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

Michel Habib

March 29, 2013

1 Questions de cours

1. Comment valider la pertinence des réponses d'un moteur de recherche ?
2. Le Web est-il une immense base de données ?
   Quelles sont les conséquences de penser le Web comme cela ?

2 Questions de base sur PageRank, sorte de suite du TP 1

1. Considérons une étoile symétrique à 4 branches. Soit x le centre de l'étoile et les a, b, c, d sommets feuilles. Les arcs sont donc ax, xa, bx, xb, cx, xc, dx, xd. En partant du vecteur (1/5, 1/5, 1/5, 1/5, 1/5), que se passe-t-il dans le calcul du PageRank de base ?
2. Expliquer votre résultat.
3. Si l'on ajoute un sommet a' sur l'arc xa qui devient xa', a'a. Refaire le calcul du PageRank. Expliquez votre résultat.
4. Variante: on ajoute à l'étoile symétrique à quatre branches de la question 1, une boucle en a.
5. Reprendre les questions précédentes avec un facteur ZAP on nul.

3 Autour de PageRank

1. On considère un graphe orienté sur lequel le PageRank de base converge (rappelez les conditions sur le graphe).
2. Notons $pg_i(x)$ $(resp. pgz_i(x))$ la valeur du Pagerank du sommet x (resp. du PageRank avec un facteur de ZAP non nul) après i itérations de calcul.

   A-t-on $\forall x$ et $\forall i \geq 0$, $pg_i(x) \leq pgz_i(x)$? (on donnera soit une preuve, soit un contre-exemple, une réponse OUI /NON ne suffit pas !)

3. A-t-on pour $\forall x, y$ et $\forall i \geq 0$: $pg_i(x) \leq pg_i(y)$ implique $pgz_i(x) \leq pgz_i(y)$?

4. Ces inégalités sont-elles valides à la limite (i.e. pour les grandes valeurs de i) ?

5. Dans le cadre du calcul d'un PageRank pour un moteur de recherche de type GOOGLE que faire des boucles du graphe (i.e. des pages que se référencent elles-mêmes)?

6. Considérons le graphe orienté des citations entre auteurs informaticiens. Comment interpréter l'ordre sur les auteurs obtenu à l'aide d'un calcul de PageRank sur ce graphe ? Que mesure ce coefficient ?

   Peut-on le justifier par rapport aux deux ordres suivants:

   Celui du nombre total de publications

   Celui du nombre total de citations.

7. Dans le cas de ce graphe des citations, faut-il tenir compte des boucles, c.a.d des autocitations, si oui comment ?

8. Peut-on faire un système d'élection d'un leader à l'aide de PageRank ? Donnez quelques exemples de votes. Par exemple chacun vote pour un candidat et on construit le graphe orienté associé.

   En cas de reponse OUI, expliquez comment. Par exemple: faut-il que chacun vote pour plusieurs candidats ou encore faut-il voter en ordonnant les votes ?

4 A l'occasion d'un entretien d'embauche chez Google

On vous pose les questions suivantes:

1. Sachant que la sémantique des pages passe par les noms de pages WEB (leur chemin d'accès) et par les balises, quelle suggestion d'amélioration proposeriez-vous afin que le moteur de recherche répondre un peu mieux aux questions posées (au lieu de répondre une liste ordonnée de liens vers des pages Web qui contiennent peut-être la réponse)?

   Il s'agit de proposer quelque chose de programmable facilement et d'efficace même si ce n'est pas parfait.

2. Quelle autre suggestion d'usage ou de structure du moteur de recherche actuel ?

5 Recommandations

Etant donné un site de vente en ligne de produits de beauté (resp. de livres, de vin, de places de spectacle) il s'agit de proposer un système de recommandation pour ce site.

1. Choisissez une application et présentez votre solution. Précisez votre modélisation, l'usage de votre système ainsi que les algorithmes que vous allez utiliser pour réaliser votre système de recommandation.
2. Comment organiser une validation de votre système (avant et après installation) ?
3. Votre système peut-il s'adapter aux autres applications parmi les 3 proposées ? Quelles modifications ?