National University of Singapore
Institute of Systems Science

ISY5001
Academic Knowledge Platform

Project Proposal

Group 10

Bian Weizhen A0285814W
Goh Minhua A0285810A
Li Jiacheng A0285823W
Mao Zhihong A0285799X

Supervisor: Guzhan

FILING REF:

PROJECT PROPOSAL

Date of proposal: Oct. 2nd, 2023

Project Title:

"🥭 Mango" - An Academic Knowledge Platform

Group Members:

Goh Minhua A0285810A
Li Jiacheng A0285823W
Bian Weizhen A0285814W
Mao Zhihong A0285799X

Sponsor/Client:

As a young college student, Ken was eager and ready to plunge deep into the intricate world of PhD research. His passion was alight, but the path was obscured by a daunting volume of papers and journals spanning decades. Ken knew that to truly grasp his research field, he needed to digest its most influential and up-to-date papers. Unfortunately, there was no compass guiding his journey.

Ken spent countless nights scrolling through databases, attempting to discern which papers were the keystones, the ones that would give him the insight he craved. It wasn't just about saving time; it was about ensuring he was on the right trajectory from the start.

He craved a resource offering:

- **Broad Coverage:** To explore diverse fields.
- **Timely Updates**: Keeping her informed of cutting-edge studies.
- **Intuitive Interface:** Easy access and organization.
- **Expert Curation:** Highlighting essential papers.
- **Collaboration Tools:** Connecting with peers and mentors.
- **Summaries:** Concise paper insights.
- **Customization:** Tailoring content to her evolving interests.

Ken's quest led him to an AI-driven research platform, empowering him with the tools she needed for his academic journey. It is not only a dedicated platform, but also a virtual guide, which would streamline the most recent and pivotal papers right at his fingertips. The platform would be more than just a search engine—it would be a beacon for young scholars, a curated library of milestones essential for grounding and innovation in any research field.

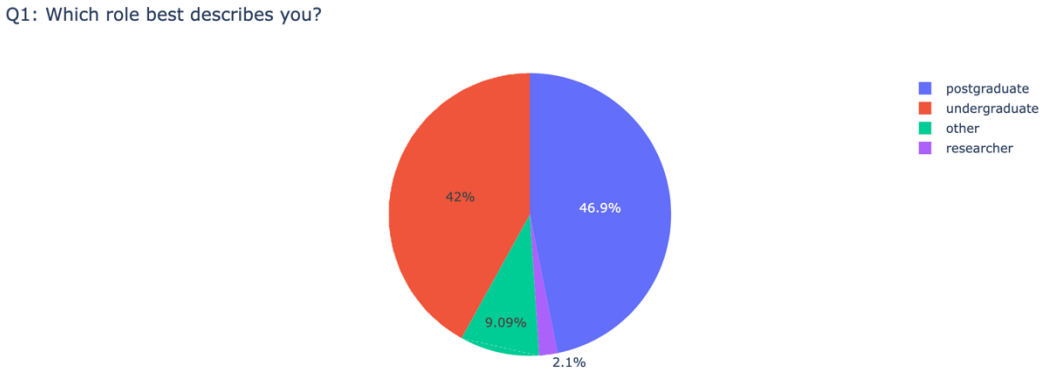
Such a tool, Ken believed, would not just assist his individual quest, but empower an entire generation of researchers to make the most out of their academic pursuits, ensuring no seminal paper went unread.

Background/Aims/Objectives:

Survey & Requirements Identification

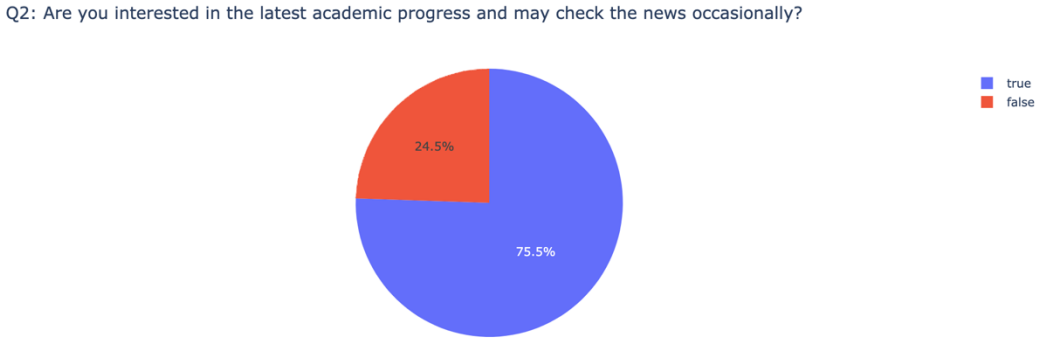
To gauge the interest and needs of the academic community regarding a dedicated platform for academic essay searching and recommendation. By understanding users' current habits, preferences, and challenges, we aim to design a tool that efficiently caters to their research demands and enhances their academic experience.

