

DACO:

Towards Application-Driven and Comprehensive Data Analysis via Code Generation

University of California, Los Angeles ByteDance



Background

- Tabular data is widely used to store information
- Tabular data analysis is crucial for informed decision making



Financial analysis



Healthcare analysis

- Existing work on table QA: focus on **factual retrieval** targeting specific cells, rows or columns

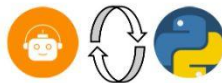
What are the name of the patient that... ?

Task Introduction

- Input: tabular data + complex queries motivated by real-world user needs
- Output: findings and suggestions supported by data



As a consumer rights advocate, I want to assess if the coffee shop is promoting age discrimination.



Required skills:

- Analytical reasoning
- Mathematical reasoning
- Strategic reasoning

Database: coffee shop

member

Member_ID	Name	Age	...
1	Ashby, Lazale	29	...
2	Breton, Robert	67	...
3	Campbell, Jessie	34	...
...

happy_hour

HH_ID	Shop_ID	...
1	1	...
2	1	...
3	10	...
...

happy_hour_member

HH_ID	Member_ID	Total_amount	...
1	3	20.9	...
4	3	20.92	...
7	9	4.9	...
...

Findings

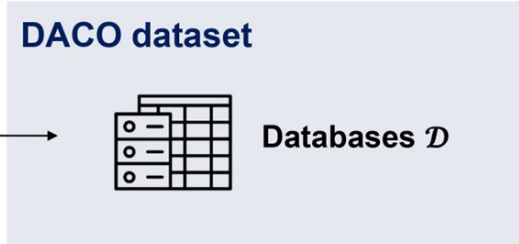
1. The age distribution of the coffee shop's members is primarily within the age range of 19-65... There are no clear evidence of discriminatory behavior based on age.
2. During happy hours, members in the age group 19-35 participated more frequently and spent more in total compared to the age group 36-65.

Suggestions

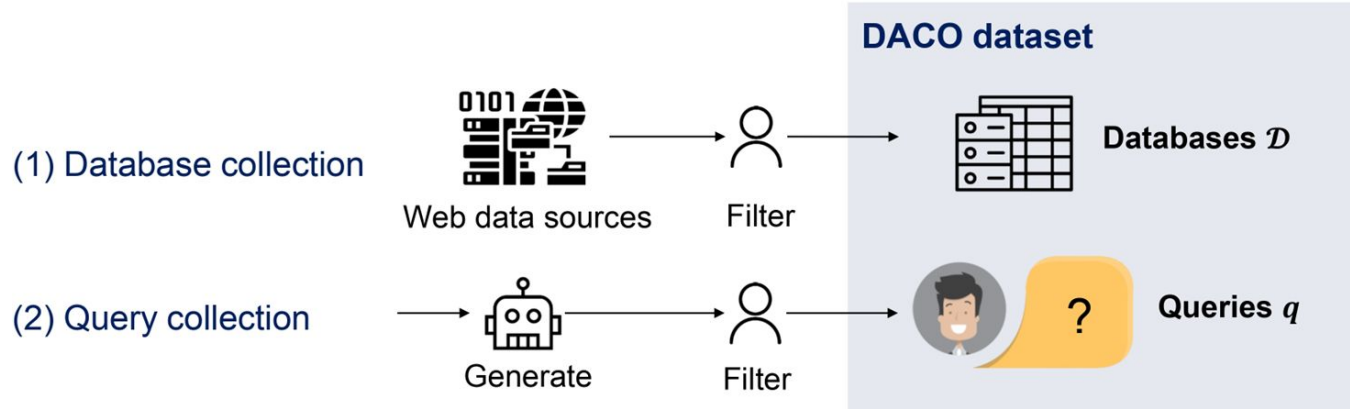
1. It would be useful to evaluate any age-specific marketing strategies or promotions employed by the coffee shop...
2. Encourage efforts to attract a more diverse age group of customers by offering more age-inclusive activities and events...

