A Framework for Analyzing Book Covers and Co-purchases using Object Detection and Data Mining Methods

Brenda Cinthya Solari Berno, Ademir Cristiano Gabardo, Leandro Takeshi Hattori,
Matheus Gutoski, Andrei de Souza Inácio, Heitor Silvério Lopes
Federal University of Technology - Paraná
Paraná, Curitiba 80230–901
Email: hslopes@utfpr.edu.br

Abstract—**Co-purchase analysis is becoming an essential task for e-commerce businesses as it enables to predict new purchases and understand customer behavior. In this paper, we propose a novel framework to analyze books sales from co-purchase and visual data. An existent dataset created from Amazon co-purchase data was extended to include visual information from book covers. Deep Learning was used to perform object detection and Data mining techniques were used to analyze relationships concerning book co-purchases and the influence of visual information on books category. Our results revealed several interesting relationships between co-purchases and predominant colors in the cover, as well as between objects and categories.**

I. Introduction

Online shopping is increasingly becoming a part of our lives. It is also pushing businesses to collect useful data to understand their customers and create products that meet their requirements. This data also contributes to the creation of recommender systems, which help customers discover items they might otherwise not have found.

Suggestions made by recommender systems affect customers decisions in different ways, such as selecting books, music, or other items [1]. Companies utilize these systems to increase their sales by offering relevant products and increasing customer satisfaction [2]. The subject has become a relevant academic research topic concerning understanding and analyzing customer purchase preferences [3].

In addition to analyzing customer preferences and co-purchasing networks of items, approaches have considered visual relationships in their studies. These methods are capable to recommend, for instance, clothes and accessories that go well together or predict which books will be co-purchased based on their cover art [4], [3].

Book covers have been studied as objects of art and as collectibles [5]. Publishers usually design covers to attract readers and inform about the book's contents. The objects and shapes depicted in the covers provide relevant information to recommend a book to a specific type of customer. For instance, covers containing animals or toys may be more attractive to children. However, extracting high-level visual features as well as objects from book covers is not a trivial task. One option is to use human labor to annotate each object. However, this becomes impractical as the volume of data increases.

Recent advances in Computational Intelligence and Deep Learning have provided means to perform the object extraction task automatically. In particular, the Darknet framework [6], [7] is well-suited for the task, since it can detect more than 9000 object categories.

Once the high-level semantic visual information is extracted from the covers, it is possible to search for several kinds of useful relationships between books. Computational approaches such as Classification Rules and Clustering allow performing this knowledge discovery task.

This work explores two main aspects of knowledge discovery concerning books: (1) the book co-purchase behavior, which uses information such as book categories, predominant cover colors, and sales data; (2) the relationship between cover objects and book categories, which uses the high-level semantic visual information extracted by the Darknet framework. To the best of our knowledge, this is the first work that uses visual information of the book cover, such as colors, shapes, and objects to discover interesting information about book purchases.

The main contributions of this work are:

- A methodology for extracting, analyzing, and integrating features of book covers;
- A dataset of books and its covers with pre-extracted objects with probabilities for each book;
- An analysis of the behavior of book co-purchases concerning predominant cover colors and categories;
- An analysis of book categories in relation to the objects and colors present in the cover.

This paper is organized as follows: Section II presents the related work, Section III describes the background methods, Section IV describes the method and all its steps, Section V present the experiments and their results. Finally, in Section VI conclusions and future directions are pointed out.

II. Related works

In [4], a dataset of products purchased from Amazon Web Store was presented and they proposed a similarity measure between images using features extracted from a Convolutional Neural Network. The authors perform clustering of the products to discover the relation between the objects based on their appearance.

Bucur et al. [8] used the same dataset proposed by [4] to measure gender homophily and analyze preferential gender association in book consumption. They used clusterization of graphs to gendered book communities detection and confirm

