Paper Title

Jaouhara Chanchaf
UM6P

Karima Echihabi
UM6P

ABSTRACT

300 word description of the project

PVLDB Reference Format:
Jaouhara Chanchaf and Karima Echihabi. Paper Title, 15(10): XXX-XXX, 2022.
doi:XX.XX/XXX.XX

1 INTRODUCTION

- **Use case 1: Keyword Query**
  A data scientist wants to analyze the impact of food cost inflation on food consumption. Initially The user decides to start the search with the keywords {”food”, ”consumption”} and $k = 10$. The search engine returns Table 1 which contains data from year 1990 to 2009 about *”Per Capita Consumption of Principal Foods (in pounds)”*. The user decides to keep Table 1 for the study and continue to search for other relevant tables.
- **Use case 2: Join Query** Table 1 is a good first result as it contains a complete list of the main food types, however the result lacks information on food prices. For that the use perform a join query on the food column to explore other tables that may have information about food prices for the years 1990 to 2009.
  Attempt 1: To speed up search the user submits the first query $Q_1 =$ (Table 1, Join column : ”Food”, $k = 10$). The search engine returned 775 tables. However, after skimming through the list of returned tables nothing seemed relevant to the user.
  Attempt 2: The user decides to increase k to get more results from the search engine. He/she submits a second query $Q_2 =$ (Table 1, Join column : ”Food”, $k = 20$). This time the search engine returned 161 tables, because the number of results is big the user could notice Table 2 ranked at position 55.
  Attempt 3: For the last attempt the user gave up on getting any fast meaningful result so he/she decide to increase k significantly in hope that a relevant table will appear in the list of results. He/she submits Q_3 to the search engine.$Q_3 =$ (Table 1, Join column : ”Food”, $k = 200$). Finally and after several attempts, the search engine returns Table 2 at position 11 which contains information on food prices from the 2007 WIC program.

2 BACKGROUND

(Important definition in the literature)

3 SYSTEM ARCHITECTURE

(...)

4 DEMONSTRATION

This is where you describe the approch that solves the problem in section **??**.

5 IDEAS AND QUESTIONS

5.1 Ideas

(1) The dataset discovery process is more likely to be iterative. At the beginning the user forms a general query for his information need. Retrieved datasets help the user better understand his need and hence better reformulate his query in the next iteration.
(2) Guarantee interactive-level response time.
(3) The relationship that we aim to capture between datasets oftentimes defines the structure of the query that search engine will accept.
(4) speed up the data science workflow.
(5) Similarity measures:
  (a) Text data: Hammin distance, Jaccard similarity, Jaccard containment ...
  (b) Numerical data: Correlation, cosine distance, euclidean distance ...
  (c) Binary data: ...
  (d) TF-IDF, Okapi BM25, LDA topic modeling
(6) Order based similarity measures (image retrieval)?

5.2 Questions

(1) If the query time is too small how can the user observe changes in query results as they arrive incrementally?
+ The user can observe changes in query results because when querying over Terabytes (or even Petabytes of data) the query time will increase significantly.

ACKNOWLEDGMENTS

We sincerely thank X, Y and Z.

REFERENCES

[1] K. Echihabi, K. Zoumpatianos, T. Palpanas, and H. Benbrahim. The Lernaean Hydra of Data Series Similarity Search: An Experimental Evaluation of the State of the Art. *PVLDB*, 12(2), 2018.

Proceedings of the VLDB Endowment, Vol. 15, No. 10 ISSN 2150-8097.
doi:XX.XX/XXX.XX

Food	1990	1995	2000	2002	2004	2005	2009
Wheat flour	135.9	140.0	146.3	136.8	134.3	134.1	134.6
Vegetables	386.0	407.6	423.0	411.8	422.8	415.4	390.9
Veal	0.9	0.8	0.5	0.5	0.4	0.4	0.3
Turkey	13.8	13.9	13.7	14.0	13.4	13.1	13.3
Tree nuts	2.45	1.94	2.57	3.24	3.62	2.7	3.7
Rice (milled basis)	15.8	17.1	18.9	19.5	20.4	21.0	21.2
Red meats2,3,4	112.2	113.6	113.7	114.0	112.0	110.0	105.7
Poultry2,3,4	56.2	62.1	67.9	70.7	72.7	73.6	69.4
...	...	...	...	...	...	...	...

Table 1: U.S. Economy and the Federal Budget Economy: Per Capita Consumption of Principal Foods

Food item	**Retail sales database selection criteria**	**Units**	**Price per unit (inflated to FY06)**
Yogurt	Quart sized containers and larger. Plain, vanilla, and fruit flavors	qt	2.068
Whole-grain	bread Wheat or grain bread	lb	1.422
Whole	Fresh dairy milk only,1/2gallon or gallon containers. Reduced fat includes skim milk and milk identified as 2% or lower milk fat	qt	0.767
Tuna	Chunk light, canned	oz	0.101
Peanut butter	All forms and varieties. Not mixed with jelly	oz	0.094
Brown rice	Instant or regular	lb	1.178
...	...	...	...

Table 2: Federal Register — Special Supplemental Nutrition Program for Women, Infants and Children (WIC), 2007

[2] PhDComics. Graduate Student Work Output. https://phdcomics.com/comics/archive.php?comicid=124, 2022.