Dataset Construction

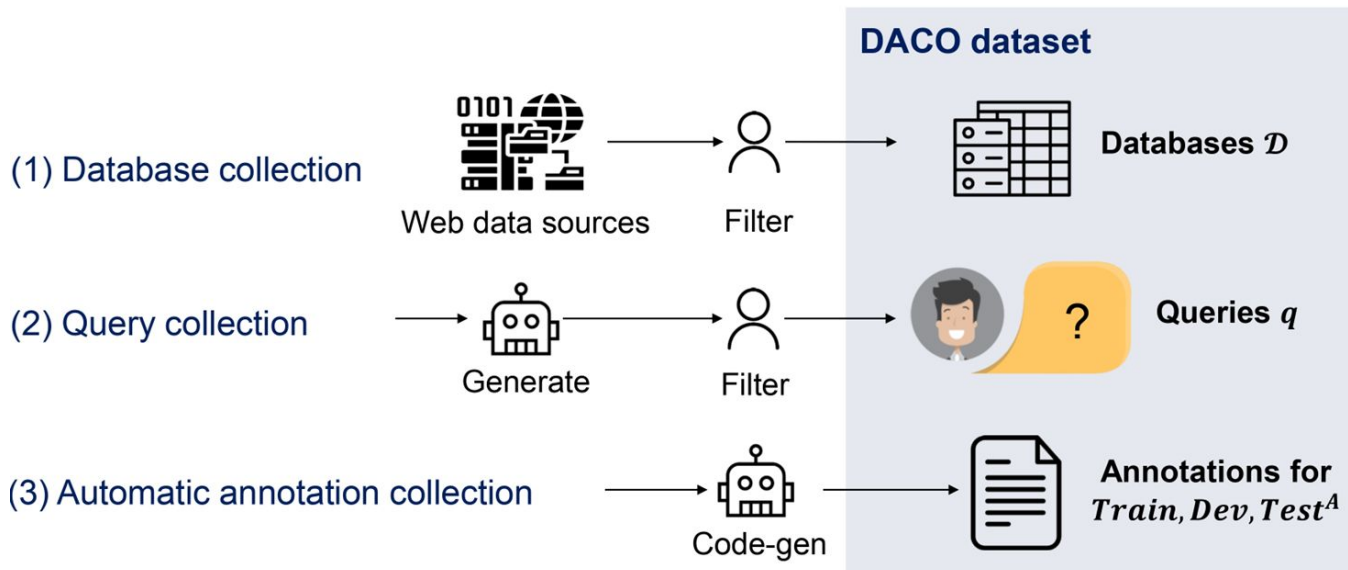
(1) Database collection



Dataset Construction



Dataset Construction



Automatic Annotation via Code Generation



As a customer rights advocate, I want to assess if the coffee shop is promoting age discrimination.

LM (t=1):

I will now analyze the age distribution among members to see if there are any anomalies or unusual patterns...

```
member_table = db["member"]
age_stats = member_table["Age"].describe()
print("Age statistics for members:")
print(age_stats)
```

Code interpreter:

```
Age statistics for members:
count    10.000000
mean     43.700000
std      11.518584
min      29.000000
...
```

LM (t=2):

I will now examine if there are any age-related discrepancies in happy hour participation and spending patterns...

```
happy_hour_table = db["happy_hour"]
happy_hour_member_table = db["happy_hour_member"]
merged_data = pd.merge(happy_hour_member_table, member_table, on="Member_ID")
```

Code interpreter:

	HH_ID	Total_amount \
0-18	0	0.00
19-35	4	50.92
36-65	2	33.82
66+	0	0.00
...		

.....

LM (final answer):