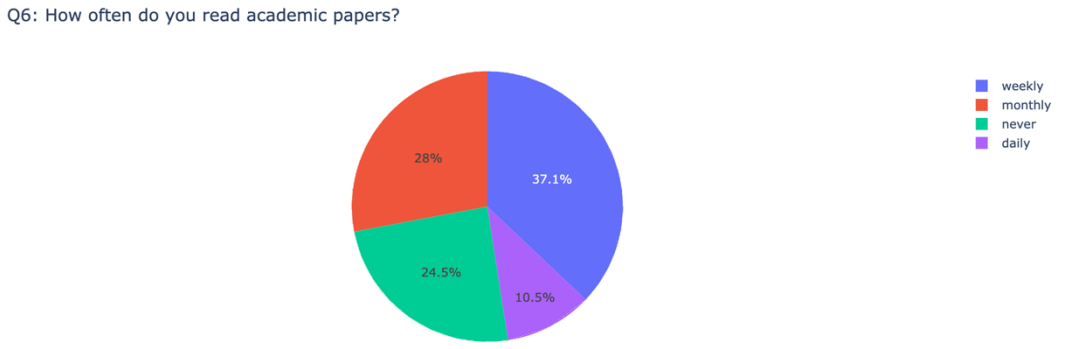
User Demography & Role



- The survey starts by determining the user's position: undergraduate, postgraduate, or researcher.
- This helps understand the academic stage and research needs of the user.
- We can tell from the result that our ***target users*** should be ***college students***.

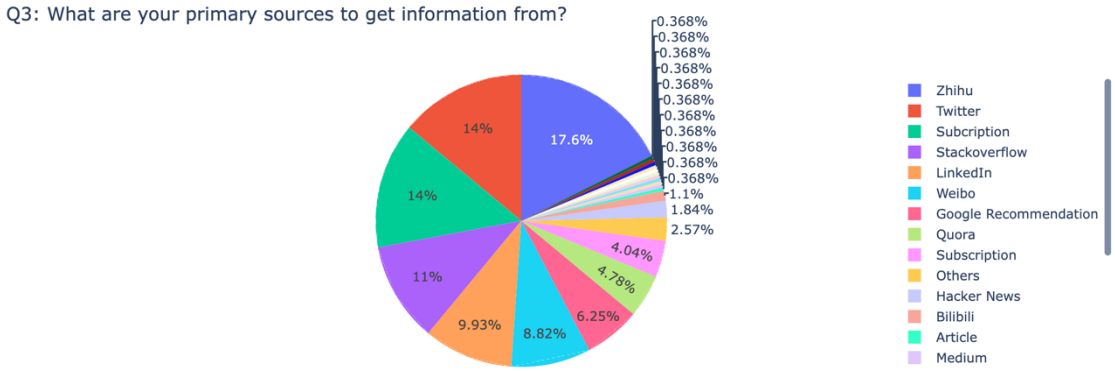
Frequency & Interest in Academic Updates

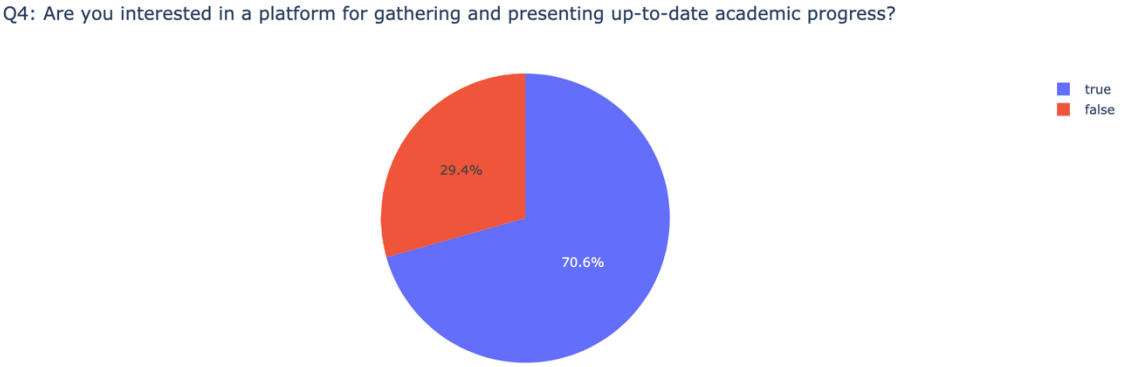




- Questions about the frequency and interest in the latest academic progress help gauge the general inclination towards being updated.
- This is fundamental for a platform focusing on up-to-date information.
- More than ¾ people state they are used to reading papers and tend to keep themselves updated on academic progress.

