GROUP 61

</> Digital Library

IR Final Project Presentation

DIGITAL LIBRARY: INFORMATION RETRIEVAL ON ACADEMIC DOCUMENTS

CSE508_Winter_2024

Problem statement and motivation

```
Relevant_Users {
 Students: Students often need access to research papers to complete assignments;
 Researchers: Academia requires you to have extensive knowledge of your field;
 Public: One needs relevant results to properly educate themselves;
}
Problem_Statement {
 Irrelevant_Papers: We often find unrelated and irrelevant papers while searching for
                    particular topics even abstracts can be confusing and misleading!;
 Information_Overload: It is difficult to digest and seek relevant information out of
                       lengthy documents;
}
Motivation {
 Users: We are the target users for this problem, our group consists of undergraduate
        researchers and students as well;
 Interest: We are keen on solving the problem at hand applying new technologies like LLMs
           along with hybrid information retrieval systems;
}
```

Literature Review

SPARSE MEETS DENSE: A HYBRID APPROACH TO ENHANCE SCIENTIFIC DOCUMENT RETRIEVAL

(HTTPS://ARXIV.ORG/PDF/2401.04055.PDF)

01

METHOD

We Combine Sparse and Dense embeddings to retrieve relevant documents

Sparse Retrieval Method: BOW embeddings +TF-IDF token weighting and cosine similarity to judge document relevance.

Dense Retrieval Method: Specter-2, a DL model used to create dense embeddings on scientific documents

02

RESULTS

By weighing the similarities and combining them we are able to create a new hybrid similarity score which in the paper, showed higher better results than just using the methods separately leading to better results

03

PROPOSAL

We believe this new hybrid approach is better and want to apply to our system to obtain the best results. The papers applicability is also high as it is in the same domain of scientific text retrieval.

Literature Review

IMPROVING TAG CLOUDS AS VISUAL INFORMATION RETRIEVAL INTERFACES

(HTTPS://ARXIV.ORG/ABS/2401.04947)

01

METHOD

The study used a large sample of 218,063 URLs tagged with 242,349 tags by 111,234 users, downloaded from del.icio.us bookmarking tool. Tag similarity was measured using the Jaccard coefficient of relative co-occurrence between tags. An alternative tag selection method based on tag usefulness was proposed, considering the tag's capacity to represent resources, the volume of covered resources, and its ability to cover less-covered resources. The bisecting K-means clustering algorithm was applied to the tag similarity matrix to group tags for the visual layout.

02

RESULTS

The proposed method selected more relevant and discriminative tags compared to traditional frequency-based selection, reducing semantic density and increasing topic diversity in the Tag-Cloud. The clustering-based layout grouped semantically similar tags together, allowing users to infer semantic relationships from the neighbors' proximity. While synonym and plural form issues remained, the authors suggested a potential solution using a predefined similarity threshold. The approach aimed to improve browsing experience and enable hierarchical navigation through sub-Tag-Clouds for specific topics.

03

SHORTCOMING

The described system assumes that tags are input by visitors/users of the system. However, issues related to the accuracy of the tags obtained with the above design exist. If the number of people using the system is few, then an insufficient amount of tags will be generated. Moreover, the semantic meaning of these tags is not judged. Thus, making it unfit for application in our system

Proposed Solution

We want to build an Information Retrieval System that improves the User experience of searching for academic papers

- Easy to use search engine UI, the results will be ranked by relevance to the users query.
- Providing a summary along side the paper so that we can view and digest the information at a glance.
- Giving user visual aids such as word clouds so that the most relevant words can be extracted and viewed.

Novelties in the Proposed Solution

Summaries generated

Abstracts can often times be misleading and may not provide the whole picture while reading a research paper. Thus, having a summary generated on the whole text which provides information about the results as well is mega helpful to the users

Word clouds for relevant words

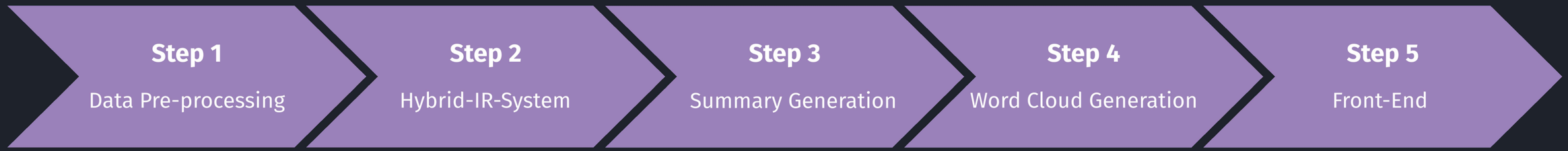
We assume that our system can have problems sometimes giving relevant results as most systems are dependent solely on the query text provided by the user. At a glance our user can look at the word cloud and see if the paper has relevant features pertaining to them

