VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

The Social Network Analysis of Amazon Product dataset

A Course Application Report

submitted as part of the course Social Information Networks CSE3021

School Of Computer Science and Engineering

VIT Chennai

FALL 2021-2022

Course Faculty : Dr. B Radhika Selvamani

Submitted By

Lokesh K (18BCE1239)

Abstract

In Real world, social networks are often temporal in nature. They develop with time as new nodes might show up, old nodes might stop to exist and the connections between them may likewise change. We have considered an openly accessible dynamic organization dataset, in particular Amazon co-buy organization dataset. We have broken down the organization information to comprehend the meaning of hubs with high in-degree and high out-degree. We dissect the development of networks in the organization by noticing the difference in relationship among hubs and by noticing networks to perceive the number of individuals they could hold after some time. We show some continuous thing sets from the co-buy market bushel as far as the thing classifications and subcategories. We at last select some focal elements and gather such substances from the organization to suggest how advancing a portion of the co-bought things might build the deals of the chosen things. Predicting co-shopping merchandise primarily based totally on preceding order records of customers can assist on-line purchasing websites to suggest the right merchandise to customers. Successful prediction of purchasing merchandise can assist the web sites growth revenue. In our project, we examine statistics from Amazon product co-shopping community metadata accumulated via way of means of crawling Amazon internet site and carries product metadata and evaluate data approximately 548,552 extraordinary merchandise (Books, tune CDs, DVDs and VHS video tapes) to discover a few key motives in the back of co-shopping and create a version to expect the co-shopping merchandise of a product primarily based totally on capabilities that relate to co-shopping.

Contents

1 Introduction

Online commercial centers have become famous because of the purchasers' accommodation to shop from any spot. There are likewise other added benefits, such as, delivering the item to home, purchasing an item without taking care of money, purchasing items that are not accessible in the neighborhood commercial center, and some more. Amazon utilizes a proposal framework to recommend other habitually co-bought things during the offer of a thing. In the event that the purchase is helpful to the client, the proposal based promotion makes it almost certain that the additional thing is additionally purchased from a similar commercial center. Giving limited cost on suggested co-available things, just when Co-purchased, may likewise give further motivator to the client. These internet based stores obtain information from their own site furthermore the information isn't restricted to co-buy information. This paper manages investigation of an enormous co-buy network information from an online commercial center. We investigate the organization information from a diagram hypothetical viewpoint. Curiously, the co-buy organization information is worldly in nature and information is accessible for four time-stamps.

2 Related Works

The Amazon co-buy network was utilized by Leskovec et al. [1] for investigating individual-to-individual suggestions in viral advertising. By and large the suggestions were viewed not as exceptionally compelling in instigating buys. Yet, they showed that viral advertising works better when the information is at first ordered dependent on specific highlights. In another paper, Amazon co-buy network was utilized to clarify online business interest. They guarantee that thing classifications with compliment request appropriation are affected more by the design of the organization [2]. Clauset et al. proposed a local area location strategy in O(mdlogn) time where, n, m are the quantity of hubs and edges separately and d is the quantity of various leveled divisions expected to arrive at the most extreme particularity esteem. Particularity [3] is a famously

known capacity for estimating generally integrity of recognized networks. This technique is famously known as CNM [4] and they have utilized the Amazon co-buy network as a benchmark information to track down the networks in the organization. They distinguish networks with a high greatest seclusion esteem, i.e., 0.745, however the size of the networks appear to be extremely huge to track down any apparent examples from them. For instance, the biggest estimated local area they get comprises in excess of 100 thousand hubs, which is over 25% of the complete number of hubs in the organizations. Luo et al. concentrated on the neighborhood networks in Amazon co-buy organization and guaranteed that proposal turns out preferable for advanced media things over books [5]. Theme examination, for 3-hub and 4-hub themes, has additionally been performed on the Amazon co-buy network [6]. Continuous themes have been distinguished, which without any help can not contribute well to understanding the personal conduct standard of the co-buy organization. Late deals with finding regular subgraphs and mining subgraphs in powerful organizations [7], [8],

3 Methodology

The main objective of this project is to analyze Amazon's product dataset and to recommend the co-purchasing products for a corresponding product. The dataset is a raw dataset in a text file. The modules in the project include first the cleaning of the dataset from the metadata to produce the edgelist. Then the next module is to analyze the network, followed by detecting communities. The last module is the recommendation system.

