



UNIVERSITÀ DI PISA

Artificial Intelligence and Data Engineering

Cloud Computing

PageRank

Project Documentation

TEAM MEMBERS:

Federica Baldi

Daniele Cioffo

Edoardo Fazzari

Mirco Ramo

Academic Year: 2020/2021

Contents

1	PageRank	2
1.1	Pseudocode	2
2	Hadoop Implementation	3
2.1	Introduction	3
2.2	Pseudocode	3
2.2.1	First Phase: Graph Construction	3
2.2.2	Second Phase: PageRank Estimation	4
2.2.3	Third Phase: Sorting	4
2.3	Implementations Details	5
2.3.1	First Phase: Graph construction	5
2.3.2	Second Phase: PageRank Estimation	6
2.3.3	Third Phase: Sorting	6
2.4	Performance	6

1 — PageRank

1.1 Pseudocode

The pseudocode presented in this chapter should be considered as a *generic* implementation of the **PageRank Algorithm**, a detailed and framework-based version of the procedures is shown in the next chapter.

Algorithm 1 PageRank

```
1: procedure COUNTING_NODES(dataset d)
2:    $N \leftarrow d.countNodes()$ 
3:   return N

4: procedure GRAPH_CONSTRUCTION - PARSING(dataset d)
5:   nodesList new AssociativeArray
6:   for all page in d do
7:     title  $\leftarrow page.getTitle()$ 
8:     outgoingEdges  $\leftarrow page.getOutgoingEdges()$ 
9:     nodeLists.add({title, outgoingEdges})
10:  return nodesList

11: procedure COMPUTE_PAGERANK(nodesList NL, numberOfNodes N, nOfIterations NI)
12:   NL.addInitPageRankToNodes( $\frac{1}{N}$ )
13:   for i in range(NI) do
14:     NL.Map()
15:     NL.Reduce()
16:   NL.sortByPagerankDescendingOrder()

17: procedure MAP(title t, {outgoingEdges oe, pagerank p})
18:   for all e in oe do
19:     EMIT(e,  $\frac{p}{oe.length}$ )

20: procedure REDUCE(title t, pagerankContributions  $[p_1, p_2, \dots]$ )
21:   damping  $\leftarrow 0.8$ 
22:   sum  $\leftarrow 0$ 
23:    $N \leftarrow numberOfNodes$ 
24:   for all p do in pagerankContributions
25:     sum  $\leftarrow sum + p$ 
26:   pagerank =  $\frac{(1-damping)}{N} + damping * sum$ 
27:   EMIT(t, pagerank)
```

2 — Hadoop Implementation

2.1 Introduction

In this section a description of the MapReduce implementation of *Page Rank* is given. The algorithm is carried out in **three distinct steps**:

1. *Graph Construction* phase
2. *Page Rank Computation* phase
3. *Sorting* phase

We have decided to represent the structure of the graph through adjacency lists, so each node (that represents a page) will keep the list of outgoing edges as status information, as well as its ranking.

2.2 Pseudocode

In this chapter we will report a pseudocode of the three phases, to better understand the implementation in Hadoop.

2.2.1 First Phase: Graph Construction

In this phase we parse the information in the input file taking the information that interests us, i.e. the title of the page and the outgoing edges. In addition to this, we take advantage of this phase to also perform the count of the graph nodes.

Algorithm 2 Graph Construction Mapper

```
1: procedure SETUP
2:   counter  $\leftarrow 0$ 
3: procedure MAP(PageId id, Page p)
4:   title  $\leftarrow getTitle(p)$ 
5:   outgoingEdges  $\leftarrow getOutgoingEdges(p)$ 
6:   counter  $\leftarrow counter + 1$ 
7:   EMIT(title, outgoingEdges)
8: procedure CLEANUP
9:   EMIT("", counter)
```

Algorithm 3 Graph Construction Reducer

```
1: procedure REDUCE(Title t, [e1, e2, ...])
2:   initialPageRank  $\leftarrow -1$ 
3:   if t == "" then ▷ If these are the counters
4:     globalCounter  $\leftarrow 0$ 
5:     for all e in [e1, e2, ...] do
6:       globalCounter  $\leftarrow globalCounter + e$ 
7:     EMIT("", globalCounter)
8:   else
9:     edges  $\leftarrow e_1$ 
10:    EMIT(t, {initialPageRank, edges})
```

In *Algorithm 3* only e_1 is considered as edges because the title is a unique identifier of the pages, so the list of input values will always consist of a single element. Only one mapper will manage one page. This is not the case when we receive the local counters from the mappers. More information can be found in the "Implementations Details" chapter.

