UNIVERSITY OF MILAN

PROFESSOR: DARIO MALCHIODI

Link Analysis

Amazon US Customer Review

DAO YEN HOA - 988008

2022-2023

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work. I understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. This assignment, or any part of it, has not been previously submitted by me or any other person for assessment on this or any other course of study.

1 Introduction

In the realm of business and marketing, web analysis or search engine optimization (SEO) have emerged as one of the most important practices in business, especially for marketing sector and information retrieval strategies. Furthermore, Link Analysis has been considered as an important approach in examining ties and interconnections between diverse web sites. This significance is further underscored by the integration of PageRank, an influential algorithm that assigns numerical weights to web pages based on the quantity and quality of incoming links. The more links a page receives from other reputable and high-ranking pages, the higher its PageRank score will be. Followed by that goal, this project is to implement a ranking system using Amazon US Customer Review by products.

2 Data Pre-processing

The dataset has been downloaded from Kaggle and composed of many several files by specific products. The data related to products in "Baby" categories has been chosen for the analysis with the aim of examine the links between every two products if they were reviewed by the same customer.

The selected columns are 'customer-id', 'review-id', 'product-id' and 'product-title'. Then, the customer that only review one product also has been reduced from the dataset in result, there are 1,070,716 reviews of product as shown in the figure 1. Finally, it was transformed into Spark Dataframe for next steps of analysis.

```
Total new rows: 1070716
Dropped the 38.79 % of the inital rows ( 1749190 )
 - the 0.0 % are duplicates
 - the 38.79 % of customers with one review
```

Fig.1

3 Methodology

As I mentioned at the beginning, PageRank algorithm, which was introduced by Larry Page and Sergey Brin, the co-founders of Google, in the late 1990s has been applied for this project. The primary idea behind PageRank is to assess the importance of web pages as an essential technique for search engine. The value of webpages has been determined by the quality and amount of links. Incoming links are treated as votes, with pages having more high-quality links being judged more significant in search results. And to avoid the term spam, the content was not judged by only the terms on that page but also terms appear on linked pages. In conclusion, PageRank function assigned a real number to each page in the Web, in which the higher the value the more important it is.

3.1 Construction of links and transitional matrix

First, after pre-processing the data, it is organized by grouping the products according to each customer, thus yielding a detailed overview of the products that have been reviewed by each individual buyers. This process enables the identification of links formed by every pair of products reviewed by same customer as showing the customer preferences and decision-making patterns. Then we explode them enabling a new dataframe in which each row is a pair of products which was reviewed by the same customer. Subsequently, the algorithm proceeds to count the links originating from each distinct product. And therefore, formulating transitional probabilities by establishing the number of links from a product to another within a network, showing the relations of diverse products.

product_a	product_b	links_product_a	probability
B0006HBS1M	B004EWGDCE	6653	1.503081316699233...
B007BEHSDU	B0034XQXB0	1769	5.652911249293386E-4
B007BEHSDU	B00HF3Y100	1769	5.652911249293386E-4
B007BEHSDU	B003LNU8PK	1769	5.652911249293386E-4
B011J5TUUU	B0126D8J9M	28	0.03571428571428571
B008BRVTLI	B00R9QNCQW	15	0.06666666666666667
B00LFB616A	B00Q29MYN4	27	0.037037037037037035
B00LFB616A	B00Q29MYOS	27	0.037037037037037035
B004YL3332	B00171WXII	735	0.001360544217687...
B004YL3332	B0071D1AKI	735	0.001360544217687...
B004YL3332	B0034G60JM	735	0.001360544217687...
B00078ZHPS	B00EPFLHVM	1185	8.438818565400844E-4
B00078ZHPS	B0038JDUBQ	1185	8.438818565400844E-4
B00078ZHPS	B00IX6PEB8	1185	8.438818565400844E-4
B00QZ3UNIU	B010RWGMOI	104	0.009615384615384616
B00QZ3UNIU	B00XB50YL6	104	0.009615384615384616
B0006HBS1M	B00G34TL2U	6653	1.503081316699233...
B0006HBS1M	B00DXXVJRO	6653	1.503081316699233...
B0006HBS1M	B00DB5F114	6653	1.503081316699233...
B0006HBS1M	B007PDHPZ8	6653	1.503081316699233...

Fig.2

3.2 Initial value and PageRank iterative computation

The algorithm determines the initial PageRank values based on the initial computation to count the distinct products, where each product's starting value is set to the equal probability over the entire count of unique elements. The initial values, serve as the starting point for iterative calculations and assessments.