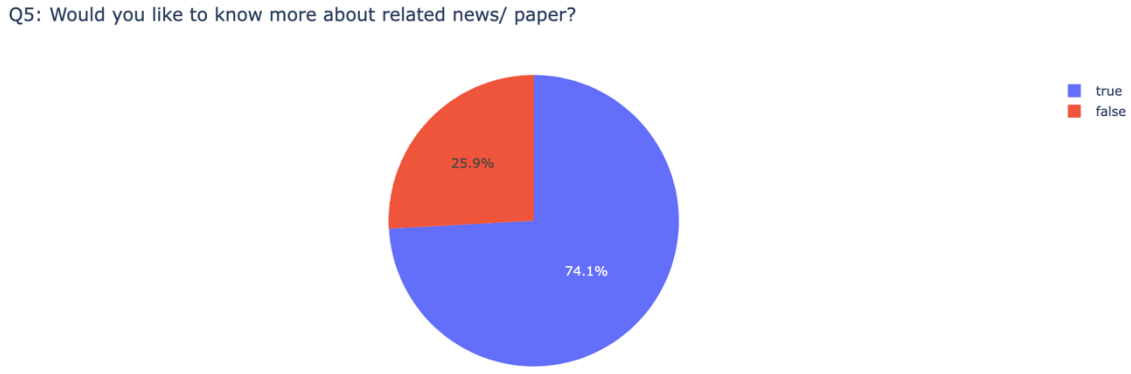
Current Information Sources

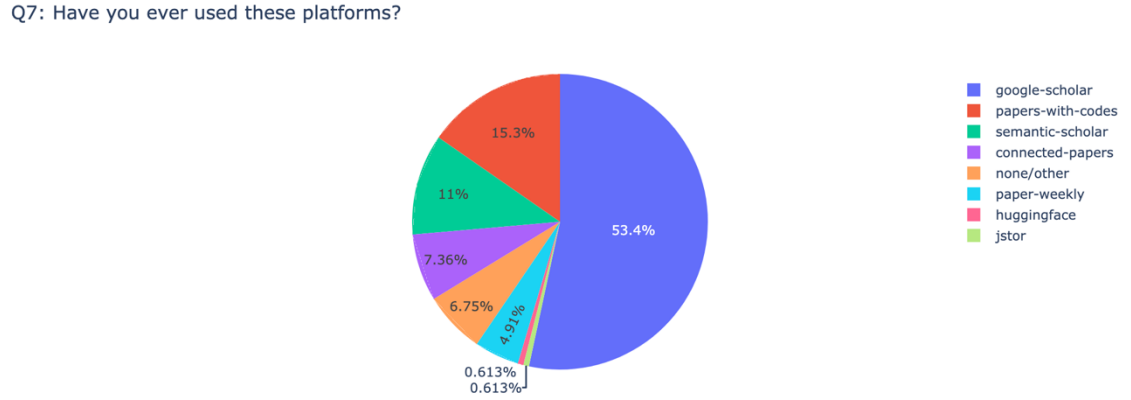


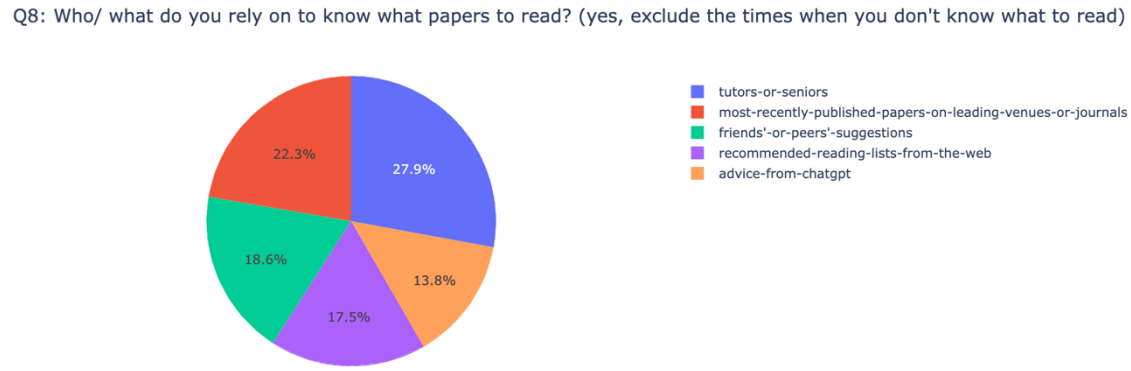


- Identifying the primary sources of information helps to ascertain current platforms' popularity and their possible shortcomings.
- Currently, students and researchers do not have a unified media platform for latest academic information, progress, and news. Most of our interviewees are interested in having such a platform.

Platform Utility & Interconnectivity







- Queries regarding interest in a platform for gathering recent academic developments and connecting related papers gauge the demand for a centralized hub for academic knowledge.
- According to the Q7 chart, we can find users usually spend much time in searching essays, but there is no such platform to provide customizable recommendation of academic readings to help them explore a certain research field.

The survey results resonated with our initial expectations. The overwhelming interest and feedback from users strongly advocate the necessity and viability of an academic essay searching and recommendation platform.

Objective

Mango is an ***academic knowledge*** platform for college students and researchers.

It has two main components -

- A ***Search Engine*** for essays.
- An ***Academic Papers Recommender***.

The platform aims to assist ***young college students*** and ***researchers*** to quickly grab a general understanding of a particular field and keep themselves up to date on the forefront development of that field.

Market Research - Competitor Research

	Techniques	How they employ
Semantic Scholar	Use the articles to analyze your preferences.	If the user hasn't searched by the engine, the website will ask you to input the area you are interested in. Otherwise, the model will just recommend the latest work or influential thesis your interested area.
Google Scholar	Determine relevance using statistical models	Models required topics of your articles, the places where you publish, the authors you work with and cite, the authors that work in the same area as you and the citation graph.
Papers With Code	Recommend the latest trending research in Computer Science.	Since PWC only contains code-related works, there's no obvious recommendation system applied.
Connected Papers	Knowledge Graph.	By a graph including Citation, Prior Works and Derivative Works.
Paper Weekly	Trending research and Personalized Thesis in AI area.	Multi-functional Recommendation by the users' history.

Potential Application

[Research Management] NUS students and professors can rely on our platform to aid in literature review by finding papers and information more easily. Some experts working in NUS research facilities may find relevant information to be scarce. With the search engine, it will only shrink the circle down to information relevant to their own fields.

[Educational Resources] NUS professors can curate educational resources, such as lectures, tutorials and guides, by reviewing academic papers and sometimes from academic news recommended to them, translating their research to NUS students' education in the classrooms.

[Research Funding Opportunities] NUS researchers belonging to facilities like NUS CRISP can look up about papers and receive news relevant to remote sensing to integrate information to craft their next upcoming research grants and funding opportunities.

Project Descriptions:

Literature Review

It is understood that some well-known search engines like Google Scholar have their proprietary algorithms kept a secret while others like Google's and Yahoo's are out in the open for the public to be studied. Since our platform closely resembles that of Google Scholar, one still should know as much as possible as to how much its algorithm works, albeit the complete algorithm is not public.