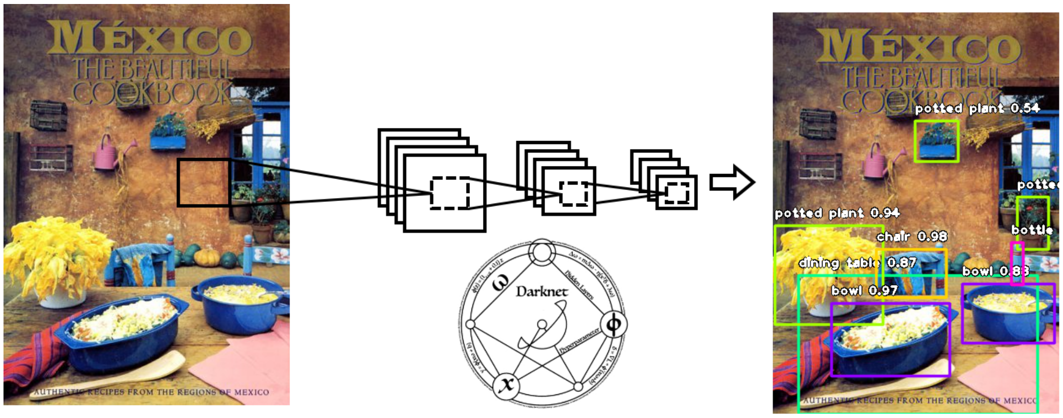


Fig. 1: Object detection in a book cover using Darknet.

their hypothesis analyzing the output of Louvain community-detection algorithm.

Shi et al. [9] analysis was about the science preferences of those who purchase liberal and conservative political books. Also, other platforms such as Goodreads.com were used to identify book preferences according to the gender of the person [10], or to understand whether the collective reading behavior on Goodreads can distinguish the Amazon best sellers from the rest of the books [11].

Some studies applied the analysis of books design in areas such as psychology and design. [12] presented a study on the influence of the book cover on the resist of a reader. The results indicate that in rapid decision making in book choice cover image and table of contents play a significant role. Another research suggests that book design can impact on the practice of children's reading and how books with more information on the cover have a higher rate than books with simple covers [13]. These results indicate the influence of the book cover analysis since they impact sales and customer satisfaction.

III. Background

This Section briefly presents the main approaches used in this work: Darknet, Classification Rules and Force Atlas 2.

Darknet [6] is an open source neural network framework written in C and CUDA, which served as a base to create the YOLO (You Only Look Once) network [14]. The version used in this work can detect 9,000 object categories. The model takes as input an image and outputs the coordinates of each detected object, along with their probability. It is important to note that, despite being a state-of-the-art object detection method, some False Positives and False Negatives can be expected from the model. Figure 1 shows a sample output of the model.

JRip (RIPPER) [15] is one of the most popular algorithms that implements a propositional rule learner. The algorithm learns Classification Rules by searching the data for frequent if-then patterns.

Force Atlas 2 [16] is a layout algorithm used for network spatialization that allows to transform networks into maps. This work uses the implementation of Force Atlas 2 provided by the software Gephi[1].

IV. Materials and Methods

An overview of our method is presented in Figure 2. First, it filters the original dataset to remove irrelevant data. Then, a Webcrawler is used to collect additional data. These steps are described in Section IV-A. The next step is to extract color features and objects from the cover images, as presented in Section IV-B, generating the extended dataset. The extended dataset is used for knowledge discovery using data mining methods such as JRip and Force Atlas 2, as presented in Section IV-C.

A. Data Acquisition and Preprocessing

The dataset presented by [4] was used as a starting point for this work. Since the original dataset has data not related to books, a preprocessing step was performed to remove the irrelevant data. The total number of books obtained afterward this step was 59,173 instances, and for each instance, it stored eight information (see Table II).

We extend the dataset by adding new features related to the book covers, in specific the cover images, predominant colors and the objects in the cover with their respective probabilities. Moreover, a WebCrawler was used to collect more information from the existing books in the database. This data collection includes updating missing data when possible and adding information such as the year of publication, categories, subcategories, and description. The WebCrawler downloads the book cover image when available. But, the number of books decreased to 39,873 since these books do not contain a cover image on their page. The data will be available at https://labic.utfpr.edu.br/datasets.html.

B. Feature and Object Extraction

This step has two main purposes: (1) extract the predominant colors; (2) extract the objects from the covers.

