TEAM DATA

Abhishek Raval
Manikantha Dronamraju

OBJECTIVE

- Using the Bibliographic Data Set to generate interesting observations
- Pre-Process nodes and edges to generate data of the form
  [PageId | PaperName | AdjList Of Citations | AdjListSize | Number of times this page is cited | List of Authors(authorId:authorName:totalPublicationsOfAuthor) | Conference(conferenceId:conferenceName:conferencePublicationCount) | List of Terms(termId:keyword)]
  From the data in following format:
  0 \tab term \tab systems \tab 0
  75568 \tab paper \tab High-speed three dimensional laser sensor \tab 2000
  1323803 \tab author \tab Phillip S. Yu \tab 550
  1318800 \tab conf \tab IEEE Transactions on Information Theory \tab 7346,
  connected by edges 1004048 \tab 568
- Find the paper with Highest PageRank ? Is it one with most citations?
- Predict the keywords of any paper, using K Nearest Neighbour Algorithm

IS PRE-PROCESSING EASY?

In the class we talked a lot about pre-processing, but does that seem a job to just process and reformat a data as per our needs?

But what happens if all the nodes contain multi-attribute values with each node having it's own id, and the only solution is lookup from edges.txt

Let's find out more about that.

WHAT DID IT TAKE TO PRE-PROCESS DATA?

- 4 Jobs with each doing MapSide Joins(For Joining paper, with other papers, authors, conference, keywords)
- 1 MR Jobs for computing Citations count for paper
- And 1 Final MR job for merging all the attributes

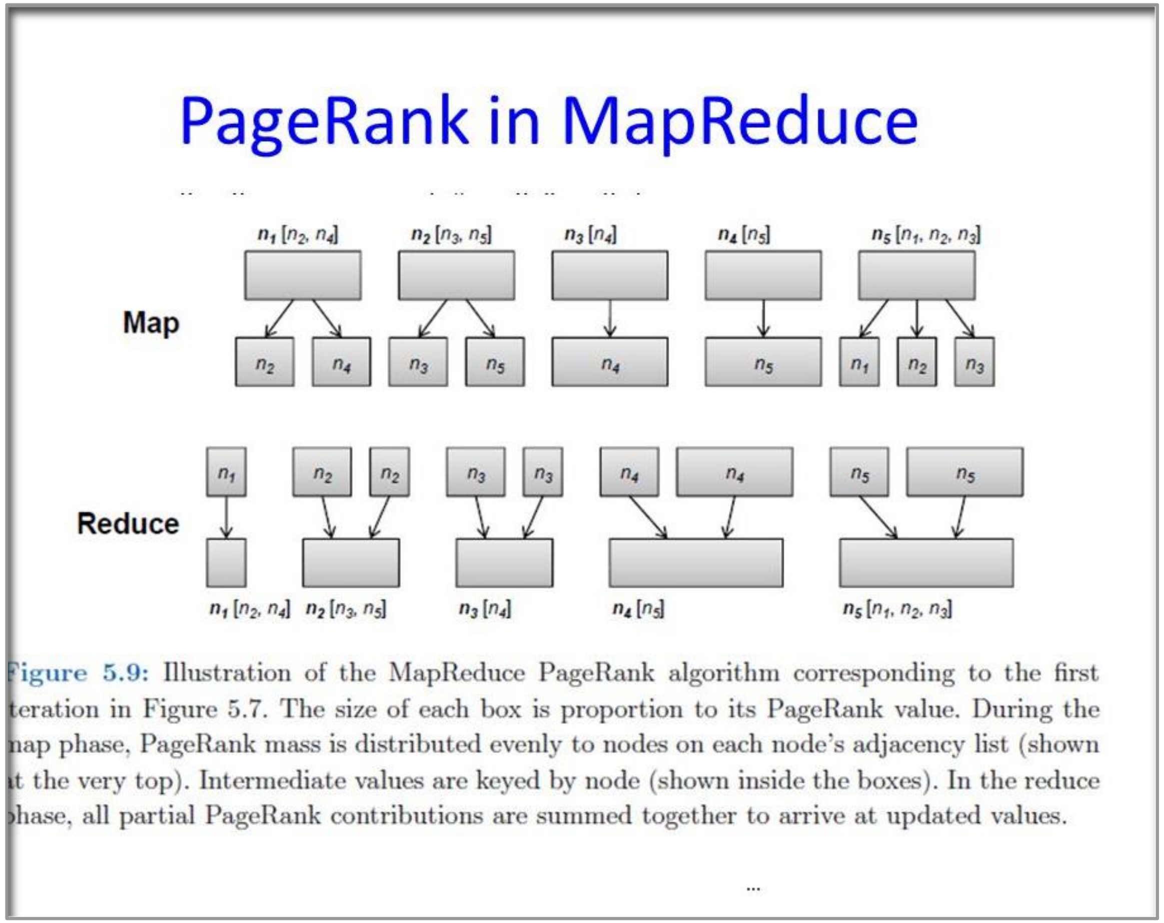
PAGE RANK

Recursive Formula:

- Each Page gets “1/Number of nodes” initial page rank
- The Page’s rank is calculated as the sum of the incoming link based on the formula PR(n) = (1-d)/N + d(∑PR(Ni)/C(Ni))
- PR(n) is the Page Rank of Node n
- PR(Ni) is the Page Rank of Ni
- C(Ni) is the number of nodes in the adjacency list
- N is the number of Pages
- d is the damping factor generally set to 0.85.

