PageRank in MapReduce*

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

September 4, 2016

Ranking Wikipedia Web pages with PageRank

PageRank is the algorithm proposed by the founders of Google (L. Page, S. Brin) for computing the "importance" of Web pages. The PageRank algorithm uses the fact that the importance of a Web page P is proportional to the importance of the Web pages containing a hyperlink to P. For each Web page, PageRank computes a score between 0 and 1. The larger the PageRank score of a Web page, the more important is such a Web page.

We are given a directed graph G where nodes represent Web pages, while directed edges represent hyperlinks. We assume there are n pages. The graph is represented as an adjacency matrix G with n rows and n columns, i.e., $G(i,j) = 1$ if and only if there is a directed link from node i to j, otherwise $G(i,j) = 0$. The PageRank score for each node can be computed as follows:

1. Normalize the rows of G such that the values in each row sum up to 1.
2. Let r^0 be the uniform vector, i.e. each element is equal to $1/n$. This is the PageRank vector of the pages (i.e. it contains one score for each page) at step 0.
3. Let r^t be the PageRank vector at step t. Iterate until the L_1 norm of the vector $r^{t+1} - r^t$ is at most 0.01:

   $v := (1-d)^t Gv + du$, where $d = 1/4$.

You will find in the archive posted on Moodle a dataset containing the simple English version of Wikipedia (`http://simple.wikipedia.org`) represented as a directed graph. Such a dataset consists of a set of Wikipedia pages (labels) and a set of links (edge_list.txt) where every line is of the kind:

$A\ B1\ B2 \dots Bn$

where A is the id of a web page, $B1, \dots, Bn$ are the pages A links to.

The goal is to compute the PageRank scores of the Web pages in MapReduce and then sort them non-increasingly (i.e. the largest scores first). In particular, you should:

- store the dataset in the $HDFS$ (Hadoop Distributed File System) in a format which is compatible with Hadoop. We suggest to use the `SequenceFile` which can be produced by using `hadoop.writetb` (see `http://hadoopy.readthedocs.org/en/latest/tutorial.html`). To this end, you should fill the missing parts in the file LoadIntoHDFS.py.

  The function `hadoop.writetb` receives the path on the HDFS and an iterator to a list of pairs (key, value). Note that a value of format 'v1 v2... vm' will be interpreted as a list. This encoding exists in order to deal with keys that have several corresponding values.

*Thanks to Pierre Senellart for helping preparing this exercise.

- Write a MapReduce job which implements the multiplication between matrices. To this end, you should fill the missing parts in the file PageRank.py; To debug the MapReduce job use the function *hadoopy.launch_local*.
- Sort the PageRank scores and try to understand the results.

What is the most important page in Simple English Wikipedia?