[1] Gephi website: https://gephi.org/

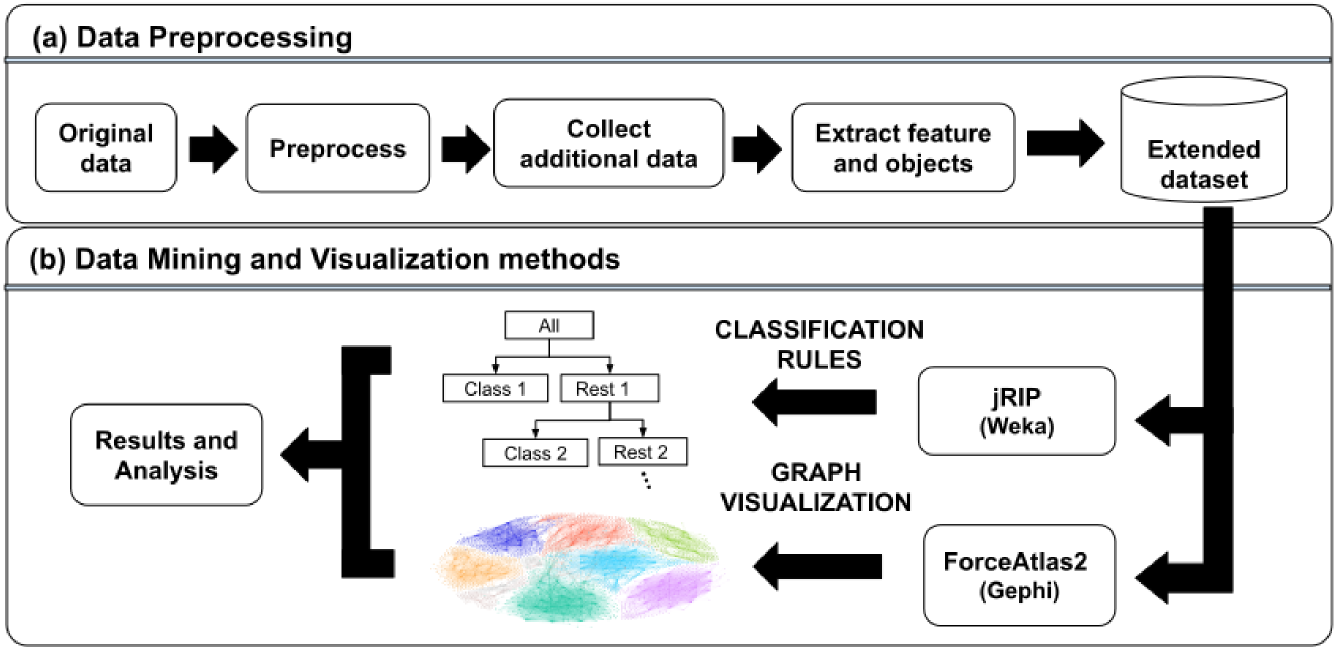


Fig. 2: Methodology Overview: (a) filtering, collecting and feature extracting (b) results and analysis from JRip and Force Atlas 2.

1) Get Colors: The goal of this step is to get the percentage of black, red, green, blue, yellow, magenta, cyan, maroon, purple, orange, gray and white (12 colors) present in each cover in RGB color space. The algorithm calculates the Euclidean distance between each pixel and each of the 12 colors and assigns the pixel to the closest color. We reduced images to a maximum area of 80,000 pixels for reason of computational complexity. This value was chosen empirically as a trade-off between computational time and image quality.

2) Get Objects: In this step, the objects that appear in each cover book are detected by the Darknet [7]. The network recognized 1,080 different object types in all books. This data (object, probability) was stored in the extended dataset. Finally, some objects were excluded since they were detected as the book itself, such as “book”, “notebook”, “comic book”, “book jacket”, etc.

C. Data Mining and Visualization methods

This section presents the method used for two distinct analyses, the co-purchase of books and the analysis regarding the book covers and their categories. The co-purchase analysis employs the Force Atlas algorithm implemented in Gephi, while the book cover analysis employs the JRip algorithm implemented in Weka. The communication between different tools is indirect, since they share the same input.

1) Co-purchase Analysis: The co-purchase analysis was performed using visualization tools provided by the Gephi implementation of Force Atlas 2. After several experiments with different parameters, we empirically defined the parameters presented in Table I, since they revealed relevant relationships between co-purchased books. The attributes used for generating the graphs are shown in Table II. The result of this algorithm is a graph where each node represents a book, and an edge $A \rightarrow B$ represent that a book A is purchased together with a book B. The width of the edge represents the number of times this co-purchase occurred. In order to detect only important relationships, we remove every edge with fewer than 100 input edges. This filter results in a total of 2971 nodes and 10910 edges.

