Discrete Information Retrieval: Search over Structured/Unstructured Data with LLMs

Pablo Martín Rodríguez Bertorello, Jean Rodmond Junior Laguerre

Conversational Assistants , Stanford University

Stanford Computer Science

Problem

Traditionally, **Dense Information Retrieval** relies on fine-tuned language models for document ranking. Alas, the result is both low **Query Recall**, because of the approximate nature of the methods employed. And, with the advent of Large Language Models, increasingly specific user queries result in low **Query Precision**.

Solutions like **ColBERT** represent text with dense embedding vectors. Each query is scored against an index of documents. To minimize compute complexity, a late interaction model is employed. However, the underlying **Approximate Nearest Neighbors** algorithm fails to take advantage of implicit text regularities in documents, where it exists.

Enter **Discrete Information Retrieval**, which takes advantage of text regularities, for example in product descriptions.

It achieves high **Query Precision** and **Recall**, by utilizing LLMs both for text discretization and SQL Semantic Parsing, as validated in 33 different domain datasets.

Stanford University

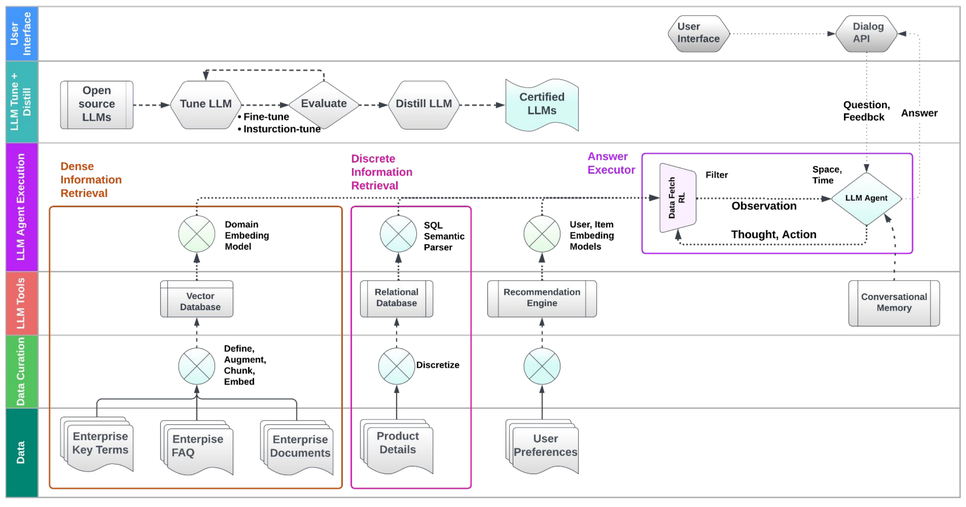
Background

Dense Information Retrieval:
- **Approximate Nearest Neighbors** Does not Understand Implicit Text Regularity
- Has **Low Query Recall**.
- Implies **Low Query Precision**:
  - Yelp dataset reflects the past
  - Users unable to search for exactly what they want

Structured and Unstructured Query Language (**SUQL**):
- Generates SQL DB fields
- Discovers **ENUMS**
- **Summary** and **Answer** Functions

Methods

Discrete Information Retrieval:
- Like **SUQL**:
  - Optimizable **Discretize** generates SQL DB fields
  - Optimizable **ENUM**
  - **Semantic Parsing** translates natural language into SQL
- Unlike **SUQL**:
  - A data pipeline rather than a query language
  - No dense IR (i.e. **ColBERT**)
  - External **Answer** explicitly **Reinforcement Learns to Reason and Act (ReAct)**



Experiments

- Created novel **Discretize** LLM prompt capable of generating categorical information from free text
- Experimented specific and exploratory user query, responses and corresponding **Dialog State**
- Observed method effectiveness across 33 different ecommerce product domains
- Estimated relative LLM performance: GPT 3.5, Palm2, GPT 4.0
- Tuned **ENUM** function for domain and cross-domain grounding
- Evaluating dataset query Recall and Precision, relative to **ColBERT**
- Prototyped multi-hop hybrid **Answer** module with dense and discrete data source tools

Analysis

The **Discretize** function needs to be optimized to generate database fields. An LLM operating over a dataset with text datasets rich in regularity, such as product descriptions, may discover thousands of fields per domain. This appear's to exceed what may be expected from **SUQL**'s **Summary** function. It could easily over-run any SQL database's maximum number of columns. As database tables are map-reduced, **ENUM** values are identified. Their size, at **Semantic Parsing** time, could itself out-run LLM maximum input size.
Therefore, **Discretize** should be considered on a per-domain basis. This requires prompts to ensure grounding across domains.