EXCEUTION

- Internal Execution



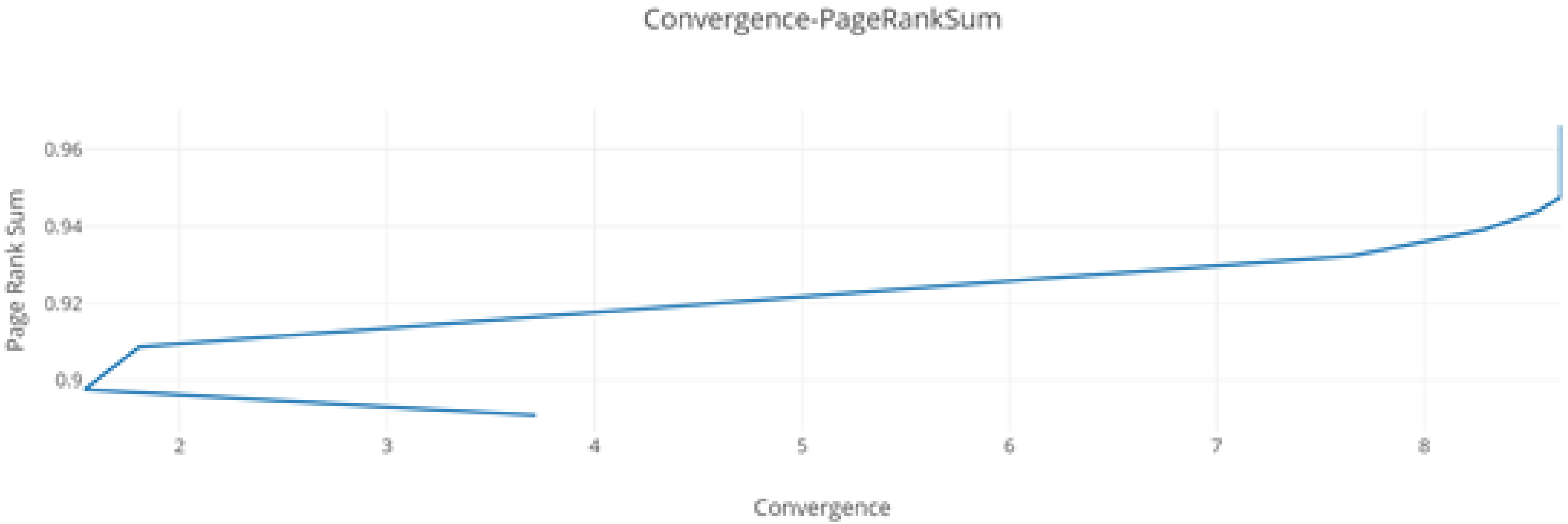
PageRank in MapReduce

Figure 5.9: Illustration of the MapReduce PageRank algorithm corresponding to the first teration in Figure 5.7. The size of each box is proportion to its PageRank value. During the nap phase, PageRank mass is distributed evenly to nodes on each node's adjacency list (shown t the very top). Intermediate values are keyed by node (shown inside the boxes). In the reduce hase, all partial PageRank contributions are summed together to arrive at updated values.

...

DANGLING NODE

- A few nodes have no outbound link and are known as Dangling Node
- Since, there is no outgoing link the PR mass accumulates at that node and is lost at every iteration.
- To handle that and distribute the mass equally among all nodes we modify the formula as below:
- $PR(n) = (1-d)/N + d(\sum PR(Ni)/C(Ni) + \S/N)$
- Where "§" is the Dangling Mass
- To solve this in one MR job and avoid the unnecessary overhead we used a Global counter to accumulate the mass at every reducer.