Google Scholar weighs heavily on words in the title, author and journal names, on articles' citation counts and takes into consideration of words directly included in an article but not synonyms of those words [1]. Google's search engine's first and most well-known ranking algorithm was PageRank, in which the number of links pointing to the web page influences its rank and links coming from high-quality sites are given more weight [2].

Yahoo's search engine utilizes Core Ranking function, which filters between good and bad URLs with Gradient Boosting Decision Tree, along with Logistic Loss [3].

More well-known recommendation systems like Google News developed a Bayesian framework for predicting users' current news interests from activities of the user like past click behavior and the news trends as seen from user's activity. This framework of information filtering is combined with collaborative filtering mechanism to generate its recommendation system [4].

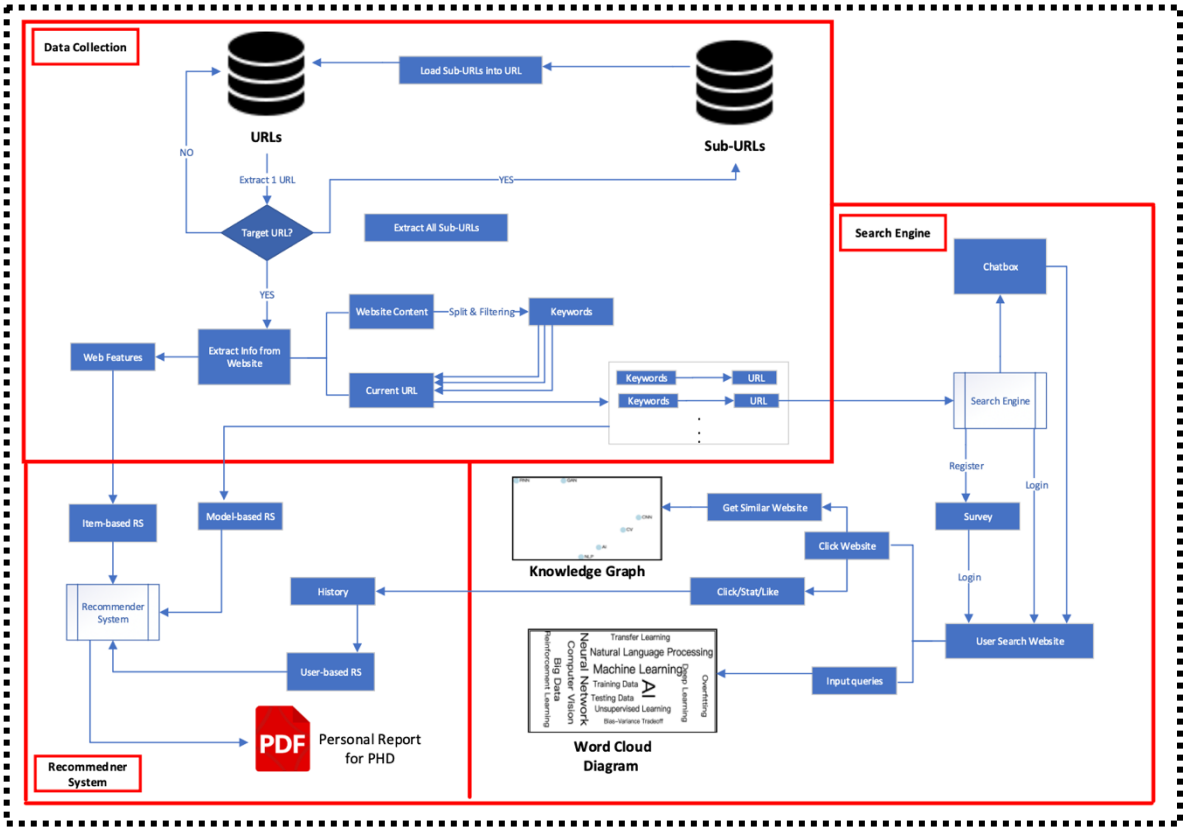
GroupLens, another news recommendation platform, adopts collaborative memory-based algorithm to gather the ratings of users and to predict scores of news articles for individual users, based on the heuristic rule that users who agreed in the past will probably agree again [5].

Amazon book recommendation website system also utilizes item-to-item collaborative filtering method to analyze users' book purchases and then recommend books purchased by other customers based on similar book(s) between the two customers have purchased. Other parameters like topics of interest, demographic characteristics aid to suggest books which the user may like [6].

Digg, another news curating platform, makes use of collaborative filtering to find interesting similarities between users and their search history, then recommending similar users' favorite articles to target user [7].

System Overview

The system collects academic articles, maintaining a capacity-limited URL pool for efficiency. It enhances Google Scholar with personalized searches based on user interests and activity, integrating features like word clouds, knowledge graphs, and an assisting chatbot. Additionally, a recommendation system offers tailored article suggestions and allows users to generate a PhD recommendation report, leveraging user and item-based algorithms.



Algorithm/Technique	Explanation
Web Scraping and Data Collection	Breadth-first or Depth-first search for navigating URLs efficiently.
Recommendation System	Hybrid recommendation combining content-based (using keywords, year, abstract) and collaborative filtering (likes, favorites).
User Profiling	Algorithms analyzing user activity and preferences, like K-means for segmenting users based on interaction patterns.
Word Cloud Generation	TF-IDF (Term Frequency-Inverse Document Frequency) for highlighting significant words.
Knowledge Graphs	Graph databases and algorithms, leveraging technologies like Neo4j, for relations between entities and concepts.
Chatbot	Sequence-to-sequence model with attention mechanisms for accurate and context-aware responses in an academic context.
Filtering and Interactivity	Decision trees or rule-based systems for user-specific website filtering.

