PageRank Report

Big Data Platforms Lab 1

Hippolyte Jacomet
Karl Bou Abboud

OUR APPROACH

As for the TFIDF, the PageRank is computed using several successive MapReduce jobs organised in the PageRank.java file. SortPRResults is called separately and allows to process the output.

First Step

A first simple MapReduce job allows to transform the input file to the format {node -> [list of linked nodes]} with PreMapper and PreReducer.

We tried to use also this step to count the number of nodes with a counter, since it is required to compute the PageRank. Unfortunately, it failed, yielding only 60,000 nodes instead of the desired 75,879. Indeed, here it fails to count nodes with no outbound links. We didn't find any satisfying solution to this problem in time, and hard coded the number of nodes.

Second Step: PageRank Initialisation

The second MapReduce job initializes the PageRank with PRMapper and PRReducer.

Iteration

Then a MapReduce job, which is only really a reduce operation is iterated over for a given number of times, passed in argument. SimpleMapper (since it doesn't do anything but map) and PRReducer are used, up until the last iteration which uses FinalReducer to have as output one node and its PageRank.

Convergence is not evaluated here since we didn't find any satisfying mean to do so, but it can be achieved either on the resulting ranking of nodes, or on the coefficient themselves.

RESULTS

Using only 1 mapper and 1 reducer for each job, and after 40 iterations.

Using several reducers would require concatenating

the different final outputs in a single file to sort the results.

Top 10 nodes:

```
18 : 0.003250714618219084
737 : 0.002258261805993395
118 : 0.0015210605870605413
1719 : 0.0014896446775919943
136 : 0.0014243407734474162
790 : 0.0014112945320872378
143 : 0.0014026730313270155
40 : 0.001308103350814045
1619 : 0.0011011908520745534
725 : 0.0010723736693660842
```

Monitoring (Last few jobs only)

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
44	0	0	44	0	0 B	8 GB	0 B	0	8	0	1	0	0	0	0

User Metrics for dr.who

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Containers Pending	Containers Reserved	Memory Used	Memory Pending	Memory Reserved	VCores Used	VCores Pending	VCores Reserved
0	0	0	0	0	0	0	0 B	0 B	0 B	0	0	0

Show 20 entries   Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Progress	Tracking UI
application_1495024755187_0044	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:59:58 -0700 2017	Thu May 18 06:00:26 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0043	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:59:27 -0700 2017	Thu May 18 05:59:56 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0042	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:58:56 -0700 2017	Thu May 18 05:59:25 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0041	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:58:24 -0700 2017	Thu May 18 05:58:52 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0040	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:57:51 -0700 2017	Thu May 18 05:58:21 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0039	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:57:18 -0700 2017	Thu May 18 05:57:48 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0038	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:56:48 -0700 2017	Thu May 18 05:57:16 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0037	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:56:14 -0700 2017	Thu May 18 05:56:45 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0036	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:55:39 -0700 2017	Thu May 18 05:56:12 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0035	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:55:06 -0700 2017	Thu May 18 05:55:36 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0034	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:54:33 -0700 2017	Thu May 18 05:55:04 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0033	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:53:58 -0700 2017	Thu May 18 05:54:30 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0032	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:53:26 -0700 2017	Thu May 18 05:53:56 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0031	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:52:51 -0700 2017	Thu May 18 05:53:23 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0030	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:52:20 -0700 2017	Thu May 18 05:52:49 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0029	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:51:47 -0700 2017	Thu May 18 05:52:18 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0028	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:51:12 -0700 2017	Thu May 18 05:51:44 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0027	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:50:36 -0700 2017	Thu May 18 05:51:08 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0026	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:50:01 -0700 2017	Thu May 18 05:50:32 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History
application_1495024755187_0025	cloudera	Step : iterate pagerank	MAPREDUCE	root.cloudera	Thu May 18 05:49:26 -0700 2017	Thu May 18 05:49:58 -0700 2017	FINISHED	SUCCEEDED	N/A	N/A	N/A		History