

Notes de cours Moteurs de recherche : Master 2 Pro, Université Paris Diderot

Michel Habib and Antoine Meyer

22 janvier 2009

1 Introduction

Ce document a été rédigé à partir des trois mémoires de thèses : [6, 1, 2] et de l'ouvrage [5].¹

Définition 1 *Le Graphe du Web est le graphe orienté dont les sommets sont les pages html et les arcs sont les liens hypertextes. Notons $G = (X, U)$ ce graphe.*

Il a été beaucoup écrit sur la structure de ce graphe, pas toujours très sérieusement (cf. la fameuse structure en noeud papillon de ce graphe), néanmoins nous devons constater que même si le nombre de sommets est énorme (plusieurs milliards) le graphe est peu dense. En fait une page contient peu de liens hypertextes. Cela veut dire que les degrés sortants d'un sommet sont bornés ($d^+(x)$ = le nombre de pages référencées par la page x), mais par contre les degrés entrants ($d^-(x)$ = le nombre de pages pointant sur la page x) ne le sont évidemment pas.

La matrice d'incidence du graphe est dite creuse, et $|U| \in O(|X|)$.

On dit que les degrés entrants vérifient une loi de puissance (i.e. la probabilité qu'un sommet soit de degré entrant k est en $\frac{1}{k^\alpha}$, où α est une constante).

Il aurait été mesuré que $\alpha = 2,5$. Ceci signifie que les sommets dont le degré entrant est important sont peu nombreux.

1. Merci de nous faire part de toute remarque, concernant ce texte, erreurs, passages trop rapides ...

En outre ce graphe est dynamique et on ne peut travailler que sur des extraits de ce graphe calculé par des "crawlers", avec éventuellement un biais important [1].

Le fonctionnement d'un moteur de recherche actuel est grossièrement le suivant :

Données : une chaîne de caractères

1. Extraire des mots clés (analyse syntaxique de la phrase)
2. Trouver toutes les pages de la base de données qui contiennent ces mots clés.
3. Extraire et classer les pages les plus pertinentes.

Réponse Une liste ordonnées d'URL.

Bien souvent pour une requête donnée il y a des centaines de pages (voire des centaines de milliers) qui contiennent cette chaîne de caractères. Il s'agit alors de trier, d'ordonner ces réponses possibles afin de mettre les plus "pertinentes" en tête de liste.

Pour ce faire il existe de nombreuses techniques, souvent les tris sont multi-critères et la démarche globale est heuristique (il est possible d'utiliser le nom de domaine de la machine qui a posé la requête, un profil de l'utilisateur, voire de trier les réponses en fonction des redevances perçues ...).

Afin d'ordonner les réponses, plusieurs auteurs ont eu l'idée de tirer parti de la structure du graphe du Web **indépendamment des contenus des pages**. On peut fabriquer le néologisme **syntaggraphiquement**.

En s'inspirant des théories développées par les sociologues sur les calculs des indices de citations scientifiques, on peut interpréter un lien $A \rightarrow B$, entre deux pages Web A, B comme A "vote" pour B , et en déduire des mesures de popularité.

Plus précisément, Il s'agit d'associer à chaque page un score qui vérifie les règles suivantes :

- Selon **PageRank**, d'après S. Brin, L. Page les fondateurs de Google et R. Motvani, T. Winograd 1999 [3].²

Chaque page est d'autant plus importante qu'elle est référencée par d'autres pages importantes.

2. PageRank est un jeu de mots sur les pages du Web et le nom de L. Page

- Selon **Hits**, d'après J. Kleinberg [4]
Une page mérite un haut score d'annuaire (en anglais hub) si elle référence des pages ayant un haut score d'autorité et un mérite un haut score d'autorité si elle est référencée par des pages ayant un haut score d'annuaire.

Ces deux définitions sont circulaires, mais c'est justement ce qui va permettre de faire un calcul efficace en termes de points fixes d'un calcul matriciel.

2 Calculs effectifs de rangs

Nous allons tout d'abord proposer une formalisation mathématique de l'idée de PageRank. Supposons que la répartition se fasse linéairement sur les arcs du graphe (une sorte de flot traversant le graphe) et si l'on note $R_n(p)$ la valeur du Pagerank de la page p à l'étape n du processus, et $R_{n+1}(p, q)$ la quantité qui passe sur l'arc pq entre les étapes n et $n+1$ l'on obtient l'équation :

$$R_{n+1}(p) = \sum_{q \rightarrow p} R_{n+1}(q, p)$$

où $q \rightarrow p$ signifie arc de la page q vers la page p .