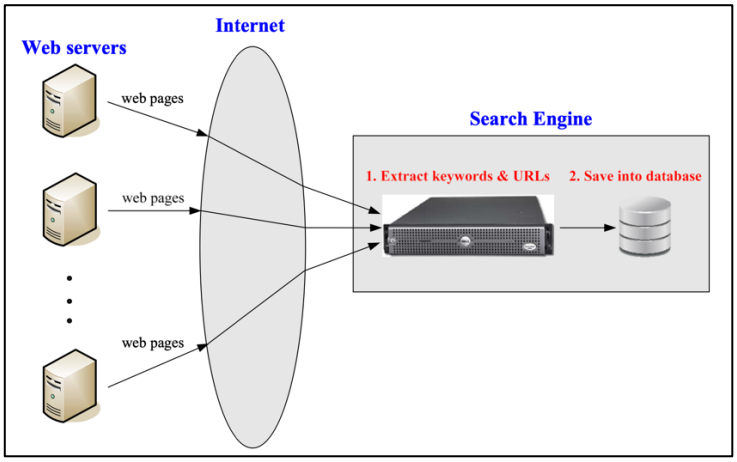
Data Collection

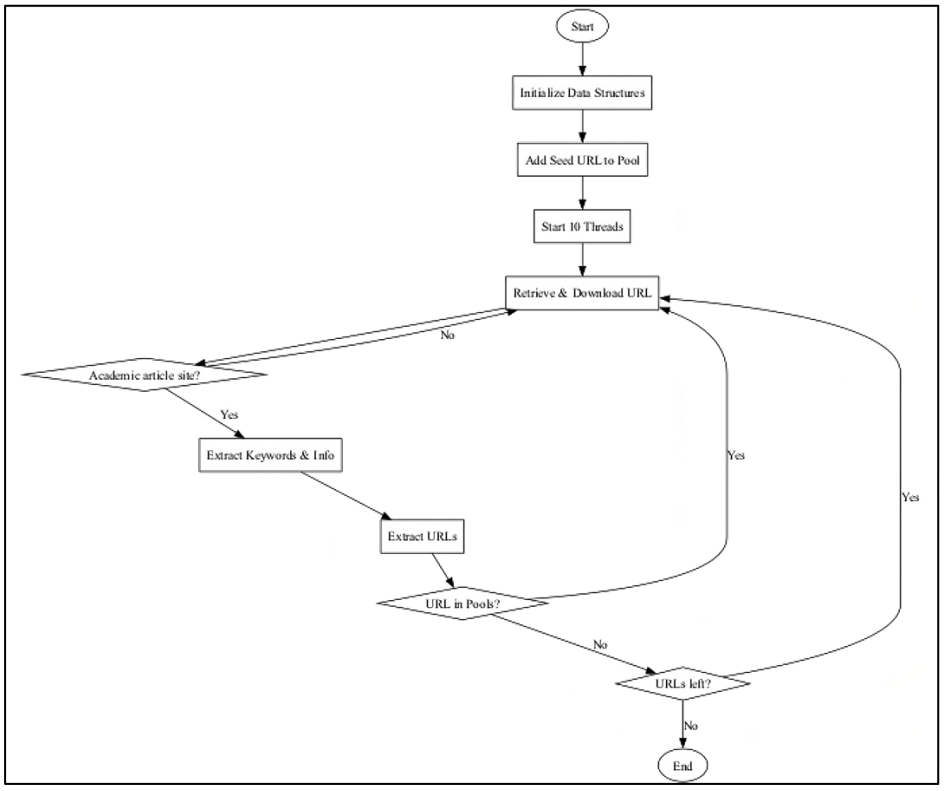
(1). Web Scraping

The process begins with the initialization of data structures and adding a seed URL to the pool. Ten threads are then initiated simultaneously for data capture. Each thread retrieves and downloads a URL.

- If the downloaded site is recognized as an academic article site, the system extracts keywords and other relevant information like year and abstract. Subsequent URLs found on the page are extracted. For each extracted URL, the system checks if it's already in the pools.
- If not, it is processed further. This loop continues until there are no URLs left to process.

The process concludes once all URLs are processed. The procedure is shown as follow:





(2). User Profiling

This survey will be initiated to users when they sign up for an account with our platform. This is to narrow down the scope of the recommendation system to users' interests as much as possible before relying on the feedback received from using the search engine. These questions will be in the form of multiple options for users to select. The link to the complete survey is provided below.

1. What is your primary field of study or academic interest? Are there any specific subfields or topics within your primary field that you are particularly interested in?
2. What are the industries you are interested in?
3. Are there specific academic journals or conferences that you prefer or frequently read?
4. Do you prefer papers published within a specific timeframe?

https://drive.google.com/file/d/128V3_tbBkTgJTHkU6k5EPkQXTDsgzT20/view?usp=drive_link

(3). User Activities - Implicit Feedback

Frequency Analysis:
Parameter required: Frequency of searches and page reads.
Analyze the frequency of searches and page reads related to specific topics, authors, or keywords. Higher frequency indicates higher interest in those topics or keywords.

Time Spent:
Parameter required: Time spent on pages or search results.
Analyze the amount of time users spend on different pages or viewing search results. More time spent can imply more interest or engagement with that content.

Scroll Depth:
Parameter: Depth of page scrolling.
Analyze how far users scroll on a page. Deeper scrolling may suggest more interest in the content.

