Università degli studi di Milano

Course: Algorithms for Massive Data

Exploring Link Analysis through the Application of Page Rank Algorithms on Amazon Customer Review Data.

Academic Year 2022-2023

Professor: Dario MALCHIODI

Student: Ly NGUYEN

Matriculation number: 988858

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 offenses 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

The e-commerce industry has witnessed remarkable growth and transformation over the past 15 years, resulting in the generation of massive amounts of data. For businesses, it is crucial to discover what this data reveals, particularly regarding customer behavior, patterns, and trends. By gaining insights from the data, businesses can make informed decisions and retrieve valuable information for their operations.

The primary objective of this project is to implement the page rank algorithm on an Amazon customer review data set. The algorithm will be applied to analyze the links among products that have been reviewed by the same customer. In other words, when a customer reviews at least two different products, a link is established between those products. Consequently, there are no dead ends to be handled in this scenario.

The underlying principle of the page rank algorithm is that a page's importance is determined by its connections to other important pages. By applying this principle, the project aims to identify and showcase the significant products that possess connections to other noteworthy products.

The project workflow commences with data set pre-processing. Once the dataset is properly prepared, the Methodology is applied, guiding the step-by-step calculation of a page rank value for each product. As a result, the Experimental Result and Conclusion section will present the ranking scores of the top five and bottom five products based on the calculated scores.

The project concludes with a comprehensive summary that provides an overview of the work conducted. It ties together the findings, insights, and outcomes, offering a holistic perspective on the project's contributions and implications.

2 Data Pre-processing

The selected dataset, called **Beauty** is a subset of a large, compressed file owned by Amazon that contains information about customers, products, and reviews. The dataset was obtained from Kaggle and was processed to extract the specific Beauty dataset required for this project.

The extracted data was transformed into a data frame consisting of three selected vari-

ables: customer id, product id, and product title. The primary focus of the analysis is on the customer and product identification variables, while the product title variable is optional and can be used for additional analysis and understanding when necessary.

To ensure the data frame's quality, two steps were taken as part of the data pre-processing phase. First, duplicate entries were removed to eliminate any redundancy. Secondly, customers who had fewer than two reviews were excluded from the analysis. This decision was based on the understanding that there is a meaningful connection between two products if they have been reviewed by the same customer. Customers with only one review do not contribute to establishing such connections. These data pre-processing steps resulted in a final data frame containing 3,143,789 rows.

There were no missing data, or any other manipulations required for the data set. Exploratory data analysis was not performed as it falls outside the scope of this project. However, it is possible that further analysis could be conducted in the future to derive deeper insights and inform strategic decision-making in areas such as marketing, sales, and finance.

3 Methodology

3.1 Constructing links between a pair of two products

In this phase, the construction of links between pairs of products plays a critical role in establishing the network structure required for the PageRank algorithm. Links are formed based on the existence of shared customer reviews between two products. If two products have been reviewed by at least one common customer, a link is established between them.

This process enables the identification of relationships and connections between products that exhibit similarities in terms of customer preferences or characteristics. By constructing these links, we create a network that facilitates the flow of influence and importance throughout the product ecosystem.

3.2 Computing the transition probability

The computation of transition probabilities entails the categorization of product1 into distinct groups, followed by the determination of the likelihood of product1 contributing to product2. This probability is derived by dividing 1 by the aggregate sum of the grouped

product1.

Through this analytical procedure, we are able to quantitatively assess the probability of transitioning from product1 to product2, thereby establishing the transition probabilities governing the interconnected products.

This calculation assumes a pivotal role within the PageRank algorithm, facilitating a comprehensive analysis and hierarchical ranking of various products based on their respective transition probabilities. By leveraging this approach, we gain valuable insights into the significance and influence of diverse products within the network structure.

3.3 Initiating the initial value

The initialization of the initial value is a crucial step in the PageRank algorithm. At the outset of the computation, each product is assigned an initial value, often represented as an equal probability for all products.

This initial value serves as a baseline probability distribution, representing the likelihood of a random customer reviewing a particular product. It sets the foundation for subsequent iterations of the PageRank algorithm, allowing for the refinement and convergence of ranking scores. The initial value accounts for the inherent importance of each product before incorporating the influence and importance propagated through the network.

3.4 Implementing page rank calculation

The implementation of the PageRank calculation involves a series of iterative steps to determine the importance or ranking score of each product within the network. This calculation is based on the connectivity and transition probabilities established in the previous steps.

