Amazon Co-purchase Network

Professor: Dr. JAMES ABELLO
TA: PRANEETH CHANDRA THOTA

Akash Patel Yash Barapatre Shreyans Gupta

December 2020

1 Abstract

Graph datasets are a new and upcoming forms of datasets that carry unstructured but very useful information. Amazon Co-purchasing Dataset can be used to draw inference on the product network, as in how the products are related to each other, the communities/clusters they form and how important certain products are in a sub-group or in the whole graph. In this project, we have implemented such operations and have drawn some major inferences from the data.The results may provide interesting insights into user behavior and a better understanding of marketing techniques and logistics management.

2 Data Description

The Amazon Product Co-Purchasing Network is a graph dataset, where nodes represent products and edges represent links between commonly co-purchased products. This network was collected by crawling the Amazon website and is based on "customers who bought this item also bought" feature of the Amazon website. If a product i is frequently co-purchased with product j, the graph contains a directed edge from i to j. There are 548,552 different products information is available in this dataset. These products belong to the following groups – Books, DVDs, Videos and Music CDs. 17,88,725 pairs of products are considered co-purchased products. Also, we have a total of 7,781,990 product reviews. The total data size is approximately 1.3 GB.

1. Title
2. Salesrank
3. List of similar products (that get co-purchased with the current product)
4. Detailed product categorization
5. Product reviews: time, customer, rating, number of votes, number of people that found the review helpful

3 Project Description

3.1 Products to be recommended when a person is viewing a certain product.

1. When a user is viewing a certain product, what all products can be recommended to the user will be answered. Input will be the ASIN number of the product and output will be the product recommendations.

a. Product recommendation based on clustering: Recommending products from the same cluster closest to the center of the cluster. b. Product recommendation based on user reviews: Recommending products based on user reviews that are helpful to more than 25 people using Bipartite graphs. c. Product recommendation based on Top Products in each group. Recommending products in the product group, sorted by PageRank.

3.2 Logistic management for products

There are certain products in the data/graph that lie at very critical positions and may need replenishment at a faster rate than others. Such products can be identified by computing cliques. Products that are parts of a large number of maximal cliques can be considered as being in high demand. Such information is very useful for stock management and future predictions.

3.3 Top 20 most bought products from the whole data set

Pagerank is used to compute this list of products. This helps in finding the most active products in the network.

4 Type of users

Our application is very easy to use and can be used by all types of users. It has a basic drop down feature which helps the user to select what action is needed to be performed. It can be used by an inventory manager who would want to check the products that are needed to be replenished and also by a technical analyst who considers various aspects for recommending a product to a user. According to the situation he can choose between clustering, bipartite and pagerank to get a list of recommendations. A potential user could also visually analyze each product group based on certain parameters/filters.

5 Interactivity

We have created an interactive application in which the user can choose from a list of drop down menus to select what action he wants to perform. There are different actions our application can execute which will be talked about in next sections. For example, to get a recommendation for a particular product,

the user has to enter the ASIN number of the product and choose what type of recommendation he wishes to perform to get the results.

Also we have integrated Tableau in our application to show how the network looks like for each category. This integration helps in having user interaction with graph nodes and see each node's connections very easily.

5.1 Web App Snippet

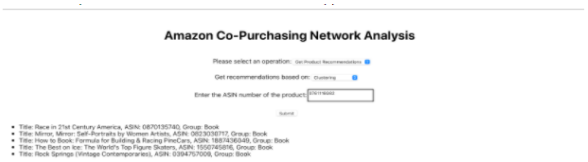

Figure 1: Recommendation based on clustering

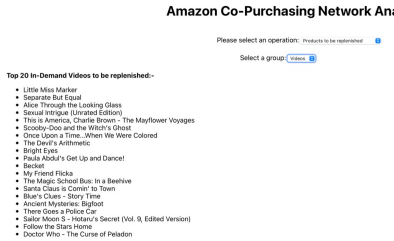

Figure 2: In-Demand Items

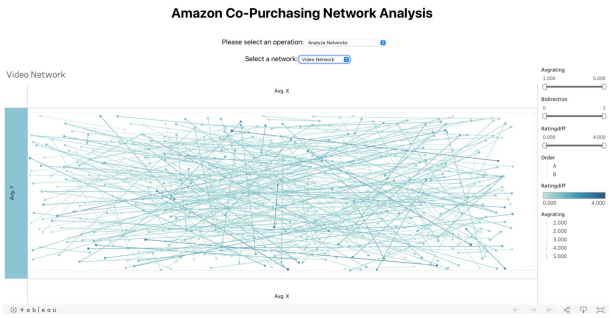

Figure 3: Top Video network visualisation

6 Mode of Processing and tools

The data was initially in a raw text format. This raw text data was processed in various formats suitable for further computations as per the needs. It was used to form graphs in PySpark GraphFrames using Pyspark Dataframes to perform computations such as Clustering and graphs in Networkx to perform computations such as identification of Cliques and PageRank. The results from these computations were stored and displayed to the user via a Web Application. The visualization of the network was done on Tableau, from which the results were exported to the Web Application. The application was implemented using the React.js framework.

7 Data Flow

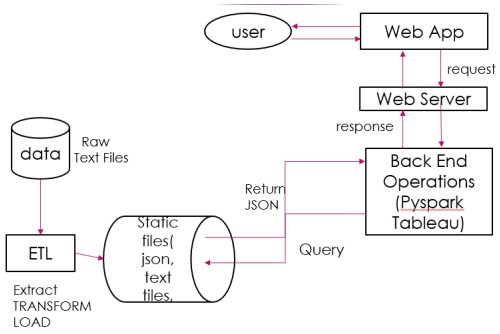

Figure 4: Data flow

8 Gantt Chart

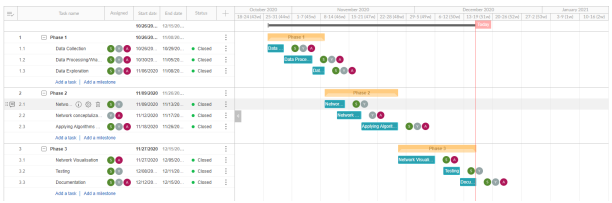

Figure 5: Gantt chart

9 Future Work

Improving use interface aesthetically. Also, Implementing user based Feedbacks and recommendations. Appending Existing data with data sets with similar

data structure and use. Visualising Cliques and groups in top to bottom approach(Collapsing View) to cover all the nodes in the network.

10 References

1. https://snap.stanford.edu/data/amazon-meta.html
2. https://spark.apache.org/docs/latest/mllib-clustering.html
3. https://networkx.org/documentation/stable/reference/algorithms/clique.html
4. https://networkx.org/documentation/stable/reference/algorithms/link_analysis.html
5.https://ladataviz.com/2019/12/15/build-a-network-graph-in-tableau-in-three-steps/
6.https://gwu-libraries.github.io/sfm-ui/posts/2017-09-08-sna