The PageRank algorithm is an iterative process in which the ranking of each item is updated step by step. At each iteration, the preceding iteration's probability values are multiplied by the relevant transitional probabilities, resulting in the dynamic changes of product rankings, with their values constantly updated until convergence is achieved.

In order to determine if the iterative process has reached a stable result, a convergence criterion is set on the distance between the values of the i_{th} iteration and the ones of

the i_{th-1} one. For this particular case, the Squared Euclidean Distance is preferred over the standard one because of its computational efficiency, allowing to reach convergence faster by increasing the sensitivity to smaller value changes. A tolerance value was set as the threshold for the distance as the first condition used in the loop for checking the convergence. In order to avoid the code running for an excessive amount of time, in case convergence would not met in a reasonable number of iterations, a maximum number of allowed computational steps is also added as a second condition to the loop.

After running 10 iterations to converge, given a tolerance value on the Squared Euclidean Distance of $5 * 10^{-9}$, returning the result as figure below.

product	page_rank
B00699FWD6	2.331197611075329E-6
B003TJ9PDC	1.753911420481256E-5
B00JLI73ZM	3.807141442074022...
B00O3GVSZI	5.114570327617026E-5
B0013HI718	3.424252421109331E-6
B001QXCF1C	1.633640633632889...
B00AHVR3ES	2.616634045486809...
B00E9RMHD8	1.715410809635548...
B00G6R6KVO	6.891847679417847E-7
B0006HBS1M	9.138603552230531E-4
B004V23YJM	1.137411902645144...
B00U0DD6EO	2.063240206200156...
B007BEHSDU	2.424804641068821...
B00G3XR8PS	9.568021643431567E-5
B00FP1XKVK	5.756984433978033...
B004GCJML6	1.011628999485941...
B0035EQKF2	2.843061383011291E-6
B004EL0G9Q	7.548233381351409E-6
B00F7PXANW	3.881080727479839E-5
B003HAI7CI	4.128463486203096...

Fig.3

4 Result conclusion

The figure 4 then clearly shows us the products with the highest probability of connection to other items and what they are, as well as the lowest ones. We can see the diversity of top products varying from toys for the babies to the helpful items for the parents in taking care of their kids. The application of the PageRank algorithm in the context of Amazon US customer review project has yielded valuable insights into the influence of various products across the network and their relationships, as well as analysis of customer preferences. Subsequent iterations, guided by the PageRank algorithm, systematically updated the product rankings, resulting in a dynamic representation that clearly highlighted the most influential and top-performing products while effectively identifying those with comparatively lower significance. In conclusion, PageRank has helped us in understanding the dynamic links between products, and therefore can be used for the creation of specialised product offers and targeted marketing strategies that resonate with the consumers database.

Top products:

product	page_rank	product_title
B000YDDF6O	0.002834902848013501	Baby Einstein Take Along Tunes Musical Toy
B000IDSLOG	0.002743799720272408	Vulli Sophie la Girafe
B00171WXII	0.0027176925737394217	FridaBaby NoseFrida The SnotSucker Nasal Aspirator
B002QYW8LW	0.0019224379509214986	Baby Banana Toothbrush
B009EDSWJA	0.001425089010029274	Summer Infant 5-Piece Essentials Diaper Changing Kits
B0052QYLUM	0.0013910880737661613	Infant Optics DXR-5 Portable Video Baby Monitor
B001WAJVZM	0.001328465103660608	Skip Hop Moby Bath Spout Cover Universal Fit, Blue
B00295MQLU	0.001218435568982723	Simple Wishes Hands Free Breastpump Bra

Least products:

product	page_rank	product_title
B00P8WLEJG	1.3217452530130463E-7	Bazzle Baby Banda Bib, Machine Washable Bandana Bib with Adjustable Snaps, 0-36 Months (Lime Green Star)
B00CXVM606	1.322775242320851E-7	Woombie Organic Cuddle Towels
B00RMKQGIQ	1.3257257182530577E-7	Disney Frozen Elsa Anna Olaf Figurine Playsets
B00BH291GY	1.3257780227376137E-7	Koala Baby 4-Piece Super Soft Crib Bumper - Blue
B00LNYT7VU	1.327675361055372E-7	DwellStudio Burp Cloth
B00GSMU1BC	1.329074503649127E-7	JOOVY Qool Silver Single Stroller
B00IRRRCYA	1.3321527629086446E-7	Itzy Ritzy Reusable Snack Bag (tiny dancer)
B001EHW2WW	1.3323145330702857E-7	Baby Angels

Fig.4