Il est possible de faire l'hypothèse que la répartition se fasse équitablement sur les différents liens sortant d'une page q . Ce qui permet d'exprimer $R_{n+1}(q, p)$ comme suit :

$$R_{n+1}(q, p) = \frac{R_n(q)}{\text{degre}(q)} \text{ pour chaque arc } q \rightarrow p.$$

On obtient donc :

$$R_{n+1}(p) = \sum_{q \rightarrow p} \frac{R_n(q)}{\text{degre}(q)}$$

Vectoriellement cette équation peut s'écrire :

$R_{n+1} = A^T R_n$ où A est une sorte de matrice d'incidence du graphe du Web vérifiant :

$$A[i, j] = \frac{1}{\text{degre}(i)} \text{ si l'arc } ij \text{ existe et } 0 \text{ sinon.}$$

Lorsque la suite R_n converge, elle le fait vers le vecteur propre associé à la valeur propre 1 de la matrice A^T .

La plupart des calculs de PageRank procèdent de manière itérative à partir d'un vecteur initial R_0 . En général $R_0(p) = \frac{1}{\# \text{ total de pages}}$, ce qui revient à répartir de manière équitable sur toutes les pages.

3 Les données

Même si l'on veut classer les pages à l'aide d'une mesure simple telle que le degré entrant d'un sommet ($d^-(p)$), cette donnée n'est accessible qu'à l'aide d'un parcours entier du graphe du Web. Car la seule manière de découvrir les liens existants c'est de lire l'ensemble des pages, i.e. déployer un "crawl" ou chalutage sur l'ensemble du Web, avec toutes les difficultés techniques associées et tous les problèmes de distorsion de modèle [1].

Ainsi la difficulté est la même pour calculer $d^-(p)$ pour une page p donnée que de calculer la matrice A du graphe du Web en entier.

Cette difficulté se retrouve si l'on veut calculer une communauté associée à une page p .

Les moteurs de recherche maintiennent un crawl permanent sur le Web et périodiquement mettent à jour leur classement (pour Google le rythme est mensuel).

4 Le classement ne fait pas tout

Un moteur de recherche fonctionne à l'aide d'une table inversée précalculée qui associe à chaque mot clé la liste ordonnée (suivant PageRank?) des pages "associées" à ce mot clé.

Le traitement d'une requête contenant plusieurs mots clés, revient alors à faire des intersections de listes ordonnées.

Revenons sur cette méthode d'association.

1. le mot est dans le titre de la page,
2. le mot figure x fois dans la zone Meta de description du contenu de la page,
3. Le mot figure x fois dans la page,

L'importance des mots figurant dans le titre de la page est testable par la requête :

French Military Victories pour laquelle Google répond en première ligne une page dont c'est exactement le titre.

Mais il semble que dans Google intervienne aussi le fait que le mot figure dans le chemin d'accès au fichier de la page html.

Plus étonnant encore, Google utilise les textes des balises des pages qui pointent la page p pour indexer cette page. Ceci est révélé par le phénomène du **Google bombing**.

En effet il suffit que quelques pages pointent sur la page Web, biographie officielle de G.W. Bush avec pour balise : "Miserable Failure" pour qu'à la requête : Miserable Failure,

Google réponde en 1ère position la page officielle de G.W. Bush.

À ma connaissance cette page ne contient pas l'item Miserable Failure!

Parmi les autres bombes célèbres on trouve :

N. Sarkozy versus Iznogood

Ministre Blanchisseur versus Renaud Donnedieu de Vabres ...

Il n'est pas facile de gérer ce problème syntagmiquement, car dans la plupart des cas le texte de la balise qui réfère une page est pertinent. La réponse de Google est éloquente sur ce sujet qui persiste depuis 2001.

"Nous ne ferons pas de traitement manuel des liens. Après tout si des sites réfèrent la page officielle de G.W. Bush avec pour balise : "Miserable Failure", nous nous devons d'en tenir compte!"

En janvier 2007, le discours officiel a changé chez Google par l'annonce d'une méthode algorithmique de détection des bombes. Effectivement "Miserable Failure" ne permet plus de retrouver la biographie officielle de G.W. Bush. Cependant la requête Nicolas Sarkozy retourne toujours la page du film Iznogood! On peut légitimement se poser des questions sur l'existence d'un tel algorithme général, et se demander si les modifications n'ont pas été faites à la main?

Ces bombes sont à retardement (au mieux 1 mois) et sont très difficiles à effacer.