Refined Hybrid Retrieval System

Implementation of this new age Hybrid IR system which has dense DL derived dense embedding along with traditional BOW embeddings lead to much better results

SUBHEADLINE

Methodology and Timeline



Technologies Used:

Python: Data-pre-processing, Backend of system

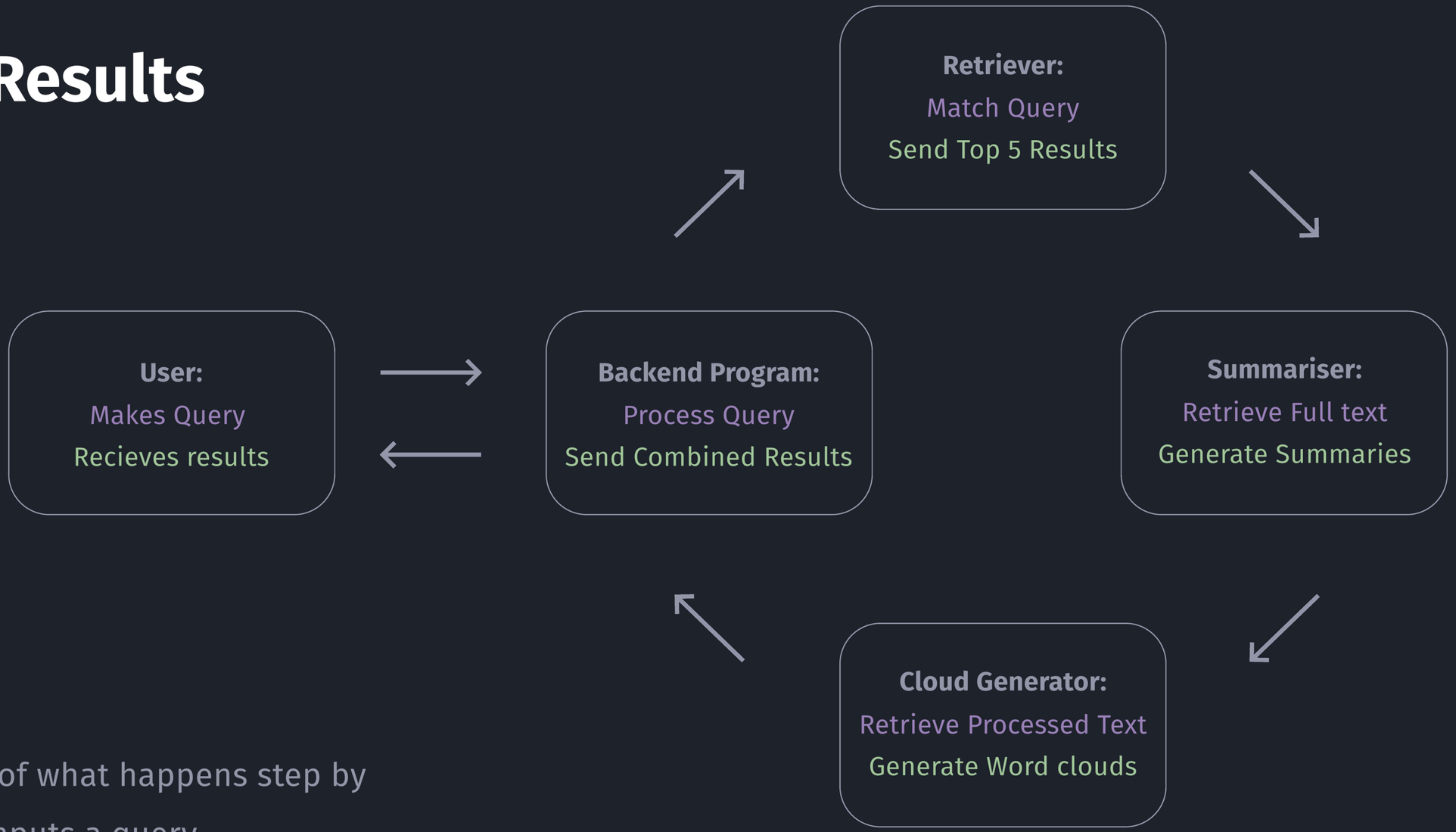
Vertex Ai: API used to convert text of documents

