Assignment 06- Computing PageRank using Hadoop Report

Xiaozhi Yu 46921411

1. What is PageRank

Google describes PageRank:[1]

PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important".

Mathematically the web graph can be treated as a Markov chain. States are web pages (nodes), transition probabilities are decided by links. PageRank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page, which has a starting probability. If the Markov Chain is ergodic, the PageRank converges.

Pagerank is very important. Successful crawling depends on size of crawl and crawl policy[2] . crawl should based on some quality metrics derived from last crawl, among the four , PageRank is a very important crawl selection policy. PageRank is also useful for serving queries. In Google, Relevancy is determined by over 200 factors, one of which is the PageRank for a given page.

2. Difficulties in PageRank computing and problem solving

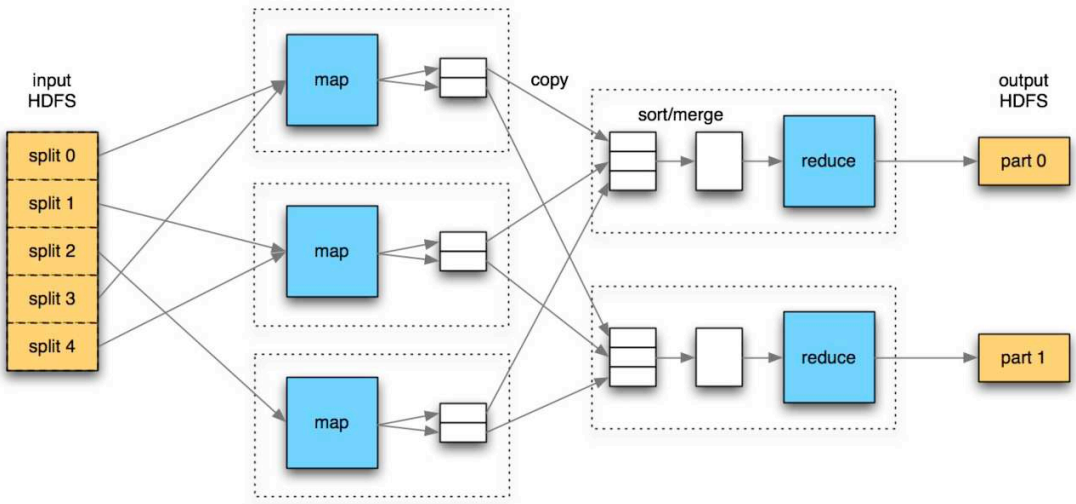
Computing wiki pages' PageRank has huge storage challenge. Starting from a 500000 URL-docID list, finally ends in 6,152,829 URLs. If we use uncompressed matrix to store the connection information, the storage used would be TB. Fortunately the matrix is very sparse, each wiki page has around 100 outgoing links only. So we should store the link graph in a compression way.

In my implementation method, I only store the outgoing URL's docID of a page. If a URL is not an outgoing link of the page, it will not show up a zero in the page's outgoing list. In this way, large storage is saved and high efficiency is attainted.

3. What is Hadoop MapReduce

Hadoop Map/Reduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.[3]

Processing sequence in Hadoop:



Five programming parts[4]:

InputFormat-Creates splits, one split is assigned to one mapper. A split is a collection of <K1, V1> pairs.

Mapper-Takes a <K1,V1> pair as input, produce <K2,V2> pairs.

Partitioner- takes a <K2,V2> pair as input, produces a bucket number as output.

Reducer- Takes a <K2,list<V2>> pair as input, produce <K3,V3> as output, output is sorted.

OutputFormat- Does something with the output(like write it to disk)

4. Compute Pagerank Using Hadoop

In My project, I ***start from*** Top 500000 wikipedia links. Using Hadoop to crawl those pages to get out going links from them. Here only English contain Wikipedia links count, other links are just discarded. This is implemented in *WikiLinkCrawler.java*. Output of this program is "source URL {'out link1','out link2'}".

