Utrecht University

Dataset Discovery using Semantic Matching

Enas Khwaileh

Yannis Velegrakis

DBDBD
November 22, 2024

Utrecht University

Task: Analyzing COVID-19 vaccine data to assess effectiveness and potential side effects.

Keyword: “Covid-19 vaccine”.

CDC

Region	Date	Trade Name	Dosage
North Amercia	01-03-2020	Pfizer_BioNTECH	First
Europe	01-06-2020	AstraZeneca	Second
Asia	01-05-2020	Moderna	First
Africa	01-02-2020	Pfizer_BioNTECH	Second

WHO

Country	Date	Vaccine	Disease
Germany	01-01-2020	Comirnaty	COVID-19
France	01-02-2020	Vaxzevria	COVID-19
Spain	01-01-2020	CoronaVac	COVID-19
Italy	01-02-2020	Covaxin	COVID-19

ECDC

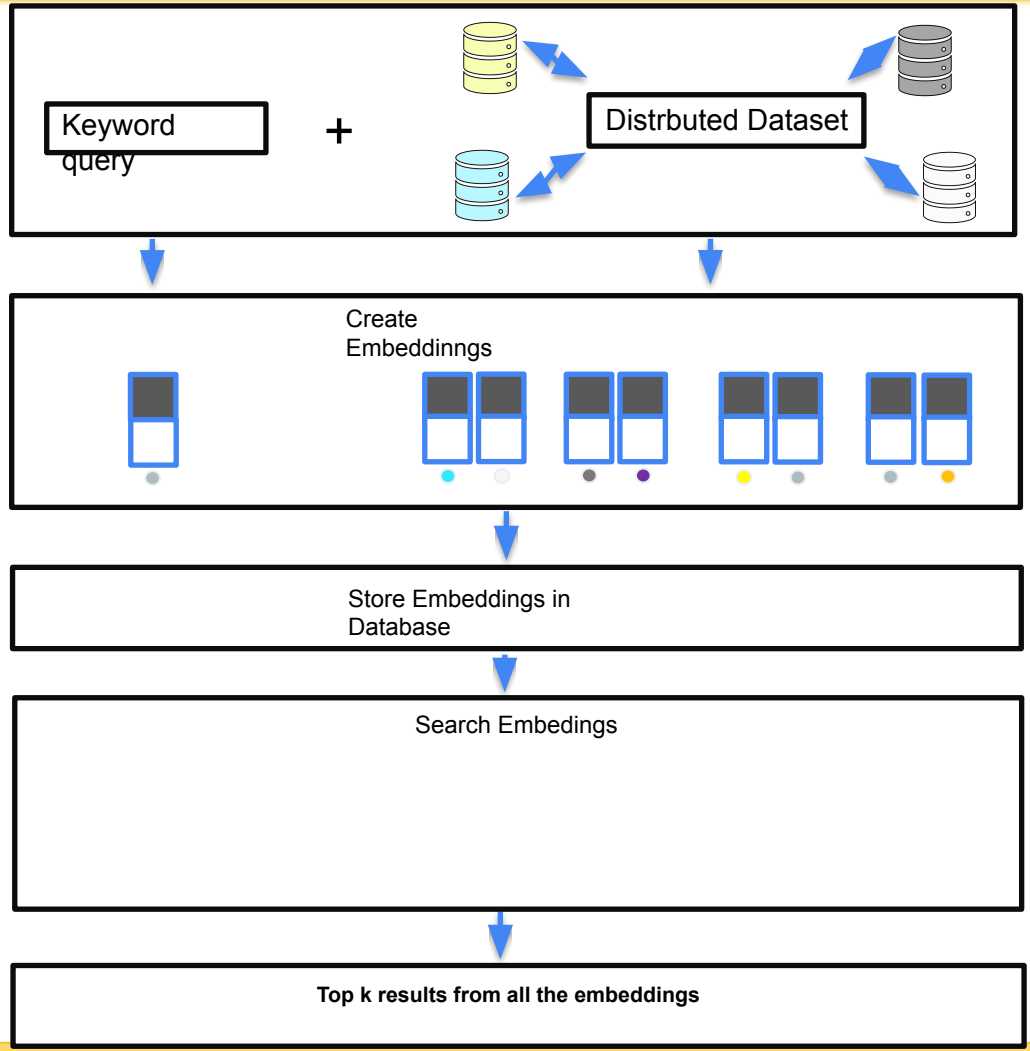
State	Date	Immunogen	Manufact
California	01-01-2021	mRNA	Moderna
Europe	01-04-2020	Vector Virus	Janssen
Asia	01-05-2021	mRNA	Pfizer
Africa	01-02-2020	Protien Subunit	Novavax

