UNIVERSIDADE
LUSÓFONA

Analysis of joint purchasing patterns for recommender systems in e-commerce

Final Course Work

Final Report

Student Name: Joana Okica de Sá

Supervisor's name: Sofia Fernandes

Final Course Work | Degree in Computer Science | 28/06/2024

www.ulusofona.pt

Copyright

Abstract

Recommender systems (RS) have a pivotal role in enhancing user experience by delivering personalized service support based on individual preferences, thereby significantly impacting businesses. The success of these systems is undeniable and intricately linked to the strategic management and utilization of both the size and diversity of databases.

The present undergraduate study revels in the analysis of the relations among items within an e-commerce dataset, exploring how different types of interactions impact the similarity between products and thus open avenues for enhancing recommender systems.

Furthermore, the research goes into the multifaceted realm of RS with a focus on leveraging product metadata to transcend the traditional "Who bought X also bought Y" functionalities. Thus, it is by recognizing the imperative role of RS in today's digital age and exploring the field of social network analysis (SNA) that the research is initiated.

Accordingly, for the implementation and better understanding of the problem, this study utilizes Python as the programming language and the software Gephi to analyse a dataset from Amazon, dated to 2006, characterized by more than 500,000 products, covering various topics relevant to the proposed analysis, such as co-purchasing book genre patterns, and the connections formed by the interactions between products.

Keywords: E-Commerce, Recommender Systems, Social Network Analysis, Data analysis

Resumo

Os sistemas de recomendação (RS) desempenham um papel fundamental na melhoria da experiência do utilizador, fornecendo um serviço de apoio personalizado com base nas preferências individuais, tendo assim um impacto significativo nas empresas. O sucesso destes sistemas é inegável e está intrinsecamente ligado à gestão estratégica e à utilização da dimensão e da diversidade das bases de dados.

O presente trabalho de licenciatura debruça-se sobre a análise das relações entre itens de um conjunto de dados de comércio eletrónico, explorando a forma como diferentes tipos de interações têm impacto na semelhança entre produtos e abrindo, assim, caminhos para melhorar os sistemas de recomendação.

Para além disso, a investigação entra no domínio dos RS, centrando-se na utilização de metadados de produtos para transcender as tradicionais funcionalidades "Quem comprou X também comprou Y". Assim, é reconhecido o papel imperativo das RS na atual era digital e explorado o campo da análise de redes sociais (SNA) que a investigação é iniciada.

Assim, para a implementação e melhor compreensão do problema, este estudo utiliza Python como linguagem de programação e o software Gephi para analisar um conjunto de dados da Amazon, datado de 2006, caracterizado por mais de 500.000 produtos, abrangendo vários tópicos relevantes para a análise proposta, tais como padrões de compra conjunta de géneros de livros, e as conexões criadas pelas interações entre produtos.

Palavras-chave: Comércio eletrónico, Sistemas de Recomendação, Análise de Redes Sociais, Análise de dados

Table of Content

Table of Figures

1 Problem Definition

1.1 Contextualization

Given the volumes of information generated globally every second, advanced methods such as data analysis become essential in effectively managing the current data magnitudes.

By delving into the complexities of data management, it is imperative to recognize the multifaceted nature of the objects embedded within this expansive data realm. These objects, whether representing individual data points (entities) or clusters, serve as the foundation for understanding the underlying structures present in our society.

Within these structures, these entities play a crucial role in revealing and providing information about trends, patterns, or even changes that have occurred over time. By exploring the relations between these entities, we can not only comprehend the structure of the database but also gain a deeper insight into the dynamics shaping the studied landscape.

However, the recognition of interconnections among these objects and structures extends beyond social contexts. These connections represent dependencies, correlations, or influences between different elements. Whether they be neuronal connections, interactions between proteins, or associations among products, the concept of a network, which allows for the representation of relations between entities, has evolved beyond its conventional interpretation. It has emerged as a potent tool for comprehending intricate interdependencies across diverse domains, offering a holistic perspective that enhances our comprehension of our world.

1.2 Motivation

Currently, large e-commerce companies dominate the market, leveraging extensive databases that enhance their recommender systems, which makes the entire purchasing process more appealing to customers for various reasons, fostering customer loyalty.

Amazon's recommendation engine, which is frequently praised as the best in its class, considers a variety of characteristics, including previous purchases, browsing history, ratings and reviews, and connections with other Amazon services. A stunning 35% of Amazon purchases are the result of recommendations.[16]

This research explores patterns of purchasing behavior and interactions between products by utilizing a compact dataset, unlike larger companies, to deepen the understanding of recommendation dynamics. This targeted approach allows for the identification of specific patterns within a company's niche, providing personalized and adaptable insights that consequently will present an opportunity for smaller businesses to improve their systems and compete more effectively.

Accordingly, the study seeks to find a balance in the playing field between large and small companies and provide a strategic advantage to smaller enterprises.

1.3 Dataset Description

Leveraging sources from a dataset of Stanford University's SNAP repository, the data utilized in this research [1] contains product metadata and rating information on 548,552 different products dated of the summer of 2006, covering categories such as Books, CDs, DVDs, and VHS of the Amazon platform.

The dataset's structure contains various attributes as illustrated in Figure 1. To reach the established purpose the study narrows its focus to a specific group of interest, Books, which compose a quantity of 250,823 products, and takes from each: the product identification number, the title, similar products, the categories for each product, and the IDs of customers who reviewed the product.

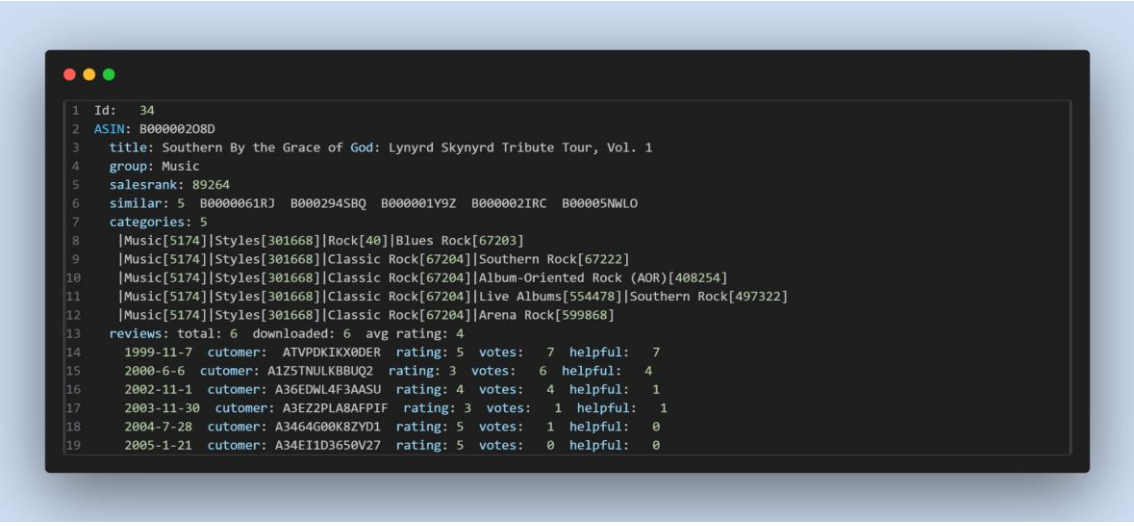

Figure 1. Segment of research dataset

A further analysis showed a total of 27 valid book categories from an initial quantity of 61. Valid categories refer to book genres, such as Fantasy, Science Fiction, or Romance, while non-valid categories include items that do not represent book genres,

such as Holiday Greeting Cards. Figure 2 illustrates a bar graph displaying the percentages of the top 10 categories with more products, these were the top categories in the initial dataset which also influenced the choice of the study group.

