RELATREE

PROJECT REPORT | CMPE 256-01: LARGE SCALE ANALYTICS

TEAM - RELATREEVS

AASHISH SUBRAMANIAN
CHINTAN VACHHANI
PAVANA SRINIVASADESHIKA ACHAR

INTRODUCTION

Duta Inc. is a company that provides services such as wiki search, news, stock information and so on through the Whatsapp platform. To provide context, a user on Whatsapp adds a number to his/her group to subscribe to channels that provide these different services. The problem we want to solve using our model is to provide personalized recommendations of services that are more relevant to users in a group. This relevance is based on the fact that users in a group often share similar interests, and a graph approach would help in filtering and categorizing the most interesting services for the users in a group based on the services used by other group members in different groups. This would help in improving engagement, and discovery of services for the users. Our model compares both the traditional approach to recommendation using a utility matrix and the graph approach to understand the which would work better in the given context.

SYSTEM DESIGN AND IMPLEMENTATION

ALGORITHM

The recommendations were obtained using two techniques - Graph based and Item based. In the graph based technique, efficient querying methods in GraphFrames are used to extract information of all the connections in a group. The edges from those connections are then checked to find the other groups the users are present in. Groups from first connects of the users are also added to this set. The channels from those groups are compared to current subscribed channels, and k out of the channels are recommended to the group based on a weighted approach where the weight is calculated using the number of common groups between users and channels. In the traditional method, Item-based Collaborative filtering is used to obtain the top k recommendations. This algorithm was selected as we look to obtain a likeliness value for the other available channels out of which the top k can be selected.

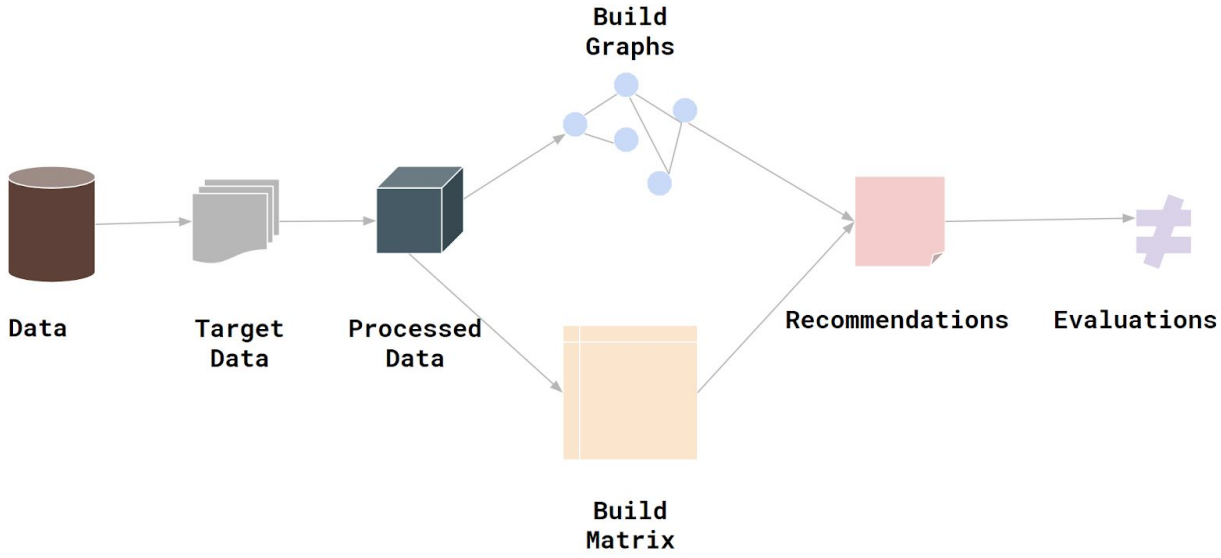
TECHNOLOGIES AND TOOLS USED

The primary tool used for development was Jupyter Notebook. The graph and respective structures were created using PySpark which is a wrapper on the Apache Spark project. PySpark provides the GraphFrames package which uses Hadoop HDFS and HQL. It stores and retrieves the data efficiently using the Spark Dataframe. These were selected due to its strong developer community and support for querying large scale distributed data.

