Exploratory Web Usage Mining for E-commerce: A Multi-approach Analysis Using Descriptive Learning Techniques

Jefferson G. M. Lopes[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Rua Reitor Píres Albuquerque, ICEx - Pampulha, Belo Horizonte - MG, 31270-901

jeffersong@ufmg.br

Abstract. *This paper explores web usage mining techniques applied to e-commerce, utilizing a multi-approach analysis with descriptive learning methods. By analyzing a dataset sourced from Harvard Dataverse, containing 3.5GB of Nginx access log data from "zanbil.ir," the study aims to uncover actionable information to improve marketing strategies, website usability, and profitability. Key objectives include identifying frequently co-viewed or co-purchased products, understanding navigational patterns, and distinguishing characteristics of buyers versus non-buyers. The study employs itemset mining and subgroup discovery algorithms to extract non-trivial information, providing a robust methodology for iterative data mining and business decision support in e-commerce environments.*

1. Introduction

Web usage mining has emerged as a technique in the analysis of user behavior on websites. This field leverages data mining algorithms to process extensive web server logs, enabling the extraction of potentially valuable patterns and knowledge that can inform business strategies.

The application of web usage mining in e-commerce is particularly significant due to its potential to reveal hidden patterns in user navigation and purchasing behavior. By analyzing web server logs, businesses can gain a deeper understanding of how users interact with their websites, which pages are most frequently visited, and what factors influence purchasing decisions. This understanding is crucial for improving website design, personalizing user experiences, and implementing targeted marketing strategies that can increase conversion rates and customer satisfaction.

Another significant advantage is the use of a data source, web access logs, that is not usually directly tied to the application database. With this, it is possible to perform extensive data analysis without consuming infrastructure resources that are directed to actual customers.

This study utilizes a dataset of access log data from the e-commerce website "zanbil.ir," collected over a four-day period. The primary objective is to employ descriptive learning techniques, specifically itemset mining and subgroup discovery, to extract meaningful information that can support business decisions aimed at enhancing marketing conversions, website usability, and profitability. The iterative nature of the proposed methodology ensures that the analysis can be refined and repeated to achieve optimal results.

This paper presents a detailed exploration of web usage mining techniques applied to e-commerce data, focusing on identifying frequently co-viewed or co-purchased products, understanding navigational patterns, and distinguishing characteristics of buyers versus non-buyers. The next sections are organized as follows: Related work presents past progress in this field, Methodology describes the data mining process and particularities of the dataset, Data Analysis presents and discuss the results of the mining process and, finally, Conclusion discuss the overall process and next steps.

2. Related Work

Web usage mining has been a significant area of research since the late 1990s, focusing on extracting meaningful patterns from web data to enhance user experiences and business outcomes. [Cooley et al. 1997] laid the groundwork by introducing methodologies for information and pattern discovery on the World Wide Web, setting a foundation for subsequent studies on web mining techniques and their applications in various domains .

[Srivastava et al. 2000] further expanded on these concepts, emphasizing the importance of discovering usage patterns from web data. Their work provided an overview of the processes involved in web usage mining and its potential applications in optimizing web structures and content delivery . These early contributions have been important in shaping the field and have influenced numerous studies aimed at understanding and leveraging user behavior.

In the context of e-commerce, web usage mining has been employed to enhance website design and functionality. A study on OrOliveSur.com [Carmona et al. 2012] demonstrated how mining web usage data could lead to improvements in website design, ultimately enhancing user satisfaction and increasing conversion rates . This research displays the practical benefits of applying web usage mining techniques to real-world e-commerce platforms.

The discovery of navigational patterns using self-organizing maps (SOM) has also been explored as a method for analyzing user behavior on websites. [Etminani et al. 2009] used SOM to uncover users' navigational patterns, providing relevant information into how users interact with web content and identifying areas for optimization . This approach has proven effective in visualizing complex user behaviors and aiding in the design of more intuitive web interfaces.

3. Methodology

The following section describes the data and the original format, the mining task objectives, the methodology applied to pre-process the data and analyze it.

3.1. Dataset

The dataset utilized in this study was sourced from the Harvard Dataverse repository, specifically titled "Online Shopping Store - Web Server Logs" [Zaker 2019]. It comprises 3.5GB of Nginx access log data from the e-commerce website "zanbil.ir," collected over the period from January 22, 2019, to January 26, 2019.

