Hadoop + Índice invertido + PageRank

Ricardo Manuel Lazo Vásquez - TheReverseWasp (GitHub)

PageRank y el Índice Invertido

$$PR_{t+1}(P_i) = \sum_{P_j} \frac{PR_t(P_j)}{C(P_j)}$$

Vocabulary	n_i	Occurrences as inverted lists
to	2	[1,4],[2,2]
do	3	[1,2],[3,3],[4,3]
is	1	[1,2]
be	4	[1,2],[2,2],[3,2],[4,2]
or	1	[2,1]
not	1	[2,1]
I	2	[2,2],[3,2]
am	2	[2,2],[3,1]
what	1	[2,1]
think	1	[3,1]
therefore	1	[3,1]
da	1	[4,3]
let	1	[4,2]
it	1	[4,2]

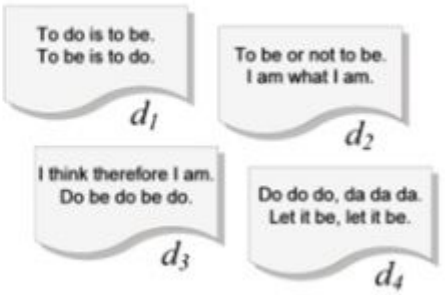

Figura 1 – Proceso de construcción del índice (Martín-Daucasa, 2012) .

Desafío

- Dados n sitios web crear una navegación basada en PageRank y utilizando índice invertido.
- Debe utilizarse el cluster de Hadoop de por lo menos 3 nodos.
- Debe tener interfaz gráfica.

Consideraciones

- Google Corpus como dataset
  - 5GB de datos
  - Presenta pares de palabras con calificación.
    - La primera palabra es la subpágina del documento.
    - La segunda palabra es considerada parte del contenido.
    - La segunda palabra apunta a otros documentos.primera_palabra.
- Se realizó la implementación en python.
  - Se desarrollaron 10 scripts de mapreduce, 1 de pagerank y 4 scripts lineales.
- Se realizaron pruebas en:
  - Un solo nodo.
  - Tres nodos.
  - Un solo proceso.
- Interfaz de escritorio.
- Problemática pesada 4M de palabras es decir 4M de posibles búsquedas.
  - Se lograron obtener búsquedas precompiladas de 5600 palabras.

Script de compilación de scripts en mapreduce y python

```
hadoop jar /usr/lib/hadoop-3.2.1/share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar \

-file map1.py    -mapper map1.py \

-file reduce1.py   -reducer reduce1.py \

-input filenames.txt -output output1
```

- En cada ejecución se debe eliminar la carpeta output.
- En el caso de Manjaro y Arch se puede instalar hadoop completamente configurado en un nodo por AUR el script seria el anterior.
- En caso de Ubuntu se tiene que cambiar el parametro de input por “file:///path_al_archivo”.

Github y Recursos

Demo Single-node: https://youtu.be/Svv4vEzYgKM

Demo Cluster: https://youtu.be/i55Ax8jok6A

GitHub: https://github.com/TheReverseWasp/PageRank_and_Inverted_Index

Hadoop + Índice invertido + PageRank

Ricardo Manuel Lazo Vásquez - TheReverseWasp (GitHub)