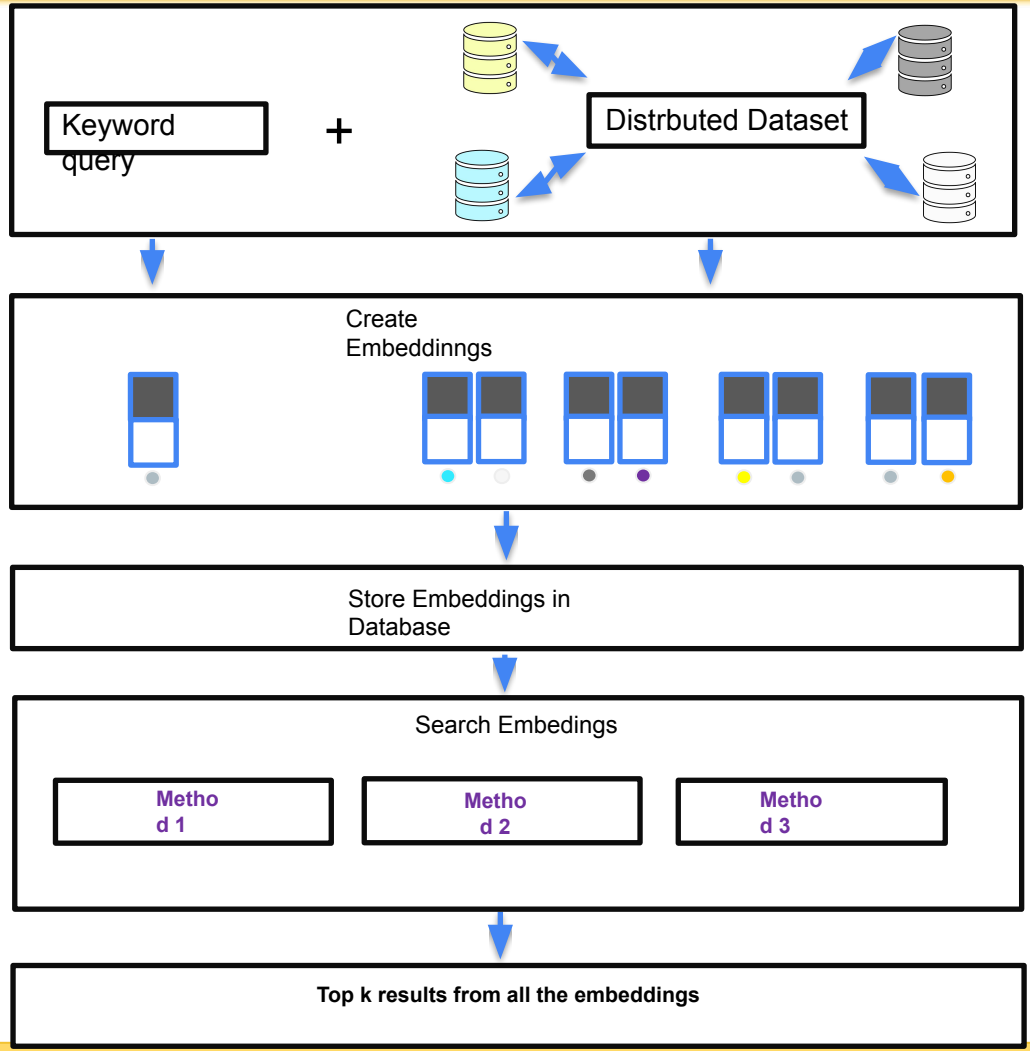
Problem statement

Assume we have a query **Q** and a datasets D we want to find the *match* function:

$$match: \boldsymbol{\mathcal{D}}\ x\ Q\ \rightarrow R$$

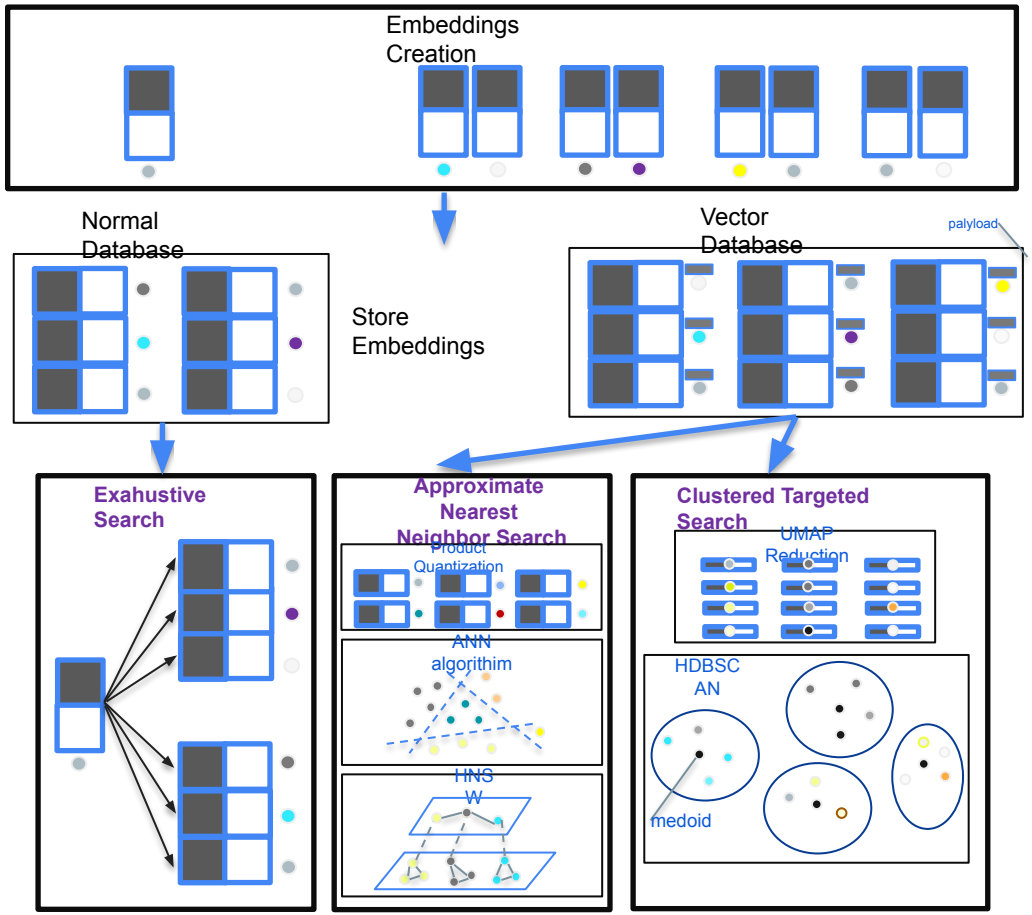
$\mathcal{D}$: set of dataset
Q: set of queries


Keyword query
+
Distrbuted Dataset
Create Embeddinngs
Store Embeddings in Database
Search Embedings
Top k results from all the embeddings