3.1 Dataset Characteristics :

The dataset we use is Amazon product co-purchasing community metadata from Stanford Network Analysis Project. The facts changed into accrued with the aid of using crawling Amazon internet site and incorporates product metadata and evaluate records approximately 548,552 one-of-a-kind products (Books, tune CDs, DVDs and VHS video tapes).For every product the subsequent records is available: Title Sales rank List of comparable products (that get co-bought with the current product) Detailed product categorization Product reviews: time, customer, rating, range of votes, range of humans that determined the evaluate helpful The facts changed into accrued in summer time season 2006.

3.2 Data Cleaning :

The entire amazon dataset in its raw format is a text file as shown in (fig 1.)

```
Id:   15
ASIN: 1559362022
  title: Wake Up and Smell the Coffee
  group: Book
  salesrank: 518927
  similar: 5  1559360968  1559361247  1559360828  1559361018  0743214552
  categories: 3
   |Books[283155]|Subjects[1000]|Literature & Fiction[17]|Drama[2159]|United States[2160]
   |Books[283155]|Subjects[1000]|Arts & Photography[1]|Performing Arts[521000]|Theater[2154
   |Books[283155]|Subjects[1000]|Literature & Fiction[17]|Authors, A-Z[70021]|( B )[70023]|
  reviews: total: 8  downloaded: 8  avg rating: 4
    2002-5-13  cutomer: A2IGOA66Y6O8TQ  rating: 5  votes:   3  helpful:   2
    2002-6-17  cutomer: A2OIN4AUH84KNE  rating: 5  votes:   2  helpful:   1
    2003-1-2  cutomer: A2HN382JNT1CIU  rating: 1  votes:   6  helpful:   1
    2003-6-7  cutomer: A2FDJ79LDU4O18  rating: 4  votes:   1  helpful:   1
    2003-6-27  cutomer: A39QMV9ZKRJXO5  rating: 4  votes:   1  helpful:   1
    2004-2-17  cutomer:  AUUVMSTQ1TXDI  rating: 1  votes:   2  helpful:   0
    2004-2-24  cutomer: A2C5K0QTLL9UAT  rating: 5  votes:   2  helpful:   2
    2004-10-13  cutomer:  A5XYF0Z3UH4HB  rating: 5  votes:   1  helpful:   1
```

Fig 1. Raw Amazon MetaData

The data is cleaned and filtered to format the data in a meaningful way using the R programming language. Being a larger dataset, only a subset of the products, i.e, only the book products are used for the project. The book has the majority of the shares among all the products in the dataset. After cleaning the data is stored in two separate files, an edgelist file as in (fig 2). to create the graph and a metadata file which contains the metadata.

```
# FromNodeId    ToNodeId
0       1
0       2
0       3
0       4
0       5
1       0
1       2
1       4
1       5
1       15
2       0
2       11
2       12
2       13
2       14
3       63
3       64
3       65
3       66
3       67
```

Fig 2. Edgelist dataset

3.3 Analysis :

Graph Creation : Based on our data analysis, the graph contains Avg-Degree, which is calculated by Total edges/Total nodes thus providing a powerful tool to analyze the network. Based on the result the avg degree is quite good in connection. Then Density which is the ratio of observed edges to the no. of possible edges for a given network. In light of the outcome the density of this graph is in decimal value. Hence the network doesn't have good density. The Diameter is the shortest distance between the two most distant nodes in the network. The result says it has a good diameter. The Mean-dist –Average Path length, i.e. of steps along the most limited way for all potential sets of the organization nodes. Thus the measure of the efficiency of information is Good. Betweenness centrality & closeness centrality ,both are the measure of other's dependence and independence respectively on a given node. Thus higher the count would indicate someone holds authority over disparate clusters.Also theEigen Centrality-Index that calculates the centrality of an actor based not only on the connection, i.e. influence level. In this network it is high.The Assortativity is a preference for a network 's node to attach to other that are similar in some way. Here the value is in negative. Hence the connections are not good.Last the Transitivity which is the overall probability for the network to have adjacent nodes interconnected. Based on the analysis the network is not good in Transitivity.