TABLE I: Force Atlas parameters of the Gephi platform

Tuning	Scaling	55.0
	Strong Gravity	True
	Gravity	0.6
Behavior Alternatives	Dissuade Hubs	True
	Prevent Overlap	True
	Edge Weight Influence	1.0

TABLE II: Informations and descriptions of each instance of the dataset.

Information	Description
asin	ID of the book
title	booktitle
related	related products (ID of books that were purchased together)
categories	list of categories the product belongs to
colors	percentage of each color present in the cover

2) Book cover analysis: The book cover analysis was performed using the JRip algorithm implemented in Weka[2] using its default parameters. Before using JRip, three data

[2] Weka website: https://www.cs.waikato.ac.nz/ml/weka/

TABLE III: Amount of books, objects, and categories after each filtering processes.

	# books	# objects	# categories
Initial Data	39873	1061	247
1º filter	12209	63	46
2º filter	9568	47	29
3º filter	9568	47	14

filtering steps were taken. The first step disregards all object categories that appeared in less than 100 book covers and also ignores any Darknet object detection with a probability of less than 50%. The second step merges similar object categories to avoid semantic duplicates (e.g., teddy and teddy bear). After the merges, categories with fewer than 70 instances were discarded. Finally, the third step reduces the number of categories 14 by grouping them by semantics. For example, the categories "Christian books & bibles" and "Religion & Spirituality" were combined. Increasing the number of instances per category allows JRip to generate more significant classification rules.

These filtering steps increase the reliability of the data at the cost of reducing the number of books, objects, and categories. Table III presents the number of instances before and after applying the filters.

Another necessary preprocessing step was to remove the object "person". This is done because of the number of objects in this class overwhelms the other classes. The "person" class appeared approximately 25,000 times in the book covers, while the second most frequent class appeared only approximately 1,200 times. This heavy unbalance causes the JRip algorithm to be biased towards the "person" object. Moreover, prior experiments have shown that this class is not informative and does not provide any interesting conclusion. Figure 3 presents the frequency of each object in the filtered dataset.

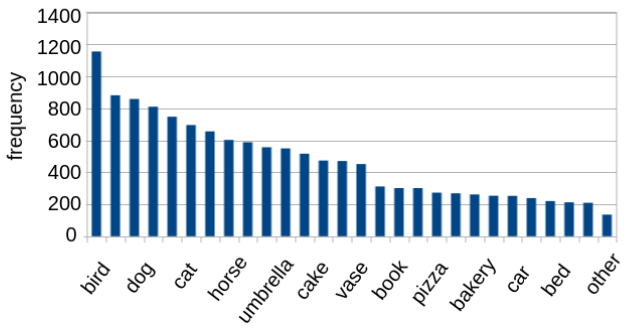
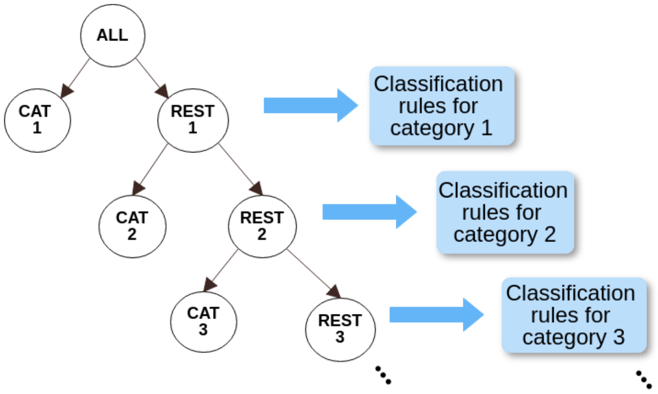


Fig. 3: Frequency of objects detected in the book covers.

The classification was performed using a hierarchical approach to account for the unbalanced classes. The process consists of classifying the majority class against all other classes to generate rules for this class. Then, the data belonging to the majority class is discarded and the second majority class is classified against the rest. This process repeats until the last two classes. Figure 4 illustrates this process.



ALL
CAT 1
REST 1
Classification rules for category 1
CAT 2
REST 2
Classification rules for category 2
CAT 3
REST 3
Classification rules for category 3

Fig. 4: Hierarchical classification in sub sets.

