Perceptron Clustering on Wikipedia Data

LI, JUNBO LARRY
VALDIVIEZO, NICOLAS ALEJANDRO

Question??????????????

1. Is there a prefered topic that people look up in Wikipedia?
2. What are the most looked up topics?
3. Is there a pattern by time of the day? Morning, afternoon, etc?

List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- “Job 1”
- “Job 2”
- Problem encountered
- Summary and learning Outcome

List of Content

- **Definitions and Clustering Technique**
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- "Job 1"
- "Job 2"
- Problem encountered
- Summary and learning Outcome

Definitions and Clustering Technique

- Perceptron clustering:
  - Clustering technique for Big Data and AI.
  - Linear Binary Classifier.
  - IT IS A DOT PRODUCT!.
  - What needs:
    - Per class there is a threshold ϑ.
    - Dictionary with Scores
    - Input: Set of words **W**
    - Process: max { (**W . St1**), (**W . St2**), (**W . St3**), … , (**W . St9**)}
    - Output: Class and score IF score > ϑ.

List of Content

- Definitions and Clustering Technique
- **The Data - WikiMedia**
- Big picture of high performance: Data downloading, wrangling and analysis
- "Job 1"
- "Job 2"
- Problem encountered
- Summary and learning Outcome

The Data - WikiMedia

- WikiMedia - Open Data
- *file* -> file with pages views in an hour of a day (yymmdd-hh)

Size for 2015:

24 files/ day x 365 days/year = 8760 files

1 file ≈ 450MB

8760 year ≈ **3.7594 TB**

```
919013 en Environmental_Biology 1 1
919014 en Environmental_Choice_Program 1 1
919015 en Environmental_Control_and_Life_Support_System 2 2
919016 en Environmental_Data_and_Information_Service 1 1
919017 en Environmental_Defence_Canada 2 2
919018 en Environmental_Defence_Society 3 3
919019 en Environmental_Defense 1 1
919020 en Environmental_Engineering 1 1
919021 en Environmental_Foundation_for_Africa 2 2
919022 en Environmental_Impact_Assessment 1 1
919023 en Environmental_Information_Regulations_2004 2 2
919024 en Environmental_Kuznets_Curve 3 3
919025 en Environmental_Law_Institute 1 1
919026 en Environmental_Management 1 1
919027 en Environmental_Measurements_Laboratory 1 1
919028 en Environmental_Media_Awards 1 1
919029 en Environmental_Modeling_Center 1 1
919030 en Environmental_Performance_Index 7 7
919031 en Environmental_Protection 1 1
919032 en Environmental_Protection_Agency 16 16
919033 en Environmental_Psychology 1 1
919034 en Environmental_Robots_Inc. 1 1
```

List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- **Big picture of high performance: Data downloading, wrangling and analysis**
- "Job 1"
- "Job 2"
- Problem encounteted
- Summary and learning Outcome

Downloading Data, Data Wrangling, & Data Analysis

Hadoop	MPI
	Download raw data .gz
Unzip the jar of raw files	
Clean the raw data (get rid off non-english search)	
	Job 1
Parse each category list	
	Job 2
Run data analytics	

List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- **"Job 1"**
- "Job 2"
- Analysis
- Summary and learning Outcome

“Job 1”

Why? Need of something more general than a page name. Categories associated with this page.