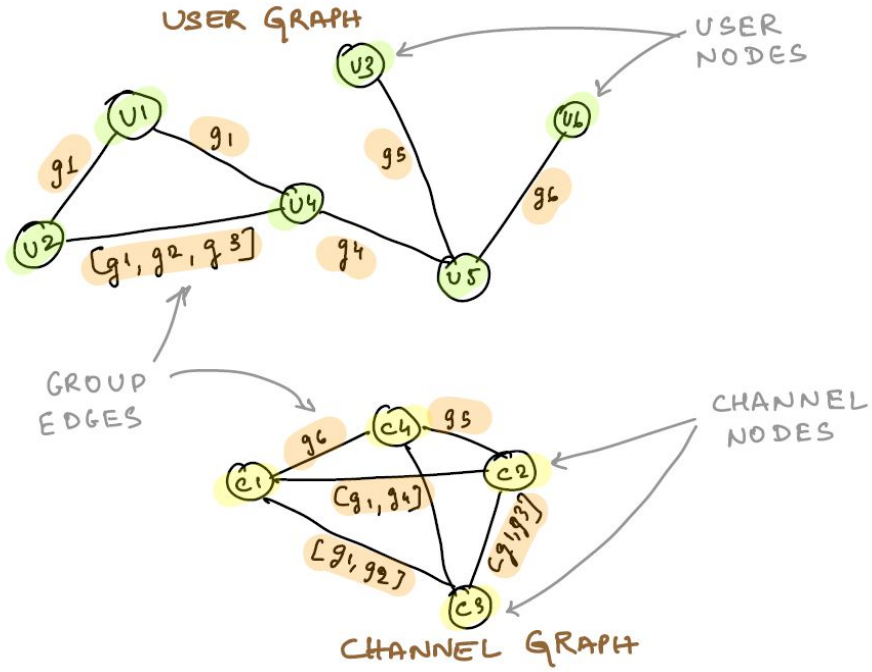
DESIGN

As depicted in the model below, the data was obtained from different tables in the source database. The data was then explored, and anonymized using standard hashing techniques. Required columns were then selected to generate the graph. The build graph phase builds 2 graphs - user graph and channel graph, using the packages available in the GraphFrames package provided for PySpark. For the traditional method, the utility matrix is build and stored in a dataframe. In the traditional model, Item-based collaborative filtering is used to determine the interest of the group in other channels out of which k are selected. In the

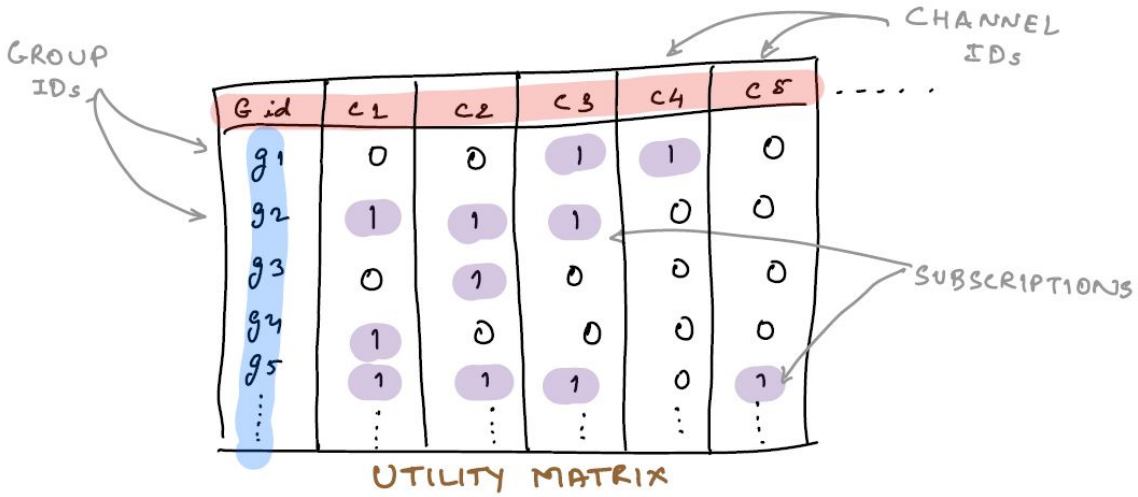
graph based model, the graph is queried to get user, group and channel information to obtain the recommendations based on the channel subscribed to in other groups by the users in the same group. Evaluation has been carried out by using a small sample from the available data that represents the statistics of entire data.



