LiveRank : comment faire du neuf avec du vieux ?

The Dang Huynh, Fabien Mathieu, Laurent Viennot

LiveRank : comment faire du neuf avec du vieux ?[†]

The Dang Huynh[1,2], Fabien Mathieu[2] et Laurent Viennot[1]

[1] *Inria*
[2] *Alcatel-Lucent Bell Labs France*

Une capture du Web n'est valable qu'à l'instant où elle est faite et se périme ensuite petit à petit. Dans cet article, nous cherchons à savoir comment récupérer d'une ancienne capture un maximum de pages toujours vivantes en un minimum de requêtes. Plus précisément notre contribution est la suivante : nous posons le problème sous la forme du calcul d'un ranking, le LiveRank, qui essaie de séparer les pages mortes des pages vivantes ; nous proposons plusieurs LiveRanks basés sur le PageRank, avec ou sans apprentissage ; nous validons notre approche sur un graphe réel et évaluons numériquement le gain que peut apporter un bon LiveRank.

Keywords: Graphes du Web, ré-actualisation, PageRank

1 Le problème du LiveRank

Capturer un graphe du Web, ou même d'une partie du Web est une opération longue et fastidieuse qui nécessite de récupérer un très grand nombre de pages (plusieurs milliards). Mais le Web est un graphe naturellement dynamique par nature : des pages sont créées et détruites en permanence.

On considère un graphe du Web $G = (V,E)$ (V : ensemble des pages ; E : ensemble des hyperliens) de taille n obtenu dans le passé. Au moment présent, seul un sous-ensemble V^+ de n^+ pages subsiste. Si l'on veut rafraîchir notre vision du Web, il est naturel de vouloir partir d'un sous-ensemble de pages encore vivantes dans G, mais on aimerait ne pas gaspiller trop de ressources à essayer de récupérer des pages qui n'existent plus. Tout serait facile si l'on connaissait V^+. Dans ce contexte, nous proposons le concept de *LiveRank*. Un LiveRank d'un graphe du Web est une fonction sur les pages qui *essaie* de donner un score élevé aux pages de V^+ et un score bas aux autres.

Une manière d'évaluer la performance d'un LiveRank est de considérer sa fonction d'efficacité f, qui mesure en fonction d'un nombre de pages i la proportion de pages vivantes parmi les i premières du LiveRank. Dit autrement, un LiveRank permet de récupérer, après avoir visité i pages de l'ancien graphe, $f(i)i$ pages encore vivantes.

Un LiveRank idéal (I) a pour fonction $f(i) = 1$ pour $i \leq n^+$, $f(i) = n^+/i$ ensuite. Un tel classement est évidemment impossible à produire avec certitude tant que l'on a pas testé toutes les pages, ce qui est précisément ce que l'on veut éviter. Notre objectif est donc de proposer des approximations de LiveRank, avec une efficacité raisonnable tant que l'on ne cherche à récupérer qu'une partie de V^+.

1.1 LiveRanks non-adaptatifs

Les LiveRanks les plus simples que l'on peut considérer n'utilisent que l'information contenue dans G.

L'ordre aléatoire est proposé afin de servir de référence. En l'absence de corrélation, son efficacité moyenne vaut $\frac{n^+}{n}$.

Le degré entrant est un LiveRank simple dont on peut espérer une efficacité supérieure au simple hasard : intuitivement, plus une page a haut degré, plus elle est vieille (au moment de la mesure) ; plus une page a vécu, plus elle a des chances de vivre encore longtemps.

[†] Le travail présenté ici a été effectué au LINCS (`http://www.lincs.fr/`).

Le PageRank [PBMW99] est une version raffinée du degré entrant. On espère (avec raison) qu'il donne un meilleur LiveRank. Pour rappel, le PageRank utilise G pour inférer une importance définie récursivement : *une page est importante si elle est référencée par des pages importantes*. Formellement, le PageRank est généralement défini comme la solution X d'une équation du type

$$X = dPX + (1-d)Z. \tag{1}$$

où P est une matrice sous-stochastique issue de G, $d < 1$ un terme d'amortissement (fixé empiriquement à $d = 0.85$), et $Z \gneqq 0$ un *vecteur de zap*. Z représente une importance attribuée par défaut aux pages, qui est propagée de page en page selon P avec un amortissement d. Pour le PageRank par défaut, on prendra Z uniforme sur V.

Remarque : il est très difficile d'évaluer la valeur du classement renvoyé par un PageRank en tant que mesure d'importance, car la notion d'importance est très subjective. Il est au contraire facile d'évaluer la valeur du PageRank en tant que LiveRank, par exemple en mesurant l'efficacité.

1.2 LiveRanks Adaptatifs

Un LiveRank étant évalué en le testant sur un graphe, il est naturel de chercher à réinjecter l'information obtenue. Pour rester simple, on considère dans cet article une approche minimaliste en deux étapes : partant d'un LiveRank non-adaptatif (par exemple le PageRank), on se fixe un seuil d'apprentissage z et on teste les z meilleures pages du LiveRank non-adaptatif. On obtient ainsi un ensemble Z^+ de pages vivantes et un ensemble Z^- de pages mortes, qui vont nous permettre de mettre à jour le LiveRank des pages restantes.