Revisits:
Parameter required: Number of revisits to a particular page or topic.
Count the number of times a user revisits a particular page, topic, or keyword. Frequent revisits may suggest the user finds the content valuable or of interest.

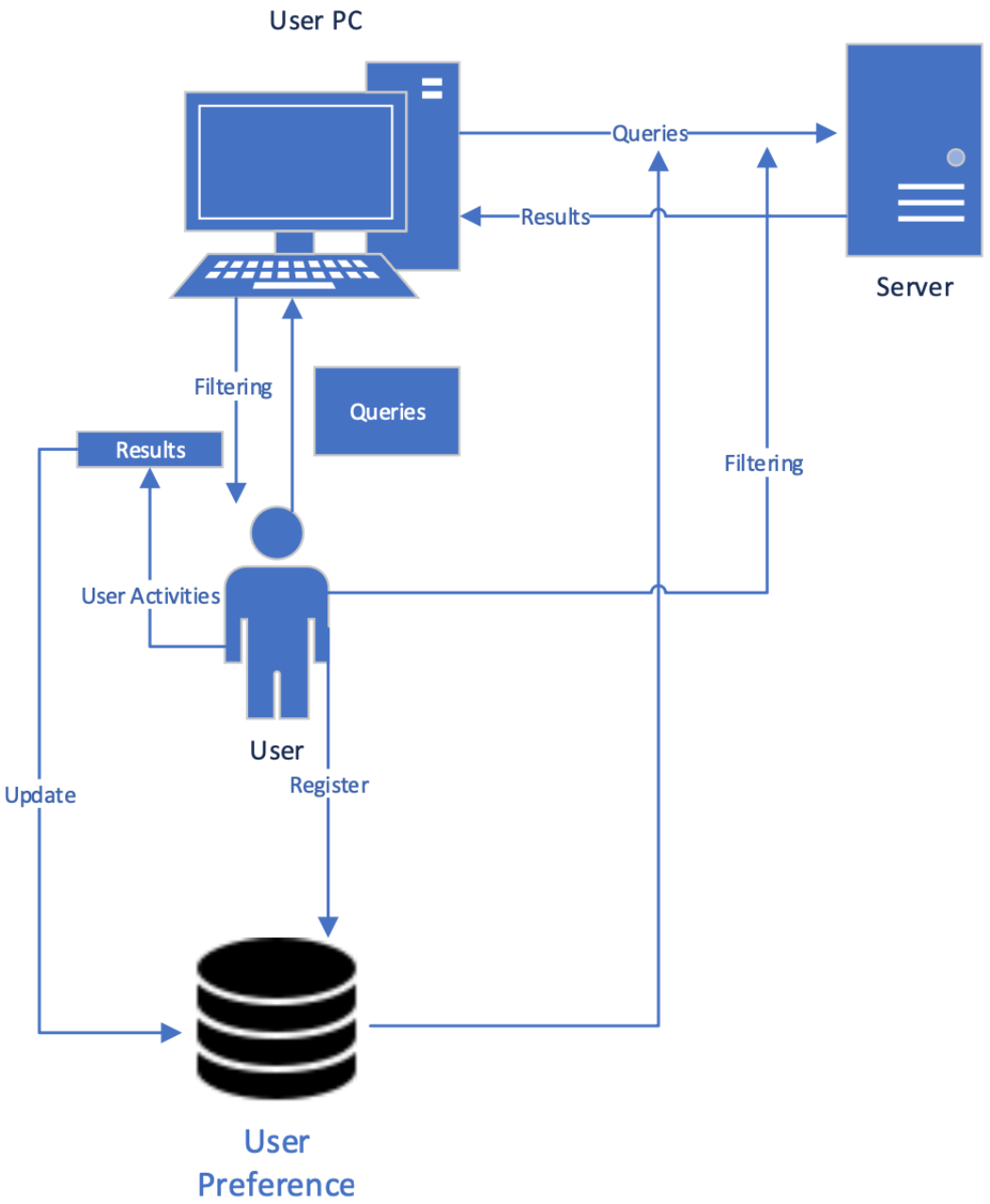
External Links:
Parameter: Clicks on external links or references.
Track clicks on external links or references within the content. Clicking on external links can indicate interest in exploring related content or topics.

Search Engine

The user begins by executing a search, and in response, the search engine displays the relevant results. Recognizing the importance of a personalized search experience, our system allows the user to manually filter out certain websites from these results.

After narrowing down the results with these filters, the user can further interact by clicking on a result to delve deeper, expressing interest through liking a specific result, or saving a particular result to their favorites for future reference.

