Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Analysing User Navigation Patterns through Association Rule Mining and Clustering: A Case Study on Web Server Logs

Mukesh Kumar[1], Dharminder Kumar[2]
[1]Department of Computer Science & Engineering,
Guru Jambheshwar University of Science & Technology, Hisar (Haryana)
mukesharora@gjust.org
[2]Retired Professor, Department of Computer Science & Engineering,
Guru Jambheshwar University of Science & Technology, Hisar (Haryana)
dr_dk_kumar_02@yahoo.com

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Web usage mining improves the efficiency and personalization of Web-based services by using data mining techniques to identify patterns in user activity from Web log data. Preprocessing, pattern finding, and pattern analysis are its three primary stages. Among other methods, association rule mining is essential for identifying connections between user actions or viewed web sites. In addition to offering a thorough taxonomy of current research and commercial systems in the field, this study investigates the application of association rule mining within the larger framework of Web usage mining. This study examines user navigation behavior by analysing web server logs. Using association rule mining and clustering techniques, we identify frequent usage patterns and group similar webpages. Key findings highlight connections between webpage visits and offer insights into user behaviour over time.

Keywords: Association rule mining, KMeans clustering, web server logs, heatmap visualization, behavioural analysis, Apriori algorithm, webpage frequency matrix

1. Introduction

The 21st century has entered the Internet era, propelled by advancements in science, technology, and socio-economic standards. With each passing year, the Internet's influence has expanded, becoming an integral part of people's daily lives. Its impact spans various domains, encompassing social networking, sports, food, entertainment, education, and even the medical field [1]. The rapid development of the Internet has led to a swift increase in data across diverse topics, fostering closer social information interaction, enhancing user experiences, and enabling the digitization and replication of people's behaviours and activities. Big data plays a pivotal role in corporate decision-making, organizational structure, and operational procedures. In a different context, we are now in the era of big data. This term refers to datasets that are too extensive for conventional software to efficiently acquire, extract, and process. To effectively contribute to discovery, decision-making, and process optimization [2], a novel processing approach is necessary to manage the rapid growth and abundance of diverse information assets. Given the substantial informational capabilities of big data, scholars worldwide have extensively researched it from theoretical, technological, and application standpoints. Through the analysis of large-scale data, it has been revealed that big data has a wide range of applications. Besides its role in public transportation and security, big data can be applied to the fields of transportation, people's safety, medicine and health, social management, and more. This utilization leverages the diverse, nonlinear, parallel, and real-time characteristics inherent in big data for comprehensive analysis and insights [3].

Since the beginning of the big data age, data mining has become increasingly popular in the computer industry. Data mining involves the task of revealing hidden and interesting patterns within extensive datasets stored in databases, integrated data platforms and other sources of information. Remarkably, the size of these datasets can exceed terabytes. Knowledge discovery in databases (KDD), an alternate term for data mining [4], represents a multidisciplinary field that integrates methods from information retrieval, database technology, machine learning, statistics, and neural networks, among other disciplines. The objective of KDD algorithms is to efficiently and swiftly extract compelling patterns. The KDD process encompasses various stages, including data cleaning, data selection, data transformation, data preprocessing, data mining, and pattern evaluation, all conducted to extract patterns for the user [5]. The design of a data mining system needs several steps to retrieve the relevant data from data warehouse, database etc. A server is employed to retrieve data from these repositories based on user requests. The system also incorporates a knowledge base that directs searches in accordance with predefined constraints. The data mining engine primarily operates through modules that handle tasks like summarization, predictive classification, relationship discovery, continuous value prediction, and temporal trend analysis. [6]. An additional module is responsible for pattern evaluation, interacting with the data mining modules to identify interesting patterns. Lastly, graphical user interfaces are provided to enable user interaction and communication with the data mining system.