The data is provided in a simple text format, with each line representing an individual server access event. Table 1 describes the availble information per-line of the log file.

Field	Description
IP Address	The unique identifier of the user's device.
Timestamp	The date and time of the interaction.
Request Type	The type of HTTP request made (e.g., GET, POST).
URL	The specific URL requested by the user.
HTTP Response Code	The status code returned by the server (e.g., 200 for success, 404 for not found).
Referrer	The URL of the webpage that referred the user to the current page.
User Agent	Information about the user's browser and operating system.

Table 1. Description of fields in the dataset

3.2. Data Mining Task Description

This work aims to explore sources of non-trivial information within the web log file data. The extracted information can then be used to support business decisions aimed at improving marketing conversions, enhancing website usability, and increasing profit margins of an e-commerce company. The work of [Srivastava et al. 2000] laid multiple potential applications for this kind of data, such as system improvement, business intelligence, usage categorization and etc. This study is particularly interested in obtaining the following information:

- **Shopping Cart:** Identifying products often viewed or bought together can reveal opportunities for promotions and other marketing techniques.
- **Navigational Patterns:** Optimization of website layout, internal linking structures, and caching can be derived from this information.
- **Subgroup discovery among buyers and non-buyers:** User segmentation can help understand specific characteristics and behaviors, leading to targeted investments to improve conversion.

3.3. Data Mining Process

From the objectives, two main descriptive learning tasks can be obtained: itemset mining and subgroup discovery (SD). This work developed a interactive method to classify, execute algorithms and analyse the extracted information. This way, the process can be repeated multiple times until the executor judges that the information presented is sufficiently descriptive and, ultimately, useful for the organization. Figure 1 describes the overall data extraction and analysis process, in witch both descriptive learning tasks were applied to. The parsing, cleaning and session identification steps can be executed only once and the following ones can be performed one or more times. The following sections describes in detail each step of of the process.

3.3.1. Parsing

To properly use the file, it was necessary to parse it as a CSV file to allow operations upon specific fields. A regex expression with patterns specific to each field was utilized and more than 99% of the dataset could be parsed, resulting approximately 1.6 million lines of access entries.

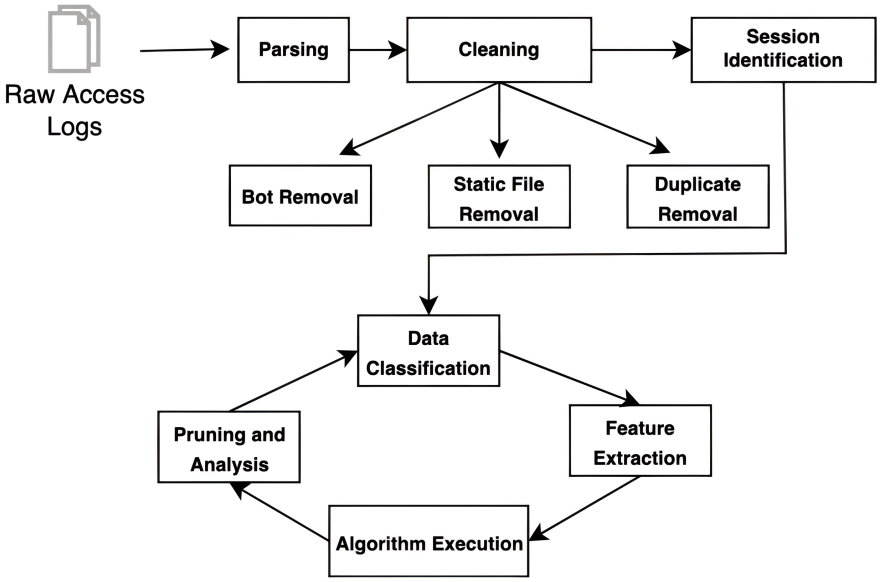

Figure 1. Data Mining Process

3.3.2. Cleaning

It is imperative to clean the data to remove common issues of web logs, such as invalid or not relevant entries. Firstly, it is removed all entries that contains any identification of bots or spiders in the user agent, as they are not real users or potential clients. To perform this task an initial search was made in the database for the works 'bot' and 'spider', then they are manually inserted in a filter function. Secondly, static files were removed as they are not particularly interesting for the research questions. In a modern web application, such as the one studied, there can be many static files, such as Javascript library files, image files, text files and other resources that are not directly contained in the html of the page, but get assembled together to it in the client side. As those requests are, in fact, part of the same page, they were excluded. Finally, duplicate entries, with the same IP address and timestamp were removed as bugs or issues with the server application can cause repeated entries to be logged.