Then ***assign document Ids*** for all the links got from first step. Which is implemented in *LinkDataURLIDConverter01.java*. Input is "source URL {'out link1','out link2'}", out put two files first one is URL-ID mapping file, the second file is "source Id [1.0,{outID1,outID2}]". The second file is what Mapper in Hadoop is going to take.

PageRankComputer.java is the class contains Mapper and Reducer classes that follow the Hadoop framework, which implement map and reduce function compute the Pagerank.

Map function in Map class take ***input*** is “source IdY [pr(Y),{outID1,outID2}]”. Map function ***output*** is (outID1, pr(Y)/2),(outID2, pr(Y)/2), (sourceIdY,{outID1,outID2}).

Reduce function will gather new values from this iteration for each doc. Eg. for outID1,reduce process maybe: outID1 [0.15+0.85*(pr(Y)/2+pr(Z)/3),{outID4,outID5}]

Suppose outID1 is one of three page IDZ out going links and outID1 has two out going links itself.

The reduce output can be directly taken as input to map function in next iteration.

Usually need 20-30 ***iterations*** to compute PageRank. I use 20 iterations in total. The driver class is PageRank.java. Which prepares the input file (the output from last iteration), delete the output file and run the ***PageRankIteration.java*** for 20 times.

PageRankIteration.java prepares the configure for one iteration and the interface for PageRank.java to let one iteration run.

5. Some Top Results : By 20 iterations

Portal:Featured_content 1074.941700

Main_Page 1074.932600

Portal:Contents 1074.931900

Special:Random 1074.930400

Wikipedia:About 1074.930000

Wikipedia:Community_portal 1074.927900

Help:Contents 1074.927100

Wikipedia:General_disclaimer 1074.925800

Wikipedia:Contact_us 1074.925700

Special:SpecialPages 1074.918200

Special:RecentChanges 1074.914800

Wikipedia:Text_of_Creative_Commons_Attribution-ShareAlike_3.0_Unported_License 1074.914700

Portal:Current_events 1074.910200

Wikipedia:Upload 1074.900600

Special:Categories 1068.403100

Wikipedia:Stub 312.518800

United_States 185.414660

Geographic_coordinate_system 169.821320

Wikipedia:Verifiability 154.355760

Wikipedia:Reliable_sources 149.049180

Wikipedia:Citing_sources 131.356690

Wikipedia:Verifiability#Burden_of_evidence 127.691920

Category:All_articles_with_unsourced_statements 124.986084

Wikipedia:Citation_needed 121.785355

Category:Living_people113.907684

Category:All_articles_lacking_sources 108.819084

Template:Citation_needed 85.655040

United_Kingdom 83.641860

France 82.900760

England 64.096740

International_Standard_Book_Number 63.634120

Population_density 62.915073

Italy 61.115025

Internet_Movie_Database 60.911705

Help:Disambiguation 59.231037

File:Disambig_gray.svg 59.041885

Category:All_disambiguation_pages 59.039375

Category:All_article_disambiguation_pages 59.039375

Germany 57.001083

Wikipedia:Citing_sources#Inline_citations 54.582127

Category:All_articles_needing_additional_references 52.996590

World_War_II 51.737526

File:Question_book-new.svg 50.612396

Japan 50.114326

Area 48.863840

English_language 48.452312

Canada 44.957787

World_War_I 23.192709

Scotland 22.968071

Category:All_articles_to_be_expanded 22.579340

Public_domain 22.238787

UTC%2B2 22.217169

China 22.175457

List_of_sovereign_states 22.018793

Many pages's PageRank are very small.

References:

[1] **http://en.wikipedia.org/wiki/PageRank**

[2] The impact of Crawl Policy on Web Search Effectiveness.

[3] **http://hadoop.apache.org/common/docs/current/mapred_tutorial.html**

[4]http://www.ics.uci.edu/~djp3/classes/2010_01_CS221/Lectures/Lecture07_Slides_CS221.html