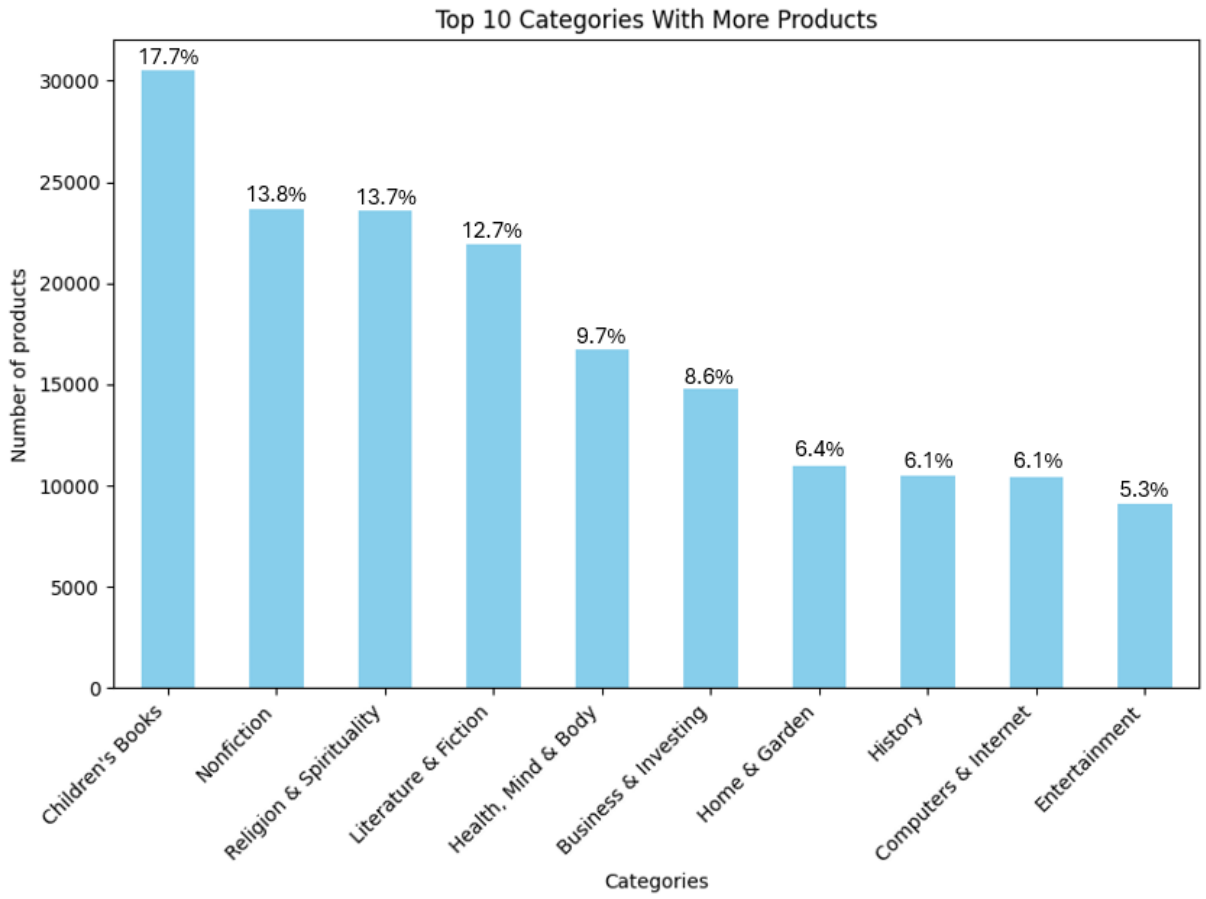

Figure 2. Top 10 categories with more products

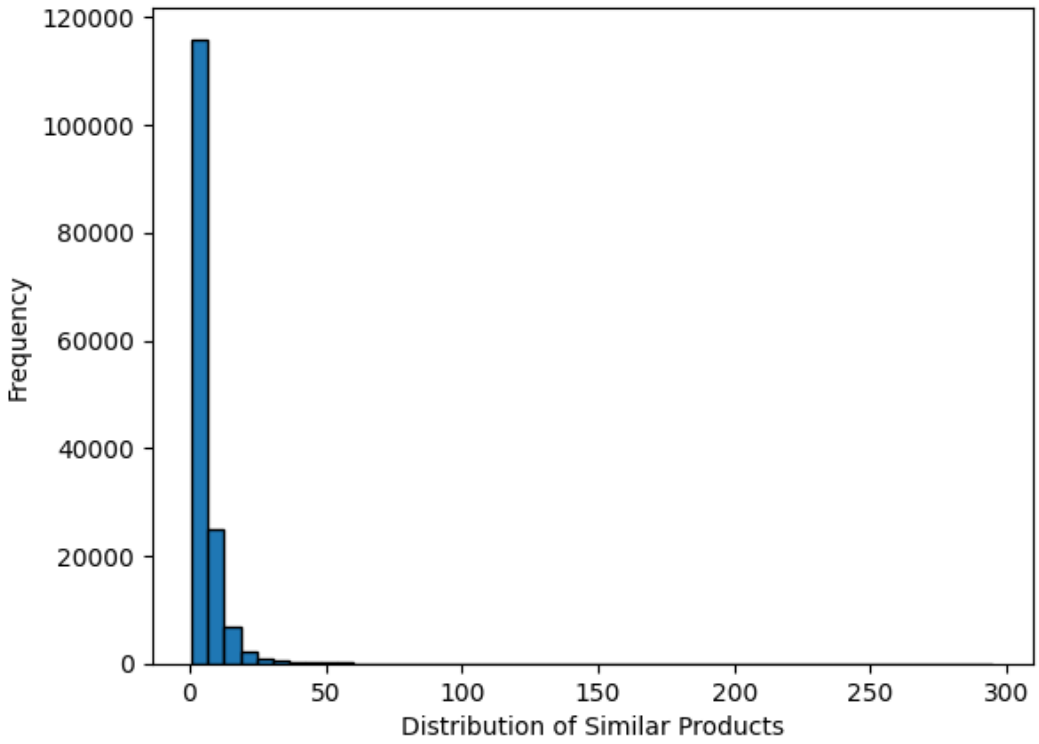

Figure 3. Distribution of Similar Products

While examining the dataset structure it is noted a maximum of five similar books per book, a more thorough examination of the sample, as observed in Figure 3, showed that, on average, a product is deemed similar 3 times and, at most, 295 times.

By employing social network analysis techniques and analyzing how the similarity between products may fluctuate based on the chosen network representation, we seek to gather insights into the dynamics of the relations that link products, thus providing a nuanced understanding of their interrelations within the broader context of network representations and ultimately contributing to the achievement of the research goals.

2 Benchmarking

To gain a comprehensive understanding of the principles underlying social network analysis and the intricate workings of recommender systems, thus providing a solid foundation for an analysis and interpretation of their roles in the broader context of this study, this chapter revels in detail these two subjects.

2.1 Social Network Analysis

In today's world the existing relations between data made it possible to create a vision of a network of interconnected objects. Accordingly, an area of study has emerged to discuss these connections, called Social Network Analysis (SNA).

As stated by Tabassum et al., 2018, “A social network consists of a finite set of vertices and the relations, or ties, defined on them (Wasserman and Faust, 1994).”.

In this context, these relations can be demonstrated by graphs (Figure 2). A graph is defined by the following elements: vertices, which represent a variety of individual entities (e.g., people, organizations, countries, products, animals, etc.) and an edge, which is the link between the vertices and can represent a panoply of relations between individual entities (e.g., communication, cooperation, friendship, etc.) [2, p.2].

