Sravani Beeram

CS6240 Parallel Data Processing using Mapreduce
Homework 3

Design Discussion

Pseudo-code:

```
map(key,values)
{
        delta = value retrieved from counter
        recordCount = value retrieved from counter
        alpha = 0.15
        for each value in values
         Split value to get pageName,linkPages,pageRank
        //setting initial value of pageRank after preprocessing to -0.0
        If first iteration of pageRank
          pageRank = 1 / recordCount
        else
          pageRank = split value from above

        If delta != -0.0
         pageRank += (1-alpha) * (delta/recordCount)

        If linkPages == empty
          emit("dummy",pageRank)
        else
          emit("linkPages",pageRank/recordCount)

        obj = linkPages,pageRank
        emit(pageName,obj)
}

reduce(key,values)
{
        alpha = 0.15
        recordCount = value from counter
        If key == "dummy"
         for each value in values
```

```
            delta += pageRank
            Set delta value in counter
        else
            for each value in values
              Split value to get pageName,linkPages,pageRank
              pageRank = (alpha/recordCount) + (1-alpha) * s

        Obj = linkPages + pageRAnk
        emit(key,obj)
}
```

Top - k:

```
map(key,values){

TreeMap<Integer, Text> repToRecordMap

Split the value into pageName, pageRank
repToRecordMap.put(pageRank,key+pageRank)
If repToRecordMap >  k
  remove top key
}

Cleanup
{
  For each value in repToRecordMap.values
    emit("dummy",value)
  }

reducer(key,values)
{
   For each value in values
       Split value to get pageName, pageRank

        If repToRecordMap >  k
          remove top key
     For value in repToRecordMap.descending().values
          emit(null,value)
 }
```

I followed preprocessing steps discussed in the document and also removed duplicate self pages and duplicate link pages. In pagerank algorithm,three approaches were mentioned to calculate delta.
I used solution 2 to calculate page rank.In this we merge the computation of delta into the previous reduce and use it in next map phase.Compared to solution 1 which needs extra 10 job runs and solution 3 which sends pagerank of dangling nodes to all reducers, solution 2 does the job best.

Data Transfer:

6 m4.large machines (1 master and 5 workers)

Iteration	**Mappers to Reducers(bytes)**	**Reducers to S3(bytes)**
Pre-processing	1115433596	1130216010
PageRank - run 1	1532022295	1181464806
PageRank - run 2	1999094537	1181462889
PageRank - run 3	2000576568	1181460700
PageRank - run 4	2000687109	1181460948
PageRank - run 5	2000991015	1181452384
PageRank - run 6	2000836038	1181453240
PageRank - run 7	2000697657	1181446961
PageRank - run 8	2000374446	1181445998
PageRank - run 9	2000730618	1181445695
PageRank - run 10	2000315904	1181446677
Delta Add	774506139	1183205338
Top k	56957	3404

11 m4.large machines (1 master and 10 workers)

Iteration	**Mappers to Reducers(bytes)**	**Reducers to S3(bytes)**
Pre-processing	1139591934	1130216010
PageRank - run 1	1568557082	1181468297
PageRank - run 2	2044287876	1181465468
PageRank - run 3	2045506564	1181463944
PageRank - run 4	2045637202	1181457031
PageRank - run 5	2046032393	1181460310
PageRank - run 6	2045950217	1181453402
PageRank - run 7	2045776466	1181455481
PageRank - run 8	2045546725	1181457645
PageRank - run 9	2046396159	1181448114
PageRank - run 10	2046066616	1181447555
Delta Add	787478035	1183217358
Top k	61246	3408

Performance Comparison:

6 m4.large machines (1 master and 5 workers)

	Time (Minutes)
Pre-processing	36.47
To Run Ten iterations of PageRank	26.77
To find the top-100 pages	0.73

11 m4.large machines (1 master and 10 workers)

	Time (Minutes)
Pre-processing	24.38
To Run Ten iterations of PageRank	16.17
To find the top-100 pages	0.55

I expected that as number of worker machines increases time to run the program should reduce.By comparing the run times it confirms the same.Pre-processing phase showed a good speedup as the data is distributed over several machines in the second case.

