Page Rank

Haoyu Jia (hj342)
Ke Wang (kw427)
Yan Huang (yh553)

Data Preprocessing

Single Node

For each node, we preprocess the data in the following format:

<from_node, pagerank, to_node >

Blocked

For each node, we preprocess the data in the following format:

<from_node+from_block, pagerank, to_node+to_block>

Simple Computation (node by node)

5 MapReduce Passes residual data:

Pass 0: 2.3387482810589146
Pass 1: 0.3228494055864454
Pass 2: 0.1920276605957123 8
Pass 3: 0.0940220933555156 7
Pass 4: 0.0627171763933277 9

Far from converged!!!

Blocked Computation of PageRank

- **Map Input/Reducer Output:**

<u_id+block_id "\t" PR "\t" v0_id+block_id0,v1_id1+block_id1,.....>

- **Map Output/Reducer Input:**

1. Current Rank Value: *<Block_id : ! u_id "\t" PR>*
2. Links: *<Block_id : | u_id "\t" v0_id0+block_id0,v1_id1+block_id1 , …….>*
3. *<Block_id0 : v0_id "\t" PR "\t" from_u_id+Block_id_i "\t" degree_of_u >*

   <Block_id1 : v1_id "\t" PR "\t" from_u_id+Block_id_i "\t" degree_of_u >

* What if node u has no out-links ("Sink node")?

  Add an edge to itself with degree "0" in Map.

  Discard the tuple with "0" degree in Reduce.

Blocked Computation (Cont.)

- **In-block residual:**

  block convergence
- **Global average residual:**

  global convergence

Implemented by Hadoop Counter

Blocked computation result

- Avg. # of iterations per block performed by Reducer:
  - *Pass 0: 17*
  - *Pass 1: 7*
  - *Pass 2: 6*
  - *Pass 3: 4*
  - *Pass 4: 3*
  - *Pass 5: 1*
- Global Average Residual Error for each pass:
  - *Pass 0: 2.815*
  - *Pass 1: 0.03818*
  - *Pass 2: 0.02395*
  - *Pass 3: 0.009886*
  - *Pass 4: 0.003847*
  - *Pass 5: 0.0009584*

Blocked computation result (Cont.)

- highest-numbered nodes id in each block

Highest Rank Node in Block 0 is 8354
Highest Rank Node in Block 1 is 16563
Highest Rank Node in Block 2 is 25567
Highest Rank Node in Block 3 is 33067
Highest Rank Node in Block 4 is 40655
Highest Rank Node in Block 5 is 50463
Highest Rank Node in Block 6 is 60843
Highest Rank Node in Block 7 is 70647
Highest Rank Node in Block 8 is 80124
Highest Rank Node in Block 9 is 94237
Highest Rank Node in Block 10 is 100696
Highest Rank Node in Block 11 is 110643
Highest Rank Node in Block 12 is 129662
Highest Rank Node in Block 13 is 139525
Highest Rank Node in Block 14 is 140574
Highest Rank Node in Block 15 is 161328
Highest Rank Node in Block 16 is 167683
Highest Rank Node in Block 17 is 174682
Highest Rank Node in Block 18 is 184001
Highest Rank Node in Block 19 is 192982
Highest Rank Node in Block 20 is 211611

Blocked computation result (Cont.)

- highest-numbered nodes id in each block

Highest Rank Node in Block 21 is 222760
Highest Rank Node in Block 22 is 232579
Highest Rank Node in Block 23 is 237032
Highest Rank Node in Block 24 is 245877
Highest Rank Node in Block 25 is 258804
Highest Rank Node in Block 26 is 265974
Highest Rank Node in Block 27 is 280153
Highest Rank Node in Block 28 is 289833
Highest Rank Node in Block 29 is 297951
Highest Rank Node in Block 30 is 309302
Highest Rank Node in Block 31 is 319654
Highest Rank Node in Block 32 is 323550
Highest Rank Node in Block 33 is 343322
Highest Rank Node in Block 34 is 345482
Highest Rank Node in Block 35 is 355915
Highest Rank Node in Block 36 is 370257
Highest Rank Node in Block 37 is 374642
Highest Rank Node in Block 38 is 390739
Highest Rank Node in Block 39 is 396871
Highest Rank Node in Block 40 is 406300

Blocked computation result (Cont.)

- highest-numbered nodes id in each block

Highest Rank Node in Block 41 is 418216
Highest Rank Node in Block 42 is 431942
Highest Rank Node in Block 43 is 437330
Highest Rank Node in Block 44 is 446565
Highest Rank Node in Block 45 is 462310
Highest Rank Node in Block 46 is 466044
Highest Rank Node in Block 47 is 481196
Highest Rank Node in Block 48 is 490478
Highest Rank Node in Block 49 is 499366
Highest Rank Node in Block 50 is 512248
Highest Rank Node in Block 51 is 514131
Highest Rank Node in Block 52 is 524510
Highest Rank Node in Block 53 is 534709
Highest Rank Node in Block 54 is 545088
Highest Rank Node in Block 55 is 555467
Highest Rank Node in Block 56 is 574139
Highest Rank Node in Block 57 is 586313
Highest Rank Node in Block 58 is 589179
Highest Rank Node in Block 59 is 605111
Highest Rank Node in Block 60 is 610392

Blocked computation result (Cont.)

- highest-numbered nodes id in each block

Highest Rank Node in Block 61 is 625356
Highest Rank Node in Block 62 is 633930
Highest Rank Node in Block 63 is 640499
Highest Rank Node in Block 64 is 651680
Highest Rank Node in Block 65 is 657785
Highest Rank Node in Block 66 is 674796
Highest Rank Node in Block 67 is 678618

Extra credit

Jacobi *vs* Gauss-Seidel

- **Jacobi**
  - w_k+1 = (1-d)*Z + d*B*w_k
  - compute new PageRank values into temporary variables
  - write over the old values at the end of each pass
- **Gauss-Seidel**
  - using the new values as soon as they are available
  - can improve convergence rate

Random Block Partition

- **random hash function:**
  - **(id<<1)|(id+rand)%68**

Gauss-Seidel Output

- Avg. # of iterations per block performed by Reducer:
  - *Pass 0: 12*
  - *Pass 1: 6*
  - *Pass 2: 5*
  - *Pass 3: 3*
  - *Pass 4: 2*
  - *Pass 5: 1*
- Global Average Residual Error for each pass:
  - *Pass 0: 3.161*
  - *Pass 1: 0.03867*
  - *Pass 2: 0.02436*
  - *Pass 3: 0.008956*
  - *Pass 4: 0.003919*
  - *Pass 5: 0.0009203*

Converges faster than Jacobi within the block

Random Block Output

6 MapReduce Pass residual data:

Pass 0: 2.341
Pass 1: 0.3219
Pass 2: 0.1901
Pass 3: 0.09266
Pass 4: 0.06124
Pass 5: 0.03294

21 iterations to converge!

Thank You !