It is through SNA techniques, that it is possible to identify consumer profiles and thus make personalized recommendations. A customer's tastes and patterns of actions can be inferred, and it is possible to gain a perception of how diverse actors collaborate, share resources, and communicate across a network. [3] Therefore, the following points highlight selected SNA techniques aligned with the study's objectives.

2.1.1 Community Detection

Communities represent aggregations of nodes characterized by shared specific attributes. Within the network architecture framework, these communities assume a crucial role, elucidating patterns and linkages that contribute to an enhanced comprehension of interrelations within a networked system.

Thus, the value of community detection [2, p.19] lies in its ability to reveal patterns of relations, influence, information flow, and potential areas of interest or concern. In this way, it facilitates comprehending the underlying organization and structure of complex networks.

When detecting communities there are two possible leading information: the network structure and the traits and attributes of nodes. As a result, it is possible to locate logically related objects, investigate relations between modules, infer missing attribute values, and forecast undiscovered linkages.

2.1.2 Node-level Statistical Measures

Node-level Measures [3] determine the relevance of a network actor or node, indicating which has the most essential relations and providing insights into their power among their peers.

Degree or Valency

The degree of valency [3] of a node v, measures the node's engagement in the network. It is calculated as the number of edges incident on a given node or as the number of node neighbors (nodes directly connected to v).

The degree is an effective metric for determining an actor's prominence and impact in a network.

Betweenness Centrality

The betweenness centrality [4] role in network analysis allows to identify nodes that function as crucial connectors and facilitators of communication within a network. Nodes with a high betweenness centrality score are more likely to act as 'bridges' in the communication between other nodes. They create the network's shortest communication paths, which normally, would suggest important information gatekeepers between groups.

Closeness

Closeness [2, p.9] is an approximate measure of the overall position of an actor in the network, indicating how long it will take to reach other nodes from a given beginning node.

Closeness algorithm calculates the shortest path between each node and then provides a score to each based on the sum of all the paths. “Nodes with a high closeness value have a lower distance to all other nodes.” [4]. They'd be effective information disseminators.

2.2 Recommender Systems

Within the theme of data management, several sub-themes emerge, one of which is the analysis of recommender systems (RS). Several definitions of RS have been formulated, and over the years the focal aspect that has most led to its adaptation has been the fact that RS is used in several areas and consequently has different uses, therefore a definition would have to cover the entire set of applications in which it is used.

Goldberg pioneered the development of the first recommender system, aiming to address the issue of email saturation experienced by users [5, p.15]. An agreement on the purpose is that generated recommendations are intended to improve one user's experience, not to represent collective consensus. It is also meant to assist the user in deciding among discrete possibilities. [6, p.13]

User demographics, item qualities, and user preferences are frequently used to create recommendation selections [6, p.13]. In addition to these criteria, as reported in Recommender Systems in E-Commerce, presented at the 2014 World Automation Congress, significant types of data enrich the foundation upon which personalized recommendation systems are built, enhancing their accuracy and relevance to the user experience. Attributes such as behavioral data which shows our interaction with technology, such as clicks and downloads; transaction data which reflects purchasing patterns; and production data which gives us insights into the consumption of content.

Recommender systems utilize various algorithms and methodologies some of them are: Collaborative filtering, content-based filtering, hybrid filtering, and knowledge graphs. [7, p.3]

Collaborative filtering

Collaborative filtering (CF) [7, p.3] operates autonomously. It relies exclusively on the analysis of prior observations of users' collective behavior to formulate and furnish novel recommendations. In other words, as written by Gonçalves-Sá & Pinheiro, 2023 in *Societal Implications of Recommendation Systems: A Technical Perspective*, if there is a similarity between user A and user B, the previous choices made by user B can provide insights into what recommendations would be suitable for user A.

Content-based filtering

Content-based filtering (CB) [7, p.3] approach's fundamental idea is to use past knowledge about users' preferences and item attributes to provide the most

relevant recommendations. Recommendations can be generated instantaneously, even for things that have never been recommended to a user before.

Hybrid filtering

Hybrid filtering (HF) [8, p.3] addresses the issues raised by the lack of interpretability in classic collaborative filtering systems. It results from a combination of different filtering techniques, both collaborative and content-based. [9]

Knowledge Graphs

Knowledge Graphs [8, p.3] provide a structured manner to represent and connect information about entities and their relations. It is through them that it is possible to create systems that provide both high-quality suggestions and interpretable explanations for user-item correlations by including knowledge graphs in the recommendation process.

2.3 Theoretical Framework

The present analysis takes a step in exploring different representations of data in networks and different types of information, aiming to understand the impact of these representations on the similarity between products.

Joint purchasing patterns refer to the phenomenon where multiple users make similar product choices, reflecting shared preferences or interests. This collaborative decision-making process can be influenced by various factors such as social connections, demographic similarities, or contextual circumstances. Unraveling the complexities of these patterns holds the potential to refine recommender systems, making them more attuned to the collective needs and behaviors of users.

This is where the present study goes beyond the existing work. The preexistent research already uses networks in recommender systems, particularly in knowledge-based approaches. However, these approaches often do not delve into the various possibilities for modeling network data. Thus, it is not clear how these representations can influence the recommender system. Therefore, this study seeks to contribute by elucidating and filling this gap, offering a more comprehensive analysis of the implications of these representations in the context of recommender systems.

3 Feasibility and Relevance

As more people utilize websites and digital services recommender systems play a larger role in the decisions that customers make daily.

3.1 The economic impact

Recent research on the subject by Adomavicius, et al shows that recommender systems are now considerably impacting customer behavior from an economic and decision-making standpoint.[13]

Beyond enhancing user interactions, recommender systems play an important role in driving sales. Tailored product suggestions based on individual preferences not only increase conversion rates but also elevate the average order value. In a fiercely competitive market, the continual analysis and optimization of these systems are essential for staying relevant and gaining a strategic edge.[14]

Thus, it is possible to conclude that understanding the impact of network modeling on the similarity between products is a key element in making relevant decisions when presenting information, as it becomes possible to make more insightful choices.

Furthermore, this study aims to establish a valuable foundation for improving existing approaches, such as strategies based on knowledge graphs. By applying the findings of this research, specific enhancements in these approaches can be targeted, such as optimizing recommendation capacity and therefore contributing to the continued advancement in the field of recommender systems. Therefore, this study not only hopes to add knowledge to the understanding of similarity in recommender systems but also offer practical guidelines for improving and innovating existing approaches.

4 Proposed Solution

The proposed solution focuses on applying network analysis to the task of recommending items in an e-commerce system. In particular, the interest is in studying how product metadata can be used to improve recommender systems and go beyond the "Who bought X also bought Y" functionality. Specifically, the interest is in analyzing how product similarity varies when considering, in addition to co-purchase patterns, product category.

Thereby, the development of a solution was divided into two main parts as shown in the following Figure 4:

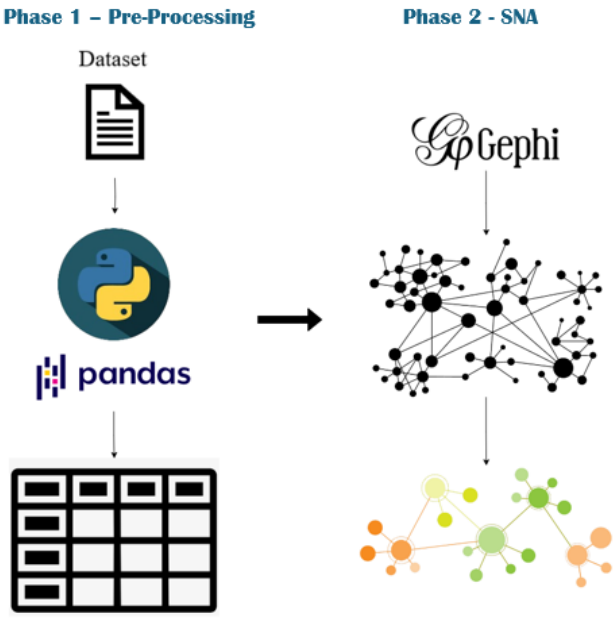

Figure 4. Network Graph

4.1 Architecture

4.1.1 Phase 1 - Pre-Processing

This phase encompasses the establishment of two distinct data networks: one illustrating co-purchased book relations based on similarities, and another reflecting the categorization of products in comparison to their counterparts.