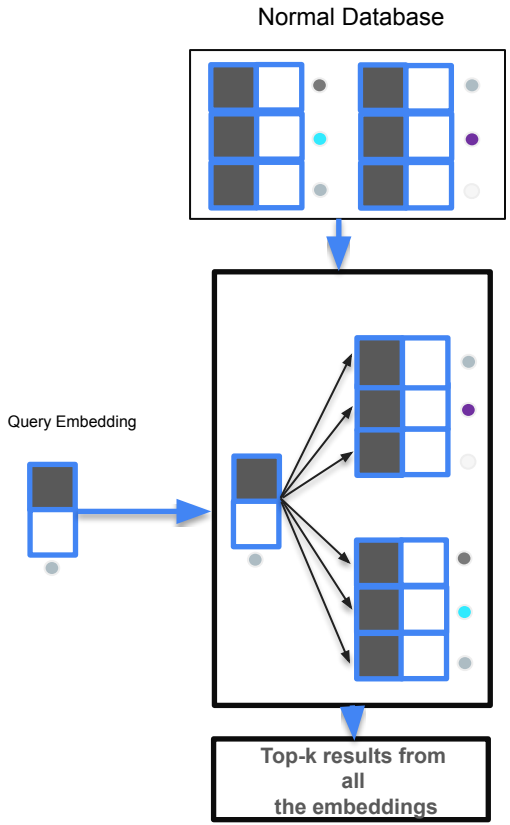

Keyword query
+
Distrbuted Dataset
Create Embeddinngs
Store Embeddings in Database
Search Embedings
Metho d 1
Metho d 2
Metho d 3
Top k results from all the embeddings

Our Solution: Semantic Matching Techniques

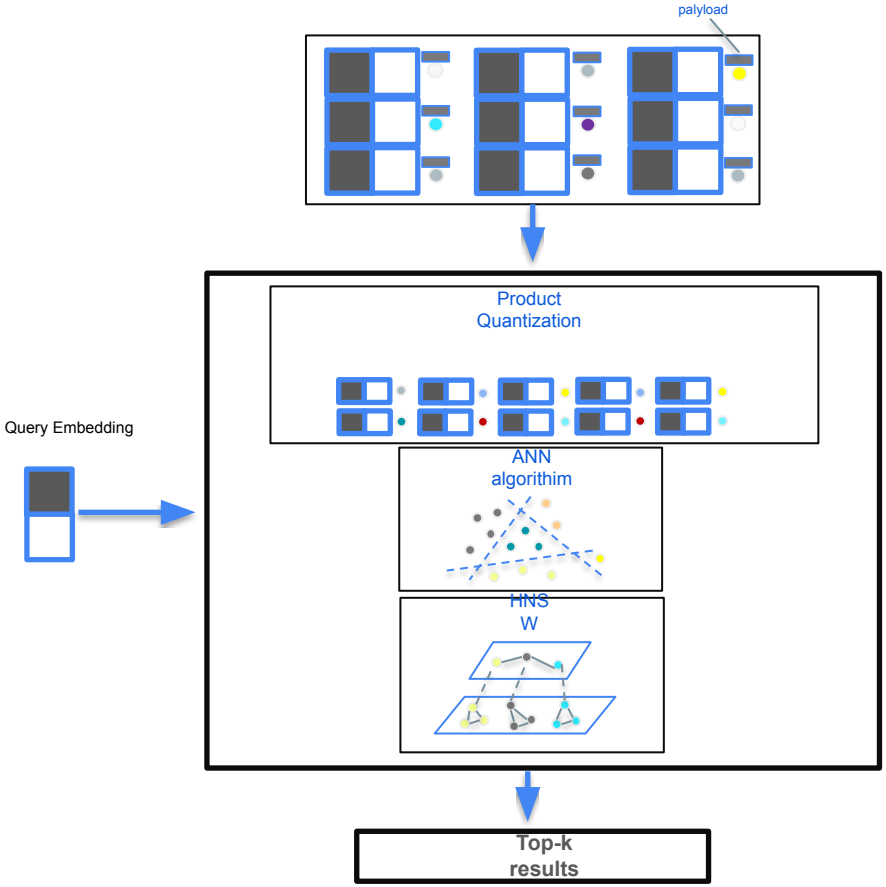
1. Exhaustive Search (ExS).
2. Approximate Nearest Neighbors Search (ANNS).
3. Clustered Targeted Search (CTS).


