Report

Class Number: CS6240

HW Number: 5

Name: Gautam Vashisht

Design Discussion

Please find the algorithm design that I followed for each step –

Firstly, one Indexing file is created for each node with its index. A separate MR job has developed for it which will take the input as Adjacency List Graph and will write the index for each node.

Step1: XML file will be parsed as Adjacency List Graph and indexing file is created for it. Then, other MR job will create the M matrix, it will use the index file from distributed cache. It will write non-zero values only in file.

Mapper will emit –

For each edge, (edge, {indexOfParent, share} and (edge, {index})

Then one emit for node, (node, index).

Extra 2 emits per map call is needed for including dangling nodes in M matrix.

Then Reducer will emit – M, rowNum, colNum, value where rowNum will be node's index and colNum will be index of parent of node.

Step 3:

For multiplication part, there are two design algorithms used.

1. Row By Column
    - For row by column, R matrix is transferred to each reducer task as Distributed cache file and then M matrix is taken as input file. Mapper will emit (rowNum, record) for each record in mapper.
    - Reducer will receive one complete row from mapper and then from distributed cache, it will take the particular column value of R and sum the each multiplication of row value and col value. Finally, for each reducer it will emit (rowNum, value).
2. Column By Row
    - For column by Row, Mapper will emit (colNum, {M-record}) for M matrix while (rowNum, {R –record}) for R matrix.

- Reducer will receive the record for same column of M and same row of R. Two lists will be created – one for M and other for R. Each value of m record is multiplied with record of R and is emitted with (rowNum, colNum, value).
- Finally, one extra MR job is needed to sum up all the values for same (rowNum, colNum) in previous job.

For both the approaches, after this dangling share is calculated and added to each value in intermediate R matrix created as discussed below.

Step 4: To find Top100 records, same TreeMap approach is used to emit Top 100 records from local mapper and then finding the final Top 100 in one reducer task. In reducer, program needs to load indexing file in distributed cache to find the exact pageNames from the rowNum of R file.

As Matrix was sparse, therefore while creating the matrix M, only the records with non zero values are added to M. For both the version A and B, as M values and R values are non zero only, so program has to write all the computed values after multiplication.

For Row by Column, R matrix is loaded to Distributed Cache while for Column by Row both the matrices are read as input files.

Handling Dangling Nodes

I have developed additional jobs to compute and add dangling share as discussed below -

-To add dangling share, program is creating D matrix containing all the dangling node and their share.

-For this, mapper will read the Adjacency List Graph and will emit the nodes with no edges. Reducer will emit (node, 1/count) where 1/count will be the share of each dangling Node.

-Then there is other MR job to multiply the D matrix with R(t) and emit the dangling share value of each dangling node.

-One other MR job will run to calculate the dangling share for each dangling node after the end of iteration.

- One other job will sum up the dangling share to add to intermediate R matrix created after M * R(t)

-Then, finally one Last job will add the danglingSum to each value of matrix to create output matrix R(t+1)

Performance Comparison

	6 m4.large Machines	11 m4.large Machine
Time to Complete step 1 & 2	44 minutes	29 minutes
For Step 3 per iteration(Row by Column)	6 minutes 30 seconds	5 minutes 30 seconds
For Step 3 per iteration(Column by Row)	8 minutes	5 minutes 53 seconds
For Step 4	2 minutes	2 minutes 20 second
Total Time(Row by Col)	1 hour 52 minutes	1 hour 36 minutes
Total Time(Col by Row)	2 hour 35 minutes	1 hour 47 minutes

Step1 and 2 includes starting from reading the input xml file.

For Adjacency List, With 6 slave nodes it has taken 3 minutes per iteration while for 11 machines with 2 iteration per minutes which is very less compared to Matrix Multiplication. Matrix Multiplication has used few extra jobs to compute Dangling share and summing up the values in case of column by row. So more number of jobs will definitely have more transfer of data and will take more time. From log file, it is evident that more data is transferred for creating matrix multiplication as compared to normal implementations which caused it to take more time. Also, few of the jobs needed to run on single reducer(index file creation and summing the dangling share), so that also increases the time.

Top 100 web pages name looks similar to the ones for Hadoop implementation. Only the difference is with the values not the pages. This difference might be because of precision across 10 iterations as in this algorithm I used float instead of double.