The corresponding code, accessible on Github, was subsequently adjusted to extract details such as product ID, title, similar products, frequent categories, and customer review IDs.

Utilizing the Pandas library, the dataset was organized into a structured DataFrame. To reach the goal, which is to facilitate a comprehensive analysis of the data networks in Gephi and ultimately contribute to the refinement of system effectiveness, this phase involved processing the dataset and constructing a data network based on the obtained results. The data processing was performed using the Python language within the PyCharm application.

Accordingly, the information was organized into a DataFrame as shown in Figure 5.

	id	group	title	similarities	categories	reviews
0	0827229534	Book	[Patterns of Preaching: A Sermon Sampler]	[0804215715, 156101074X, 0687023955, 068707423...	Religion & Spirituality	[A2JW67OY8U6HHK - 5, A2VE83MZF98ITY - 5]
1	0738700797	Book	[Candlemas: Feast of Flames]	[0738700827, 1567184960, 1567182836, 073870052...	Religion & Spirituality	[A11NCO6YTE4BTJ - 5, A9CQ3PLRNIR83 - 4, A13SG9...
2	0486287785	Book	[World War II Allied Fighter Planes Trading Ca...	[]	Home & Garden	[A3IDGASRQAW8B2 - 5]
3	0842328327	Book	[Life Application Bible Commentary: 1 and 2 Ti...	[0842328130, 0830818138, 0842330313, 084232861...	Religion & Spirituality	[A2591BUPXCS705 - 4]
4	1577943082	Book	[Prayers That Avail Much for Business: Executive]	[157794349X, 0892749504, 1577941829, 089274956...	Religion & Spirituality	[]
...	...	...	...	...	...	...
516528	B000059TOC	DVD	[The Drifter]	[630366704X, B0002ERXB8, B0001932ZU, B0001VTPU...	Special Features	[A32PCPZL40G5N8 - 5]
516529	B00006JBIX	DVD	[The House Of Morecock]	[B0002HOE6C, B0002I84JO, B00004WZQN, B00069CQ8...	Genres	[A24IFZUH8NLISK - 1, A2SVXZKU40G7N - 5, A3HM5G...
516530	0879736836	Book	[Catholic Bioethics and the Gift of Human Life]	[1931709920, 188187110X, 081890643X, 158051046...	Nonfiction	[A2PD80S1N7920J - 4]
516531	B00008DDST	DVD	[1, 2, 3 Soleils: Taha, Khaled, Faudel]	[B00012FWNC, B0002UNQQI, B00069FKLO, B0000CNTH...	Genres	[A3NKS7CVEJVTQ6 - 5, A3EQ4YAZ5OEVK9 - 5, A3HRK...
516532	B00005MHUG	Music	[That Travelin' Two-Beat/Sings the Great Count...	[B00008OETQ, B00005O6KL, B00006RY87, B0002OTI9...	Broadway & Vocalists	[ABTSEEYVYQ52M - 5]

516533 rows × 6 columns

Figure 5. Sample of dataset

Given the size of the dataset, for the possibility of studying the network within the Gephi there was a need to restrict the scope of the analysis, in this manner, the study focused on the book products. Accordingly, the data was processed starting by, excluding information that was irrelevant to the study, such as genres that do not classify literary genres such as “Special Features” as observed in Figure 5.

Furthermore, when analyzing similar products, it was observed that just because one product is designated as similar to another, it does not imply reciprocity. This insight led to the refinement of the data, eliminating references to similar products that were not reciprocally indicated.

```
#limpeza de generos redundantes
unwanted_categories = ['jp-unknown1', 'jp-unknown2', 'jp-unknown3', 'Genres', 'Reference', 'Books on Tape','Book Clubs',
'Bargain Books','Books on CD',"O'Reilly",
                       'New & Used Textbooks','Home & Office','Boxed Sets','McGraw-Hill','Sheet Music & Scores','Book Accessories'
,'Holiday Greeting Cards','Accessories',
                       'Books, Music & More','Gift Categories','Classical','Camera & Photo','Calendars', 'Libros en español',
'Foreign Language Books', 'Parenting & Families','Engineering'
                       'Large Print', 'Stephen E. Ambrose', 'Miscellaneous', 'HOWdesign Studio','John Wiley & Sons',
'Opera & Vocal']
book = book[~book['categories'].isin(unwanted_categories)]
book
```

Figure 6. Process non-existent book genres

Hence, through SNA techniques, this study sets out to explore different types of interactions and networks, in particular, co-purchased books and co-purchases between genres interactions to gain a deeper understanding of the impact of these relations on the similarity between products.

Co-purchased book interactions refer to the analysis of products that are often purchased together, providing insights into associations between specific items. Co-purchases between genres interactions involve the investigation of product categories that are purchased together, providing a broader understanding of consumer preferences at more abstract levels.

Co-purchased books and co-purchases between genres:

The primary criterion for determining similarity between two products is whether they belong to the same genre. This implies that grouped products share similar characteristics or purposes, suggesting an intrinsic connection between them.

Additionally, the research adopts a more dynamic approach by considering consumer purchasing patterns. The identification of products as similar when their genres are frequently bought together. This observation is crucial as it reflects consumer trends and preferences, providing valuable insights into purchasing behavior.

Finally, the identification of products as similar if they were purchased together. This co-purchase suggests a strong correlation in consumer choice, indicating that these products are often viewed as complementary.

4.1.2 Phase 2 - Network Analysis

By leveraging the community detection capabilities of Gephi to analyse and identify communities within diverse networks, in this phase the aim is to gain insights

into the complex dynamics of the present networks, contributing to a more nuanced interpretation of the data.

With the data duly prepared, the study was segmented with the aim of examining existing interactions and identifying potential improvements. Two main interactions were analysed: the co-purchased book relation and the co-purchases between genres relation.

4.2 Technologies and Programs Utilized

To achieve the purposes established previously the coding process is executed utilizing Python programming language. Additionally, for network analysis, the Gephi platform is employed, facilitating the visualization, analysis, and manipulation of networks and graphs. Hence, this systematic approach sets the stage for an insightful analysis as part of this report.

4.3 Research Scope

The development of the research was propelled by critical disciplines, specifically Programming Languages I, Programming Languages II, and Data Science.

Marked by the imperative task of data processing, accomplished through the use of Python programming language and Pandas's library.

Within the Data Science curriculum unit, valuable knowledge in areas such as Data processing was acquired and effectively applied, further enriching the research development process.

5 Results

5.1 Co-purchases - Book Genre Analysis

In the analysis of co-purchasing networks, a relation is the existence of a purchase between products or genres.

The co-purchases between genres analysis focused on investigating the various literary genres, including which genres tend to be purchased together and the frequency of purchase of books belonging to the same genre compared to books of different genres. In doing so, the research tries to find answers to the following questions, to enhance the understanding of co-purchases between genres relations:

- How often does one buy books in the same genre compared to books in different genres?
- Which different genres are often bought together?

For a better understanding of the purchasing frequency of books within the same genre versus those of different genres, a strip plot was constructed as shown in Figure 7. Thus, there's a need to know that a strip plot is a type of graph used in statistics and data analysis to represent the distribution of data in one or more categories, and that 'Weight' represents the number of books in the given relation.

