PageRank*

Oana Balalau (TA), Luis Gallaraga (TA), Mauro Sozio
(firstname.lastname@telecom-paristech.fr)

29 novembre 2015

Calcul de PageRank dans Simple English Wikipedia

PageRank est la technique qu'ont proposé les fondateurs de Google, Brin et Page, pour associer un score aux pages du Web. L'idée de PageRank est la suivante : *les pages importantes sur le Web sont les pages étant pointées par des pages importantes.* Plus généralement, le score de PageRank est une mesure utile à calculer dans tout graphe orienté, et peut révéler des informations sur l'importance et la centralité des nœuds de ce graphe.

PageRank est défini comme la probabilité qu'un *surfeur aléatoire* effectuant une marche aléatoire sur le Web en suivant les liens uniformément au hasard (et, avec une faible probabilité, effectuant un saut vers une autre page du Web choisie uniformément au hasard) se retrouve sur une page donnée dans un point distant du futur (une fois que la *mesure d'équilibre* de la chaîne de Markov a été atteinte).

Étant donné un graphe orienté de matrice d'adjacence G ($G(i,j)$ vaut 1 s'il y a un lien de i vers j, 0 sinon), le PageRank des nœuds du graphe peut être calculé de la manière suivante :

1. Normaliser G pour que chaque ligne somme à 1.
2. Soit u le vecteur uniforme de somme 1, soit v égal à u.
3. Répéter jusqu'à convergence (par exemple différence relative inférieure à 1% entre les versions successives de v) :
   — $v := (1-d)^t Gv + du$ (avec par exemple $d = \frac{1}{4}$).

Vous trouverez dans l'achive un jeu de données formé du graphe de la version Simple English (voir `http://simple.wikipedia.org/`) de Wikipedia. Ce jeu de données est formé d'un ensemble de titre d'articles (labels) et d'un ensemble d'arêtes(edge_list.txt), décrit par un fichier dont chaque ligne est de la forme :

`A B1,C1 B2,C2 ... Bn,Cn`

où `A` est l'index d'un article (le fichier des titres d'articles les donnant dans l'ordre), `B1`, ..., `Bn` sont des index d'articles pointés par `A`, et `C1`, ..., `Cn` sont le nombre de liens de `A` à l'article en question.

En utilisant MapReduce, le but est de calculer le PageRank de l'ensemble des nœuds du jeu de données, et de trier le résultat par PageRank décroissant.

En particulier vous devrez :

— charger le jeu de données dans HDFS, sous un format lisible par Hadoop (le format `SequenceFile` est recommandé, il peut être produit avec `hadoop.writetb`, cf. `http://hadoopy.readthedocs.org/en/latest/tutorial.html` : complétez le fichier LoadIntoHDFS.py.

*nous remercions Pierre Senellart (`pierre.senellart@telecom-paristech.fr`) pour avoir préparé ce projet.

— écrire la multiplication matricielle sous la forme d'un job MapReduce : complétez le fichier PageRank.py ;

— écrire la structure générale du programme faisant appel à ces jobs jusque convergence : complétez le fichier PageRankDriver.py ;

— trier et interpréter le résultat.

Quel est l'article le plus important dans Simple English Wikipedia ?