V. Experiments and Results

A. Experiment 1: Behavior of books co-purchased

This experiment has the aim to answer the following question: Does the co-purchases of books present unusual patterns that affect book sales? To tackle this question, we generate graphs using the method presented in Section IV-C1, in which two of them presented interpretable results.

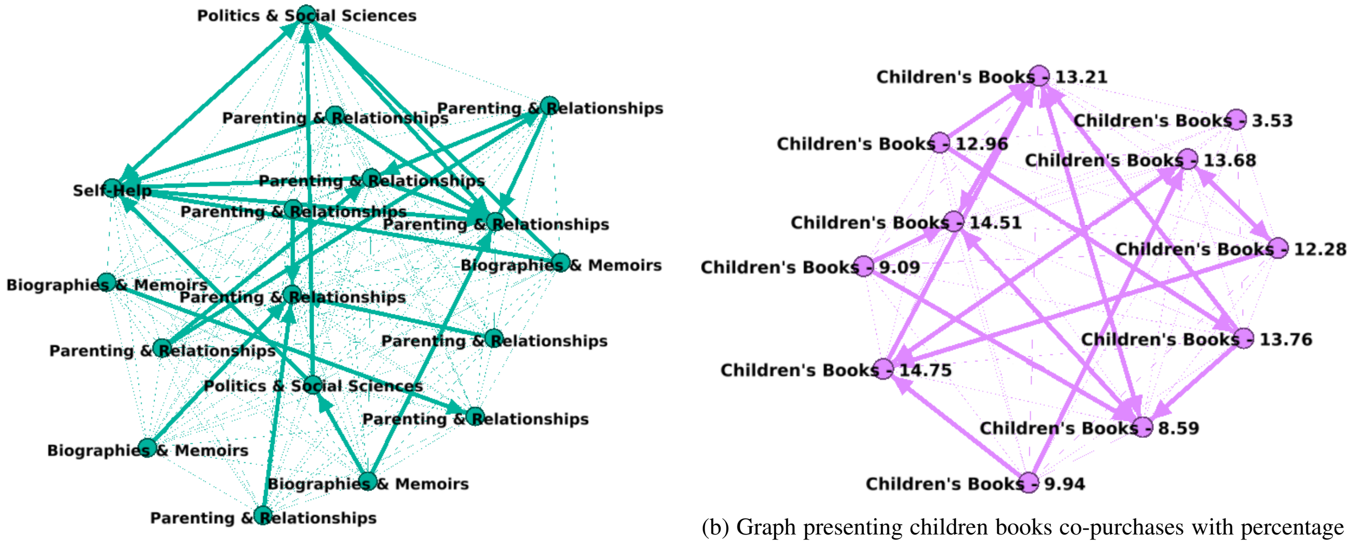
In Figure 5a, it was observed a relationship between Politics & Social Sciences, Parenting & Relationship, Biographies & Memoirsand Self-help books. This last category has a strong relationship with the other categories, which could mean how interest in self-help books influences books about relationships in society, whether in politics or parenting. The categories of Biographies & Memoirs and Parenting & Relationship also have an important connection in the graph. This result suggests that customers perceive reading about experiences to be more impact than reading a list of normative steps.

Another interesting relationship discovered concerns the category of Children Books in relation to the predominant colors present in the cover. Figure 5b shows that books with a larger amount of co-purchases present a predominance of the red color. This suggests that if the cover has a high percentage of the red color, this book may be more likely to be purchased. While a percentage of approximately 14% may seem low, it is significantly higher than the other colors.

B. Experiment 2: Behavior of books belonging to a category

This experiment pursues to understand the relationship between book categories and the objects present in their cover.

Using the method described in Section IV-C2 provided two different types of results. The first set of results is presented in Table IV. This table presents the classification quality metrics (sensitivity and specificity) obtained by predicting the book category using the objects detected in the cover as attributes. The results indicate that most categories are not predictable using only the objects in their cover since the classification metrics are low. However, the categories Children's books and Cookbooks, Food & Wine presented a significantly higher sensitivity and specificity compared to the other classes. This



(a) Graph presenting the relation between Politics & Social Sciences, Parenting & Relationship, Biographies & Memoirs and Self help books

(b) Graph presenting children books co-purchases with percentage of red color.

