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

This is where you motivate the problem.

2 PRELIMINARIES

This is where you put definitions, including a formal problem definition.

- **Use case 1: Keyword Query**
  An analyst wants to study (a topic related to yield production)
  Query = {"crop", "yield"}
  Results: Table **??** and **??**
- **Use case 2: Dataset Query**

3 RELATED WORK

This is where you survey and cite the related work. Add all referenced work in the paper to the ref.bib file (using meaningful names not the default ones) and to cite it as follows [1].

4 THE XYZ APPROACH

This is where you describe the approch that solves the problem in section 2.

5 EXPERIMENTAL EVALUATION

This is where you describe the experimental setup and discuss results. Save plots in the img directory and make sure they look nice(Figure 1)

6 CONCLUSIONS AND FUTURE WORK

One concluding paragraph summarizing the contributions and pinpointing any future work directions.

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

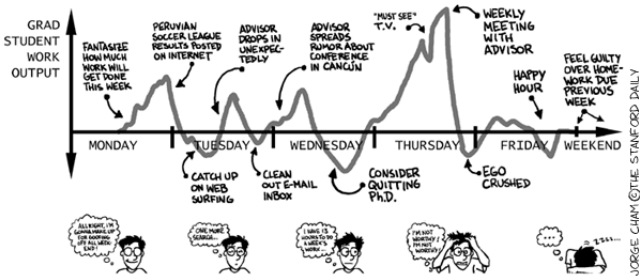

Figure 1: I promise to be productive today! [2]

7 IDEAS AND QUESTIONS

7.1 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)?

7.2 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.

Crop	Net Acres	Liability
Wheat	44,113,027	4,001,440,669
Sunflowers	1,815,262	197,029,917
Soybeans	60,724,474	10,832,278,232
Rice	1,977,293	341,003,283
...	...	...
Corn	62,104,041	16,764,714,958

Table 1: 2006 Top ten insured crops by net acres

Crop	Green maize	Grain maize	Tomatoes	Leafy vegetables	Total
Gross income (Z$)	16 400	4 000	62 950	16 050	99 400
Costs (Z$)	200	200	300	900	-
Total Costs (Z$)	15 000	2 600	56 250	11 250	85 100
Energy Costs (Z$)	-	-	-	-	79 000
Labour	300	300	2 000	2 000	4 600
Seed	600	600	1 500	1 500	4 200
Gross margin (Z$)	40 000	6 933	150 000	30 000	3 600
Fertilizer	100	100	2 000	400	2 600

Table 2: Cropping pattern for a 1.5 ha plot at Wenimbi irrigation scheme (Source: Extension Worker's record book, 1998)

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