Top-100 Wikipedia pages :

Full Datasets

United_States_09d4	{0.0026228819063981094}
2006	{0.0012284958960365547}
United_Kingdom_5ad7	{0.0012031348356531671}
Biography	{9.82092533370611E-4}
2005	{9.170679629211132E-4}
England	{8.802034155013219E-4}
Canada	{8.559101054800394E-4}
Geographic_coordinate_system	{7.716575264523922E-4}
France	{7.25022863297641E-4}
2004	{7.198826828067749E-4}
Australia	{6.804760970122956E-4}
Germany	{6.543564654161378E-4}
2003	{5.8740567597147E-4}
India	{5.834182294230754E-4}
Japan	{5.828440485752622E-4}
Internet_Movie_Database_7ea7	{5.335066455319993E-4}
Europe	{5.092715553063083E-4}
Record_label	{4.9145507619617E-4}
2001	{4.8701121553507924E-4}
2002	{4.828809380696294E-4}

World_War_II_d045	{4.7804790150092697E-4}
Population_density	{4.703359028901132E-4}
Music_genre	{4.671967104885918E-4}
2000	{4.6467310443859295E-4}
Italy	{4.4578721465219983E-4}
Wiktionary	{4.362100943060581E-4}
Wikimedia_Commons_7b57	{4.352922482287458E-4}
London	{4.347924762540636E-4}
English_language	{4.1850189877228874E-4}
1999	{4.059399256712982E-4}
Spain	{3.629363942550768E-4}
1998	{3.5631239461452315E-4}
Russia	{3.439051949368919E-4}
1997	{3.3728875883151194E-4}
Television	{3.3629725352112363E-4}
New_York_City_1428	{3.3462889097300583E-4}
Football_(soccer)	{3.2615292157846617E-4}
1996	{3.236279457176917E-4}
Census	{3.235548579660166E-4}
Scotland	{3.221925491786915E-4}
1995	{3.1015438331368807E-4}
China	{3.086429141770705E-4}
Population	{3.0429773026491243E-4}
Square_mile	{3.040553930706499E-4}

Scientific_classification	{3.0401163949627194E-4}
California	{3.0166593177180436E-4}
1994	{2.90692728678802E-4}
Sweden	{2.876214644399713E-4}
Public_domain	{2.8741327202268653E-4}
Film	{2.8626856055247244E-4}
Record_producer	{2.841117233658785E-4}
New_Zealand_2311	{2.83101486893104E-4}
New_York_3da4	{2.788864306286084E-4}
Netherlands	{2.766736487685457E-4}
Marriage	{2.7581360675260483E-4}
1993	{2.748036927109333E-4}
United_States_Census_Bureau_2c85	{2.746671842039841E-4}
1991	{2.7189182502329113E-4}
1990	{2.683260417786711E-4}
1992	{2.6636906929413843E-4}
Politician	{2.6489478420230244E-4}
Album	{2.60553575311202E-4}
Latin	{2.6045674608350314E-4}
Actor	{2.5833937831538586E-4}
Ireland	{2.5810380987814147E-4}
Per_capita_income	{2.556430009475822E-4}
Studio_album	{2.518566386208085E-4}

Poverty_line	{2.511652879592252E-4}
Km²	{2.495068485871728E-4}
1989	{2.4689427169380656E-4}
Norway	{2.4099176323767531E-4}
Website	{2.390120334044441E-4}
1980	{2.353218564669126E-4}
Animal	{2.2937871942032206E-4}
Area	{2.2919060581376874E-4}
1986	{2.270331499860329E-4}
Personal_name	{2.2626259154537277E-4}
Poland	{2.2613835093759316E-4}
Brazil	{2.2570374098478116E-4}
1985	{2.2402926455276594E-4}
1987	{2.233054189296557E-4}
1983	{2.2175654565569005E-4}
1982	{2.211080635996189E-4}
French_language	{2.193792430008506E-4}
1981	{2.1934742447232488E-4}
1979	{2.193286292881624E-4}
1984	{2.1879020471940756E-4}
World_War_I_9429	{2.186883187422462E-4}
1988	{2.1857679884743355E-4}
Paris	{2.1801988048787357E-4}
1974	{2.1797480004607495E-4}

Mexico	{2.1566802986057083E-4}
19th_century	{2.1185510632180566E-4}
1970	{2.1132389988537734E-4}
January_1	{2.108743910187393E-4}
USA_f75d	{2.107090789344143E-4}
1975	{2.0860204677863251E-4}
1976	{2.0846726271630202E-4}
Africa	{2.0780099007039016E-4}
South_Africa_1287	{2.0736101132623775E-4}