Django: Framework used for development

CSS, HTML: Used to design our front end

A LOOK INTO

Query to Results



A Process diagram of what happens step by step when a user inputs a query

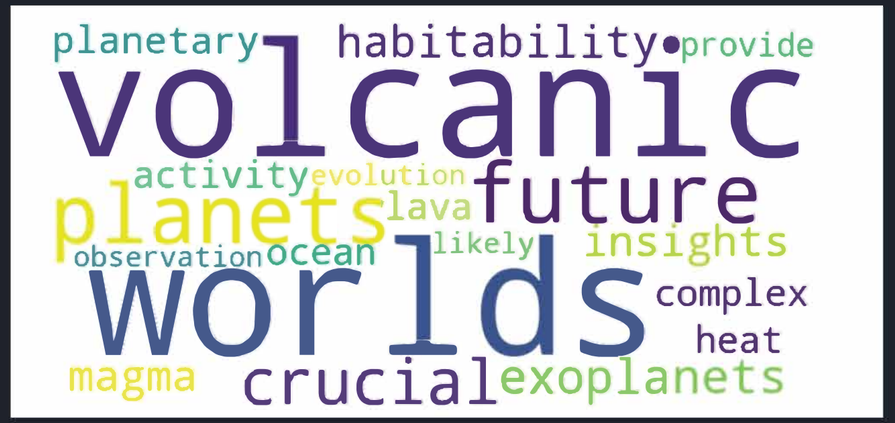
Results

```
{'1703.02156': 0.658124469158372,
 '2312.06528': 0.6259806352933099,
 '2103.04893': 0.6148378161536002,
 '1705.10363': 0.5972843489759834,
 '1709.04609': 0.591434515281865}
```

RETRIEVAL SYSTEM

Here are the hybrid similarity scores and the relevant documents generated by the sample query, "Machine Learning in computer vision".

Firs



WORD CLOUD

Generated word cloud for the paper "Highly Volcanic Exoplanets, Lava Worlds, and Magma Ocean Worlds", this helps us see that we are discussing the habitability of these volcanic exoplanets

Example Generated summary:

Summary of "Highly Volcanic Exoplanets, Lava Worlds, and Magma Ocean Worlds

Introduction: This paper argues that highly volcanic exoplanets, encompassing lava worlds, magma ocean worlds, and super-Ios, are crucial targets for future research. These planets are likely common and offer unique advantages for observation due to their proximity to their stars and bright infrared signatures. Studying them will provide insights into planetary formation, geodynamic processes, habitability, and the composition of planetary interiors.

Key Science Opportunities and Observations:

Energy Sources: Volcanic activity is fueled by internal heat from radionuclides and tidal forces, and external heat from intense stellar radiation. This creates diverse volcanic environments with varying compositions, temperatures, and eruption styles.

Observational Opportunities: Techniques like transit spectroscopy, phase curve analysis, and direct imaging can be used to detect volcanic plumes, magma oceans, and lava lakes. JWST and future telescopes will be instrumental in characterizing these worlds.

Habitability: Volcanism plays a complex role in habitability. It can provide essential volatiles for life but also create extreme conditions. Understanding this interplay is crucial for assessing the potential for life on these planets.

Relevance to Other Worlds: Studying volcanic exoplanets offers insights into the early history of Earth and other rocky planets, which likely experienced a magma ocean phase.

Dynamical Role: Intense volcanic activity can influence a planet's rotation and orbital evolution, impacting its long-term stability and climate.

Results and Future Directions:

The paper highlights the need for:

Continued observation and characterization* of volcanic exoplanets with current and future telescopes.

* Dedicated missions like the Io Volcano Observer to study volcanic moons in our solar system as analogs.
* Theoretical models to understand the complex interplay of volcanic activity, planetary interiors, atmospheres, and orbital dynamics.
* Cross-disciplinary collaborations between exoplanet scientists, geologists, and volcanologists.
* Public engagement to raise awareness of the exciting discoveries in this emerging field.

The authors envision a future where exoplanet volcanology becomes a thriving field, providing crucial insights into the diversity and evolution of planets beyond our solar system.

Thanks for sticking around :)

MEMBERS:

Aditya Daipuria

Ankit Kumar

Harshit Sharma

Mohammed Kaif

Navvrat Rao

Pavit Singh

GROUP 61 CSE_508_INFORMATION_RETRIEVAL