LiveRank adaptatif simple Quand une page est vivante, on peut penser que cela augmente les chances que les pages qu'elle pointe dans G le soient, et que la vie se transmet en quelque sorte à travers les hyperliens. Suivant cette idée, un LiveRank adaptatif possible consiste à remplacer dans (1) Z par Z^+ (plus précisément par un vecteur uniforme sur Z^+ et nul ailleurs). Cette diffusion à partir d'un ensemble de départ peut être vue comme une sorte de parcours en largeur à partir de Z^+, mais à la sauce PageRank, c'est-à-dire en pondérant selon la structure de G.

LiveRank adaptatif double Le LiveRank adaptatif simple n'utilise pas l'information donnée par Z^-. Une manière de le faire est de calculer un anti-PageRank basé sur Z^- au lieu de Z^+. Ce ranking mesurerait une sorte de diffusion de la mort. On combine alors les deux, par exemple en considérant le rapport des deux valeurs [‡].

2 Évaluation sur un graphe réel

2.1 Description du graphe

Nous avons testé les différents LiveRanks proposés sur plusieurs graphes issus du projet WebGraph [BV04, BRSV11], de différents millésimes et tailles. Nous présentons ici un graphe représentatif de nos essais, `uk-2002`, capture du domaine `.uk` de 2002 réalisée avec UbiCrawler [BCSV04]. Ce graphe comporte plus de 18 millions de pages pour près de 300 millions d'hyperliens.

Pour pouvoir calculer l'efficacité, il a été nécessaire de déterminer V^+. Nous avons pour cela effectué des requêtes `curl` en parallèle sur 4 machines. Diverses limitations, en particulier au niveau des serveurs interrogés, font que le débit moyen traité est de l'ordre d'un million de requêtes par jour. Le résultat de l'opération est résumé dans le tableau 1.

On remarque en particulier que :

– Un tiers des pages ne sont plus disponibles aujourd'hui, le serveur renvoyant une erreur 404 ;

– Pour un quart des pages, la résolution DNS a échoué (le site web n'existe plus a priori) ;

‡. Empiriquement, le rapport donne de bons résultats – en pensant à remplacer les valeurs nulles dans l'anti-PageRank par une valeur petite mais strictement positive. Cela est probablement lié au fait que les PageRanks ont souvent une répartition en loi à aile lourde.

Statut	Description	Nombre de pages	Pourcentage
Code HTTP 404	Pages introuvables	6 467 219	34,92%
Erreur curl 6	Hôte introuvable	4 470 845	24,14%
Code HTTP 301	Redirections	3 455 923	18,66%
Cible 301	Cibles des redirections	20 414	0,11%
Code HTTP 200	Pages existantes	2 365 201	12,77%
Vraie 200	**Pages réellement existantes**	1 164 998	**6,29%**
Autres (403,...)	Erreurs diverses	1 761 298	9,51%
Total	Taille du graphe	18 520 486	100%

TABLE 1: Statut des pages de `uk-2002`, mesuré en décembre 2013.

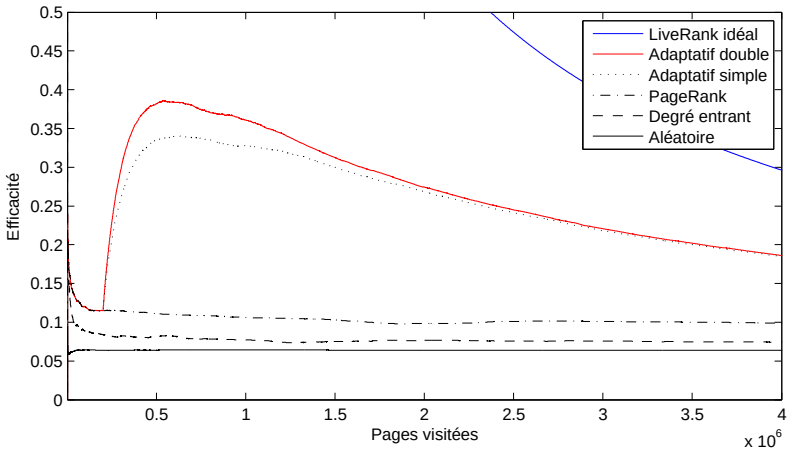

FIGURE 1: Efficacité de différents LiveRanks ($z = 200000$)

– Dans près d'un cinquième des cas, le serveur renvoie un message de redirection (code 301). Mais la plupart du temps, la redirection renvoie l'ensemble des pages d'un ancien site vers la racine d'un nouveau site. Si l'on regarde le nombre de pages vivantes distinctes obtenues à l'arrivée, on tombe à environ 0,1%.
– Restent un peu moins de 13% de pages renvoyant un code 200 (page valide). Mais certaines de ces pages affichent juste une erreur 404. Nous avons donc exclu les pages dont le contenu renvoie *Page Not Found* ou encore *Error 404*.