Convergence-PageRankSum
Page Rank Sum
0.96
0.94
0.92
0.9
2
3
4
5
6
7
8
Convergence

EXECUTION TIME

Global Counter

Iterations	Machine	Execution Time(Min)	
20	1	23	
20	5	22	
20	10	23	

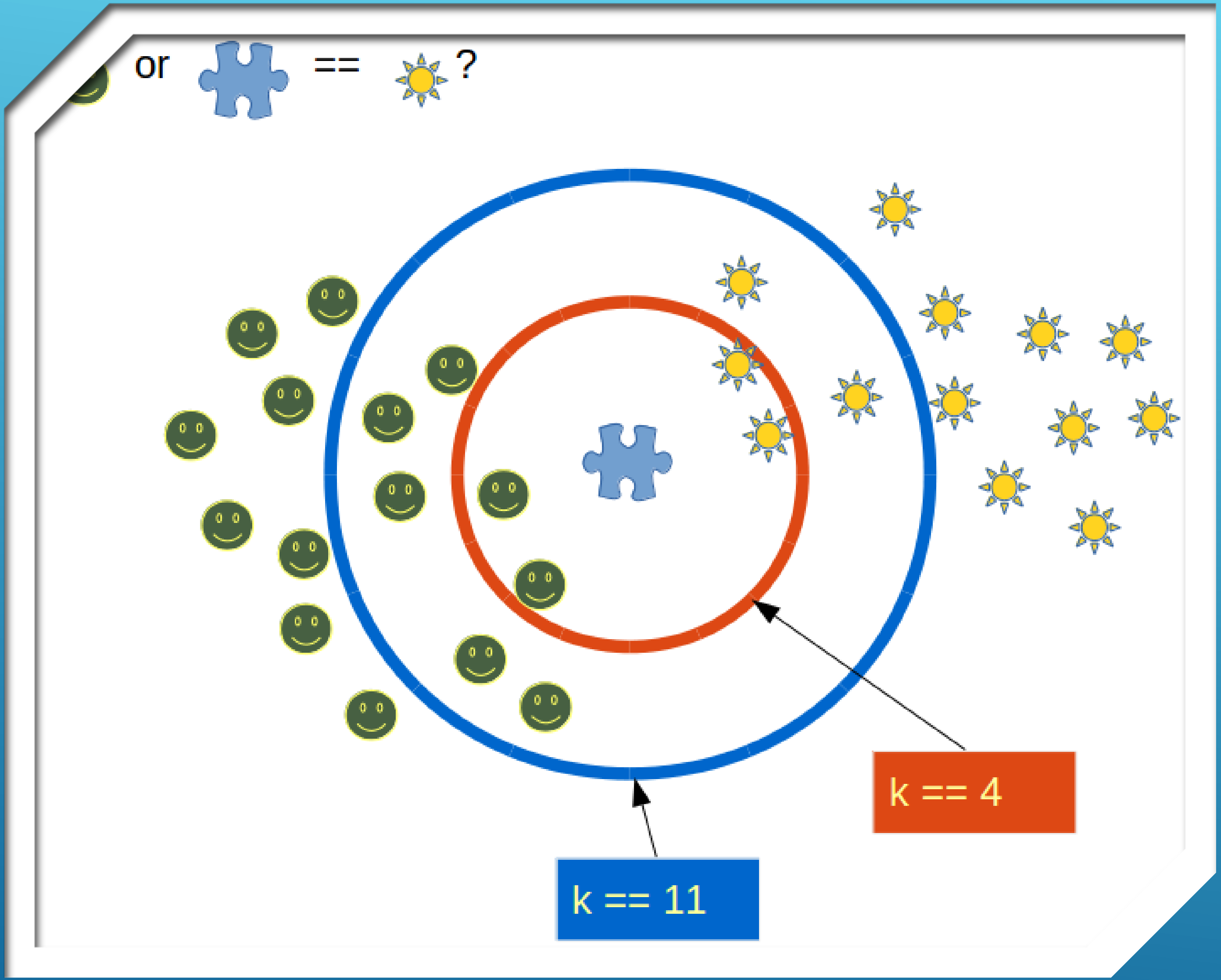
MR Job

Iteration	Machine	Execution Time	
20	1	37	
20	5	36	
20	10	37	

K NEAREST NEIGHBOUR

Recursively Iterate through each testcase in testDataSet:

- Compute distance metrics between testPaper and each node from our dataset and emit K nodes at each mapper task of format(null,(distance, Keyword))
- Distance metrics can be computed using Euclidean distance, Manhattan distance, etc. Since we were dealing with Strings, we used Edit Distance.
- We Considered AuthorName, ConferenceName, PaperName, AdjacencyList of Paper, Keywords, to build our metrics.
- On the basis of k, the most frequent keyword, would be the majority and output, as it will be the closest to the test dataset.



EXCEUTION

Internal Execution

EXECUTION TIME

#Test Cases	Machine	Execution Time	
15	1	35	
15	5	23	
15	10	11	

- Time taken on Aws

CONCLUSION

- The paper with highest Page Rank is different then the paper that has been cited the most.
- The most cited paper is not distributed across all nodes that are in the data set
- Where as the Paper with Max Rank is distributed across many adjacency lists.
- On the basis of selected value of K, the predictions for KNN deviates a lot.
- Selecting the value of K, can tune the prediction for KNN significantly.
- Pre-processing could be a challenging task, handle enough time for it, or else be very careful when selecting a dataset.