3.3.3. Session Identification

This step aims to identify sessions, that are a collection of pages made by the same device, with the same IP address at a given time. As this work aims to primarily understand user behaviour, the sections were not limited to an arbitrary time constraint, so that even unusual patterns can be observed. As such, a grouping by IP address and user agent was performed. Finally, sessions with duration time less than 1 minute were not considered. The result was a dataset with 29.496 valid sessions. Figure 2 displays the final distribution of session per duration time obtained.

With this transformation, each line of the dataset contains: user agent, pages visited, time of visit of each page and IP address.

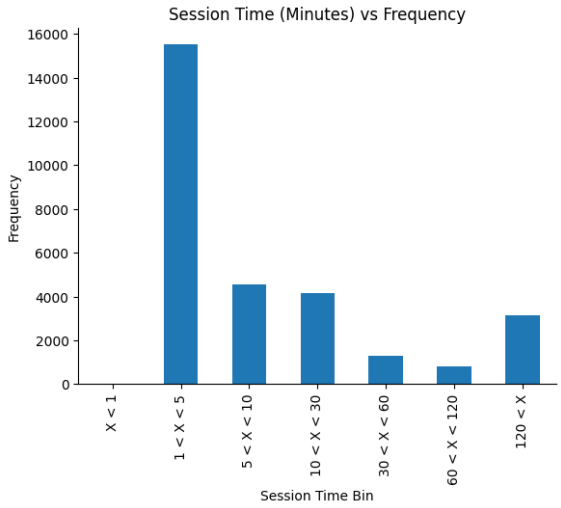

Figure 2. Session time distribution

3.3.4. Data Classification

In the dataset, the page visited set for each section contains specific URLs, that can greatly vary depending on product, filters applied, device used to access and so on. For example, the website adds the prefix */m/* to requests made from a mobile device, making difficult the direct comparison between sessions from different devices. With that, the granularity of the items can be too great to find any relevant frequency between URLs, leading to very low supports.

To mitigate this problem, the work semi-automatically classified each URL of the dataset using a list of known regex patters. With that, 54 classes of URLs were created. This step can be executed multiple times, depending on the abstraction level required, to extract relevant information. For example, it is possible to classify product page visits by product id, by category of product or just aggregate all page visits into a single category. Table 2 exemplifies how the URLs were transformed by this classification step in one of the iterations.

Pattern	**Class**
/product/(.+?)/(\d+)/	PR-[ID]
/article/(.+?)/	AR-[ID]
/order/create/	ORDER-CREATE

Table 2. Example of patterns used for URL classification

To allow more abstract patterns to emerge, this work also classified each product URL into a category, like home appliances, fashion and etc. In this particular dataset, most of the product URLs contain the full name and brand of the product, allowing it to be parsed and processed. This was done using a transformer trained in Persian [Searchwise 2023], the e-commerce default language, to classify the product into 86 classes of products. The result was stored in a new feature of each session called *product categories*. Other features were similarly build with this strategy, like *cart item categories*.

A peculiarity of modern web applications is that actions performed in the website, such as filtering or searching, do not necessarily lead to a new page load or navigation. This characteristic implies that not every URL in the pages set is a different page, but it can rather be an action taken in a page. Because of that, some of the URL classes are indeed actions, such as SEARCH, COMPARISION-ADD, GIFT-CARD and etc.

3.3.5. Feature Extraction

To better investigate sessions with certain characteristics, a number of features can be extracted from the available data. This step is particularly important for the task of subgroup discovery, in which discriminators between sessions are essential to segregate subgroups. As this step heavily depends on the available data, it can be re-executed whenever the nature of the data changes. Some of the boolean features extracted include: if a user searched, filtered, purchased any product, performed check out and etc. It is also possible to extract other features, such as products visited, cart items and etc.

3.3.6. Algorithm Execution

At this step the dataset is ready to be processed by data mining algorithms. To accomplish the goals of the study, the Apriori algorithm [Agrawal and Srikant 1994] used to find relevant association rules between pages, sections, products and product categories. Additionally, for subgroup discovery, the CN2-SD [Lavrač et al. 2004] with WRacc (Weighted Relative Accuracy) [Lavrač et al. 1999] was utilized. This algorithm perform better with a large number of features, as it does not perform an exhaustive coverage of possibilities, that is essential given an explosive number of possible combinations of binary and nominal parameters.