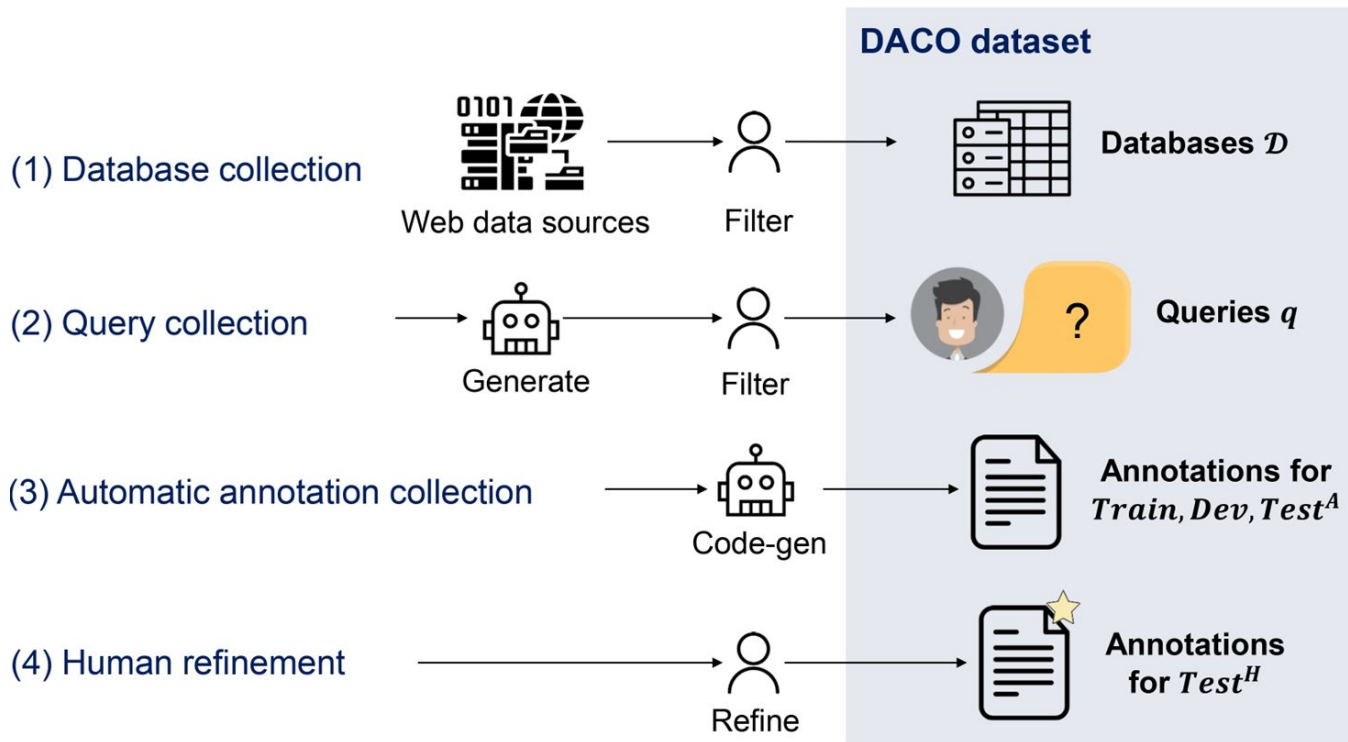
Findings

1. The age distribution of the coffee shop's members is primarily within the age range of 19-65, with 40% of members aged 19-35, 50% aged 36-65, and 10% aged 66 or older...

Suggestions

1. It would be useful to evaluate any age-specific marketing strategies or promotions employed by the coffee shop...

Dataset Construction



Dataset Statistics

Human-refined
test set



	Train	Dev	Test ^A	Test ^H	Total
<i>Input Statistics</i>					
# db	353	22	65	17	440
# queries	1558	100	284	100	1942
<i>Annotation Statistics</i>					
# bullets	14.8k	996	2728	980	19.5k
# tokens	575k	36.6k	106k	42.3k	760k
# code steps	5086	346	948	-	6380
# code lines	3.0M	208k	555k	-	3.7M

Top 5 topics:

-
1. Financial
 2. Sports
 3. Healthcare
 4. Temperature
 5. Education

Experiments

	Method	# para.	Code gen	Help.	Entail.
<i>TableQA</i> <i>Baselines</i>	TAPAS	337M	✗	16.50	3.67
	TAPEX	406M	✗	9.00	3.50
<i>Prompt- based LLMs</i>	ChatGPT	20B [†]	✗	13.50	2.07
	GPT-4	175B [†]	✗	20.50	4.36
	ChatGPT	20B [†]	✓	21.38	2.59
	GPT-4	175B [†]	✓	43.92	3.26
<i>Finetuned LLMs</i>	SFT	6B	✗	11.33	2.65
	SFT	6B	✓	9.83	4.47
	RLHF	6B	✓	7.51	3.13
	FG-RLHF	6B	✓	12.50	5.98

- Evaluation: helpfulness evaluated by LLM; entailment with ground truth
- Code generation helps data analysis
- Data analysis capabilities can be effectively distilled to 6B LLMs
- Fine-grained RLHF further improves performance on top of SFT

Thanks!



Scan to check our project page!