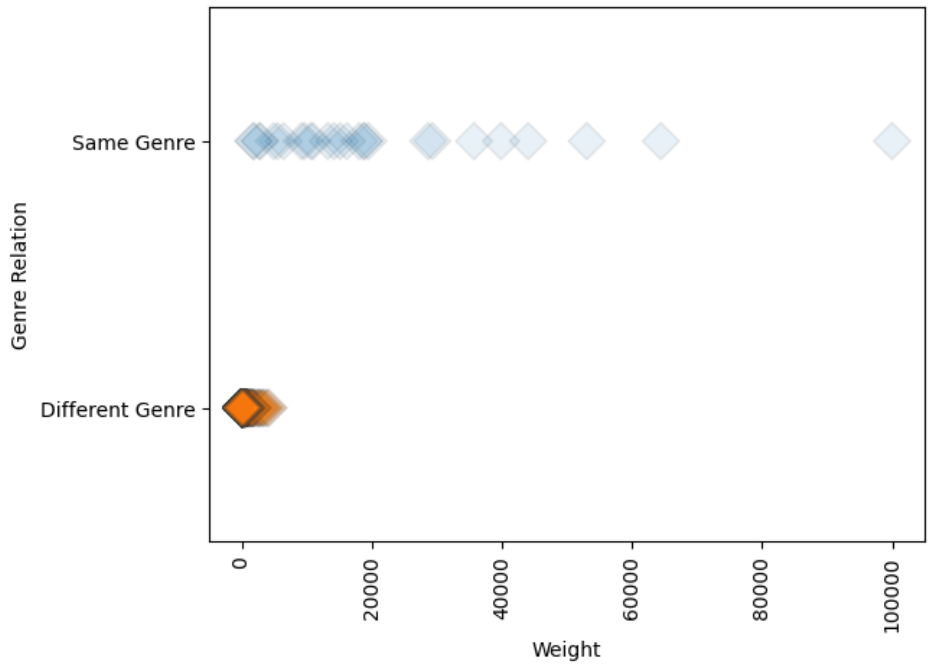

Figure 7. Comparative Distribution of Book Purchases by Genre Relation and Weight

Through an analysis of the data, comparing the results of the grouping of points, the concentration of points closer to the lowest value on the x-axis, for the 'Different Genre', suggests that, for books in different genres, the frequency of purchase is generally lower. For the 'Same Genre', the points also show some concentration, however there are points that extend further along the x-axis, which means that books of the same genre have higher purchase frequencies when compared to books of different genres.

Statistics were observed to better understand the results presented: the maximum weight, the minimum weight, the average weight, the percentage of books of the same genre, the percentage of books of different genres, and the relative frequency. For relations between different genres, the maximum recorded weight is 4173, while the minimum is 1, and the average settles at approximately 300.94. In contrast, for books purchased within the same genre, the data portrays a considerably higher maximum weight of 99.954, a minimum weight of 1778, and a notably higher average weight of 22058.11. The much higher values in the same genre category imply that books within the same genre are more frequently bought together and consequently strongly connected than books from different genres.

Furthermore, statistics showed that 75% of products purchased together are of the same genre, and by analyzing the relative frequency of the genres, it is noted that books of the same genre are bought more frequently than books of different genres, additionally, it is possible to point that two books of the same genre are purchased 3.1 times more frequently than books of different genres.

Given the much higher weight of same-genre relations, when compared to different-genre relations, it was necessary to exclude these from the scope of the study of co-purchases between genres relations to ensure an impartial analysis.

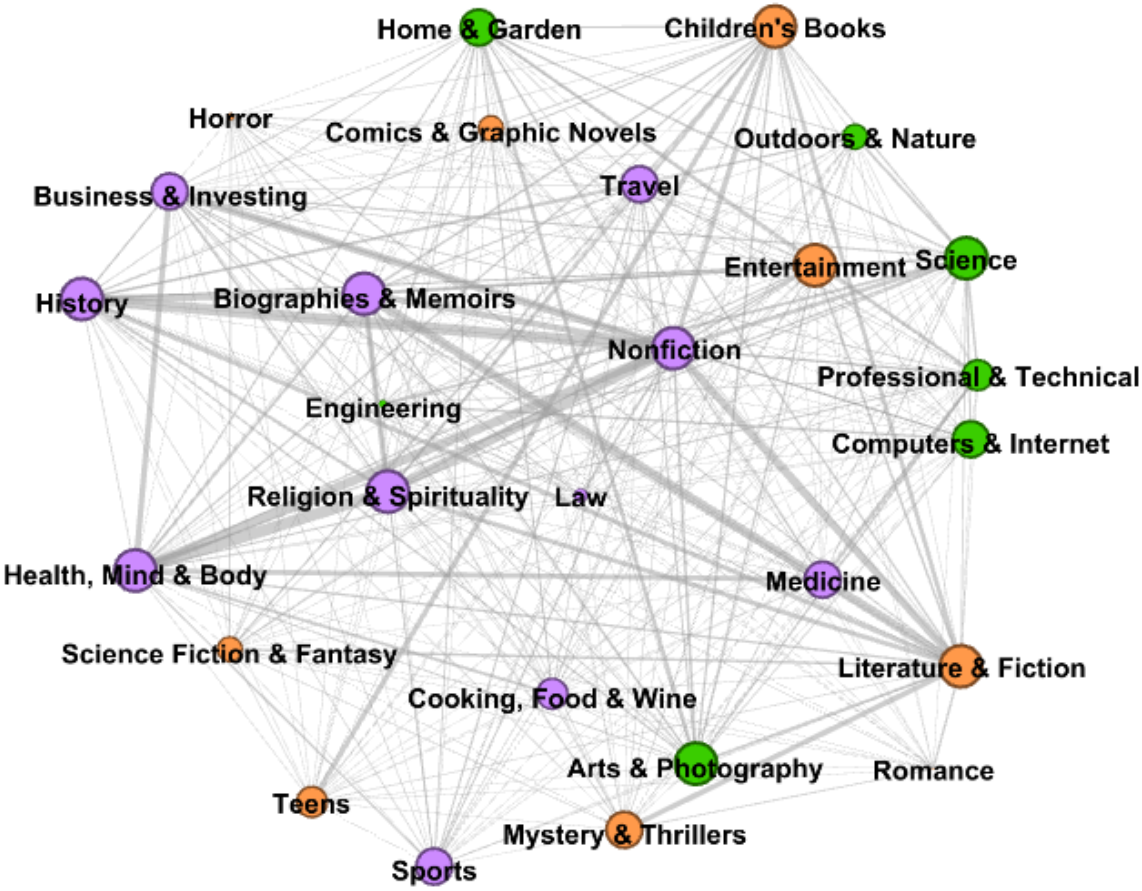

Figure 8. Network category-category

Figure 8 produced using Gephi, where a node represents a genre, and an edge represents a joint purchase with another genre, illustrates the relations between literary genres, highlighting how they are interconnected.

Additionally, through community detection, it is possible to identify existing communities within this network and to infer through the thickness of some edges, which provide insights into the proximity and frequency of interactions between different book genres, that buyers who are interested in one of these themes also tend to be interested in the other. Thus, identified by orange, there's a community characterized by literature, which includes genres aimed at young audiences. Identified by violet, a community involving human knowledge and cultural themes, and lastly identified by green covers technical and scientific themes.

Further analysis made it possible to understand which literary genres played a more significant role in the network, identify those purchased most frequently, and delineate their relational dynamics with other genres. The analysis of these relations allowed to

discern that books belonging to the same genre tend to be bought together more regularly, showing that such connections have considerable relevance in the structure of the network.

