application : les moteurs de recherche

Les moteurs de recherche (Google, Altavista, Yahoo...) permettent de chercher sur le Web les pages traitant d'un sujet donné. Un moteur de recherche classique demande à l'utilisateur un ensemble de mots et lui donne en retour toutes les pages Web qui contiennent ces mots. Toutefois, le nombre de pages répondant à la requête peut être énorme et l'utilisateur risque vite d'être noyé sous le flot d'informations. Par exemple, si on cherche des informations sur la théorie du noeud papillon décrite précédemment, la requête en anglais "bow tie theory" donne plusieurs milliers de réponses. Le moteur de recherche doit alors identifier les pages les plus

pertinentes concernant la requête. Pour atteindre cet objectif, il semble intéressant de se baser sur les avis des internautes qui ont déjà effectué la même recherche. Mais comment peut-on connaître ces avis ?

Si une personne ajoute un lien hypertexte de sa page vers une autre page, on peut raisonnablement supposer qu'elle estime que cette page est intéressante. Si beaucoup de personnes font de même, il semble naturel de dire que cette page est vraiment pertinente. Ainsi le nombre de pages qui pointent sur une page donnée peut être pris comme un critère de qualité pour celle-ci. En allant plus loin, on peut considérer qu'une page pointée par peu de pages, mais qui sont toutes de grande qualité, a de fortes chances d'être aussi de bonne qualité. Ainsi, les *bonnes* pages sont les pages vers lesquelles pointent beaucoup de pages, sur lesquelles à nouveau pointent beaucoup de pages, etc.

Ce principe est utilisé par le moteur de recherche Google pour classer les pages [1]. Initialement chaque page se voit attribuer un certain nombre de points puis, au fur et à mesure, les pages vont partager leurs points entre les pages vers lesquelles elles pointent. Par conséquent, une page beaucoup référencée va recevoir beaucoup de points, ainsi que les pages qui sont référencées par des pages ayant beaucoup de points (*i.e.* elles-mêmes bien référencées). Si on itère suffisament longtemps ce processus, on arrive finalement à un état où le nombre de points sur chaque page reste stable. Ce nombre final de points donne une mesure de la qualité des pages Web. En résumé, Google fonctionne comme suit : il rassemble dans un premier temps les pages qui contiennent les mots demandés, puis il présente celles qui ont le plus grand nombre de points en premier.

4 Conclusion

L'étude du graphe du Web n'en est aujourd'hui qu'à ses balbutiements. Comme nous l'avons vu, il est déjà difficile de définir des méthodes rigoureuses pour en obtenir des parties représentatives. De plus, de nombreuses propriétés du graphe ne peuvent pas être calculées dans la pratique, à cause de sa très grande taille. De nombreuses questions restent donc des défis pour la recherche en informatique de ces prochaines années : comment détecter effectivement des communautés, comment trouver les pages répondant au mieux à une requête, comment obtenir des très grandes parties du Web, etc.

Un aspect du graphe du Web, qui apparaît de plus en plus comme un élément central pour son étude, est sa *dynamique* : la façon dont les liens apparaissent, disparaissent et se ré-organisent au cours du temps peut permettre de comprendre en profondeur la structure du Web, et de le traiter beaucoup plus efficacement. Si on reprend l'exemple des moteurs de recherche, on imagine aisément que des pages qui sont modifiées souvent n'ont pas le même *sens* que celles qui ne sont jamais modifiées. De même, des pages très récentes mais sur lesquelles pointent déjà beaucoup d'autres pages peuvent être considérées comme particulièrement intéressantes. On peut par ailleurs imaginer qu'il est possible, en étudiant la dynamique du graphe du Web, de déceler (voire de suivre) des pages Web qui bougent très souvent. On peut considérer que ces pages cherchent à se cacher, ce qui pourrait être le cas par exemple de sites illégaux (pédophiles, terroristes, etc.). La dynamique du graphe du Web reste cependant largement à étudier, et on peut s'attendre à ce que les recherches qui seront menées sur ce sujet dans les prochaines années apportent de nombreuses réponses, tant fondamentales qu'appliquées.

5 Bibliographie

Références

[1] S. Brin, R. Motwani, L. Page, and T. Winograd, *The pagerank citation ranking: Bringing order to the web*, http://dbpubs.stanford.edu:8090/pub/1999-66.

[2] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener, *Graph structure in the web*, http://www9.org/w9cdrom/160/160.html.

[3] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, http://www.cs.cornell.edu/home/kleinber/auth.ps.

[4] Brian H. Murray and Alvin Moore, *Sizing the internet, a white paper*, http://www.cyveillance.com/web/us/downloads/Sizing_the_Internet.pdf.