Embeddings Creation
Normal Database
Store Embeddings
Vector Database
payload
Exahustive Search
Approximate Nearest Neighbor Search
Product Quantization
ANN algorithim
HNSW
Clustered Targeted Search
UMAP Reduction
HDBSCAN
medoid

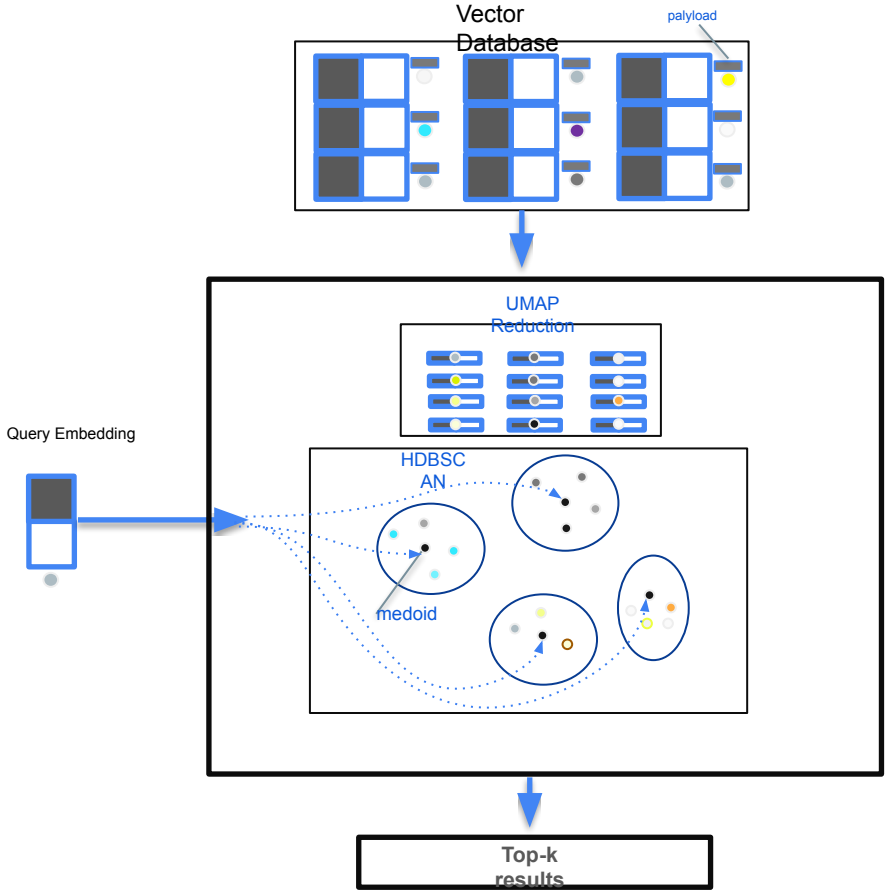
Exahustive Search (ExS)