When examining the frequency of purchases by genre, it is evident that data normalization is crucial, as the weight can disproportionately affect the results.

To normalize the data, a table was created showing the relationships between categories. It includes the weight of the relation between two genres bought together and the number of books in each genre. To get the normalized value, the relationship weight is divided by the minimum value, which is the lower quantity of the two genres.

Thus, by analyzing the normalized and non-normalized weight data, it is possible to see that there are differences in the most prevalent relationships in the dataset. Firstly, Figure 9 shows that genres with the largest number of books also have the most prevalent relationships when considering the non-normalized weights. This means that, without normalization, the genres with the most books dominate the relations due to their volume.

Source	Target	Weight
Religion & Spirituality	Health, Mind & Body	4173
Nonfiction	History	4057
Nonfiction	Religion & Spirituality	3445
Nonfiction	Health, Mind & Body	3375
Biographies & Memoirs	Literature & Fiction	3226
Nonfiction	Business & Investing	2797
Nonfiction	Literature & Fiction	2689
Biographies & Memoirs	History	2558
Biographies & Memoirs	Nonfiction	2176
Biographies & Memoirs	Religion & Spirituality	2066

Figure 9. Non-normalized weight data

However, by normalizing the data, it becomes clear which relationships are actually more prevalent regardless of the absolute number of books in each genre. Normalization adjusts the values so that we can identify relations between genres proportionally, revealing pairs of genres that, although they may have fewer books in absolute terms, are more significant in relative terms.

For example, in the normalized data, in Figure 10, we see that the relation between “Children's Books” and “Teens” has a very high normalized weight, indicating a strong

relative prevalence, although "Teens" does not have the most books in absolute terms. This shows that normalization is essential to understanding the true dynamics between the genres in the dataset, avoiding biases caused by differences in quantity.

Source	Target	Normalized_weight
Children's Books	Teens	1.425926
Science	Outdoors & Nature	0.479279
Law	Nonfiction	0.471572
Nonfiction	History	0.387081
Engineering	Science	0.365537
Biographies & Memoirs	Literature & Fiction	0.357570
Literature & Fiction	Horror	0.341187
Medicine	Health, Mind & Body	0.333395
Law	Business & Investing	0.284950
Biographies & Memoirs	History	0.283529

Figure 10. Normalized weight data

5.2 Co-purchased Books Analysis

In analyzing co-purchased book relations, unlike the co-purchases between genres approach, there is no need to exclude products with same-genre relations, as these do not affect result integrity.

Thus, to better understand the existing relations, this analysis focuses on finding answers to the following questions:

- Network modeling allows us to go beyond recommendations: who bought X also bought Y?
- What is the genre variability within communities?

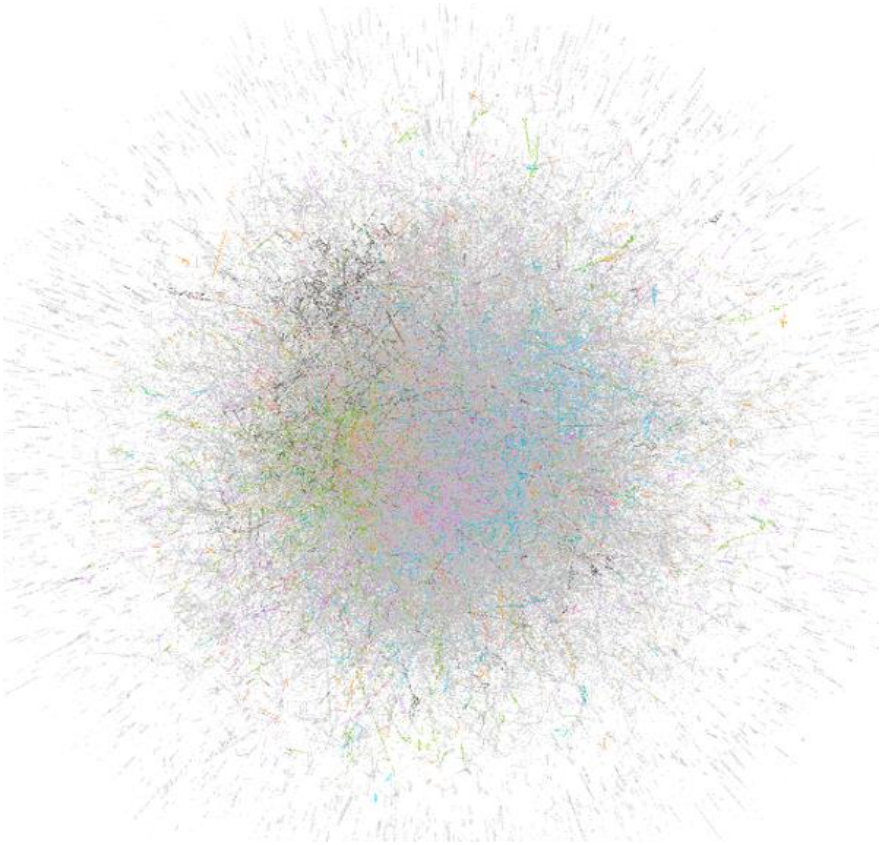

Figure 11. Network co-purchased products

Through Gephi, the corresponding network, shown in Figure 11, was created. The analysis of this network provides data such as modularity, which can range from -1 to 1, and in the present case is 0.944, which is justified by the presence of isolated communities that have dense internal connections but sparse external connections.

3957 distinct communities within the network were identified and it was revealed that, on average, each contains around 38 products, with a minimum of 2 and a maximum of 4736 products per community.

Furthermore, of 3957 communities, 1291 are communities with more than five elements, and 1054 contain more than 6 elements. These communities gather a total of 142.150 products from the initial 151.266 studied throughout this segment of the co-purchased book analysis. Consequently, it can be concluded that for 93.97% of the total number of products, it is possible to increase the number of recommendations initially assigned.

Given this, it is possible to gain deeper insights into the diversity within communities which aids in tailoring content to better meet customers' interests. By observing the number of unique genres in each community, the variability suggests that while 62% are

communities with only one genre, some communities exhibit a wider genre diversity, where multiple interests intersect, while others are highly specialized niches, which allows to recommend books from different genres, providing customers with a broader range of options.

This suggests that personalized recommendations within these communities can be targeted, catering to specific interests. Moreover, the presence of highly diverse communities, although less common, provides an opportunity to offer broader recommendations in those segments.

Thus, to expand the recommendation lists for each book, a check was conducted. For each element, it was verified whether any of the five initially assigned similar elements were also integrated into the same community. If a similar book was found within the community, the elements assigned to it as similar were then analyzed. If these products were not already present in the main book's community, they were designated as new similar products. This process allowed for the expansion of the similarity network, ensuring that new relevant elements were incorporated into the communities, thereby increasing the recommendations.

6 Conclusion

This study explored the existing dynamics of joint purchasing patterns to enhance recommender systems within the e-commerce domain. By employing data analysis and social network analysis techniques, it was possible to gain deeper insights into existing product relations and understand how these can improve a recommendation algorithm.

The study sought to obtain answers to several key questions to achieve the established goal. Specifically, aimed to understand how often individuals buy books within the same genre compared to books in different genres, which different genres are frequently bought together. The research also explored how network modeling can go beyond simple recommendations like "who bought X also bought Y," through the analysis of patterns of co-purchase of books, thus expanding the initial number of recommendations based on data from the communities of book genres and products. Lastly, examined the genre variability within communities to understand the diversity of reading preferences.

Findings and Analysis

The analysis demonstrated that products within the same genre tend to be purchased together more frequently than those from different genres. Statistics showed that 75% of products purchased together are of the same genre, and the frequency of purchasing books from the same genre is 3.1 times higher than those from different genres. Additionally, the network constructed based on co-purchases revealed which genres are often bought together, providing insights into genre relations and their influence on purchasing patterns.