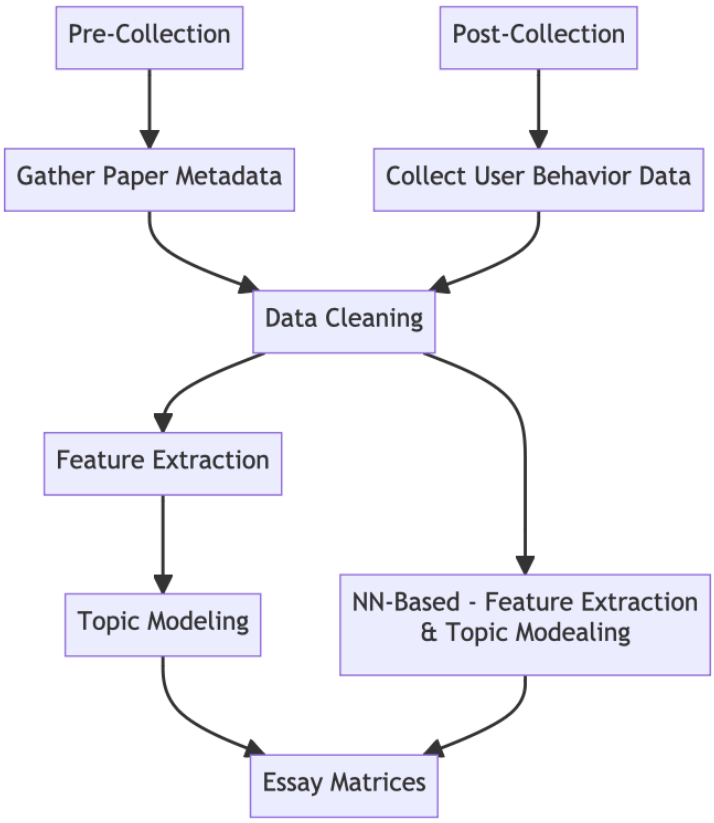
As these interactions occur, the system seamlessly captures them. This captured data is then processed in the backend, updating the database to reflect the user's actions and preferences, ensuring a continually improving and tailored user experience.



Recommender System

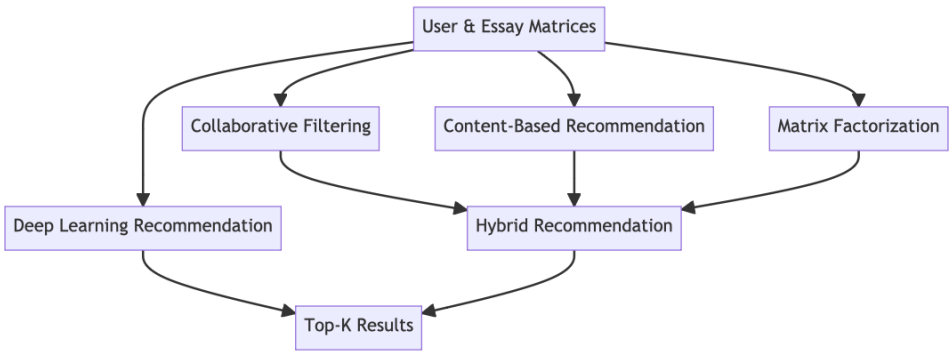
In this part, we will briefly elaborate how we plan to design our recommendation system –

Data Collection and Processing Process



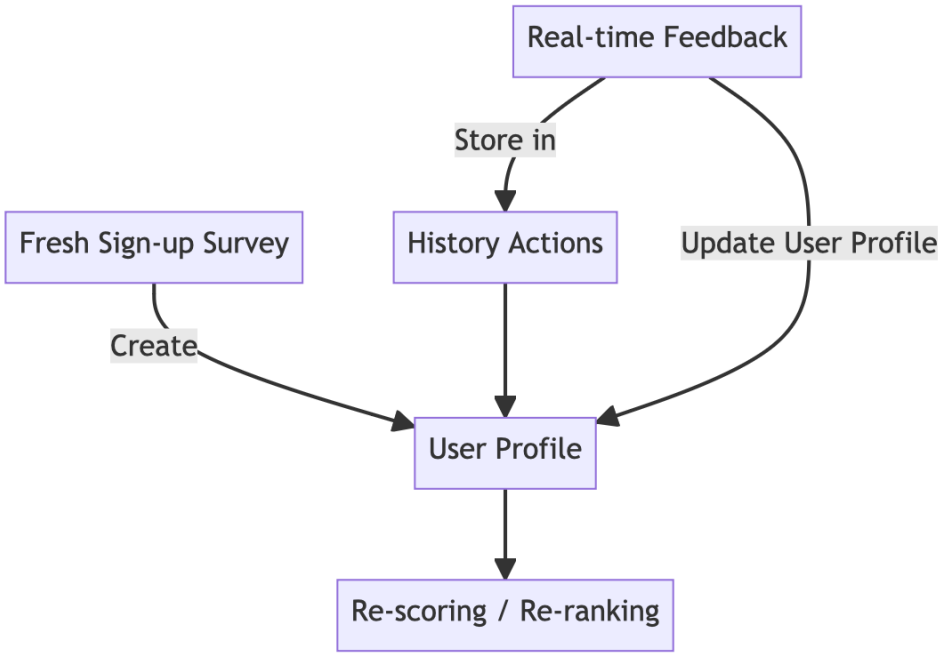
- **Paper Metadata**: Acquire details like title, abstract, keywords, authors, citation relationships, etc.
- **User Behavior Data**: Such as click-through rate, reading time, favorites, likes, etc.

Recommendation Algorithms



- **Collaborative Filtering**: Uses user's past behavior and preferences to predict what else they might like. There's User-Item and Item-Item collaborative filtering.
- **Content-Based Recommendation**: Recommends papers based on their content, like keywords, abstract, etc.
- **Hybrid Recommendation**: Combines the techniques of collaborative filtering and content-based recommendations.
- **Deep Learning**: Leveraging deep neural networks like Neural Collaborative Filtering （NCF）, DeepFM, etc.

Personalized Recommendations



- **User Profile**: Keep track of a user's research direction, browsing history, favorites, citations, etc., to create a detailed profile and enhance recommendation precision.
- **Real-time Feedback**: Allow users to give feedback on recommended papers, indicating "Interested" or "Not Interested".

System Features

- **Real-time Updates**: As new papers get published, the system should update the recommendation pool promptly.
- **Interdisciplinary Recommendations**: For cross-disciplinary research, recommend papers from related fields.