Simple Dataset:

United_States_09d4	{0.005189009000274023}
Wikimedia_Commons_7b57	{0.0048067664747098665}
Country	{0.00394028468771356}
England	{0.0027524814361112}
Water	{0.0026878096234471504}
Animal	{0.0025540875651497573}
City	{0.0025108240807830222}
United_Kingdom_5ad7	{0.0023586470936127644}
Germany	{0.002350401697711985}
Earth	{0.0023247348599551624

	}
France	{0.0023236079471425 98}
Europe	{0.0020380970371681943 }
Wiktionary	{0.0017538842142764 56}
English_language	{0.0017496771217548172 }
Government	{0.0017323446521036983 }
Computer	{0.0017168404847137419 }
India	{0.0017131709183852964 }
Money	{0.0016673836980231748 }
Japan	{0.0015516905685357748 }
Plant	{0.0015235595093602637 }
Italy	{0.0015074330904983294 }
Canada	{0.0014814073434532137 }
Spain	{0.0014711236922238 53}
Food	{0.0014246868489679735 }
Human	{0.0014120970062699572 }

China	{0.0013967150612732326}
People	{0.0013822485250560843}
Australia	{0.001329854240750792}
Asia	{0.0012844361711364016}
Capital_(city)	{0.0012742684212522298}
Television	{0.0012649972257606486}
Sun	{0.0012602100811782994}
Number	{0.0012432362289290998}
State	{0.0012403756814549102}
Sound	{0.0012352116672222234}
Science	{0.0012325431753597135}
Mathematics	{0.0012310566392958499}
Metal	{0.0011923046237497061}
Year	{0.0011770925835108738}
2004	{0.0011733573137687524}

Language	{0.0011501658848580064}
Russia	{0.0011461817792128412}
Wikipedia	{0.0011233302809884633}
Religion	{0.0010985666999662922}
19th_century	{0.0010965391417803404}
Music	{0.0010874313232146716}
Scotland	{0.001054800735006553}
20th_century	{0.0010537049832591231}
Greece	{0.0010492227329348604}
Latin	{0.0010298606131876836}
London	{0.0010273554428515458}
Greek_language	{0.0010043572566505261}
Energy	{9.990118103796353E-4}
World	{9.863508479979013E-4}
Centuries	{9.759058651368046E-4}
Culture	{9.452039652115214E-4}
History	{9.364696034256484E-4}

Liquid	{9.145230968002287E-4}
Netherlands	{9.057245076491691E-4}
Planet	{9.049322622392135E-4}
Light	{9.016763526865948E-4}
Society	{9.014920621454207E-4}
Atom	{8.900226406531586E-4}
Wikimedia_Foundation_83d9	{8.884400707763214E-4}
Scientist	{8.883836105736989E-4}
Image	{8.876884860222185E-4}
Law	{8.862908055986251E-4}
Geography	{8.788451614551062E-4}
List_of_decades	{8.785742942839089E-4}
Uniform_Resource_Locator_1b4e	{8.618845063634342E-4}
Africa	{8.605699671526473E-4}
Turkey	{8.448863678892073E-4}
Inhabitant	{8.304794882325051E-4}
Capital_city	{8.23048814043934E-4}
Plural	{8.215155955104306E-4}
Electricity	{8.137230016666796E-4}
Poland	{7.972379043155126E-4}
Building	{7.971238925722221E-4}
Car	{7.946540606240838E-4}
Sweden	{7.917125562342898E-4}

Book	{7.914884705321294E-4}
Biology	{7.869328964315903E-4}
War	{7.708172945482241E-4}
Chemical_element	{7.681607959198536E-4}
God	{7.609357218915552E-4}
North_America_e7c4	{7.562868644168604E-4}
September_7	{7.547781812642616E-4}
Website	{7.462973500605918E-4}
Nation	{7.426671526407808E-4}
Politics	{7.397103787590716E-4}
2006	{7.332900172260933E-4}
Fish	{7.322371112911321E-4}
Species	{7.308711176294926E-4}
Mammal	{7.216744135950775E-4}
Island	{7.178090203037447E-4}
Portugal	{7.171070596607482E-4}
Gas	{7.155515366540748E-4}
River	{7.115777513010685E-4}
Switzerland	{7.061075074386623E-4}
World_War_II_d045	{7.020304931583193E-4}