3.3.7. Pruning and Analysis

In subgroup discovery a pruning phase is essential as many similar subgroup descriptions can be generated for the same population. This problem was addressed by this work by utilizing Jaccard coefficient [Tan et al. 2006] for similarity to score and create a clusters of subgroups that cover a similar population subset. With this, all subgroups have been compared with each other and linked with an edge if the similarity score was above the defined threshold of 50%. After this, one representative is selected among the cluster of subgroups, that is the one with the highest quality score. This process can greatly reduce the number of relevant candidates from dozes or even hundreds to usually ten or less candidates. This is done to visually understand overlaps between populations subgroups and it does not eliminate the need to analyze subgroups with potentially lower quality. This can happen, mainly, if a subgroup has a lower quality but its attributes are more actionable from a business perspective. With that, the analysis phase for subgroups should understand the cluster representative and its high score subgroups. The approach is very similar to what has done by [Umek and Zupan 2011], with the difference that this work does not eliminate subgroups, but just suggests a potential representative.

An overall utility analysis is required for all research questions as frequent rules or subgroups does not necessarily imply usefulness for the company. If no relevant information is retrieved from an execution of the process, the analyst can decide to abstract or fine grain the dataset to retrieve the desired information. As such, this step is crucial to decide if the cicle should be executed again or if the results should be reported.

4. Data Analysis

The following section analyse and presents the data mining results and answers the research questions previously defined.

4.1. Shopping Cart

By extracting product classes from URLs and creating a new feature for the session, it is possible to mine data of cart items that were bought together. Table 3 displays a frequent itemset mining of product classes that were bought together.

Category 1	Category 2	Support
Home and Kitchen - Audio and Visual	Home and Kitchen - Home Appliances	0.0377
Fashion and Clothing - Clothing	Home and Kitchen - Home Appliances	0.0235
Beauty and Health - Personal Beauty Tools	Home and Kitchen - Home Appliances	0.0283
Home and Kitchen - Home Appliances	Tools and Spare Parts - Building Supplies and Materials	0.0141

Table 3. Frequent itemsets between product classes bought together

The discovered frequent itemsets can be used to better position items, bundle items in a sales, give coupons and other marketing strategies. Those items can be further explored with association rule mining, that can reveal a precedence order between the categories. For example, Fashion and Clothing - Clothing is revealed as the antecedent of Home and Kitchen - Home Appliances with lift 1.25 and confidence of 0.67.

4.2. Navigational Patterns

To find navigational patterns the technique of classification of visited URLs was utilized. Table 4 shows 3 selected navigational patterns. The BR class represents a page in the analyzed e-commerce that displays items from a given category. The FILTER-BRAND category means that the user filtered the page results by brand name.

Category 1	Category 2	Support
BR-big-kitchen-appliances	BR-home-appliances	0.0363
BR-cell-phone	FILTER-BRAND	0.0357
BR-cell-phone	BR-digital-supplies	0.0168

Table 4. Frequent URLs accessed in a same session

From the results, it is possible to hypothesize that the brand is relevant for users that seek for a new cell phone. Moreover, other digital products, such as storage cards, cables, headphones can be interesting to someone that is looking for a cellphone.

Additionally to tabular data, this work created an interactive visualization of the most common accessed pages in a graph, in which the vertices represent pages and edges an access in either direction. Only associations that are above a support threshold are shown and red edges represent a more common navigation path. This way, it is possible to analyze specific pages and their relationships. Figure 3 shows a graph of users that inserted at least one item into the shopping cart.

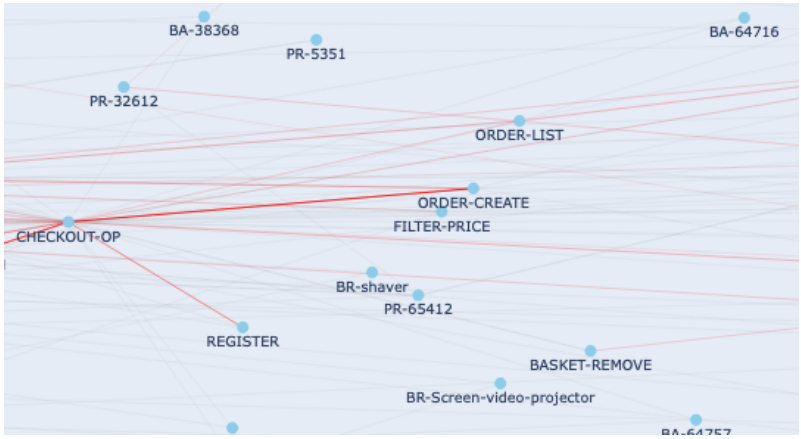

Figure 3. Frequent page visits between users that inserted at least one item in the shopping basket