2.2.2 Second Phase: PageRank Estimation

In this section, the pagerank iteration is presented. The number of iteration is fixed at the start of the execution. So the termination condition is the number of iterations.

Algorithm 4 PageRank Computation Mapper

```

1: procedure SETUP
2:    $N \leftarrow \text{numberOfNodes}$ 
3: procedure MAP(Key k, Node n)
4:   if n.pagerank == -1 then ▷ First time
5:      $n.pagerank = \frac{1}{N}$ 
6:   EMIT(n.title, {-1, n.outgoingEdges}) ▷ Pass along the graph structure
7:    $massToSend = \frac{n.pagerank}{n.outgoingEdges.length}$ 
8:   for all outgoingEdge in n.outgoingEdges do
9:     EMIT(outgoingEdge, {massToSend, []}) ▷ Pass PageRank mass to neighbors

```

Note: *outgoingEdge* is a title itself.

Algorithm 5 PageRank Computation Reducer

```

1: procedure SETUP
2:    $N \leftarrow \text{numberOfNodes}$ 
3:    $damping \leftarrow 0.8$ 
4: procedure REDUCE(Title t, Nodes [ $n_1, n_2, \dots$ ])
5:    $s \leftarrow 0$  ▷ Ingoing mass
6:    $n \leftarrow \text{null}$ 
7:   for all node in nodes do
8:     if node.pagerank == -1 then
9:        $n \leftarrow \text{node}$ 
10:    else
11:       $s \leftarrow s + \text{node.pagerank}$ 
12:  if n != null then ▷ If it is a node
13:     $n.pagerank \leftarrow \frac{(1-D)}{N} + D * s$ 
14:    EMIT(t, n)

```

2.2.3 Third Phase: Sorting

The final step is sorting the webpages by decreasing rank, this is done making advantage of the sorting mechanism of MapReduce.

Algorithm 6 Sorting Mapper

```

1: procedure MAP(Key k, Node n)
2:    $title \leftarrow n.title$ 
3:    $pagerank \leftarrow n.pagerank$ 
4:   EMIT(pagerank, title)

```

Algorithm 7 Sorting Reducer

```
1: procedure REDUCE(Pagerank rank, Titles  $[t_1, t_2, \dots]$ )
2:   for all title in titles do
3:     EMIT(title, rank)
```

2.3 Implementations Details

In these following chapters is briefly described how we have implemented the pseudocode. After that, the performance evaluation of the computation is reported.

2.3.1 First Phase: Graph construction

Parsing is a job that doesn't need to create a conglomerate value, but it creates records which, each of them, is treated as a standalone piece of information. We decide to use three reducers since we have three working nodes.

Initially we had thought of inserting an initial step dedicated exclusively to the node count, but later we came up with a better solution, complicating the parsing phase. In the section on performance the differences between the two approaches will be shown, showing the improvement obtained.

First we consider the Mapper, in which each record is parsed, looking for the title and all the outgoing links. So for each record we are going to emit a pair (**title, outgoing edges**). In addition to this we must also count the number of nodes, to do this we use an **In-Mapper Combiner**, a global counter inside the Mapper. It will be initialized to 0 in setup, and incremented at each execution of the map function; during the cleanup we will transmit a single cumulative value for that Mapper, transmitting the pair (**""**, **N**), with N the value of the counter. As a key we use the empty string **""**, which cannot belong to any title and which is always placed first in lexicographic order. Since we send two different information between Mapper and Reducer, we have decided to implement our own **partitioner**, to always send partial counters to Reducer 0 and therefore be sure to have this value in the file part-r-00000. All other keys will be split into the other reducers, using an hash function, like the default behavior. As for the reducer, it takes care of adding the partial values of the counters to obtain N, the number of nodes; in addition it will get the title and outgoing links values of each record. The initial page rank value is set to -1, and will be calculated at the first iteration of the next step. All this information will be saved in the filesystem, to be processed in the next step. The next step will take in input all the output file containing the Node information. All this process has been schematized in 1