Top 100 Full Data Set Web Pages (Row by Column)

United_States_09d4	3.6723968E-3
2006	3.2621485E-3
United_Kingdom_5ad7	1.7419201E-3
2005	1.5076274E-3
Biography	1.22372E-3
Canada	1.1435737E-3
England	1.1352980000000001E-3
France	1.1156673E-3
2004	1.0511532E-3
Germany	9.6095656E-4
Australia	9.3497883E-4

Geographic_coordinate_system	9.220969999999999E-4
2003	8.4673986E-4
India	8.2459557E-4
Japan	8.147197E-4
2001	6.809051999999999E-4
Italy	6.7981569999999995E-4
2002	6.722869E-4
Internet_Movie_Database_7ea7	6.714009E-4
Europe	6.4660922999999994E-4
2000	6.371207E-4
World_War_II_d045	6.115853E-4
London	5.9149286E-4
Population_density	5.7386805E-4
Record_label	5.720874E-4
1999	5.626371E-4
Spain	5.5657915000000005E-4
English_language	5.537749999999999E-4
Russia	5.2411685E-4
Race_(United_States_Census)_a07d	5.2304915E-4
Wiktionary	5.1199709999999995E-4
1998	4.8655111E-4
Wikimedia_Commons_7b57	4.850377E-4
Music_genre	4.821091E-4
1997	4.6377716E-4
Scotland	4.569271E-4
New_York_City_1428	4.5492552E-4
Football_(soccer)	4.5070122E-4
1996	4.3521103E-4
Television	4.309033E-4

Sweden	4.297812E-4
Census	4.1572732E-4
Square_mile	4.1476623E-4
1995	4.0970737000000006E-4
California	4.0782863E-4
China	4.0017607E-4
New_Zealand_2311	3.947208E-4
Netherlands	3.9372596E-4
1994	3.9107204000000005E-4
1991	3.7248177E-4
1993	3.6987027999999997E-4
1990	3.6716825E-4
Public_domain	3.6614242999999997E-4
New_York_3da4	3.660734E-4
1992	3.5432847999999997E-4
Film	3.5378232E-4
United_States_Census_Bureau_2c85	3.5372187E-4
Scientific_classification	3.5311884E-4
Actor	3.52859E-4
Norway	3.4675133E-4
Ireland	3.4390073E-4
Population	3.4252706E-4
Poland	3.4112756999999997E-4
1989	3.3194935999999996E-4
Marriage	3.2551089999999996E-4
1980	3.239511E-4
Politician	3.2310714999999997E-4
Brazil	3.22859E-4
January_1	3.2076042E-4

Mexico	3.2008678E-4
Latin	3.160693E-4
1986	3.1525205E-4
Album	3.0820136E-4
1985	3.077184E-4
Per_capita_income	3.0740660000000003E-4
1979	3.0687249999999997E-4
1982	3.0649907E-4
1981	3.0606394E-4
French_language	3.0335710000000002E-4
1974	3.0274055E-4
Record_producer	3.026842E-4
1984	3.006873E-4
Switzerland	3.0059277E-4
1987	3.005718E-4
South_Africa_1287	3.00557E-4
1983	3.0033145000000002E-4
1970	2.945359E-4
1988	2.9368396E-4
Km²	2.9242493E-4
1976	2.9133529999999998E-4
1975	2.8793059999999996E-4
Personal_name	2.872155E-4
Paris	2.8426325E-4
1969	2.8370292E-4
Greece	2.8348582999999997E-4
1972	2.8210559999999997E-4
1945	2.8147411999999997E-4
Poverty_line	2.8116156999999997E-4

1977	2.8008205E-4
1978	2.7892273E-4

Top 100 Full Data Set(Column by Row)

United_States_09d4	0.0036724447
2006	0.003262324
United_Kingdom_5ad7	0.0017418963
2005	0.0015076791
Biography	0.0012237031
Canada	0.0011435756
England	0.0011353085
France	0.0011156794
2004	0.0010511328
Germany	9.6096157E-4
Australia	9.3494647E-4
Geographic_coordinate_system	9.2208944E-4
2003	8.4676384E-4
India	8.245889E-4
Japan	8.147172E-4
2001	6.8097666E-4
Italy	6.798111E-4
2002	6.7228347E-4
Internet_Movie_Database_7ea7	6.7140075E-4
Europe	6.466128E-4
2000	6.371494E-4
World_War_II_d045	6.1158056E-4
London	5.914917E-4
Population_density	5.7387084E-4
Record_label	5.7208654E-4

1999	5.6263356E-4
Spain	5.5657414E-4
English_language	5.537761E-4
Russia	5.2411563E-4
Race_(United_States_Census)_a07d	5.2305224E-4
Wiktionary	5.119971E-4
1998	4.8654774E-4
Wikimedia_Commons_7b57	4.8503978E-4
Music_genre	4.821081E-4
1997	4.6377446E-4
Scotland	4.569213E-4
New_York_City_1428	4.5492715E-4
Football_(soccer)	4.5069912E-4
1996	4.352116E-4
Television	4.309048E-4
Sweden	4.2978296E-4
Census	4.1572284E-4
Square_mile	4.147661E-4
1995	4.0970635E-4
California	4.078301E-4
China	4.0017616E-4
New_Zealand_2311	3.9471904E-4
Netherlands	3.937263E-4
1994	3.91072E-4
1991	3.7249035E-4
1993	3.69869E-4
1990	3.6716764E-4
Public_domain	3.6614257E-4
New_York_3da4	3.6607505E-4