The algorithm iterates through the linked product network, updating the ranking scores for each product based on the contributions from their linked counterparts. Through this iterative process, the algorithm converges towards a stable ranking distribution, effectively capturing the relative influence and importance of products within the network. The PageRank calculation continues until a predefined convergence criterion is met, ensuring that the final ranking scores accurately reflect the inherent characteristics and interrelationships of the product network

3.5 A description of the considered algorithm

In the case of implementing the PageRank algorithm, the commonly employed approach involves utilizing the Map Reduce function. However, for this project, the Map Reduce approach was not employed due to several reasons. Firstly, it is well-suited for smaller datasets or scenarios where there is sufficient memory capacity, which, unfortunately, is not supported by Google Colab in this particular case. Secondly, the process of transforming a regular data frame into Resilient Distributed Datasets (RDD) for Map Reduce usage and subsequently converting it back to the data frame format is possible but it is inefficient and time-consuming. Empirical evidence has shown that this approach took longer to produce the desired output compared to the alternative approach employed in this project.

The proposed solution revolves around translating the mathematical principles underlying the PageRank algorithm into concise code snippets. As previously described, the well-prepared transition probabilities and initial page ranks are multiplied, and the sum of their product is computed based on the given key. This summation result is then used to update the initial page rank, serving as the new starting point for the subsequent iteration.

In the context of a web network, it is typically preferred to iterate the algorithm for approximately 50 to 75 iterations. However, the iteration process may terminate earlier if a predefined convergence criterion is satisfied. To determine convergence, the Euclidean distance is computed, and a comparison condition is set to terminate the loop. Specifically, if the calculated distance falls below the convergence criterion, the ith iteration is halted.

The proposed solution demonstrated a high level of sophistication and effectiveness when applied to the massive dataset comprising over 3 million rows. Despite the significant scale of the dataset, the solution proved to be efficient in producing the desired PageRank outputs within a reasonable timeframe

4 Experimental Result and Conclusion

The implementation of the page rank algorithm on the Amazon customer review data set has yielded insightful results in determining the relative importance and ranking of products based on their page rank scores. By analyzing the top 5 and bottom 5 products, as determined by their page rank scores, significant patterns and trends can be observed within the data set, offering valuable implications for businesses operating in the e-commerce industry.

```
Top five products:
Row(product='B0049LUI9O', page_rank=0.0005670479421548141)
Row(product='B0043OYFKU', page_rank=0.0006307481388758013)
Row(product='B00DPE9EQO', page_rank=0.0006551653014446416)
Row(product='B001MA0QY2', page_rank=0.000676126380261458)
Row(product='B0014P8L9W', page_rank=0.0007073837925858547)
Bottom five products:
Row(product='B00GBEUZGI', page_rank=1.361561776350261e-08)
Row(product='B00CUYZ5Y0', page_rank=1.3716519558940939e-08)
Row(product='B00V5FMWIM', page_rank=1.4124060951145538e-08)
Row(product='B0006IVMG2', page_rank=1.4148257192682262e-08)
Row(product='B000E9BZGU', page_rank=1.4197607818895075e-08)
```

Figure 1: The top five and top bottom products

The top 5 products, characterized by higher page rank scores, suggest that these items have received significant attention and feedback from customers. The higher page rank scores indicate strong connections and associations with other important products, reflecting their popularity and perceived quality within the market. These products may possess desirable features, competitive pricing, effective marketing strategies, or exceptional customer experiences, which have contributed to their elevated page rank scores. Businesses can leverage this information to identify successful product attributes, understand customer preferences, and develop strategies to enhance their market positioning and customer satisfaction.

Conversely, the bottom 5 products, with lower page rank scores, imply a relatively lower level of customer attention and feedback within the data set. These products may face challenges in terms of market competitiveness, customer perception, or other factors that have resulted in their diminished page rank scores. Analyzing the characteristics and customer sentiments associated with these products can provide insights into potential areas of improvement, such as product quality, customer experience, pricing, or marketing strategies. By addressing these areas, businesses can enhance the performance and competitiveness of these products in the market.

The implementation of the page rank algorithm demonstrates the interconnectedness of customer reviews and their influence on product rankings. By considering the page rank scores and associated insights, businesses can gain a comprehensive understanding of customer preferences, identify opportunities for growth, and make informed decisions to improve

their product offerings and overall market strategy.

It is important to note that while the page rank algorithm provides valuable insights, it should be considered alongside other factors such as product quality, market trends, pricing, and customer needs. Additionally, the specific interpretation and action-ability of the page rank scores will depend on the individual business context and objectives.

The page rank algorithm serves as a valuable tool in understanding the dynamics of customer reviews and their impact on product rankings, providing businesses with opportunities to optimize their strategies and succeed in the competitive e-commerce landscape.