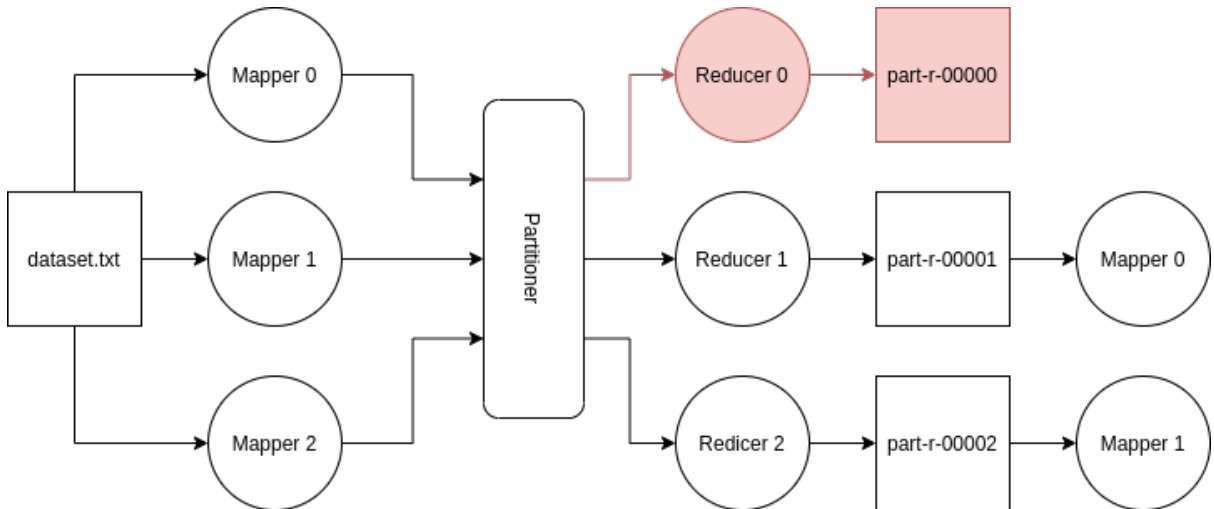


Figure 1: Parsing phase

2.3.2 Second Phase: PageRank Estimation

Since the previous phase constructs multiple outputs, this phase has to take multiple inputs. As in the *Graph construction* phase multiple reducers are used for accelerating the computation. Thus, from the first iteration to the last one this MapReduce computation takes multiple inputs and produces multiple outputs.

Since we have to output Node-type values from the mapper, the mass must also be transmitted through this data structure. Since the Pagerank value for a certain node depends only on the amount of mass entering the node and not on the value at the previous iteration, we thought to exploit the fact that the pagerank field is still transmitted, and when we are transferring the graph structure we put -1 (impossible value). This allows us to discriminate the situation in which the graph structure is transmitted, in an optimized way (without the need to insert an additional field). When we send the mass instead we put the value to be transmitted as the pagerank value and an empty list as the list of outgoing edges.

2.3.3 Third Phase: Sorting

Since the previous phase constructs multiple outputs, also this phase has to take multiple inputs. Here a single reducer is used in order to take advantage of the automatic sorting done over the keys by Hadoop. Keys are in this case the pagerank value of each page, and are passed by the mapper to the reducer as **DoubleWritable** objects. We have implemented a **WritableComparator** for getting the descending order.

2.4 Performance

For measuring the performance we considered three files with a different number of nodes:

- **wiki-micro.txt**: number of nodes 2427. Most of the nodes points to sites that are not present in the initial set.
- **dataset5.txt**: number of nodes 5000. This dataset is synthetic dataset created using the same structure of *wiki-micro.txt*. Each node has a random value of outgoing edges between 0 and 10.
- **dataset10.txt**: number of nodes 10000. This dataset is synthetic dataset created using the same structure of *wiki-micro.txt*. Each node has a random value of outgoing edges between 0 and 10.

and we test them over different numbers of iteration: 5, 10, 15.

In order to prove the performance gain in *In-Mapper Combiner* instead of counting the nodes in a MapReduce, we provide the result of both the approaches. Those results are computed as the mean of 5 iterations.

Table 1: Performance with MapReduce Node Counter

File	N Iter 5	N Iter 10	N Iter 15
wiki-micro.txt	184417	314290	436651
dataset5.txt	186320	310445	467191
dataset10.txt	206945	339101	455945

Table 2: Performance with counting in Parsing

File	N Iter 5	N Iter 10	N Iter 15
wiki-micro.txt	164519	295216	419528
dataset5.txt	168450	331793	430656
dataset10.txt	172402	294239	437840

The values in the **tables** are expressed in milliseconds.