MODEL - GRAPH BASED

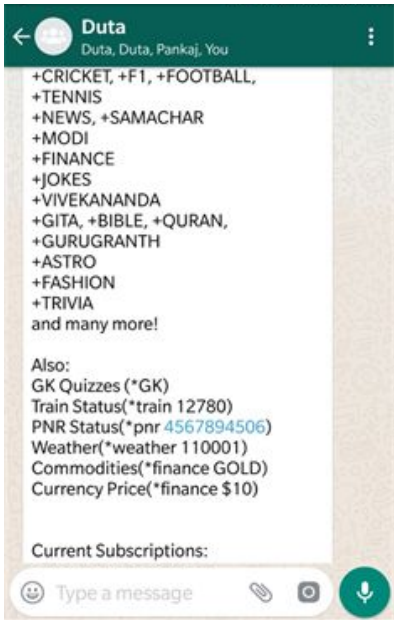


MODEL - ITEM BASED



SCREENSHOTS

The following screenshot provides a sample of how the recommendations would be provided to the user through the Whatsapp UI:



Note: This module has not been deployed and the above screenshot is only a sample of the existing method of displaying services to the user.

EXPERIMENTS

DATASET AND PRE-PROCESSING

The dataset used was obtained from Duta Inc. Different tables were queried and anonymized to finally obtain the data described below:

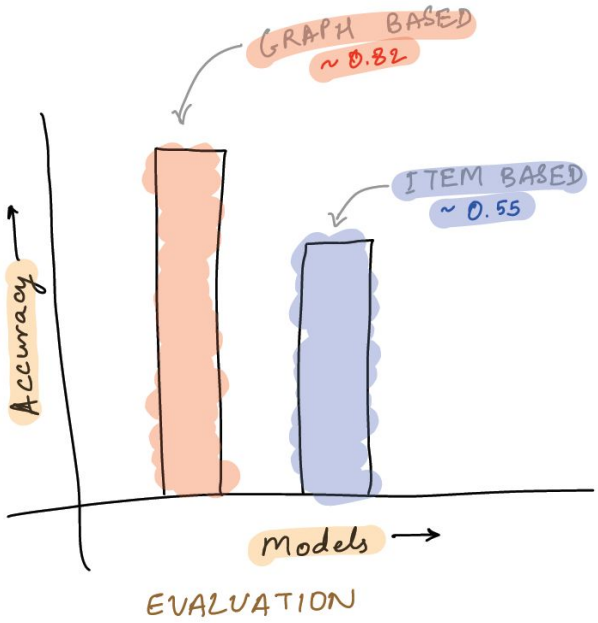
Table Name	**# of Columns**	**# of Rows**	**Features considered**
group_members	7467	20506988	create_time | timestamp without timezone update_time | timestamp without timezone group_id | character varying | not null user_id | character varying | not null is_admin | boolean added_by | character varying
channel_subscriptions	2728	12652163	create_time | timestamp without timezone update_time | timestamp without timezone channel_id | character varying | not null group_id | character varying | not null

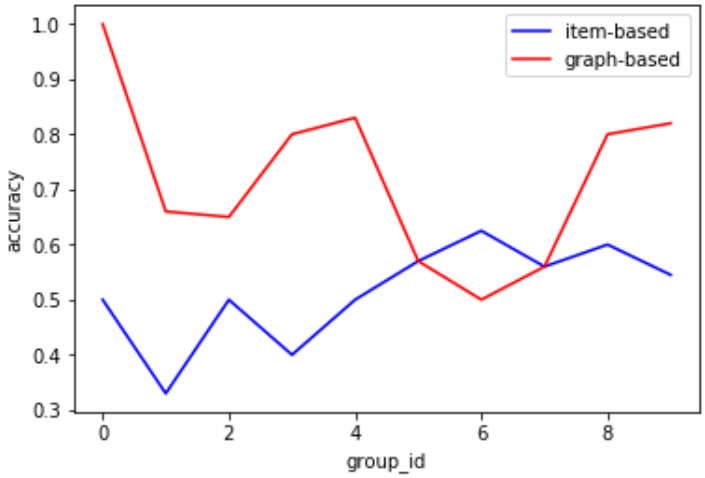
EVALUATION METHODOLOGY

- Both the models were evaluated using 10 diverse groups, out of which 3 were weakly connected groups, 5 were heavily connected groups and 2 were lone groups.
- Reason for choosing a very small number of groups for end evaluation is because of the large data. It takes approximately 45 minutes to run for a single weakly connected group.
- But the used set of test sample represent the entire data statistics well. Also, the actual evaluation would be how the live users react to these recommendations.
- Accuracy was used as the metric since subscription to a service is a binary value and Jaccard similarity was used.