Conclusion

Discrete Information Retrieval is proposed, which outperforms traditional **Dense Information Retrieval**. This is for datasets including dense texts that embody field regularity.
This makes multi-hop search over structured and unstructured data via a data pipeline, without modifying of the SQL standard. It is recommended to use an **Answer** function external to query tools, capable of hybrid multi-hop **Reasoning and Acting** across them.

References

SUQL: Conversational Search over Structured and Unstructured Data with Large Language Models, arXiv:2311.09818
ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT, arXiv:2004.12832

Discrete Information Retrieval: Search over Structured/Unstructured Data with LLMs

Pablo Martín Rodríguez Bertorello, Jean Rodmond Junior Laguerre

Conversational Assistants , Stanford University

Stanford
Department Name

Sample Data

{**"title"**: "Wildcraft Toss 39 Ltrs Black & Red Backpack",
"price": 1080,
"uri": "https://www.bigco.com/MP001000011879890",
"product_id": "8903338115243-Wildcraft",
"**description**": "Bag Type : Backpacks, Capacity : 39litres, Closure Type : Zip, Collection Name : Toss, Color : Black & Red, Height : 15cm, Length : 47cm, Material Type : Polyester, No of Compartments : 2, Size : Medium, Strap Type : Adjustable, Width : 33cm, styleNote: Add a quotient of style to your casual attire by carrying this unisex from Wildcraft. It is fashioned from premium quality polyester and is sure to last for a long time. This backpack with a color block pattern flaunts spacious compartments that can accommodate your valuables easily. This black and red backpack comes with an adjustable strap and a zip closure that makes sure your essentials are safe and secure.",
"brand": "Wildcraft"}

Sample Direct User Query

User Query: what non-black 15 liter under $400 bags do you have?

Dialog State:
{"product_capacity = '15 liters' AND product_color != 'black' AND price < 400",
"product_capacity = '15 ltrs' AND product_color != 'black' AND price < 400",
"product_size = '15 ltrs' AND product_color != 'black' AND price < 400"},
Query Response:
[{**'id'**: ZEUSHBLU-Aristocrat',
'price': 324.0,
'title': 'Aristocrat Zeus 15 Ltrs Blue Small Backpack',
'closure_type': 'zip',
'height': '48',
'length': '31',
'material_type': 'polyester',
'product_brand': 'aristocrat',
'product_capacity': '15 ltrs',
'product_collection': 'zeus',
'product_color': 'blue',
'product_size': 'small',
'product_type': 'backpack',
'strap_type': 'adjustable',
'sub_domain': 'backpacks-men.json',
'width': '15.5'}]

Sample Exploratory User Query

User Query: retirement gift for dad

Dialog State:
{"ideal_for = 'gifting' AND product_occasion = 'retirement'",
"occasion = 'retirement' AND gender = 'dad'",
"product_category = 'retirement gift' AND gender = 'dad'",
"product_type = 'retirement gift' AND gender = 'dad'"}
Query Response:
[{**'id'**: 'TATACLIQ-PRIMARY-D243-Casio',
'price': 2247.0,
'title': 'Casio WS-2100H-1AVDF Youth Unisex Digital Watch',
'dial_color': 'grey',
'dial_shape': 'round',
'functionality': 'chronograph',
'model_number': 'ws-2100h-1avdf',
'movement': 'quartz',
'product_collection': 'youth',
'product_gender': 'unisex',
'product_type': 'watch',
'special_features': 'water resistant',
'strap_color': 'black',
'strap_type': 'rubber',
'sub_domain': 'watch-men.json',
'watch_type': 'digital',
'water_resistance': '100 m'}
]