L'algorithme PageRank

Présentation

L'algorithme PageRank, utilisé par le moteur de recherche Google, a pour objectif d'évaluer une métrique de la « popularité » des pages web. Cette mesure de popularité se présente comme une probabilité de visite, associée à chaque page web. Cette probabilité est construite en distinguant deux possibilités pour visiter une page :

— soit en suivant les liens hypertexte présents sur les pages ;
— soit spontanément, ce qui correspond au cas où l'URL est connue par avance.

L'algorithme PageRank évalue alors la probabilité de visite d'une page au cours d'une marche (navigation) aléatoire combinant ces deux possibilités : suivre les liens ou se « téléporter » sur une page arbitraire. L'algorithme considère par ailleurs que la probabilité de choisir l'une ou l'autre de ces actions est la même à tout moment de la navigation (soit : s pour le suivi des liens, et $t = 1 - s$ pour la « téléportation »).

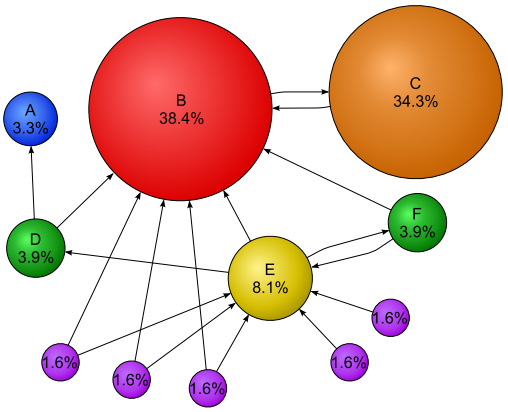
L'algorithme PageRank estime enfin [1] la probabilité de visiter une page donnée dans le cas où l'on suit les liens hypertexte comme la somme des probabilités de visiter les pages qui la référencent, divisé, pour chacune, par le nombre de pages qu'elles référencent.

$$PR(p_i) = s \times \sum_{p_j \ reference \ p_i} \frac{PR(p_j)}{nbL(p_j)} + (1 - s) \times E(p_i) \quad (1)$$

avec

— $PR(p_k)$ PageRank (probabilité de visite) de la page p_k
— $nbL(p_k)$ nombre de liens de p_k, autrement dit nombre de pages (distinctes) référencées par la page p_k
— $E(p_k)$ probabilité de visiter p_k en cas de « téléportation »
— s probabilité de choisir de suivre un lien à tout moment de la navigation.

L'ensemble des pages à évaluer est ainsi vu comme un graphe dont les nœuds ont un poids correspondant à leur probabilité de visite, et dont les arcs correspondent aux liens hypertexte.



(source : wiki commons)

L'algorithme considère un état initial où tous les nœuds sont équiprobables, puis itère l'évaluation des probabilités selon la formule (1), jusqu'à convergence [2].

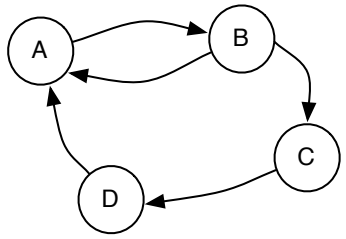
1. L'algorithme réellement utilisé par Google utilise davantage de critères, afin d'une part d'affiner la modélisation des navigations sur le web, et afin d'autre part d'éviter/limiter les stratégies visant à fausser la mesure (en particulier, tous les critères ne sont pas explicités).
2. La convergence est garantie par des considérations simples d'algèbre linéaire [1] [2].

Remarques

- le terme $E(p_k)$, qui représente la « téléportation », a aussi l'intérêt technique d'accélérer la convergence, en forçant la navigation à sortir des cliques dans lesquelles elle peut se trouver prise. L'article initial [1] propose 0,85 pour valeur de s, ce qui s'avère une bonne valeur empirique.
- $E(p_k)$ est considérée par défaut comme égale pour toutes les pages (autrement dit, pour tout k, $E(p_k) = 1/nb\ total\ de\ pages$), mais elle offre la possibilité (intéressante et largement utilisée) d'être modulée, pour traduire le fait qu'il existe des pages « favorites » vers lesquelles ont revient plus souvent, indépendamment de l'état courant de la navigation.
- ce modèle doit être complété, car il ne considère pas le cas des pages pour lesquelles il n'existe pas de lien sortant : en effet, dans ce cas, l'action de suivre un lien n'est pas définie. Le graphe obtenu doit donc être retraité pour éliminer ces situations. Une manière simple de procéder est de considérer que ces pages « puits » sont liées de manière équiprobable à l'ensemble des autres pages.

Exemple

Cet exemple omet, pour simplifier, la possibilité de « téléportation ».