4.3. Description of Buyers vs. Non-Buyers

Firstly, a general search for subgroups was executed to find users that inserted items in the shopping cart and finalized the purchase. Table 5 shows 5 relevant subgroups discovered. It is possible to derive multiple user behaviours from those subgroups. Interestingly, one of the subgroups conditions is visit of a blogpost, that in this particular e-commerce, usually contain product reviews and associated discounts.

The redundancy of the subgroups can be visualized in Figure 4, that compares each representative of cluster with each other. The coverage of itemsets are below 0.41 for all representatives, showing a relevant difference between them. To further analyze and potentially find more useful representatives of a given cluster, it is possible to visually choose another subgroup, as displayed in Figure 5, that is similar to its pairs.

Subgroup	**WRAcc**
session time bin==’120 <X’	0.0344
search==True	0.0342
basket add bin==’1 <X <4’	0.0290
is mobile==False	0.0227
blogpost==True AND search==True AND session time bin==’120 <X’	0.0179

Table 5. Subgroups of users that purchased

Lastly, the algorithm was executed to understand better the profile of users that abandon their cart. Table 6 displays the results. It is possible to derive that mobile users

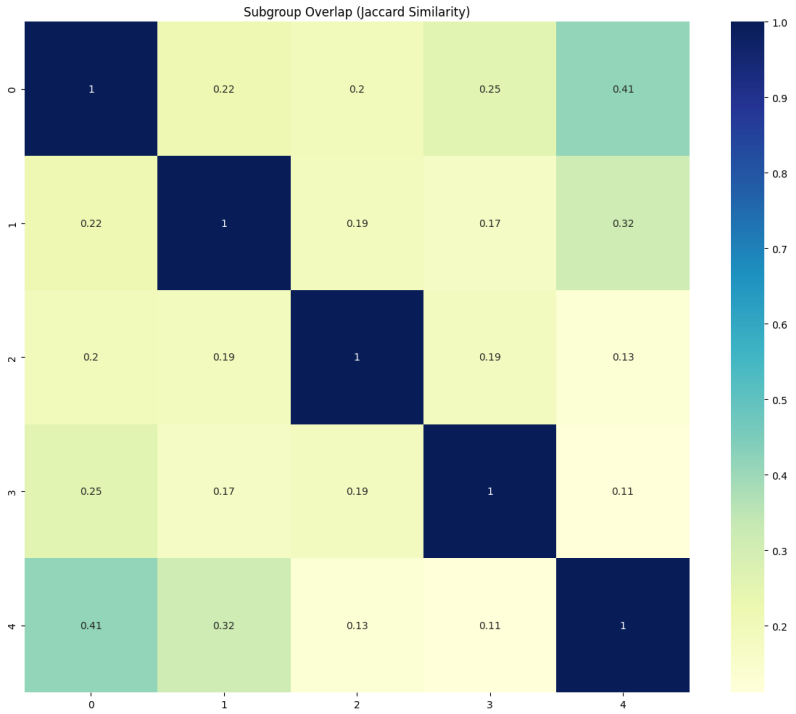

Figure 4. Coverage overlap between users that purchased, measured with Jaccard simmilarity score

are less prominent to finishing a purchase, in concordance with the previous results. More evidently, users that do not buy spend less time in the website.

As the process of subgroup discovery previously defined is iterative, an analyst may have to further expand on a given subgroup to understand better its sub-patterns. For example, it is possible to execute the algorithm only for sessions that have the operational system Android to find what characteristics better describes sessions that buy or not.

Subgroup	**WRAcc**
session time bin==’1 <X <5’	0.0441
basket add bin==’1’ AND operational system==’Android’	0.0332

Table 6. Subgroups of users that abandoned the shopping cart

5. Conclusion

This study presented a process to mine potentially relevant information from web usage data aimed at improve the efficiency of e-commerce operations. By employing descriptive learning methods, specifically itemset mining and subgroup discovery, it was possible to identify patterns in user behavior that can inform strategic business decisions.