Challenges and Roadblocks

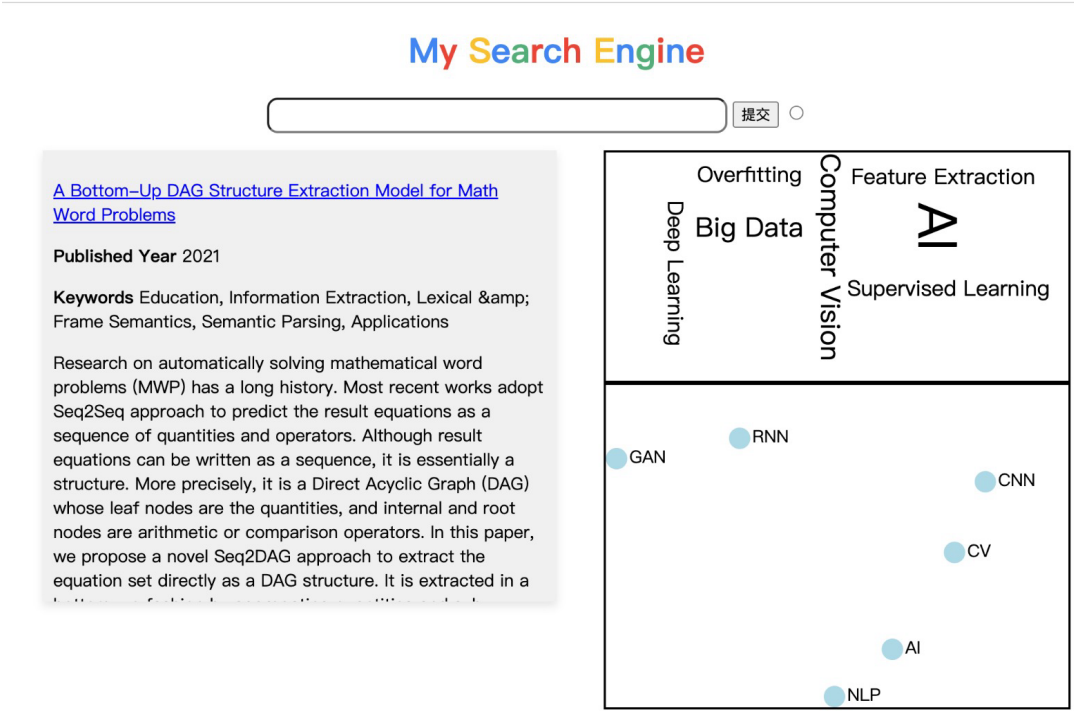
Challenge:

1. Chatbot's difficulty in secondary information extraction.
2. Sparse feedback due to the vastness of academic fields.
3. Limited data crawling scope.

Strategy:

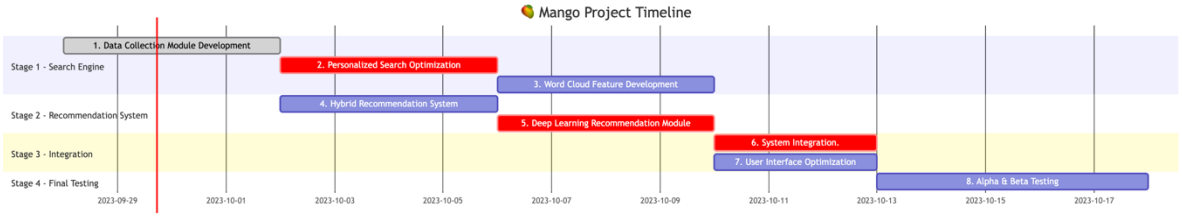
1. Enhance with advanced NLP techniques. Use feedback loops for iterative learning.
2. Diversify feedback channels and employ matrix factorization for sparse data.
3. Expand source websites and consider partnerships with academic databases.

UI Design



Future

Based on the tasks above and the estimated workload, we have drawn the following Gantt Chart to keep the development work on schedule:



References

[1] Beel, J., & Gipp, B. (2009). Google scholar's ranking algorithm: The impact of articles' age (an empirical study). *2009 Sixth International Conference on Information Technology: New Generations*. https://doi.org/10.1109/itng.2009.317

[2] Bar-Ilan, J. (2007), "Manipulating search engine algorithms: the case of Google", Journal of Information, Communication and Ethics in Society, Vol. 5 No. 2/3, pp. 155-166. https://doi.org/10.1108/14779960710837623

[3] Yin, D., Hu, Y., Tang, J., Daly, T., Zhou, M., Ouyang, H., Chen, J., Kang, C., Deng, H., Nobata, C., Langlois, J.-M., & Chang, Y. (2016). Ranking relevance in Yahoo Search. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939677

[4] Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. Proceedings of the 15th International Conference on Intelligent User Interfaces. https://doi.org/10.1145/1719970.1719976

[5] Y. Xiao, P. Ai, C. -h. Hsu, H. Wang and X. Jiao, "Time-ordered collaborative filtering for news recommendation," in China Communications, vol. 12, no. 12, pp. 53-62, December 2015, doi: 10.1109/CC.2015.7385528.

[6] Han, K. (2020). Personalized news recommendation and simulation based on improved collaborative filtering algorithm. *Complexity*, *2020*, 1–12. https://doi.org/10.1155/2020/8834908

[7] Shan Liu, Yao Dong and Jianping Chai, "Research of personalized news recommendation system based on hybrid collaborative filtering algorithm," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), Chengdu, 2016, pp. 865-869, doi: 10.1109/CompComm.2016.7924826.