$PR(A) = PR(B)/2 + PR(D)$, $PR(B) = PR(A)$, $PR(C) = PR(B)/2$, $PR(D) = PR(C)$

Initialement $PR(A) = 0,25$, $PR(B) = 0,25$, $PR(C) = 0,25$, $PR(D) = 0,25$
Itération 1 $PR(A) = 0,375$, $PR(B) = 0,25$, $PR(C) = 0,125$, $PR(D) = 0,25$
Itération 2 $PR(A) = 0,375$, $PR(B) = 0,375$, $PR(C) = 0,125$, $PR(D) = 0,125$
...
Convergence $PR(A) = 1/3$, $PR(B) = 1/3$, $PR(C) = 1/6$, $PR(D) = 1/6$

Algorithme MapReduce

L'objectif est d'évaluer le PageRank de l'ensemble des pages d'un site web donné. Un site web est identifié par un nom, comme www.enseeiht.fr, dugenou.perso.enseeiht.fr... Les liens entrants et sortants du site seront donc ignorés.

L'algorithme comporte 2 phases :

- analyse de l'ensemble des pages à évaluer, et construction du graphe ;
- évaluation (itérative) des PageRank, chaque itération appliquant la formule (1) à chacun des nœuds du graphe, en parallèle.

Construction du graphe

Chaque page p_i est identifiée par une URL (URL_{p_i}). La construction du graphe va consister, pour chaque page, à extra ire les liens figurant dans la page considérée, chaque lien correspondant à une URL. Les liens d'une page vers elle-même, ainsi que les liens sortant du site ne sont pas pris en compte. Les liens multiples d'une page donnée vers une même page sont réduits à un unique lien.

Le résultat de la construction est un ensemble de paires ⟨Clé,Valeur⟩ destiné au traitement de l'évaluation de PageRank selon le schéma MapReduce. Ces paires sont de la forme $\langle URL_{page}, (PR; \{URL_{lien}, ...\})\rangle$, où PR est le PageRank initial de la page, et $\{URL_{lien}, ...\}$ est l'ensemble des URL des liens de la page.

La construction du graphe peut être réalisée en suivant un schéma MapReduce, afin de paralléliser l'analyse des pages.

Evaluation des PageRank

L'évaluation des PageRank est réalisée en itérant un schéma MapReduce, pour un nombre d'térations donné, ou encore jusqu'à ce que la différence entre le résultat de deux itérations soit inférieure à un seuil donné.

Chaque itération part d'un ensemble de paires $\langle URL_{page}, (PR_{page}; \{URL_{lien\ de\ la\ page}, ...\})\rangle$,

Map produit pour chaque page p_i et pour chaque lien l_k de p_i une paire $\langle URL_{l_k}, (PR_{p_i}; nbL(p_i))\rangle$, où $nbL(p_i)$ est le nombre de liens de p_i.

Reduce évalue la formule (1) à partir de l'ensemble de paires fourni par la phase Map. Pour cela, on pourra (classiquement) prendre $s = 0,85$ et considérer que E_{p_i} a la même valeur pour toutes les pages, soit $1/nb\ total\ de\ pages$. La phase Reduce produit alors un ensemble de paires $\langle URL_{page}, (PR_{page}; \{URL_{lien\ de\ la\ page}, ...\})\rangle$ disponible pour l'itération suivante.

Le traitement peut être complété par un tri des pages selon le PageRank.

Résultats

L'algorithme de PageRank est une application de référence pour le schéma MapReduce. De nombreux résultats expérimentaux ont été publiés, qui font ressortir que

— l'effet de la parallélisation n'est vraiment sensible qu'au delà d'un certain volume de données traité (de l'ordre du Go). Le coût de l'initialisation est en effet assez élevé.

— le nombre de tâches reduce a un impact sur le temps d'exécution, d'autant plus qu eles volumes à traiter sont importants. Cependant, assez naturellement, l'augmentation du nombre de tâches Reduce est sans effet au-delà d'un certain seuil.

— la convergence de l'algorithme est rapide : une cinquantaine d'itérations suffit à obtenir une bonne précision.

Références

[1] Page, Lawrence and Brin, Sergey [3] and Motwani, Rajeev and Winograd, Terry (1999) : *The PageRank Citation Ranking : Bringing Order to the Web.* Technical Report. Stanford InfoLab.
http ://ilpubs.stanford.edu :8090/422/1/1999-66.pdf
L'article originel, présente l'algorithme et ses bases mathématiques, de façon un peu abrupte.

[2] Nielsen, Michael : *Introduction to PageRank*
http ://michaelnielsen.org/blog/lectures-on-the-google-technology-stack-1-introduction-to-pagerank/
Une présentation plus détaillée et pédagogique de l'algorithme, de ses principes et de ses applications.

[3] Gleich, David : *PageRank Beyond the Web* SIAM Review, 2015, Vol. 57, No. 3 : pp. 321-363
https ://arxiv.org/pdf/1407.5107.pdf
Développe et approfondit le cadre et le contexte d'application de l'algorithme PageRank

3. L Page et S Brin sont les fondateurs de Google