Data-Semantics-Aware Recommendation of Diverse Pivot Tables
SAGE: Adaptive Recommendation of Spreadsheet Pivot Tables [Demo]

Whanhee Cho, Anna Fariha

KAHLERT SCHOOL OF COMPUTING
THE UNIVERSITY OF UTAH

Manually exploring large spreadsheet data is cumbersome, recommendation of data summary can alleviate this

Pivot Table
Group-by & aggregation in Spreadsheets

Restaurant data: 20 attributes, 10K Tuples

171 attribute pairs to group by

	Cuisine		
City	Indian	Italian	Chinese
Mumbai	300	1200	700
Delhi	200	1300	900
Bengaluru	250	1750	500

Average Cost (₹)

5 aggregates | 20 value attributes

20,000 pivot tables possible!
Recommendations are necessary

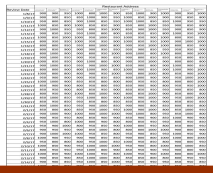
Existing spreadsheets can recommend pivot tables, but they have several shortcomings

Problem 1: Data-Semantics-unaware

City	Sum of Longitude
Mumbai	728,800
Delhi	771,000
Bengaluru	775,900

`SUM of Longitude` is **semantically invalid**

Problem 2: Interpretability-unaware



Big pivot tables are **hard to interpret**

Problem 3: Redundancy

Average Cost group by Cuisine
Average Cost group by Cuisine and Country
Average Cost group by Cuisine and City

Top-K recommendations **overlap too much**

SAGE recommends a set of diverse and high utility (data-semantics and interpretability-aware) pivot tables in real-time

Problem Definition

Objective: Maximize pivot table **utility**
Constraint: Ensure **diversity**

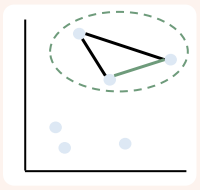
$$\max \sum Utility(\square) \text{ in } \square \text{ such that } \textit{min pairwise distance} \geq \theta$$

NP-Hard

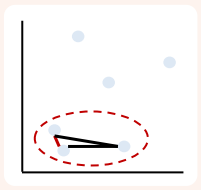
User-defined

Diversity

Maximize the minimum pair-wise distance



Diverse: uniformly spread out



Not Diverse: Too small min distance

Structural Embedding: T5 Encoder, an NL encoder fined-tuned over a Text-to-SQL dataset

Semantic Embedding: TAPEX, a pre-trained encoder trained on sentence-table pairs

Pivot Table Embedding =
Structural Embedding of the Query +
Semantic Embedding of the Content

Utility

Utility

Insightfulness

Informativeness

Variability in values offer insights

Uniform values

	Cuisine		
City	Indian	Italian	Chinese
Agra	4.5	4.7	4.4
Kanpur	4.3	4.6	4.5
Bengaluru	4.6	4.7	4.6

Average Rating

Varied values

	Cuisine		
City	Indian	Italian	Chinese
Agra	300	1200	700
Kanpur	200	1300	900
Bengaluru	250	1750	500

Average Cost(₹)

Attribute Significance

Not all attributes are insightful

`Cuisine`

`Address`

`Cuisine` is an insightful group-by attribute; `Address` is not.

Trend

Unexpected trends provide insights

	Cuisine		
City	Indian	Italian	Chinese
Mumbai	800	800	850
Delhi	500	900	1200
Bengaluru	600	800	1100

Average Cost(₹)

Expected

Surprisingness

Unexpected outliers provide insights

	Cuisine		
City	Indian	Italian	Chinese
Mumbai	800	1600	850
Delhi	1500	900	800
Bengaluru	700	800	750

Expected

Unexpected

Average Cost(₹)

Interpretability

Semantic Validity

Some aggregates don't make sense

`COUNT(ID)`

`SUM(ID)`

Density

Sparse pivot tables reduce interpretability

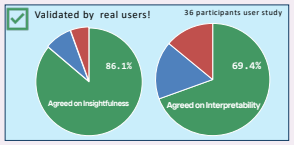
Avoids nulls caused by missing group combinations

Conciseness

Concise tables are interpretable

Penalize large pivot tables

Avoid group-by high-cardinality attributes



LLM acts as a Semantic Oracle: informs real-world expectedness

Ensuring practical response time: Four scaling methods

Prune low-quality tables, avoiding materialization

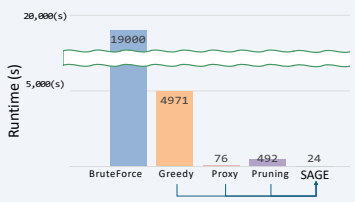
Use a **proxy** model to approximate LLM to reduce inference latency

Sample tuples and pivot tables for faster computation

Greedily select tables to ensure minimum distance from already picked tables is above threshold

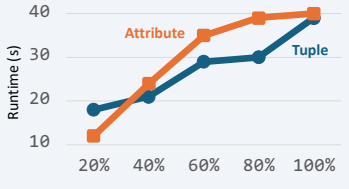
Empirical results: Ablation and Scalability

Ablation Analysis



Proxy shows the strongest speed up

Scalability Analysis (110 Attributes, 581K Tuples)



Linear with #tuples. Exponential with #attributes.

SAGE outperforms alternatives

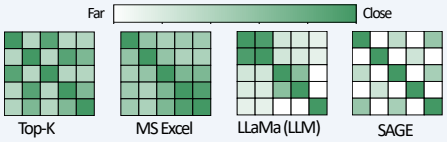
	Time (s)	Utility
SAGE	17	2.36
Top-k	64	2.67
LLM (LLaMa-3B)	20	0.52
Excel	01	1.80

SAGE strikes the best balance between response time and utility

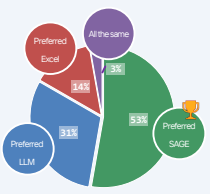
SAGE ensures diversity

Distance heatmap between pivot tables

Far — Close



Real users prefer SAGE over LLM & Excel



"SAGE seems to tell us more overall about what impacts"

"SAGE provides more meaningful and helpful queries compared to the later two"