Network modeling extended the research beyond traditional recommendations, offering a broader view of interactions and purchasing patterns between different products and genres.

Furthermore, the genre variability within communities indicated a predominant trend of purchasing books from the same genre together. However, there is also notable diversity within these communities, reflecting varied consumer interests and suggesting a blend of reading preferences among buyers.

This led to understanding several aspects, from the communities' structures, to how the quantity of books affects the analysis, and finally how to expand the number of books initially attributed as similar to each book.

This was crucial since the initial goal of the research was to gain a comprehensive knowledge of how smaller businesses can gain a strategic advantage against larger enterprises.

The analysis of co-purchases between book genres and individual products allowed us to understand how product similarities and consumer purchasing patterns can be utilized to improve recommender systems. Data normalization was crucial to avoid biases, ensuring that the most significant relations were accurately represented. The application of SNA techniques revealed community structures and node-level measures that can enhance the accuracy of recommendations.

To conclude, as it was possible to expand the existing recommendations, this research can contribute to a more enhanced recommender system, since the concrete insights extracted could be utilized not only for recommendation purposes but also, for example, to organize store items more effectively. Similar genres could be placed closer together, thus, better meeting the preferences of customers in the e-commerce environment by improving the overall shopping experience.

Method and Planning

The timetable for the development of this research project takes on Agile methodology tools, specifically, Gantt Chart as the project planning tool, which is a tool for visualizing a project timetable. The graphical representation, figure 6, depicts periods that mark the start and end of each phase as colored bars along the chart's horizontal axis.

According to the planned deliverables, four phases have been outlined to achieve the proposed solution. The initial phase involves conducting research and consolidating knowledge on SNA and RS. The subsequent phase focuses on data cleansing and organization using Python and Pandas, and additionally constructing a network dataset through Gephi. In the third phase, SNA is applied to the previously built network, and the obtained results are analysed and studied to enhance the understanding of the subject and provide a more effective solution to the problem outlined in the study.

During the current research phase, significant strides were made not only to answer key questions that enhanced our understanding of the dataset. Furthermore, DataFrames that efficiently map the co-purchased book relations and category-to-category were successfully created, a crucial step for deepening our understanding of the dataset and possibilities for a recommender system.

	Name	Duration
1	**TFC**	**192 days?**
2	**First Semester**	**80 days?**
3	**Progress Report**	**36 days?**
4	Problem Identification	2 days?
5	Research on Social Network Analysis	15 days?
6	Research on Recommender Systems	13 days?
7	Report revision and improvements	5 days?
8	Report submission	1 day?
9	**Mid-Term Report**	**41.875 days?**
10	Data processing	17 days?
11	Data analysis	10 days?
12	Creation of the networks	3 days?
13	Description of the networks	4 days?
14	Report revision and improvements	5 days?
15	Report submission	0.875 days?
16	**Second Semester**	**111.875 days?**
17	**Progress Report**	**59.875 days?**
18	**Application of social network analysis thecniques**	**59.875 days?**
19	Detection of communities	9 days?
20	Result analysis	20 days?
21	Comparison of the various networks	25 days?
22	Report revision and improvements	5 days?
23	Report submission	1 day?
24	**Mid-Term Report**	**52 days?**
25	Comparison of the various networks	25 days?
26	Final result analysis	10 days?
27	Conclusion of fiinal Report	10 days?
28	Revision and Improvements of report	7 days?
29	Final Report submission	1 day?

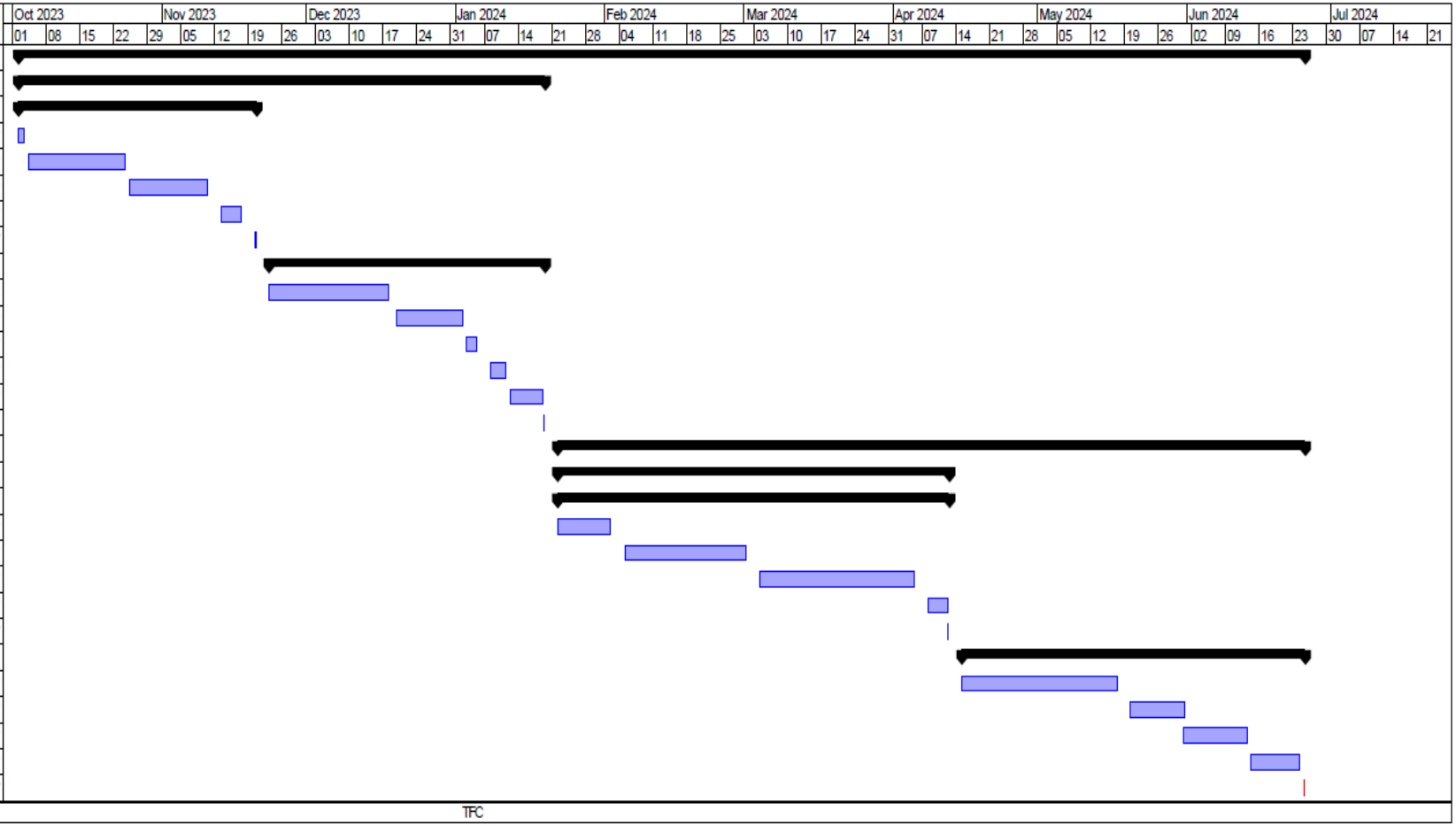

Figure 12. Gantt Chart Plan

References

[1] J. Leskovec, "Amazon product co-purchasing network metadata," 2006. [Online]. Available: https://snap.stanford.edu/data/amazon-meta.html. [Acedido em October 2023].