Fig. 5: Graphs generated by Gephi

suggests that these two categories are more correlated to the objects that appear in their covers. For instance, Children's books often contain animals and toys in their covers, while Cookbooks, Food & Wine often contain pictures of food, fruit or utensils.

The second set of results allow investigating the relationship between the cover objects and categories in a more detailed way. Table V presents rules of the type *if then else* found by the JRip algorithm. The rules can be interpreted as *if object, then category*. The table also presents the total number of instances in the category, the number of times this rule was positive (POS.) and the confidence of the rule (total number of instances divided by the number of positives).

Table V corroborates with the results of the previous experiment (Table IV), where objects like animals and food relate to Children's books and Cookbooks, Food & Wine. Some results presented in the Table are expected, such as chairs, vases and potted plants appearing in Crafts, Hobbies & Home books, and prayer rugs in Religion & Spirituality.

The table also shows unexpected relationships between objects and book categories, such as gowns and swimming trunks often appearing in Literature & Fiction books; Motorcycles in Science & Math books; cats in Mystery, Thriller & Suspense; and Crosswords puzzles in Business & Money.

VI. Conclusions and Future works

This work presented a method for extracting useful information from book sales using co-purchase data and visual attributes in the book covers using Deep Learning and Data Mining methods.

One of the main contributions of this work is the dataset introduced, which helps to carry out an analysis based on book covers and not only in the relations of purchases, which differs from other works in the literature that are based on other book attributes. The dataset contains high-level visual information extracted automatically from book covers, such as predominant colors and objects. A few difficulties appeared in the object extraction process because some cover designers use abstract objects Darknet not recognized that. Also, some covers were scanned in low quality or present stamps and stickers, which confused the object recognition algorithm. Nonetheless, this data can be used for other research purposes, thus contributing to the Data Mining and Computational Intelligence fields.

TABLE IV: Measured parameters for the generation of rules using Weka's JRip.

Category	**Sen.**	**Spe.**	**sen. x spe.**	**#Rule**
Children's Books	0.735	0.594	0.437	7
Cookbooks, Food & Wine	0.885	0.720	0.637	14
Literature & Fiction	0.776	0.375	0.291	5
Crafts, Hobbies & Home	0.847	0.287	0.243	4
Biographies & Memoirs	0.848	0.152	0.129	1
Teen & Young Adult	0.854	0.146	0.125	1
Mystery, Thriller & Suspense	0.828	0.289	0.239	2
History	0.848	0.218	0.185	3
Arts & Photography	0.763	0.241	0.184	2
Politics & Social Sciences	0.807	0.206	0.166	2
Science & Math	0.716	0.359	0.257	7
Religion & Spirituality	0.637	0.404	0.257	3
Health, Fitness & Dieting	0.567	0.550	0.312	7
Business & Money	0.567	0.550	0.312	7

The first experiment showed some interesting relationships regarding book co-purchases. It was observed strong connections between customers that buy Politics & Social Sciences, Parenting & Relationship, Biographies & Memoirs and Self-help books. Another strong relationship found is that cus-

TABLE V: Main objects in the JRip rules for each category classification, and for each object, the amount of positive classification (POS), the total amount of the instance, and confidence percentage (CONF) .

CATEGORY	OBJECTS	POS.	TOTAL	CONF. (%)
Children's books	teddy	517	591	87,48
	bird	692	1005	68,86
	cat	466	667	69,87
	dog	561	825	68,00
	horse	404	597	67,67
Cookbooks, Food & Wine	bowl	370	424	87,26
	cake	189	245	77,14
	pizza	143	171	83,63
	wine glass	104	135	77,04
	sandwich	72	77	93,51
	bakery	108	141	76,60
	donut	66	89	74,16
	fork	48	61	78,69
	orange	55	75	73,33
Literature & Fiction	gown	205	244	84,02
	swimming trunks	78	99	78,79
Crafts, Hobbies & Home	crossword puzzle	157	230	68,26
	chair and vase	11	13	84,62
	chair and potted plant	13	19	68,42
Mystery, Thriller & Suspense	cat	82	116	70,69
History	horse	76	102	74,51
	truck	20	27	74,07
Science & Math	motorcycle	17	18	94,44
	dog	25	36	69,44
	tv	23	34	67,65
	bird	85	127	66,93
Religion & Spirituality	prayer rug	15	16	93,75
	horse	13	16	81,25
Business & Money	tie	62	80	77,50
	chair	17	20	85,00
	crossword puzzle	11	12	91,67

tomers who buy Biographies & Memoirs tend to also buy Parenting & Relationship books. One interesting find was about how the predominant colors in the cover affect co-purchases. It was observed that Children's books often present a predominant red color. This may be due to that the color red is a bright color that attracts children and also psychologically expresses strength and determination.