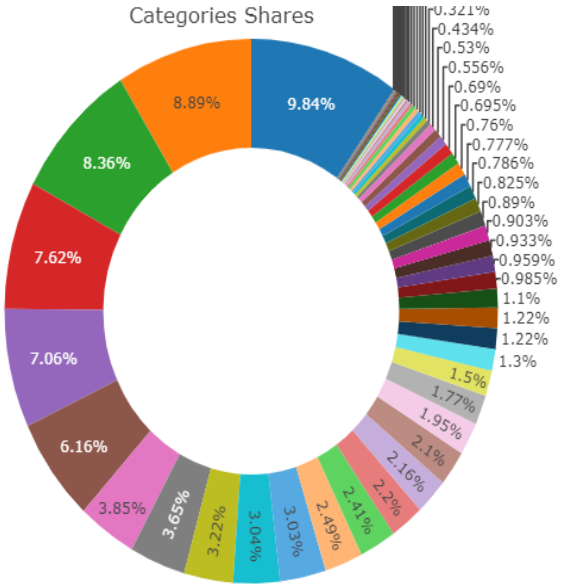

Fig 3.Categories share

Based on the above analysis (fig 3) is a pie chart which shows that the shares of various categories .From the resultant graph Home & Office has more categories than the other .Later comes the Religion & spirituality and Nonfiction which differs with less percentage .The other categories are much less than the above mentioned.

3.4 Community Detection:

Community detection algorithms that can partition the large number of networks into a couple of communities. In a massive scale network, such as an on-line social network, we should have hundreds of thousands of nodes and edges. Detecting communities in such networks will become a herculean task. In this project, we are using four algorithms such as Leading Eigenvector, Info-Map, Fast Greedy and Louvain (Multi level) algorithms. We are using these four algorithms to find the community detection and which networks belongs to which community that helps to understand the product and use it for classification of the entities.

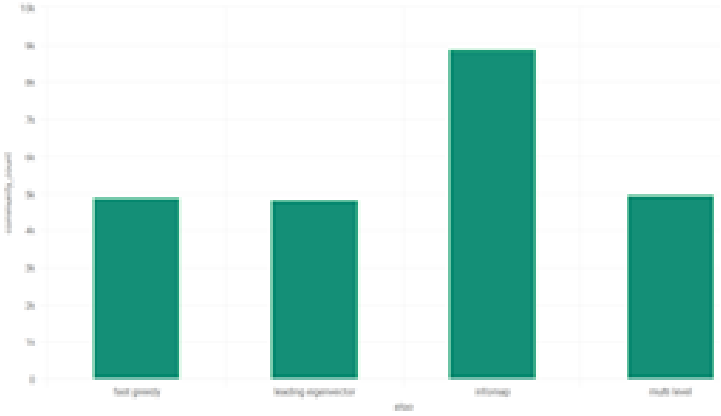

Fig 4. Community count

Here In the community detection modularity plays a major role to identify the count and importance which helps us to identify the more effective algorithm for the amazon dataset.

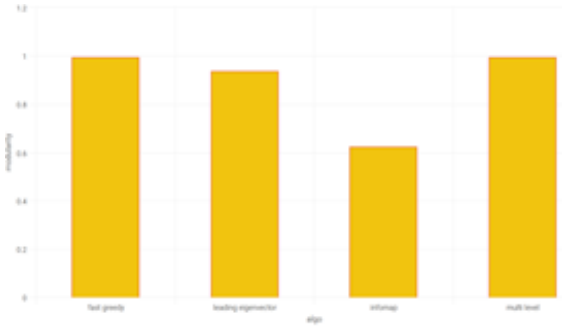

Fig 5. Community modularity

.After applying modularity and checking the community counts in (fig. 4) we can finally determine that Louvain algorithm shows the best accuracy and below graphs (fig.5) shows Louvain is the best for this dataset.

3.5 Recommendation System:

In this module, we create a function which takes a product id as a parameter and returns the top 3 recommended books for it. To do so, first we get all the nodes which belong to the same community as the passed product id as shown in (fig 6). We then create a subgraph using

those nodes, and then find the maximal cliques in that community, we go on to filter the cliques that contain the product id. The remaining nodes' product id's metadata are collected and the products are further filtered down by filtering the products which belong to the same category as the passed product, and the products are ranked based on their ratings and downloads. The top 3 ranked books are recommended for the product id.

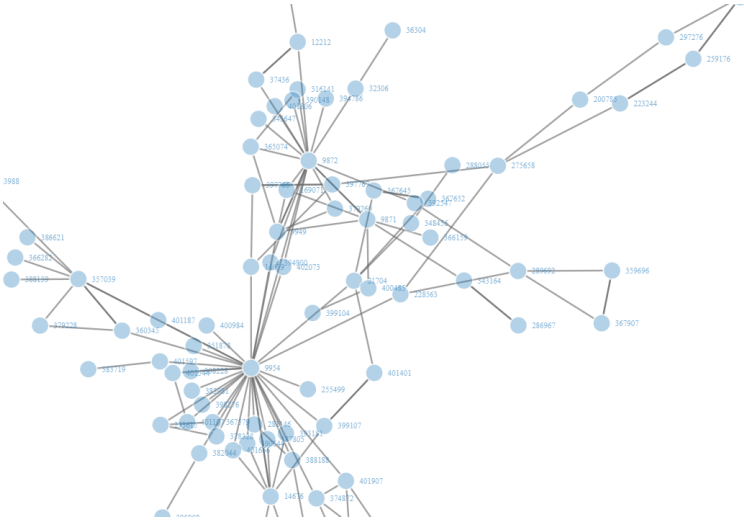

Fig 6. The community graph for node id (9954)