ANALYSIS

- It was observed that graph based recommendation works better overall. It outperforms the traditional method for weakly connected and lone groups.
- Using graph based recommendation, the cold start problem is addressed and it no longer affects the system heavily.
- Scope for improvement exists. Considering additional features such as user's activity with a particular service, more user activity data and analyzing user patterns can add to the effectiveness of recommendations.





DISCUSSIONS AND CONCLUSION

Based on the problem to be solved, there were many approaches considered while designing the graph model.

DECISIONS

- Different graph packages (GraphX, GraphFrames, NetworkX) and programming languages (Scala/Python) were compared and we decided to go with GraphFrames and Python due to it strong developer community and ease of implementation.
- Different graph structures were considered and finally, one graph with node as the users and edges containing common groups was implemented. And, another with node as the channels and edges containing weights based on common groups were implemented.
- Different recommendation techniques were explored for the traditional model and the item based technique was implemented.
- Different evaluation metrics were considered and Accuracy was finally used to evaluate the models due to simple requirement.

DIFFICULTIES

- The main issue with such a large dataset was the space. Storing the data in files and then loading them to create graphs caused low storage issues often and took more time to compute the initial graph and persist it.
- Also, setting up spark and graphframes on different OS requires

tremendous efforts and it does not easily work.

THINGS THAT WORKED

- The planning and implementation was done well as a team. We each were able to accomplish the required tasks as per the decided deadlines and support each other when required.

THINGS THAT DIDN'T WORK WELL

- Local development on large datasets was a challenging. It would take huge amount of time to perform tasks due the size of data being in millions.

CONCLUSION

We achieved our required goal of providing graph based recommendations of the services offered by Duta Inc to its users. On comparing the two models, we find that the graph based approach works well and provides more relevance when compared to the traditional approach. This idea is based on the fact that people have similar interests and are often influenced by those with whom they often converse with. Applying this to the Whatsapp platform scenario, it proves advantageous to use graph based information to select and prioritize the recommendations based on the interactors in the respective groups.

PLAN AND TASK DISTRIBUTION

The sub-modules along with their owners and respect responsibilities have been provided below:

1. Data Extraction/Preprocessing/Graph Creation:

Owner: Aashish Subramanian

Responsibilities:

- Obtain the required data for graph generation and recommendation features to be used from the Duta database.
- Perform pre-processing tasks on the data and extract relevant features.
- Transform the data to obtain structures for vertices and edges that can be used to generate the graph.
- Generate the graphs and store the constructed graphs to be used by the recommendation model.

2. Graph based recommendation:

Owner: Chintan Vachhani

Responsibilities:

- Use the stored graphs to provide channel recommendations.
- Extract user, group relationships and collect the group set with their weights.
- Extract channel information pertaining to groups and connected groups.
- Provide ordered channel recommendations considering the weights between the users in the group.

3. Traditional recommendation:

Owner: Pavana Srinivasadeshik Aachar

Responsibilities:

- Create a utility matrix between the groups and channels subscribed.

- Obtain recommendations based on channel similarity.
- Provide ordered channel recommendations based on the interests of other groups using similar channels.

4. Evaluation:

Owner: Team

Responsibilities:

- Compare the outcomes of the graph based approach and the traditional approach.
- Evaluate both the models on the test set.
- Conclude on which model works better for the given dataset.