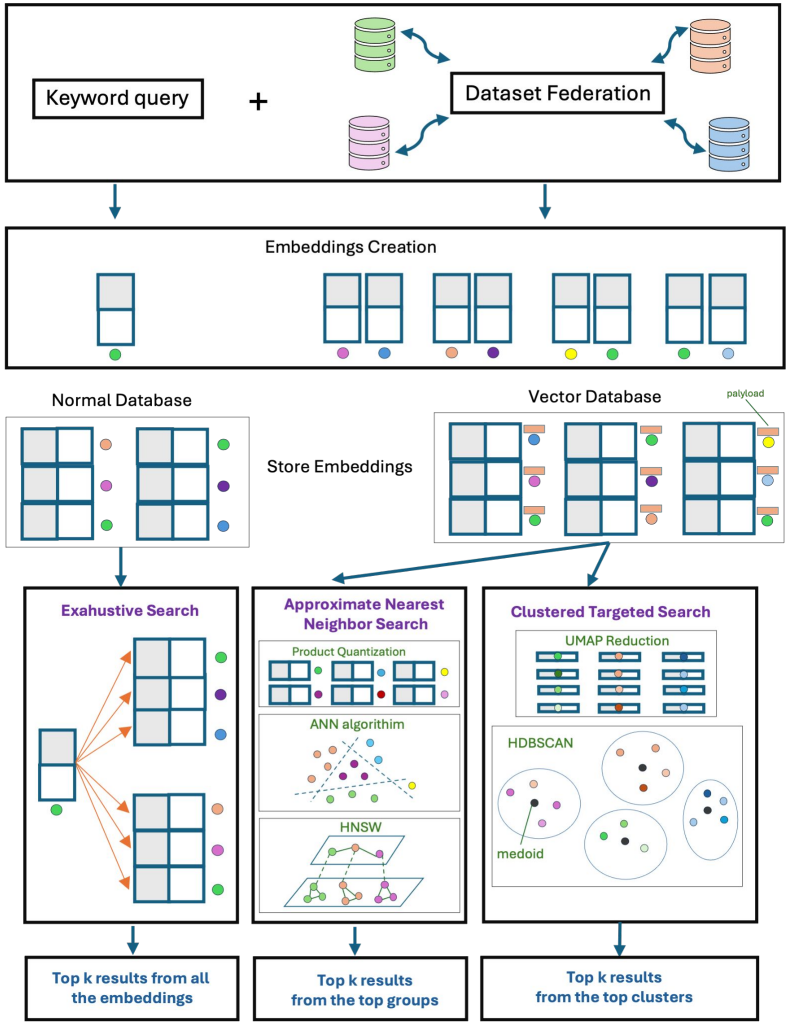


Approximate Nearest Neghbiors Search (ANNS)



Clustrered Targeted Search (CTS)




Keyword query
+
Dataset Federation
Embeddings Creation
Normal Database
Store Embeddings
Vector Database
palyload
Exahustive Search
Approximate Nearest Neighbor Search
Product Quantization
ANN algorithim
HNSW
Clustered Targeted Search
UMAP Reduction
HDBSCAN
medoid
Top k results from all the embeddings
Top k results from the top groups
Top k results from the top clusters

Experimental Setup

WikiTables (1.6M)

EDP (600K)

“60” queries
Zhang et. al 18

- Large Dataset (LD): 100%.
- Moderate Dataset (MD): 50%.
- Small Dataset (SD): 10% of the original dataset.

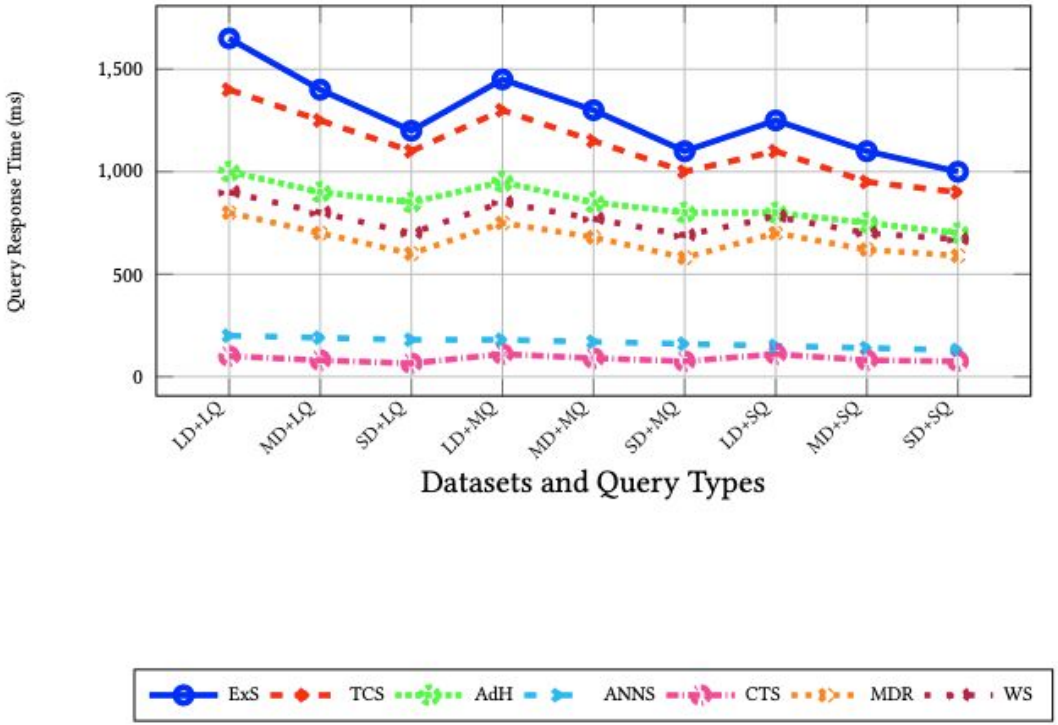
- Long Queries (LQ): Containing over 30 keywords, but not more 300 words.
- Moderate Queries (MQ): Up to 30 keywords.
- Short Queries (SQ): Not more than three keywords.