The second experiment has shown unexpected relationships between objects and the categories in which they mostly appear. Despite finding some expected relationships, for instance, ties appearing in the Business & Money category, some other interesting relationships were found, such as cats appearing in Mystery, Thriller and Suspense books. It could be assumed that this creates an environment of terror or suspense. Another interesting find was that horses often appear in books related to History, probably because of the relationship that horses have with men throughout history.

Finally, this work encourages other researchers to extend this work by using other Data Mining techniques, which can lead to a better understanding of customer behavior when buying books.

ACKNOWLEDGEMENT

B.C.S. Berno thanks the Federal University of Technology - Paraná for the scholarship; L. T. Hattori thanks CAPES for the scholarship; M. Gutoski thanks CNPq for the scholarship number 141983/2018-3; H.S.Lopes thanks CNPq for the research grants no. 311778/2016-0 and 423872/2016-8. Special thanks to NVIDIA Corporation for the donation of the Titan-Xp GPUs used in this work. A. S. Inácio thanks for support from the Federal Institute of Santa Catarina.

REFERENCES

[1] F. Ricci, L. Rokach, and B. Shapira, Introduction to Recommender Systems Handbook. Boston, MA: Springer US, 2011, pp. 1–35.

[2] Q. Abuein, A. Shatnawi, and H. Al-Sheyab, "Trusted recomendation system based on level of trust," in 2017 International Conference on Engineering and Technology (ICET). Piscataway: IEEE, 2017, pp. 1–5.

[3] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," 2018.

[4] J. McAuley, C. Targett, Q. Shi, and A. Van Den Hengel, "Image-based recommendations on styles and substitutes," in Proceedings of the 38th International Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM, 2015, pp. 43–52.

[5] S. A. Knowlton and L. N. Hackert, "Value added: Book covers provide additional impetus for academic library patrons to check out books," Library Resources & Technical Services, vol. 59, no. 3, pp. 112–119, 2015.

[6] J. Redmon, "Darknet: Open source neural networks in c," http://pjreddie.com/darknet/, 2013.

[7] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," arXiv, pp. 1–6, 2018.

[8] D. Bucur, "On the gender of books: author gender mixing in book communities," in International Conference on Complex Networks and their Applications. NY, USA: Springer, 2017, pp. 797–808.

[9] F. Shi, Y. Shi, F. A. Dokshin, J. A. Evans, and M. W. Macy, "Millions of online book co-purchases reveal partisan differences in the consumption of science," Nature Human Behaviour, vol. 1, no. 4, p. 0079, 2017.

[10] M. Thelwall, "Book genre and author gender: romance¿ paranormal-romance to autobiography¿ memoir," Journal of the Association for Information Science and Technology, vol. 68, no. 5, pp. 1212–1223, 2017.

[11] S. K. Maity, A. Panigrahi, and A. Mukherjee, "Analyzing social book reading behavior on goodreads and how it predicts amazon best sellers," in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. NY, USA: Springer, 2018, pp. 211–235.

[12] D. McKay, G. Buchanan, N. Vanderschantz, C. Timpany, S. J. Cunningham, and A. Hinze, "Judging a book by its cover: interface elements that affect reader selection of ebooks," in Proceedings of the 24th Australian Computer-Human Interaction Conference. NY, USA: ACM, 2012, pp. 381–390.

[13] K. Wright, "A comparison of children's books: Picture books versus physically and intellectually adaptive interactive children's books," Ph.D. dissertation, University of Waikato, 2015.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2016, pp. 779–788.

[15] A. Rajput, R. P. Aharwal, M. Dubey, S. Saxena, and M. Raghuvanshi, "J48 and JRIP rules for e-governance data," International Journal of Computer Science and Security, vol. 5, no. 2, p. 201, 2011.

[16] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian, "Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software," PloS one, vol. 9, no. 6, p. e98679, 2014.