1992	3.5432848E-4
Film	3.5378433E-4
United_States_Census_Bureau_2c85	3.537207E-4
Scientific_classification	3.531164E-4
Actor	3.528588E-4
Norway	3.4675066E-4
Ireland	3.4390198E-4
Population	3.4252E-4
Poland	3.4112585E-4
1989	3.3195203E-4
Marriage	3.2551127E-4
1980	3.2394944E-4
Politician	3.2310752E-4
Brazil	3.2285874E-4
January_1	3.207609E-4
Mexico	3.2008666E-4
Latin	3.1606812E-4
1986	3.1525377E-4
Album	3.0820075E-4
1985	3.0771748E-4
Per_capita_income	3.0740563E-4
1979	3.0687317E-4
1982	3.06499E-4
1981	3.0606447E-4
French_language	3.033569E-4
1974	3.0273758E-4
Record_producer	3.026827E-4
1984	3.006856E-4
Switzerland	3.0059164E-4

1987	3.0057217E-4
South_Africa_1287	3.0055828E-4
1983	3.0033028E-4
1970	2.9453667E-4
1988	2.9368504E-4
Km^2	2.9242612E-4
1976	2.913384E-4
1975	2.879443E-4
Personal_name	2.8721627E-4
Paris	2.842636E-4
1969	2.837027E-4
Greece	2.8348778E-4
1972	2.821043E-4
1945	2.8147068E-4
Poverty_line	2.8116084E-4
1977	2.8008505E-4
1978	2.7892445E-4

Top 100 Simple Data Set Web Pages

United_States_09d4	0.007004724
Wikimedia_Commons_7b57	0.0053205537
Country	0.0043243985
England	0.0030080536
United_Kingdom_5ad7	0.002911881
Europe	0.002902746
Water	0.002876313
Germany	0.002848218

France	0.002812247
Animal	0.0027442265
Earth	0.0026929332
City	0.0026388497
Week	0.0022162993
Asia	0.0021386356
Wiktionary	0.0020816354
Sunday	0.0020711091
Money	0.0020483334
Monday	0.0020413643
Plant	0.0020233123
Wednesday	0.002021296
Friday	0.001972133
Computer	0.001970985
English_language	0.0019506958
Saturday	0.0019502925
Thursday	0.0019252702
Italy	0.0019169776
Tuesday	0.0019114527
India	0.0019012514
Government	0.0018991157
Number	0.0017673097
Spain	0.0017460216
Japan	0.001696925
Canada	0.001674349
Day	0.0016240788
People	0.001612921
Human	0.001583667
Wikimedia_Foundation_83d9	0.0015328177

Australia	0.0015287233
China	0.0015265307
Energy	0.0014814109
Food	0.001472588
Science	0.001439007
Sun	0.0014361074
Mathematics	0.0014261123
Television	0.0013753652
index	0.0013667785
Capital_(city)	0.0013280393
Russia	0.0013202704
Music	0.0012995539
State	0.0012971893
Year	0.001264858
Greece	0.0012440805
Language	0.00123903
Scotland	0.0012376449
Metal	0.0012090089
Wikipedia	0.0011958167
Greek_language	0.0011833921
2004	0.0011828787
Sound	0.0011457987
London	0.0011451207
Religion	0.0011432221
Planet	0.0011431385
Africa	0.0011056933
20th_century	0.0010689863
Law	0.0010596168
Geography	0.0010499997

19th_century	0.001048407
Liquid	0.0010442017
Poland	0.0010318977
World	0.0010311649
Scientist	0.0010184827
Society	0.0010137677
Latin	9.802574E-4
History	9.766594E-4
Atom	9.757855E-4
Sweden	9.708156E-4
War	9.702359E-4
Light	9.6344197E-4
Netherlands	9.610791E-4
Culture	9.464821E-4
Building	9.3831087E-4
God	9.213127E-4
Turkey	9.1535744E-4
Plural	9.0925203E-4
Information	9.077096E-4
Centuries	9.0210635E-4
Inhabitant	8.906394E-4
Chemical_element	8.8514E-4
Portugal	8.828401E-4
Capital_city	8.6936064E-4
Denmark	8.6786353E-4
Austria	8.610523E-4
Species	8.484964E-4
Book	8.4703823E-4
Disease	8.437965E-4

Cyprus	8.4352755E-4
Ocean	8.422297E-4
University	8.4169256E-4
North_America_e7c4	8.411212E-4
Biology	8.360953E-4