

Examen du cours Moteurs de recherche M2 Pro, Université Paris Diderot, Paris 7

Michel Habib

March 21, 2011

1 Usages pratiques d'un moteur de recherche généraliste de type Google

1. Donner un exemple de requête simple Google sans réponse.
2. Donner un exemple d'une requête Google dont vous ne jugez pas les réponses pertinentes. Avez vous une explication du phénomène ?

2 Améliorations du classement des pages

1. Comment définir à la façon de PageRank, un indice EdgeRank sur les liens ?
2. Comment définir à la façon de PageRank, un indice FreshRank de Fraîcheur d'une page (définir la formule précisément en fonction de la fréquence de modifications exprimée en nombre de modifications par jour ?)
3. Quand on a calculé plusieurs indices, (par exemple PageRank et FreshRank) qui s'appliquent sur une page, comment les utiliser dans un algorithme (ou méthode) de classement ?
4. Comment faire intervenir EdgeRank ? Peut-on l'utiliser dans le calcul de Pagerank lui-même.
5. En reprenant tous les critères présentés dans le cours sur le brevet 2007 de Google, proposer une méthode pour implémenter ces critères ou du moins ceux qui vous paraissent les plus importants.
6. Certains critères de ce brevet, nécessitent de connaître le comportement des internautes (suivi des liens, temps passé sur une page ...). Ces informations ne sont pas complètes (par exemple tous les navigateurs ne fournissent pas ces informations) comment faire avec des données incomplètes ?
7. Auriez-vous ajouter d'autres critères à ceux du brevet de 2007 ?

3 Détection des fermes de contenus

Des sites comme 123.people sont actuellement bien référencés en liens normaux (i.e. non commerciaux) par Google, et pourtant ils n'apportent souvent strictement aucune information par rapport à la requête posée. La question se pose donc de modifier l'algorithme de classement de Google afin de rétrograder ces sites "inutiles".

1. Une idée simple serait de mesurer la quantité de publicité sur la page. Comment la mettre en œuvre, sachant que pour certaines pages c'est normal (SNCF par exemple) de tolérer de la publicité.

Sur la page de la SNCF, il suffit qu'il reste un rectangle dans lequel définir notre voyage, le reste ne compte pas. Il ne faut surtout pas rétrograder la page de la SNCF, car elle est très utile.

2. Quelle autre méthode proposeriez-vous ?

*Idée les anglophones ?
comparer plus de sites ?*

4 Détection algorithmique de trolls

Dans les commentaires qui suivent un article sur un journal ou un blog, on trouve souvent des commentaires qui n'ont rien à voir avec le contenu de l'article et qui parasitent les échanges d'opinion. On appelle ceux qui écrivent ces commentaires des "Trolls".

Selon Wikipedia: Un troll est une action de nature à créer une polémique en provoquant les participants d'un espace de discussion (de type forum, newsgroup ou wiki) sur un réseau informatique, notamment Internet et Usenet. Le mot désigne également un utilisateur qui a recours à ce type d'action. Par métonymie, on parle de troll pour un message dont le caractère est susceptible de générer des polémiques ou est excessivement provocateur, ou auquel on ne veut pas répondre et que l'on tente de discréditer en le nommant ainsi. Le mot "troll" peut également faire référence à un débat conflictuel dans sa globalité. Dans la majorité des cas, l'évaluation repose sur l'aspect récurrent ou caricatural de l'argumentation, les participants peuvent alors tout aussi bien être qualifiés de "trolls" que de "trolleurs".

Il s'agit d'envisager un repérage automatique (algorithmique) de ces commentaires trolls. Considérons ces exemples issus d'un article sur les économies d'énergie et l'effet rebond de la consommation (ex: quand on achète une nouvelle voiture qui consomme moins, on l'utilise plus).

L'article suivant est extrait du blog d'Audrey Garric, journaliste au Monde.fr (27 février 2011):

Des ampoules basse consommation que l'on éteint moins fréquemment. Des voitures économes en carburant que l'on conduit davantage. Un chauffage que l'on pousse plus après avoir mieux isolé son logement. Les exemples de ce genre sont nombreux. Et pourraient contrecarrer les tentatives de lutte contre le changement climatique. C'est ce que l'on appelle l'effet rebond, c'est-à-dire la

réduction voire l'annulation des économies d'énergie par des modifications du comportement de la société, directes mais aussi indirectes - par exemple en abaissant les coûts de l'énergie et donc en augmentant la demande et la consommation. Ce phénomène, observé depuis quelques années, est confirmé par de récentes études sur le sujet.