[2] S. Tabassum, S. Fernandes, F. S. Pereira e J. Gama, "Social network analysis: An overview," *Social network analysis: An overview,* p. 2, April 2018.

[3] "What is a Social Network Analysis?," [Online]. Available: https://kmb.camh.ca/eenet/sites/default/files/pdfs/General-SNA-Methodology-Overview-One-Pager.pdf. [Acedido em novembro 2023].

[4] "Social Network Analysis - Algorithms and measures to understand networks," [Online]. Available: https://cambridge-intelligence.com/social-network-analysis/. [Acedido em November 2023].

[5] A. F. Carvalho, "Movie Recommendation based on User Interests," *Movie Recommendation based on User Interests,* October 2014.

[6] R. Burke, A. Felfernig e M. H. Göker, "Recommender Systems: An Overview," *Recommender Systems: An Overview,* September 2011. .

[7] E. Frolovz e I. Oseledetsz, "Tensor Methods and Recommender Systemsy," *Tensor Methods and Recommender Systemsy,* p. 3, 18 February 2018.

[8] G. Verma, S. Simanta, H. Chen, D. Pillai, J. P. McCrae, J. A. Perge, S. Sengupta e P. Buitelaar, "Empowering recommender systems using automatically generated Knowledge Graphs and Reinforcement Learning," *Empowering recommender systems using automatically generated Knowledge Graphs and Reinforcement Learning,* 11 July 2023.

[9] F. Fessant, L. Candillier, K. Jack e F. Meyer, "Chapter 1," em *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*, 2009.

[13] D. Lee e K. Hosanagar, "How Recommender Systems Influence Customer Behavior | by Tricon Infotech," *Medium,* 28 June 2019.

[14] A. Sadeghian, A. M. Madni, H. Rahanam e S. Sivapalan, "Recommender Systems in E-Commerce," *Recommender Systems in E-Commerce,* Vols. %1 de %2Proceedings, 2014 World Automation Congress (WAC), August 2014.

[15] Gonçalves-Sá, J., & Pinheiro, F. (2023). Societal Implications of Recommendation Systems: A Technical Perspective. In Law, Governance and Technology Series (Vol. 58, p. 17). 10.1007/978-3-031-41264-6_3

[16] "BigCommerce. (Year). 'Title of the Webpage.' [Online]. Available: https://www.bigcommerce.com/articles/ecommerce/recommendation-engine/#h2_best_ecommerce_recommendation_engines"

[17] J. Primo, D. Mateus e D. Mineiro, "NORMAS PARA A ELABORAÇÃO E APRESENTAÇÃO DE TESES, DISSERTAÇÕES E OUTROS TRABALHOS ACADÉMICOS," *NORMAS PARA A ELABORAÇÃO E APRESENTAÇÃO DE TESES, DISSERTAÇÕES E OUTROS TRABALHOS ACADÉMICOS,* p. 50, June 2023.

Annex

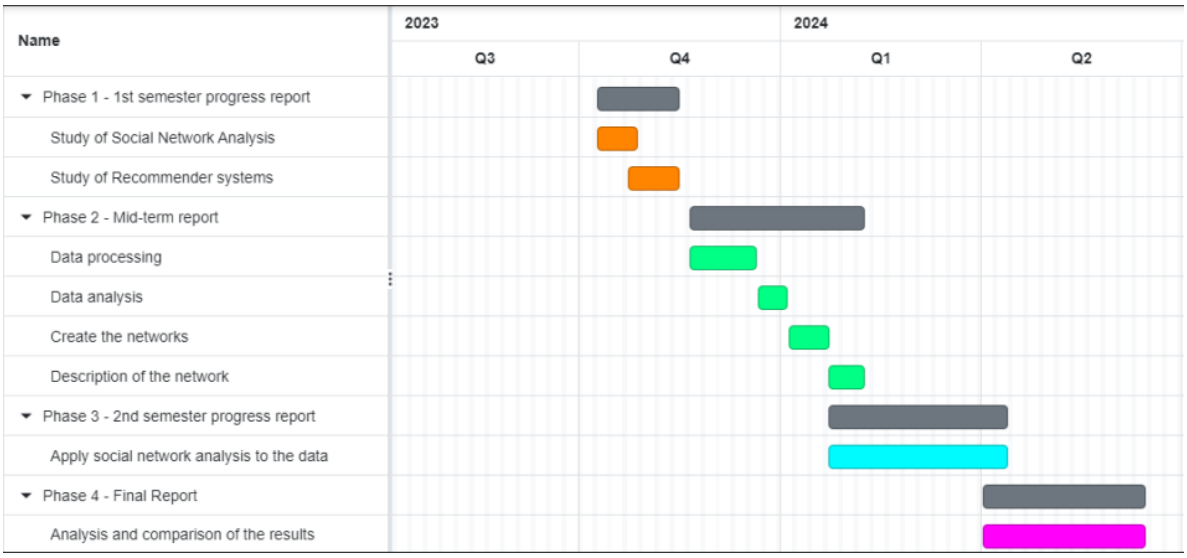

Figure 13. Preliminary Gantt Chart Plan

Throughout the research development, I encountered a few challenges. Initially, I faced difficulty in structuring the data, as I had not previously worked on a project of this nature. This lack of experience led to uncertainty about how to organize the data effectively to extract meaningful information for the study. The final structure, as depicted in Figure 14, was devised to address this challenge, ensuring that all pertinent information for the study of a given object is consolidated in a single row.

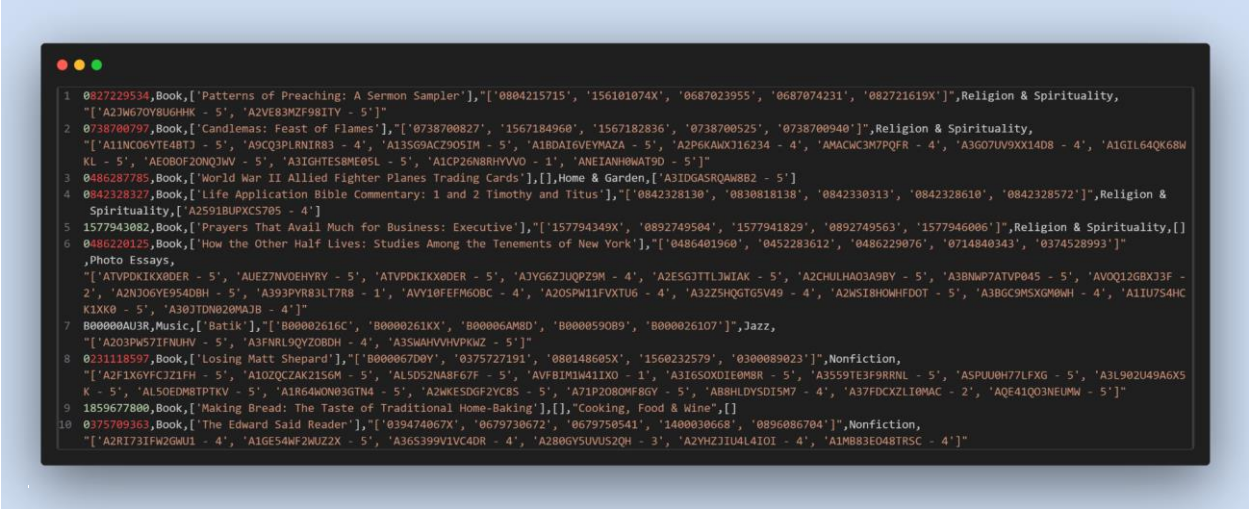

Figure 14. Initial data structure

List of Acronyms

CB	Content-Based
CF	Collaborative Filtering
FCW	Final Course Work
HF	Hybrid filtering
RS	Recommender System
SNA	Social Network Analysis