4 Results & Discussion

The product ids are looped over and passed to the recommendation function, as a result the books recommended for the corresponding product id are returned. For example, a book with product id 9954 as seen in (fig 6) has 33 cliques. The top 3 filtered and ranked books for the id 9954 are as below in (fig 7).

	A	B	C	D	E	F	G	H	I
1		NodeId	Title	Category	SalesRank	Downloads	Reviews	AverageRating	
2	770	9954	Understanding Genesis (The Heritage of Biblical Israel)	Religion & Spirituality	98821	7	7	4.5	
3	9231	140633	Night	Religion & Spirituality	16275	827	827	4.5	
4	1959	27552	Conversations with God : An Uncommon Dialogue (Book 1)	Religion & Spirituality	427	930	931	4	
5									
6									
7									

Fig 7. Csv File of id 9954 recommended books

Similarly for the rest of the ids the recommended books are stored in a file. This recommendation system can be employed in real time, by taking the data in real time and returning the recommended books.

5 Conclusion & Future Work

In light of the after effects of information examination and our expectations of various mixes of elements, we track down that class comparability is truly helpful while anticipating the co-buying sets, which implies that things inside the comparable classes will more often than not be bought together more frequently. Also, dealing with rank is additionally really supportive for expectation, recommending that well known things will be bought alongside different things all the more regularly. The investigation of co-buy networks uncovers the buying pattern of individuals and reliance on one thing being purchased with another. It can likewise be utilized to frame procedures for expanding deals. We have utilized market bushel examination for that reason. Creating incessant item sets have been utilized to comprehend the purchasing interest of the clients and connection between the item classes of interest.

Acknowledgments

We were obliged to give our appreciation to a number of people without whom we could not have completed this thesis successfully. We wish to express our sincere thanks and deep sense of gratitude to our Head of the Department B.Tech Computer Science and Engineering, Dr. Nithyanandam P, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course for the certification work. We are extremely grateful to Dr. Ganesan R, Dean of School of Computer Science and Engineering, VIT Chennai, for extending the facilities of the School towards our certification process and for his unstinting support. We express our thanks to our Dr. Geetha S, Assistant Dean of School of Computer Science and Engineering for her support throughout the course of this certification. We also take this opportunity to thank all the faculties of the School for their support and their wisdom imparted to us throughout the course. We thank our parents, family, and friends for bearing with us throughout the course of this project and for the opportunity they provided us in undergoing this project in such a prestigious institution.

References

[1] J. Leskovec, L. A. Adamic, and B. A. Huberman, “The dynamics of viral marketing,” ACM Trans. Web, vol. 1, no. 1, 2007.

[2] G. Oestreicher-Singer and A. Sundararajan, “Linking network structure to e-commerce demand: Theory and evidence from amazon co purchase network,” TPRC 2006. Available in SSRN, 2006.

[3] M. Newman, “Modularity and community structure in networks,” Proceedings of the National Academy of Sciences, vol. 103, no. 23, pp.8577–8582, 2006.

[4] A. Clauset, M. E. J. Newman, and C. Moore, “Finding community structure in very large networks,” Physical Review E, vol. 70, p. 066111,2004.

[5] F. Luo, J. Z. Wang, and E. Promislow, “Exploring local community structures in large networks,” Web Intelli. and Agent Sys., vol. 6, no. 4,pp. 387–400, 2008.

[6] A. Srivastava, “Motif analysis in the amazon product co-purchasing network,” CoRR, vol. abs/1012.4050, 2010.

[7] P. Bogdanov, M. Mongiov, and A. K. Singh, “Mining heavy subgraphs in time-evolving networks.” in ICDM, D. J. Cook, J. Pei, W. W. 0010, O. R. Zaane, and X. Wu, Eds. IEEE, 2011, pp. 81–90.

[8] M. Lahiri and T. Y. Berger-Wolf, “Structure prediction in temporal networks using frequent subgraphs.” in CIDM. IEEE, 2007, pp. 35–42.