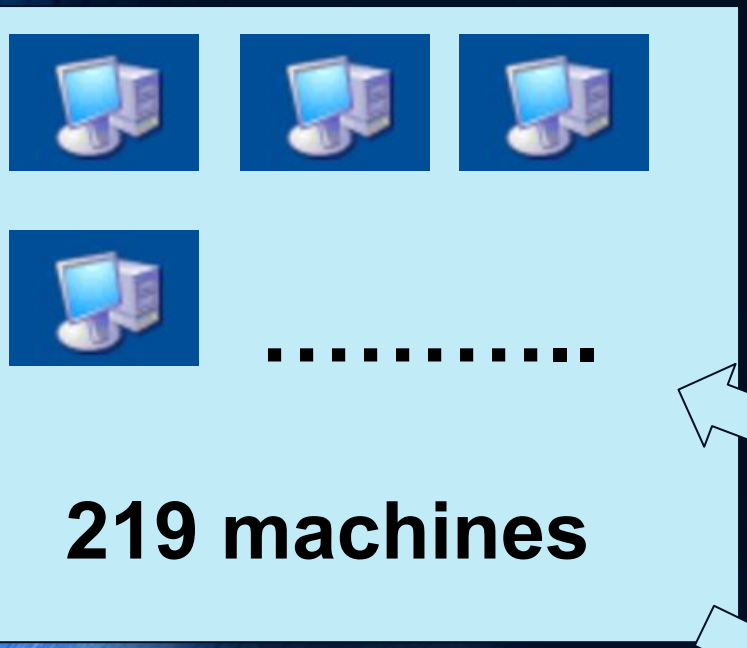
Framework: Multi-threaded MPI

MPI distribute works

- worker: works on a set of files
  - dynamic thieving:
    - Queue of files. threads thieve files and get Categories and write them out

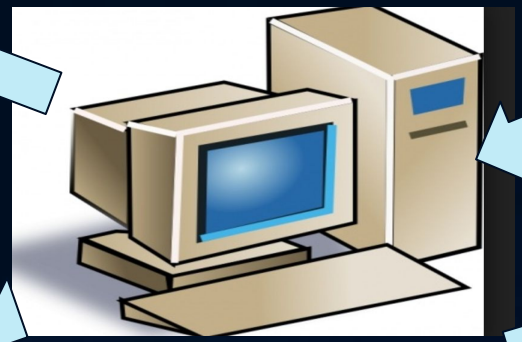
www

Parallelly query the category from the internet (MPI + Pthread)



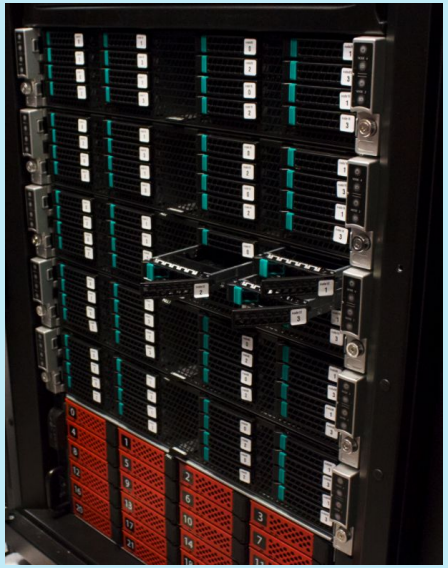
Partition the cleaned wiki data and send them to 12 219 machines

Send cleaned wiki data to Hadoop2



Hadoop2

Beo-Wulf cluster



List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- "Job 1"
- **"Job 2"**
- Analysis
- Summary and learning Outcome

“Job 2”

Why?

Clustering of the Categories into 8 topics:
Politics, CS, Natural Sci., Social Sci., Sports, FineArts, Health, Geography.
Scores are based on Oxford Learners Dictionary by Topic:

- more overlap in topics = less score across topics.
- closer to top of list = higher score in topic.

everything store in a Sqlite 3 DB.
Open lines with valid categories and compute its topic using Perceptron clustering.

WriteOut the Clustering Output (Topic with max score).

name = Barack%20Obama

W = { Presidents of the US, Democrat, Democratic party, …, 2008 elections, world leaders}

Dict:

	Politics	CS	Arts
Presidents	10	1	1
Democrat	10	0	1
UNIX	1	9	0
Renaissance	0	0	5
World	5	5	5
US	3	4	3

θ = 10*4 * 0.6 = 24

output = max {**W.**Politics, **W.**CS, **W.**Arts} = max{28,10,10} = Politics

List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- "Job 1"
- "Job 2"
- **Problem encountered**
- Summary and learning Outcome

Problems encountered:

Node	Last contact	Admin State	Capacity	Used	Non DFS Used	Remaining	Blocks	Block pool used	Failed Volumes	Version
c13.beowulf.net:50010 (172.16.1.13:50010)	1	In Service	450.05 GB	378.81 GB	29.79 GB	41.44 GB	17821	378.81 GB (84.17%)	0	2.7.3
c12.beowulf.net:50010 (172.16.1.12:50010)	2	In Service	450.05 GB	381.15 GB	29.82 GB	39.07 GB	17861	381.15 GB (84.69%)	0	2.7.3
c15.beowulf.net:50010 (172.16.1.15:50010)	0	In Service	450.05 GB	382.96 GB	29.8 GB	37.28 GB	17910	382.96 GB (85.09%)	0	2.7.3
c08.beowulf.net:50010 (172.16.1.8:50010)	2	In Service	450.05 GB	379.96 GB	29.44 GB	40.64 GB	17836	379.96 GB (84.43%)	0	2.7.3
c06.beowulf.net:50010 (172.16.1.6:50010)	2	In Service	458.32 GB	386.9 GB	29.86 GB	41.57 GB	17970	386.9 GB (84.42%)	0	2.7.3
c05.beowulf.net:50010 (172.16.1.5:50010)	0	In Service	450.05 GB	382.5 GB	29.8 GB	37.74 GB	17954	382.5 GB (84.99%)	0	2.7.3
c02.beowulf.net:50010 (172.16.1.2:50010)	0	In Service	450.05 GB	376.81 GB	29.62 GB	43.61 GB	18151	376.81 GB (83.73%)	0	2.7.3
c10.beowulf.net:50010 (172.16.1.10:50010)	1	In Service	458.32 GB	386.85 GB	29.87 GB	41.61 GB	17855	386.85 GB (84.41%)	0	2.7.3
c01.beowulf.net:50010 (172.16.1.1:50010)	1	In Service	450.05 GB	378.88 GB	29.81 GB	41.36 GB	17786	378.88 GB (84.19%)	0	2.7.3
c11.beowulf.net:50010 (172.16.1.11:50010)	1	In Service	458.32 GB	385.7 GB	30.21 GB	42.41 GB	17933	385.7 GB (84.15%)	0	2.7.3
c04.beowulf.net:50010 (172.16.1.4:50010)	2	In Service	450.05 GB	378.72 GB	29.82 GB	41.5 GB	17667	378.72 GB (84.15%)	0	2.7.3
c03.beowulf.net:50010 (172.16.1.3:50010)	1	In Service	450.05 GB	378.22 GB	29.93 GB	41.89 GB	17740	378.22 GB (84.04%)	0	2.7.3
c09.beowulf.net:50010 (172.16.1.9:50010)	0	In Service	450.05 GB	378.01 GB	29.48 GB	42.56 GB	17725	378.01 GB (83.99%)	0	2.7.3
c16.beowulf.net:50010 (172.16.1.16:50010)	1	In Service	450.05 GB	383.68 GB	29.81 GB	36.56 GB	17953	383.68 GB (85.25%)	0	2.7.3
c07.beowulf.net:50010 (172.16.1.7:50010)	1	In Service	450.05 GB	380.04 GB	29.8 GB	40.21 GB	17696	380.04 GB (84.45%)	0	2.7.3
c14.beowulf.net:50010 (172.16.1.14:50010)	2	In Service	450.05 GB	377.25 GB	29.8 GB	42.99 GB	17776	377.25 GB (83.82%)	0	2.7.3

Problems encountered:

```
<property>
  <name>
    yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-percentage
  </name>
  <value>90.0</value>
  <source>yarn-default.xml</source>
</property>
```

Problems encountered:

```
li_j8@hadoop2:~$ hdfs dfs -mkdir ./parse_test_in
mkdir: Cannot create directory /user/li_j8/parse_test_in. Name node is in safe mode.
li_j8@hadoop2:~$
```

List of Content

- Definitions and Clustering Technique
- The Data - WikiMedia
- Big picture of high performance: Data downloading, wrangling and analysis
- "Job 1"
- "Job 2"
- Problem encountered
- **Summary and learning Outcome**

Beowulf cluster and 219 lab

Difficulties:

219 computers → not good with big data

Beowulf cluster → doesn't have direct exit to WWW

Pros:

219 computers → distinct Public IPs

Beowulf Cluster + MR → AMAZING WITH BIG DATA!

Learning Outcome

1. Better understand the Hadoop.
2. Understand hardware limitations
3. Implemented our hybrid multi-parallel system to play around with big data
4. Used 3 programming languages, database knowledge, clustering (AI).
5. To be continued: answer those questions

Thank you!