

Work Progress

Demo: KNN Search with Incremental Query Answering

Jaouhara Chanchaf

October 4th, 2022

Summary

Action Items for Jaouhara

Done:

AI 2	Reproduce the error that occurs when running Kashif on a large dataset.
------	---

In progress:

AI 1	Implement Parallel Incremental Query Answering.
AI 4	Run kashif on larger datasets. Cannot run process as root. Error: "Sorry, user jaouhara.chanchaf is not allowed to execute 'bin/dstree ...' on server server150012".

Error when running kashif with large k value

Experiment on 500k tables, 2.5M columns, 25M vectors.

10 Queries of size [50 - 100].

$k = 10M$

```
jaouhara.chanchaf@server150012: ~/Projects/IQA/iqa-bf/iqa-demo/code/search-algorithms/kashif/bin
File Edit View Search Terminal Help
buffered memory = 6000.000000 MB
>>> Loading index: /home/jaouhara.chanchaf/work-dir/indexes/100k-idx/
buffered memory = 6000.000000 MB
>>> Index read successfully
>>> Index loaded successfully from: /home/jaouhara.chanchaf/work-dir/indexes/100k-idx/
curr k = 1 - curr k = 3 - curr k = 5 - curr k = 10 - curr k = 30 - curr k = 50 - curr k = 100 - curr k = 300 - curr k = 500 - curr k = 1000 - curr k = 3000 - curr k = 5000 - curr k = 10000 - curr k = 30000 - curr k = 50000 - curr k = 100000 - curr k = 300000 - curr k = 500000 - curr k = 1000000 - curr k = 3000000 - curr k = 5000000 - curr k = 10000000
.
k = 1
k = 3
k = 5
k = 10
k = 30
k = 50
k = 100
k = 300
k = 500
k = 1000
k = 3000
k = 5000
k = 10000
k = 30000
k = 50000
k = 100000
k = 300000
k = 500000
k = 1000000
k = 3000000
k = 5000000
k = 10000000
>>> Result directory name: /home/jaouhara.chanchaf/work-dir/exp-results/kashif-search/test//kashif_l100000_lq_min50_max100
Killed
jaouhara.chanchaf@server150012:~/Projects/IQA/iqa-bf/iqa-demo/code/search-algorithms/kashif$
kashif [10:bash]
```

Figure: Kashif Average Recall

Querytime, recall and precision evaluation

Experiment on 100k tables, 494k columns, 5M vectors.
10 Queries of size [50 - 100].

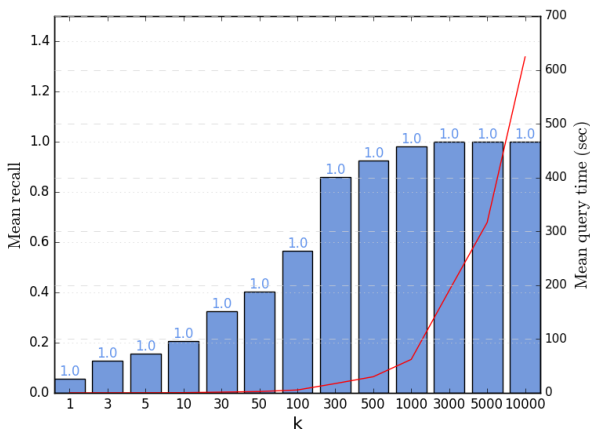


Figure: Kashif Average Recall

WDC Web Table Corpus 2015 - Statistics about English-language Relational Subset

Statistics:

#tables	50,820,165
#columns	235,087,091
datasize	69 GB

Column type distribution:

String	46.9%
Numeric	51.8%
Date	1.2%
Others	0.1%