Smart Database Retriever :

(Relevancy Searching on Structured Data)

Guide

Prof. Channa Bankapur

Team Members

Rohan Agarwal - 1PI13CS124

Srinivas Akhil - 1PI13CS164

Anirudh Agarwal - 1PI13CS199

PES Institute of Technology

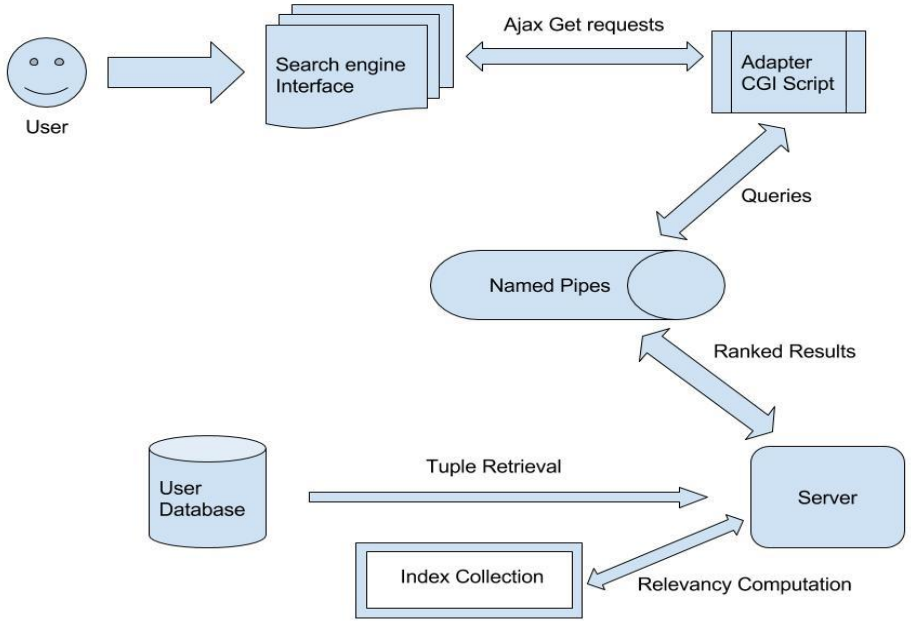
What is it ?

- ➔ It is a general application program interface to do database searches and results are retrieved in the order of their relevance.
- ➔

Interface Details ...

- ➔ User needs to attach database as a .csv file.
- ➔ Query can be entered in a general search box.
- ➔ No prerequisite querying knowledge needed, it can be done using simple language.
- ➔ Results will be prompted after writing each term.
- ➔ For a detailed result summary, user can switch to next page after completing the query.

Data flow diagram

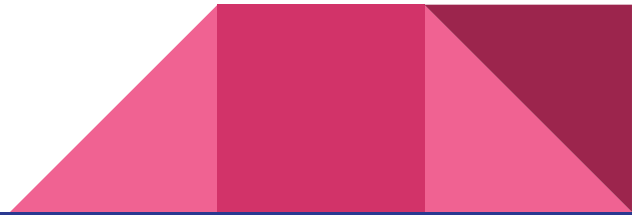


Demo ...

PES
Institute of Technology

How It Works

- ➔ It starts with server taking each database line and parsing it to create four different indexes.
- ➔ Cell index - each term (key) and tuple-IDs (value)
- ➔ Column Index - column no. (outer key), column value (inner key) and tuple-IDs (value)
- ➔ Column Cell Index - column no. (outer key), each term (inner key) and tuple-IDs (value)
- ➔ Tuple index - tuple-IDs (key) and [file-offset, tuple length] (value)



Contd ...

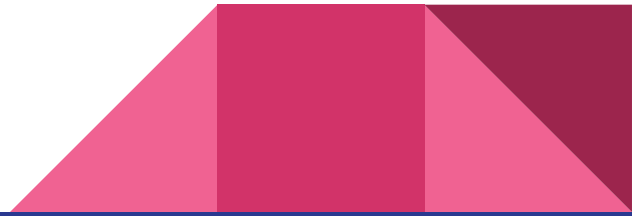
- ➔ The system then searches for workload file, if exists, goes on to create workload index.
- ➔ Workload Index - Column no (outer key), each term (inner key) and tuple-ID (value).

… The Server is now up and ready

How Indexing works

- ➔ The User query is tokenized, each term is hashed in cell index to get inverse document frequency score.
- ➔ IDF = log(total tuples / tuples having that term).
- ➔ From this we get intermediate relevancy score of database tuples for user query.
- ➔ K-box algorithm used to fill a set of n highest scored tuple.

Eg:- (Nissan Convertible)



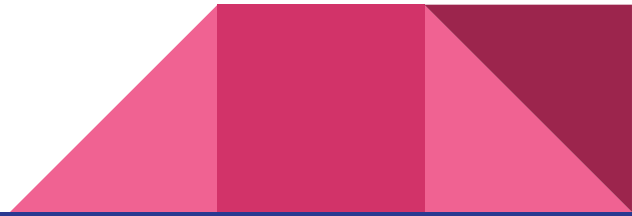
Why Workload ?

- ➔ Tuple Index is used to reduce file operation by directly fetching tuples at offsets.
- ➔ Since there would be many relevancy ties in k-box, we use ‘query frequency’ on workload to include other factors.
- ➔ Workload created and updated by user queries.

Eg1 :- (Recent homes and important locations), Eg:- (Book Search)

How Workload ?

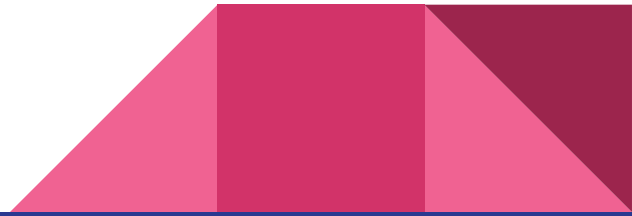
- ➔ The k-box is sent for ‘query-frequency’ calculation.
- ➔ Columns not used in user query is used for tie-break.
- ➔ QF = tuple frequency of term / Max term frequency in column
- ➔ This new variable added to previously calculated sum to get final relevancy score.



Sending back result

➔ The resultant score is final relevance score.
➔ Used to sort the tuples and send back to search engine interface.

...The client receives final tuple set as prompts and also as a detailed statistical infograph.



THANK YOU

- BTW the total LOC was 1795 :P.