Metrics: Accuracy: MAP, MRR, NDCG.
Scalability: retrieving time millisecond/per query

Performance Evaluation (Quality)

- CTS achieves MAP, MRR, and NDCG across all query types.
- ANNS offers competitive speed and reasonable accuracy.
- ExS has high accuracy but low efficiency.

Performance Evaluation (Scalability)



Performance Evaluation (Scalability) millisecond

Dataset	Query	CTS	ANNS
100%	Long	75	100
	Moderate	85	90
	Short	110	150
50%	Long	70	75
	Moderate	80	120
	Short	80	130
10%	Long	65	95
	Moderate	75	100
	Short	75	115

Table 1: Quality of long query results

Dataset	Method	MAP	MRR	NDCG			
				5	10	15	20
LD	CTS	0.705	0.680	0.720	0.700	0.685	0.668
	ANNS	0.685	0.670	0.700	0.675	0.660	0.642
	ExS	0.670	0.655	0.690	0.670	0.650	0.635
	MDR	0.655	0.640	0.675	0.655	0.640	0.625
	WS	0.640	0.625	0.665	0.645	0.630	0.615
	TCS	0.635	0.620	0.660	0.640	0.625	0.610
	AdH	0.620	0.605	0.650	0.630	0.615	0.600
MD	CTS	0.720	0.700	0.735	0.710	0.695	0.675
	ANNS	0.705	0.690	0.720	0.700	0.680	0.665
	ExS	0.690	0.675	0.710	0.690	0.670	0.650
	MDR	0.675	0.660	0.700	0.680	0.660	0.645
	WS	0.660	0.645	0.690	0.670	0.650	0.635
	TCS	0.655	0.640	0.680	0.660	0.640	0.625
	AdH	0.640	0.625	0.675	0.655	0.635	0.620
SD	CTS	0.735	0.715	0.750	0.725	0.710	0.690
	ANNS	0.720	0.700	0.740	0.715	0.700	0.685
	ExS	0.705	0.690	0.730	0.710	0.690	0.675
	MDR	0.690	0.675	0.720	0.700	0.685	0.670
	WS	0.675	0.660	0.710	0.690	0.675	0.660
	TCS	0.670	0.655	0.705	0.685	0.670	0.655
	AdH	0.655	0.640	0.695	0.675	0.660	0.645

Questions?

Thank you

Main Challenges:

- Traditional methods provide limited solutions
- Issues of fragmentation across systems.

Solution Concept:

- Semantic matching is used to uncover deeper relationships in data.

Why CTS?

- Let’s take: "Olympics Beijing" as a keyword query example.
- Estimation: ExS will perform better
- Result: CTS gave more accurate results with exceptional retrieval time.
- How? While ExS returned broader Olympic-related tables (e.g., sports data), CTS retrieved tables with more detailed information on Olympic locations, years, and regions, aligning better with the query’s intent. CTS achieves this by clustering semantically similar tables and focusing the search on the most relevant clusters, avoiding ExS’s dilution of relevance and ANNS’s approximation errors.

State-of-the-art

Table Contextual Search (TCS):

Shuo Zhang and Krisztian Balog. 2018. Adhoc table retrieval using semantic similarity. In Proceedings of the 2018 world wide web conference. 1553–1562

Ad-Hoc Table Retrieval (AdH):

Zhiyu Chen, Mohamed Trabelsi, Jeff Heflin, Yinan Xu, and Brian D Davison. 2020. Table search using a deep contextualized language model. In Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. 589–598.

WebTable System (WB)

Michael J Cafarella, Alon Halevy, and Nodira Khoussainova. 2009. Data integration for the relational web. Proceedings of the VLDB Endowment 2, 1 (2009), 1090–1101.

Conclusion

Dataset discovery is crucial in the era of big data, where finding the right dataset quickly and accurately makes all the difference. Our proposed methods achieve high performance, offering scalability without sacrificing accuracy. We believe that a good search model is one that not only delivers precise results but also meets user satisfaction by providing relevant and efficient discovery. This focus on both performance and user experience drives our approach and distinguishes our methods in the landscape of dataset discovery."