Yewon Lee

The Most Desired Product to Co-purchase: Analysis of an Amazon Product Co-purchasing Network Graph from March 02 2003

Introduction

The goal of this final project was to find the product that is co-purchased with other items the most often through the analysis of a graph from Amazon. The graph contains 262,111 nodes and 1,234,877 directed edges. Each node indicates a product, and each directed edge from i to j means item i is co-purchased with item j. By using the PageRank centrality measurement, I intended to find the node with the lowest PageRank centrality because the lower the PageRank centrality is, the more central the node is.

PageRank Centrality Algorithm

PageRank is to calculate how central a node is via counting the number of outgoing edges of a node and the node degree of the outgoing edges. I used the simplified version of the algorithm: $PR(u) = \sum_{v \in Bu} \frac{PR(v)}{L(v)}$ where v is each node of B_u, a set of nodes that the outgoing edges of the node u connect to, and L(v) is the node degree of node v. Initially, all nodes start with the PageRank value of 1 divided by the total number of nodes in a graph. As the algorithm goes through nodes one by one, it updates the PageRank value with the equation above. Since the graph is directed, I could figure out which product is co-purchased with other items most frequently and which item people purchased with other items the most. To clarify, the outgoing edges of the nodes show which items they were bought with. Conversely, the incoming edges show which item people bought it with.

Results

After calculating the PageRank centrality of all nodes in the graph with both incoming and outgoing edges, I sorted them from smallest value to largest value. To clearly see the PageRank centrality of each node, I iterated them 3 times.

Iteration	Node	PageRank Value
1	176165	1.4116156895361122e-5
2	109459	1.1086600714964268e-5
3	109459	1.3437134893232255e-5

Table 1: The node with the lowest PageRank value calculated with using outgoing edges of the nodes

Iteration	Node	PageRank Value
1	14949	0.0008398103763001296
2	14949	0.0007002740417280081
3	14949	0.0006860534399579165

Table 2: The node with the lowest PageRank value calculated with using incoming edges of the nodes

For outgoing edges, node 109459 has the lowest PageRank value, which means that this item has the highest probability of being co-purchased with other items. For incoming edges, node 14949 has the lowest PageRank value, which indicates that this item has the highest probability of people purchasing the item with other items. By calculating the PageRank centrality, I could distinguish among the nodes that have the same number of outgoing edges. Like in the table1, both node 176165 and node 109459 have 5 outgoing edges but different PageRank values. Also, the graph is a directed graph which enabled me to note the difference between the probability of being co-purchased with other items and the probability of purchasing the item with other items.

Conclusion

Node 109459 had the highest probability of being co-purchased with other items. Node 14949 had the highest probability of purchasing it with other items. In other words, when people buy items, they are most likely to buy them with item 109495. The number of items that are co-purchased with item 14949 is the highest. The most interesting part of the analysis is the difference between the results from outgoing edges and incoming edges. Node 14949 was co-purchased with 420 items. While node 14949 has 420 items bought with, node 109459, the most central item for being co-purchased, has 5 items co-purchased with.

Citation

J. Leskovec, L. Adamic and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007. https://snap.stanford.edu/data/amazon0302.html.