Dans une large enquête, le think tank américain Breakthrough Institute avait conclu que l'effet rebond pouvait avoir un impact sérieux sur la réduction des émissions de carbone. "A chaque fois que nous progressons de deux pas du point de vue de l'efficacité énergétique et donc la réduction des coûts, l'effet rebond signifie un recul en arrière d'un pas ou plus, allant parfois jusqu'à ruiner les gains réalisés au départ", assure l'auteur principal du rapport, Jesse Jenkins, cité par le Guardian. Ainsi, à une échelle micro-économique, entre 10 et 30% des économies d'énergie réalisées grâce à des voitures ou des logements performants seraient perdues.

Les causes d'un tel contre-effet ne sont pas parfaitement claires. Les raisons sont peut-être à chercher du côté du principe de la consommation, qui nous fait utiliser davantage un produit ou un service qui coûte moins cher. Ou du côté de la conscience, qui tolérerait une action néfaste pour l'environnement sous prétexte que l'on a réalisé des efforts par ailleurs. Ainsi, Steve Sorrell, expert en énergie à l'Université du Sussex, a passé au crible trois actions d'économie d'énergie : baisser son chauffage de un degré, remplacer les courts trajets en voiture par de la marche ou du vélo et gaspiller un tiers de moins de nourriture. Selon le chercheur, si vous menez ces trois actions mais consacrez l'argent économisé à vos dépenses habituelles, l'effet de rebond sera de 34%, c'est-à-dire une annulation d'autant de la réduction des émissions de gaz à effet de serre réalisée.

Pour le chauffage des logements, en particulier, de nombreuses raisons pourraient expliquer pourquoi la limitation de l'intensité énergétique ne débouche pas sur les économies attendues. "Souvent, les logements ne fonctionnent pas comme prévu, en partie parce que leurs occupants se comportent de manière plus complexe que ce à quoi s'attendaient les designers. Ils vont par exemple ouvrir les fenêtres, laisser les portes ouvertes, générer de la chaleur corporelle ou installer des écrans plasma", livre Kathryn Janda, auteur d'un récent rapport sur la consommation d'énergie pour le Centre de recherche sur l'énergie au Royaume-Uni.

Que conclure de ces études ? S'il faut bien évidemment poursuivre nos efforts pour améliorer l'efficacité énergétique de nos infrastructures et réduire nos émissions carbonées, nous devons réaliser ces efforts dans une optique plus globale de réduction de notre consommation, motivée par les retombées environnementales aussi bien que les gains économiques, pour réellement lutter contre le changement climatique.

Et vous, qu'en pensez-vous ? Avez-vous conscience de cet effet rebond et que faites-vous pour le limiter ?

Je me suis inspiré assez librement des commentaires de l'article et j'en ai extrait 7.

1. Article très intéressant, je confirme cet effet rebond de la consommation estimé à 20%.
2. Et plus on pédale moins fort, moins on avance plus vite. Bon j'arrête, les blagues les plus

courtes sont les moins longues.

3. D'accord, tous les commentaires ne volent pas haut mais perso, j'ai appris des trucs, notamment grâce à la personne qui a mis le lien vers le paradoxe de Jevons (**on trouve ici un lien vers un truc bizarre**) que je ne connaissais pas. C'est intéressant je trouve.
4. J.L. Borloo avec son Grenelle de l'environnement a tout raté.
5. Notre monarque n'y comprend rien!
6. Le PS c'est pareil, avec les 35h il a détruit la France.
7. Vive le vélo! Quand on achète un vélo neuf, on l'utilise plus et cela consomme moins !

Questions:

1. Donner "à la main" un coefficient ~~entre~~ de Troll $\in [0, 1]$ aux commentaires précédents.
2. Peut-on obtenir un résultat analogue en faisant un traitement par mots clés de ces commentaires ?
3. Comment faut-il faire pour automatiser l'analyse de commentaires ?
4. Peut-on définir un indice de troll par article, et comment le comptabiliser par rapport à la notoriété d'un journal ou d'un blog ?