# Meeting Notes

27 July 2024

|  |
| --- |
| Useful Links: |
| * [**GitHub**](https://github.com/tonyctyy/content-recommendation): This is the main GitHub repository of the project. We store all our information (e.g. logs of different research) here. Please check it regularly. * [**Perplexity**](https://www.perplexity.ai/collections/Content-Recommendation-FYP-r8AxwOpsSAyDXFh7Np00lg): Perplexity is used for research purposes, and we have a group chat. We can ask questions here. * [**Introduction to Content Recommendation Systems**](https://slogix.in/phd-research-topics-in-recommender-systems-based-on-deep-learning/): This website provides a brief intro to different content recommendation techniques, potential research directions, etc. We can kick off the project here. * [**Dataset for Content Recommendation System (GitHub)**](https://github.com/RUCAIBox/RecSysDatasets): This GitHub repo contains many different datasets for content recommendation systems. We can choose our dataset here.   # We can also find the datasets from [Kaggle](https://www.kaggle.com/datasets) or [Hugging Face](https://huggingface.co/datasets). |
| Agenda |
| * Define the objective of the project   1. Decide whether the focus will be on  i) study and practice different content recommendation techniques.  ii) solve some existing challenges/problems. iii) both.   2. Identify potential research directions.   3. How can we evaluate our work (we don’t have real time feedback)? * Define the deliverables   1. Discuss the scope and format of the research paper. (e.g. focus on Retrieval Stage)   2. Determine the scale and requirements of the content recommendation system. * Discuss the potential dataset and insights (optional)   1. Present the dataset analysis and insights gained from the exploration phase.   2. Discuss the suitability of the dataset for the project objectives. |

|  |  |
| --- | --- |
| Detail Information | |
| Content Recommendation System The Content Recommendation System includes different stages:  1. **Retrieval** (reduce results from trillions to thousands)     - It usually combines different methods (e.g., Collaborative Filtering, GNN) to retrieve results and add them up.  2. **Pre-Ranking** (reduce results from thousands to hundreds)     - This is where neural networks are used to predict evaluation metrics (e.g., click-through rate, like rate).  3. **Ranking** (a more complex mechanism to rank results compared to pre-ranking)  4. **Re-Ranking** (reduce results from hundreds to tens)     - It considers both scores and variety of the results. As a result, it sets up rules to redistribute similar results.     - It uses different sampling methods (e.g., MMR, DPP) to reduce results.     - It also adds advertisements and additional information to the results. Popular Techniques **Collaborative Filtering:** This is a widely used technique in the retrieval stage. The general applications are **Item CF** and **User CF**. We can use CF to construct multiple tunnels to retrieve related content effectively.    - **Challenges**      - It is difficult to find user behavior data in open-source datasets.        1. **Item-Item Collaborative Filtering:** This method only requires content data, making it suitable when user behavior data is unavailable.        2. **Synthetic User Behavior Data:** Another approach is to generate synthetic user behavior data from available content data (i.e. simulation data that maintains the original properties).  There will be many null data points when considering millions of items and users (e.g., User Rating to Movies Matrix). Measurements **Case Study of Social Media (e.g., Red):**  - **Key Performance Indicators:**    - Click-Through Rate: Clicks / Impressions    - Like Rate: Likes / Clicks    - Share Rate: Shares / Clicks    - Comment Rate: Comments / Clicks    - Finish Rate: (Scroll to the end / Clicks) × f(length of the content)  - **North Star Metric** (This is a more important metric):    - User Size: Daily Active Users (DAU), Monthly Active Users (MAU), etc.    - User Stats: Average Time Spent on Content (ATC), Average Number of Content Views (ANV), etc.    - Posting Stats: Average Number of Posts (AP), Post Penetration Rate (PPR), etc. Strategy Deployment 1. Offline Experiment     - This is the focus of the project. We can use the dataset to test the model.  2. AB Test     - We can't use the dataset to conduct AB tests as AB Testing requires a live environment for simulation.  3. Update Strategy Research Directions/Findings Please check [here](https://github.com/tonyctyy/content-recommendation/blob/main/logs/pre-meeting_brainstorm.md) for a more detail view of the 6 directions our Perplexity provides. Here are the most important ideas:   1. **Zero-shot or Few-shot Learning for Recommendations (i.e. Cold-Start Scenario)**   Research Question: How can we effectively recommend content to new users or for new items with minimal historical data?  Direction: Investigate techniques that can generalize from limited data to make recommendations in cold-start scenarios. This could involve:  - Developing models that can transfer knowledge from existing users/items to new ones  - Exploring meta-learning approaches for quick adaptation to new users or items  - Utilizing external knowledge bases or pre-trained language models to enhance recommendations with minimal data  # In fact, we must face the issue of cold-start so I think no matter what we do, we should have focus on cold-start.   1. **Explainable Recommendations with Limited User Data**   Research Question: How can we generate meaningful explanations for recommendations when user preference data is limited?  Direction: Explore techniques to provide transparent and interpretable recommendations. This could involve:  - Developing content-based explanation methods that don't rely heavily on user history  - Investigating how to leverage external knowledge sources for generating explanations  - Exploring visual or interactive explanation techniques for recommendations  # This is more related to our dataset. If we have a dataset with user behavior data, then we can use it to generate explanations and this will not be a problem. If the user data is limited, then we should study the content recommendation model/technique that doesn't depend on user history. Datasets  1. [Yelp Dataset](https://github.com/tonyctyy/content-recommendation/blob/main/logs/dataset/yelp.md)   Yelp is a platform for users to find, review, and recommend businesses in different categories (e.g. food, shopping, entertainment, etc.) across 8 metropolitan areas in the USA and Canada. This dataset is originally used for the Yelp Challenge and it is open for academic research.  # This dataset contains some user data which is very useful to apply many techniques. | |
| Action Items | |
| **Item** | **Owner** |
| Project Proposal (will further separate the tasks)    1. Define the objective of the project based on the discussion.    2. Define the deliverables (research paper, system implementation, etc.).    3. Finalize the dataset and document the insights and preprocessing steps. |  |
| Project Plan (will further separate the tasks)    1. Define the timeline with specific dates for major milestones.    2. Break down the project into tasks and subtasks.    3. Define the milestones and checkpoints for progress tracking.    4. Assign roles and responsibilities to team members (if applicable).    5. Identify potential risks and mitigation strategies. |  |
|  |  |
|  |  |
| Next Meeting | |
|  | |