5. Documentation/Slides:

Owner: Team

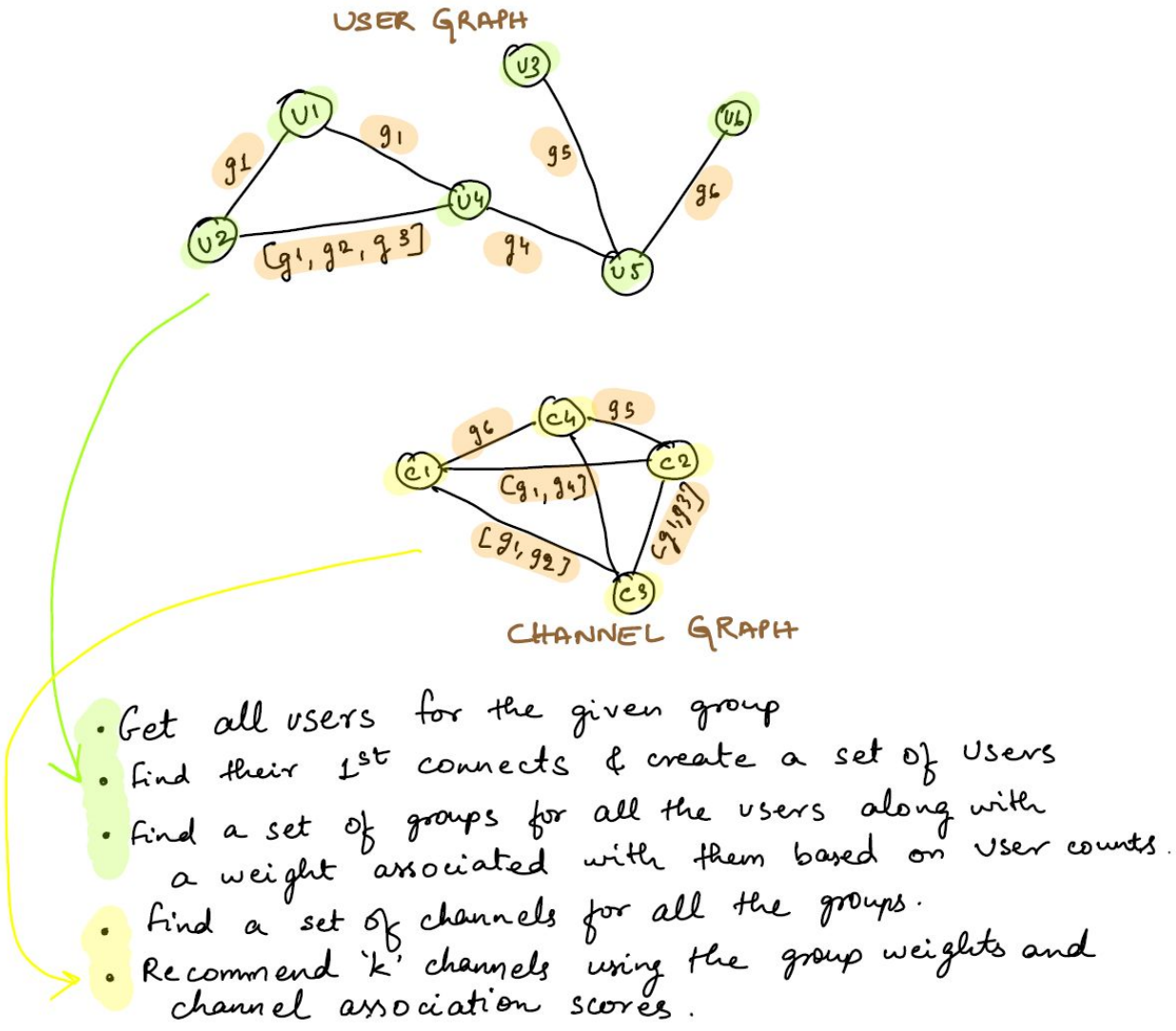
Responsibilities:

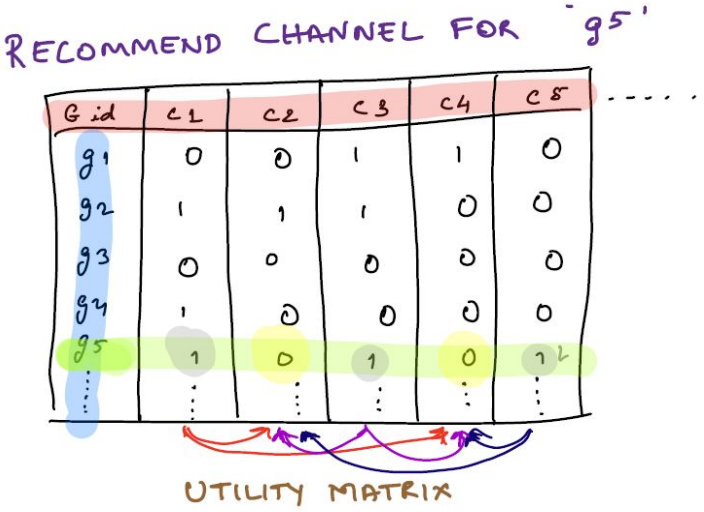
- Complete the report as per the requirement.
- Complete the presentation slides.

Each of the above tasks were successfully accomplished by the owners. In addition, the team worked together and helped each other when required.

APPENDIX

ALGORITHM - GRAPH BASED





$$Sim[C_2] = Sim(C_2, C_1) + Sim(C_2, C_3) + Sim(C_2, C_5) + \dots$$

$$Sim[C_4] = Sim(C_4, C_1) + Sim(C_4, C_3) + Sim(C_4, C_5) + \dots$$

RECOMMEND MAX (Sim[C2], Sim[C4])