Au final, il ne reste qu'un peu plus de 6% des pages de départ, même en comptant les cibles des redirections.

2.2 Analyse des performances

La figure 1 montre les efficacités des LiveRanks proposés. Nous limitons l'analyse aux 4.10^6 premières pages sur 18.10^6, qui sont les plus intéressantes.

– Comme attendu, le LiveRank aléatoire produit une efficacité quasi-constante à $6,4\%$.
– Le classement par degré fonctionne assez bien au début : $f(10^4) = 14,7\%$. Sur le long terme, le gain devient plus faible : $f(4.10^6) = 7,5\%$.
– PageRank est meilleur au début ($f(10^4) = 16,9\%$), et se maintient mieux également : $f(4.10^6) = 9,9\%$.
– Après la phase d'apprentissage (en visitant les z pages de meilleur PageRank), les deux LiveRanks adaptatifs offrent un gain réel, avec des pointes respectives à $34,1\%$ et $38,6\%$ pour les versions simple et double respectivement. Il y a un réel gain initial de la prise en compte de l'anti-PageRank, même si les deux LiveRanks adaptatifs se valent sur le long terme ($f(4.10^6) = 18,6\%$).

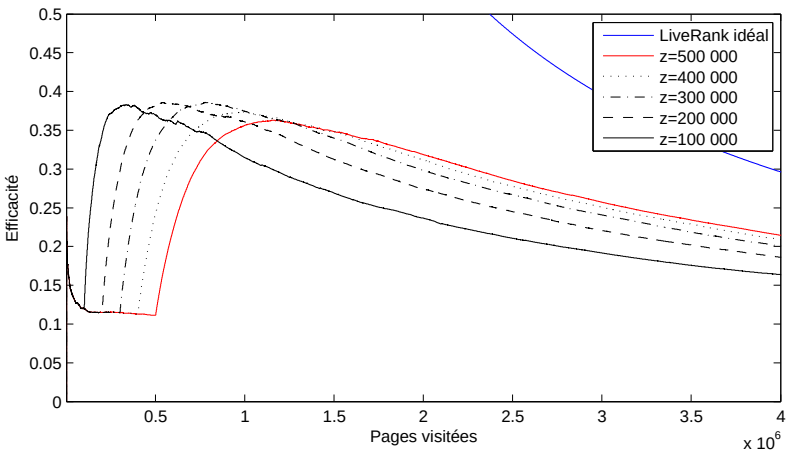

FIGURE 2: Impact de $z = 200000$ (LiveRank Adaptatif double)

Le choix du nombre z de pages visitées dans un premier temps est important. Pour l'évaluer, nous montrons figure 2 l'efficacité du LiveRank adaptatif double pour cinq valeurs différentes. On voit apparaître un compromis : si z est grand, beaucoup de pages vont être visitées selon un LiveRank non-adapté, mais l'efficacité sur le long terme s'en trouve améliorée. Empiriquement, l'adaptation est optimale semble obtenue pour z de l'ordre de 1/3 à 1/2 du nombre de pages que l'on est prêt à visiter en tout.

3 Conclusion

Il nous a fallu environ trois semaines pour récupérer un peu plus d'un million de pages vivantes à partir d'un graphe de départ de 18 millions. Avec un bon LiveRank, il est possible de récupérer le tiers de ces pages dès le premier jour, la moitié au bout de deux jours, les trois quarts au bout de quatre jours.

Le travail présenté est volontairement simplifié pour se concentrer sur la définition du problème et sur les qualités du PageRank comme solution possible. Parmi les nombreuses optimisations possibles, on peut citer : adapter le LiveRank en temps réel (ce qui pose d'intéressants problèmes de calcul différentiel de PageRank), utiliser des techniques de *machine learning*, ou rajouter diverses heuristiques (comme éliminer les pages d'un domaine disparu).

Enfin, nous pensons que ce travail peut s'étendre à divers réseaux sociaux comme Twitter ou Facebook, où les capacités de requêtes sont très limitées, ce qui renforce l'intérêt de pouvoir réutiliser efficacement une capture existante.

Références

[BCSV04] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. Ubicrawler : A scalable fully distributed web crawler. *Software : Practice & Experience*, 34(8) :711–726, 2004.

[BRSV11] Paolo Boldi, Marco Rosa, Massimo Santini, and Sebastiano Vigna. Layered label propagation : A multiresolution coordinate-free ordering for compressing social networks. In *Proceedings of the 20th international conference on World Wide Web*. ACM Press, 2011.

[BV04] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I : Compression techniques. In *Proc. of the Thirteenth International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. In *The PageRank Citation Ranking : Bringing Order to the Web.*, number 1999-66. Stanford InfoLab, November 1999.