1.1 Association Rule Mining

The term “association between data” pertains to meaningful relationships among the values of different variables within the data. This process involves exploring extensive databases to reveal interesting correlations between variables, aiming to identify robust rules present in datasets using specific metrics. The overarching objective is to aid machines in emulating the human brain's ability to extract features and abstract associations from unclassified data, provided the dataset is sufficiently large [7]. Association mining stands out as a crucial aspect of data mining, representing the most widely studied method by academics. At the core of data mining lies the extraction of association rules. The benefit of these principles is that they can reveal linkages that were previously overlooked and produce data that can be used as a foundation for forecasting and making decisions. The mining of association rules involves two distinct processes:

i. Mining frequent itemsets: This process aims to identify all frequent itemsets within the database by applying a minimal support threshold [8].

ii. Rule generation: This step involves applying a minimal confidence constraint to these frequently occurring item sets, resulting in the derivation of association rules.

If dataset I consists of k items, the 2^{k-1} frequent itemsets formed will not include any empty sets. The process of repeatedly scanning and counting datasets to determine whether candidate itemsets can eventually become frequent itemsets can be time-consuming, especially as the search for frequent itemsets generates a large number of candidates. In essence, the effectiveness of a mining association rules algorithm depends on the efficiency of handling frequent itemsets, as the mining efficiency of association rules is closely tied to the efficiency of frequent itemsets [9]. The Apriori algorithm stands out as the most popular and well-optimized among various algorithms proposed recently for constructing association rules. The breadth-first search algorithm approach is used to ascertain the support of an item set, coupled with a candidate generation function that leverages the downward closure attribute of support.

1.2 Apriori Algorithm

The Apriori algorithm is employed to extract all frequently occurring itemsets from a database and is known for its relatively simple implementation. The steps of Apriori mining algorithm are as follows:

Step 1: Frequent Itemsets Generation: The initial step involves creating frequent itemsets. In each iteration of the dataset, all datasets that meet the user's support threshold are considered as (k +1)-frequent itemsets [10].

The Apriori algorithm initiates by conducting a single database scan to determine the support count for each item. It identifies the frequently occurring 1-itemset L_1, based on the user-defined minimum support. Subsequently, a new set, referred to as candidate itemsets, is formed by establishing self-links between itemsets derived from the previously frequent ones. The process involves multiple iterations of scanning the database, determining the support count for these candidate itemsets, and assessing whether they meet the user-defined minimum support condition. If the condition is satisfied, the item is deemed frequent. This process continues through several scans of the database until no further frequent itemsets are generated [11].

The process of finding L_k from $L_{(k-1)}$ involves two steps:

Join Step - Candidate Set Generation: In this step, a new set of candidate k -itemsets is formed based on the frequent $(k-1)$ −itemsets obtained from the previous scan of the dataset. The generation of candidate itemsets involves applying $L_{(k-1)} \times L_{(k-1)}$ with itself, and then utilizing the apriori-gen function. Notably, since the $(k-2)$ items preceding them are identical, they are combined into a candidate $(k-1)$ itemset [12].

Pruning step: pruning candidate sets. The first step involves trimming candidate sets to eliminate those that do not meet specified criteria. This is achieved through the use of the support-dependent pruning technique. Simply put, candidate sets in C_k are assessed by scanning the database to determine if their support count is equal to or higher than the specified minimal support count. In other words, a candidate k -itemset is considered a superset of a frequent k -itemset. If the support count surpasses the threshold, the candidate set is deemed frequent.

Step 2: Extracting association rules from the provided frequent set is the second stage.

(1) On the basis of all the frequent itemsets L generated in the previous steps;

Pruning operation contingent upon confidence If

$$\frac{Support\_count(X \cup Y)}{Support\_count(X)} \geq min\ \_conf, X \rightarrow Y$$

When the Apriori algorithm identifies frequent itemsets, it demonstrates two distinct characteristics. The intricate calculation process comprises two steps: initially, establishing a procedure to discover all frequent itemsets, and subsequently [13], iterating over the previously determined frequent itemsets, starting from frequent 1 and progressing up to the maximum order item set.

