Dans la peau d'un moteur de recherche : le PageRank

Mise à jour du 13 décembre 2016

Rémi Bachelet

La dernière version de ce cours est ici : calcul du PageRank.

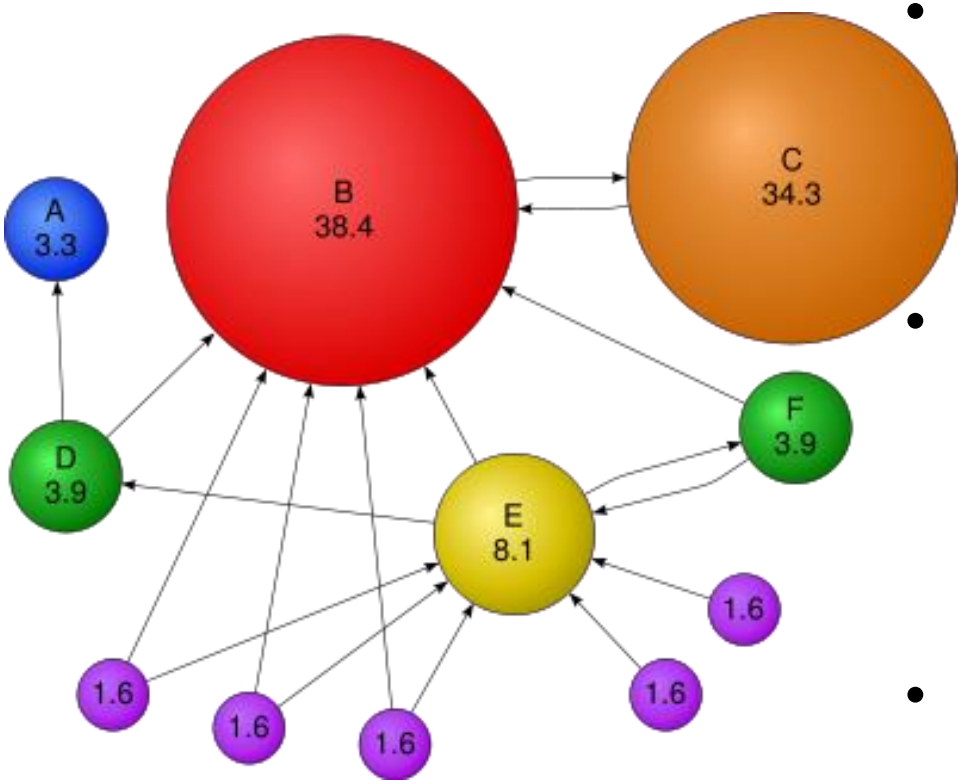
Cette formation est également enregistrée en vidéo

Source des images indiquées au-dessous ou en cliquant sur l'image

Les algorithmes de classement des pages

1. **Le PageRank**
   - **Modalités de calcul**
   - **Fiabilité : Le TrustRank**
   - **Réponse à une requête : Le SERP Rank**

Le PageRank



- « Vote » d'une page pour une autre
  - PR (C) > PR (E), même avec moins de liens !
- Échelle logarithmique 0-10
  - La barre d'outils Google permet d'en visualiser une approximation
  - Avoir un PR de 3-4 c'est déjà beaucoup
- Un processus de calcul récursif
  - Pour éviter que le PR $\rightarrow\infty$, il faut un amortissement = *Damping factor* (typ. 85% - ici il est de 90%).

Image : Source Wikimedia Commons

Le Ranking c'est fini ?

- 2009 : le ranking n'est plus disponible. Dernière mise à jour publique .. En 2013

- Les facteurs qui comptent dans l'algorithme actuel ***Hummingbird***,
  - RankBrain, basé sur l'apprentissage statistique (*machine learning*)
  - Panda, Penguin
  - Payday : antispam,
  - Pigeon : recherches locales,
  - Top Heavy : pages contenant trop de pubs
  - Mobile Friendly : Mobilegeddon
  - Pirate : copyright
  - .. Et PageRank : liens entrants

Le TrustRank

- Méthode semi-automatique pour détecter les pages de spam = classification "spam ou pas spam" (*Trust* = confiance - Le terme TrustRank vient de Yahoo!).
- Principe : une page « propre » ne propose pas de liens vers des pages de spam

1. Amorçage : établir une liste de pages « propres » de référence
   - Après une analyse « humaine ».
   - On n'a pas forcément besoin d'une grande liste (p.e 200 sites).
2. Suivi récursif des liens de la liste d'amorçage
3. Degré de confiance que l'on peut attribuer à la page : un indice
   - Plus les liens sont forts avec des pages de référence, plus leur degré de confiance est élevé
   - C'est le TrustRank (ou TR), indice **entre 0** (=spam) et 1 (=page de référence)

- Le TrustRank peut être utilisé :
  - pour filtrer l'index d'un moteur de recherche,
  - pour classer les résultats d'une recherche.