The results of the itemset mining displayed frequent co-viewed and co-purchased product categories, providing opportunities for targeted promotions and bundling strategies. The navigational pattern analysis identified common user pathways through the website, suggesting areas for optimization in website layout and internal linking structures.

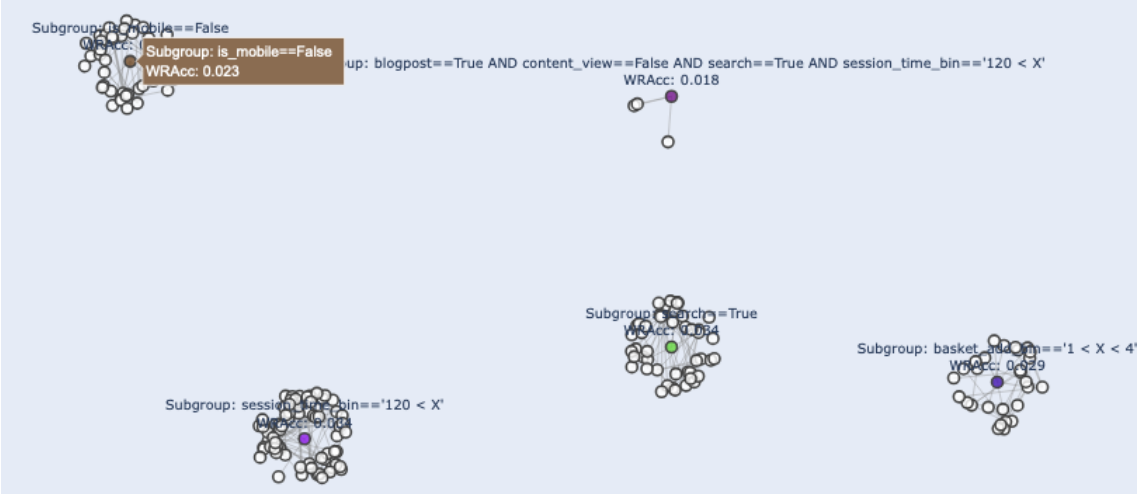

Figure 5. Clusters of subgroups that have a similar coverage

Subgroup discovery further enriched our analysis by distinguishing the characteristics of buyers and non-buyers. By identifying specific behaviors and attributes associated with successful purchases, businesses can tailor their marketing efforts and website features to better cater to different user segments. For example, in the analyzed dataset, an optimization of the mobile version of the website could potentially improve conversion rates for shopping cart abandonment.

Overall, the iterative methodology proposed in this study allows for continuous refinement of the analysis process, ensuring that businesses can adapt to changing user behaviors and emerging trends. Future work can expand on this approach by incorporating additional filtering steps, algorithms and techniques. Through ongoing analysis and adaptation, e-commerce platforms can remain competitive and responsive to the needs of their users.

References

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499.

Carmona, C. J., Ramírez-Gallego, S., Torres, F., Bernal, E., del Jesus, M. J., and García, S. (2012). Web usage mining to improve the design of an e-commerce website: Orolivesur. com. *Expert Systems with Applications*, 39(12):11243–11249.

Cooley, R., Mobasher, B., and Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. In *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, pages 558–567. IEEE.

Etminani, K., Delui, A., Yanehsari, N., and Rouhani, M. (2009). Web usage mining: Discovery of the users' navigational patterns using som. In *First International Conference on Networked Digital Technologies*, pages 200–204. IEEE.

Lavrač, N., Flach, P., and Zupan, B. (1999). Rule evaluation measures: A unifying view. *Proceedings of the 9th International Workshop on Inductive Logic Programming*, pages 174–185.

Lavrač, N., Kavšek, B., Flach, P., and Todorovski, L. (2004). Subgroup discovery with cn2-sd. In *Journal of Machine Learning Research*, volume 5, pages 153–188.

Searchwise (2023). l2-category-detection-augmented-data-epoch3. Accessed: 2024-07-28.

Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*. Pearson Addison Wesley, Boston.

Umek, L. and Zupan, B. (2011). Subgroup discovery in data sets with multi-dimensional responses. *Intelligent Data Analysis*, 15(4):533–549.

Zaker, F. (2019). Online Shopping Store - Web Server Logs.