5 Le modèle économique d'un moteur de recherche

On peut actuellement distinguer 4 sources déclarées de financement d'un moteur de recherche en prenant pour exemple Google :

1. Indexation rapide de pages (un à deux jours) au lieu du mois normal.
2. Vente des résultats de requêtes aux entreprises, c'est gratuit pour les autres.
3. Publicité, liens payants (affichage dans la colonne de droite)

Avec paiement au click (cost-per-click) (système assez dangereux, car il est possible de nuire à l'entreprise par ce biais à l'aide d'un programme).

4. Revente de profils et de logs des utilisateurs aux entreprises. Faut-il interpréter dans ce contexte, l'embauche récente de salariés de la NSA par Google ?

6 Spam versus antispam

Spam indexing : référencement abusif. Pour une fois la traduction est correcte.

Cela fait penser à la lutte millénaire entre cryptanalystes et cryptographes. À chaque avancée algorithmique des moteurs de recherche pour la détection automatique des spams, correspond une évolution dans la technologie des spams.

Citons pour mémoire quelques étapes de cette lutte. Le but premier c'est d'accroître par tous les moyens le référencement d'une page. On peut le faire par ajout de mots clés du genre : MP3, Sexe, . . . , n'ayant rien à voir avec le contenu de la page en question.

Ces mots peuvent être placés dans la partie description d'une page (i.e. non affichés), mais aussi dans la page elle-même à condition qu'il ne soient pas lisibles (par exemple de la couleur du fond d'écran). Bien entendu ces procédés sont détectables et détectés algorithmiquement. C'est ainsi qu'une nouvelle génération de pages construites comme suit : lors de leur lecture, le type du lecteur est détecté (robot de crawl ou navigateur) et suivant les cas ce n'est pas la même page qui est envoyée.

7 Améliorer son score

Afin d'améliorer la valeur de PageRank d'une page p , en utilisant le principe de l'algorithme, il faut augmenter le nombre de pages qui pointent sur p . Pour ce faire, il suffit de construire un ensemble de pages qui s'auto-référencent : on parle de "ferme de liens", ou de faire des référencement réciproques avec d'autres pages.

On peut aussi s'insérer automatiquement dans les forums des blogs en y référençant la page p .

On assiste donc à une bataille entre programmes, car il est très facile de construire des fausses pages qui ont l'air d'une vraie (générateur aléatoire de texte grammaticalement correct ou carrément récupération de morceaux de textes de pages ayant un fort coefficient d'autorité).

Et pourquoi pas, calculer pour chaque page un score de SPAM (ou spam-rank), en tenant compte de l'idée suivante :

Chaque page est d'autant plus une page de SPAM qu'elle est référencée par d'autres pages SPAM.

8 Problèmes toujours d'actualité

1. Comment détecter les bombes ?
2. Comment détecter syntagmiquement une page dite de "spam".

9 Fouille de graphes

C'est un sujet récent en anglais "Graph Mining" qui consiste en la recherche de motifs (i.e. des graphes particuliers) dans des réseaux de communication.

La première application est celle de la détection de communautés dans un réseau. Pour ce faire on peut utiliser l'un des paradigmes suivants :

- Notion de rôles développées dans le cadre des réseaux sociaux.
- Partitionner le graphe en parties possédant de nombreux liens internes et relativement peu entre deux parties distinctes. Ce problème est l'un des problèmes classiques de l'optimisation combinatoire (i.e. lorsqu'on recherche le partitionnement optimal, le problème devient NP-difficile et nombreuses heuristiques sont disponibles).
- Rechercher des bipartis presque complets.
- Utiliser un modèle particulière .

D'autres applications recherchent des motifs particuliers, par exemple lors de catégorisation d'utilisateurs des communications.

Références

- [1] T. Bennouas. *Modélisation de parcours du WEB et calcul de communautés par émergence*. PhD thesis, Université Montpellier II, 2005.

- [2] M. Bouklit. *Autour du graphe du WEB : Modélisations probabilistes de l'internaute et détection de structures de communautés*. PhD thesis, Université Montpellier II, 2006.
- [3] S. Brin, L. Page, R. Motwani, and T. Winograd. The pagerank citation ranking : bringing an order to the web. *Technical report 1999-0120 Computer Science Dept. Stanford*, 1999.
- [4] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. of ACM*, 1999.
- [5] A.N. Langville and C.D. Meyer. *Google's PageRank and Beyond*. Princeton University press, 2006.
- [6] F. Mathieu. *Graphes du Web, mesures d'importance à la PageRank*. PhD thesis, Université Montpellier II, 2004.