L’algorithme du PR : un secret bien gardé

Un nombre important de facteurs est pris en compte dans le PageRank.

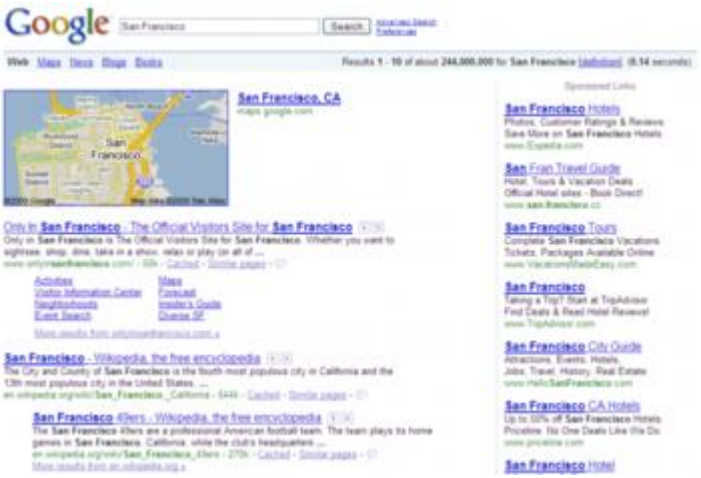
- Ces facteurs sont très nombreux (plus de 200 d'après Google).
- Leur nature et leur pondération sont secrets pour limiter les chances de manipulations (et la concurrence des autres moteurs de recherche).
- Le terme "PageRank" est une marque déposée et a été l’objet de brevets, à commencer par (U.S. Patent 6,285,999). Le brevet appartient à Stanford University et Google en a l’usage exclusif, mais l’algorithme a beaucoup évolué depuis le dépôt en 98.
- Beaucoup de spéculations sur ce sujet, voyons quelques-uns des paramètres connus…

Quelques-uns des 200 paramètres du calcul du ranking

1. Sur la page *(« onpage »)*
   - Ancienneté / Fréquence d'actualisation
   - Texte = visible sur la page / Code = Meta tags = non visibles sur la page
2. Sur le site *(« onsite »)*
   - Lien internes, arborescence, fil d'ariane (« Breadcrumbs »)
   - Paramétrage sur Google outils pour les webmasters (Sitemap..)
3. Hors du site *(« offsite »)*
   - Liens entrants `en (petite) partie visibles via une recherche Google link:http://fr.wikipedia.org`
     - Leur PageRank, Âge, TrustRank de la page
     - Social bookmaking, tweets…

Un débat : Google utilise t-il les données qu'il stocke sur le comportement des internautes pour le calcul du PageRank ?

- Temps passé sur le site, statistiques renvoyées par la barre d'outil google, annotations sidewiki, citations d'URL dans gmail, requêtes avec l'URL du site, marque-pages Google, âge/sexe/localisation des internautes, leurs recherches précédentes …. *les licences de ces services précisent souvent que non..*



Le SERP Rank

C’est l’ordre de présentation des liens lorsque l’on entre des mots-clés dans un moteur de recherche

- La page de résultats présente une liste ordonnée de liens vers des pages/images/vidéos, associés à des textes courts (*snippets*)

- Le SERP Rank est fonction du PageRank, mais aussi de facteurs liés aux mots-clés.
  - Voir le chapitre 6 sur les mots-clés et leur mise en valeur
  - SERP = *Search Engine Results Page*

La Google Dance

- Période durant laquelle Google mettait à jour le classement des pages.
- La Google Dance n'existe plus !
  - Le processus d’actualisation est désormais continu, et consultable sur GWT

Questions ?

- EdgeRank de Facebook
- Plus d’informations sur les lien-retours/backlinks
- Mathématiquement, le *PageRank* est la probabilité stationnaire d'une chaîne de Markov, c'est-à-dire un vecteur de Perron-Frobenius de la matrice d'adjacence du graphe du Web[1],[2]

Les thèmes et chapitres du cours

1. Origine du SEO, Google … et ses concurrents
2. La fréquentation d'un site : les fondamentaux
3. Dans la peau d'un moteur de recherche : le PageRank
4. Web Analytics et liens commerciaux
5. Optimisation du référencement
   - "Onpage"
   - "Onsite"
   - "Offsite"
6. Trouver et optimiser les mots-clés
7. Méthodologie de référencement et avenir du SEO