2. Literature Review

H.-B. Wang, et.al (2021) suggested a notion of parallelization and enhanced Apriori algorithm depending upon MapReduce model to tackle the blockage of conventional Apriori algorithm on enormous data set [14]. The primary task was to compute the local frequent item sets on every sub node in the cluster and incorporate all the local frequent itemsets into the global candidate ones later on. In the end, the least support threshold was considered to filter frequent itemsets which were capable of fulfilling conditions. The suggested algorithm became more effective after scanning the transaction database two times and evaluating the frequent item set in parallel. A particular instance was utilized for determining the execution procedure of this algorithm. The results demonstrated the feasibility of the suggested algorithm over existing methods.

M. M. Hassan, et.al (2023) introduced and analysed an association rule (AR) for exploring the significant rules while predicting suicidal behaviour from the given dataset [15]. The Apriori algorithm was implemented to recognize associations in databases. This algorithm was assisted in analysing the ARs of suicidal behaviour on a dataset in which 1250 cases and 27 features were comprised. This data, which included answers to cerebral assessments, family history, and routine activities, was examined in order to find correlations with self-harm behaviours. This algorithm led to provide some key rules for human suicidal behaviour. The introduced algorithm was proved effective for recognizing eight significant rules at support of 0.25 and offered confidence up to 0.90.

S. Guney, et.al (2020) devised a hybrid data mining (DM) method of clustering and association rule mining (ARM) to analyse and profile customers in video on demand (VoD) services [16]. To segment the customer, the LRFMP algorithm was implemented with k-means (KM) and Apriori algorithms for producing ARs amid recognized customer groups and content genres. The real-time dataset taken from an Internet protocol television (IPTV) operator was executed for determining whether the devised method was applicable. This resulted in recognizing 4 major customer subscribers, namely higher consuming-valuable, less consuming, less consuming-loyal and disloyal subscribers. The findings validated that the devised method was worked effectively for creating dissimilar customer segments at precise content rental features, and producing valuable ARs for these different groups.

X. Zhang, et.al (2023) emphasized on analysing the behavior of library user relied on Apriori optimization (AO) algorithm [17]. A library collection configuration optimization (LCCO) algorithm was planned on the basis of April algorithm and k-means clustering (KMC) algorithm. This algorithm was implemented for analysing the borrowing behavior of users. Moreover, the opinion and procedure of the Apriori algorithm was analysed and the data related to borrowing records was pre-processed and cleaned. The Apriori algorithm was exploited for mining the frequent itemsets and association rules (ARs). The results indicated that the presented algorithm was assisted the library in suggesting diverse books for diverse kinds of users which resulted in enhancing the user contentment and borrowing rate.

A. Ziakopoulos, et.al (2023) projected an association rule mining (ARM) for recognizing the underlying patterns in road crash injuries, recorded in Greece [18]. For this, the injuries were split into 2 subsets: mainland and island regions. The disaggregated data taken from the Hellenic Statistical Authority data related to dead, extremely and slightly injured road users, was analysed for 2017-2019. The next focus was on formulating an Apriori Rule and attaining various ARs. The results confirmed the interconnection of clear weather, urban area circumstances, and male road users which were present in injury crashes. The presented associations offered higher frequencies up to 80% above for the total injuries, and offered insights on specific patterns which would have probability of occurrence in road crash injuries because of higher exposure.

X. Ren, et.al (2021) recommended an association rule mining (ARM) approach for optimizing the Apriori algorithm [19]. The experiments were conducted on code written in the Web environment, and the optimized algorithm and the Apriori algorithm were implemented to mine data on the similar number of itemsets. This data exhibited that the enhanced algorithm offered a definite enlargement in contrast to traditional algorithm, and performed well at lower support. According to results, the enhanced algorithm consumed an execution time of 12s at support of 0.1, 17s at 0.2, 12s at 0.3, 8s at 0.4, and 6s at 0.5.

S. Haosong, et.al (2022) discussed that to diagnose power equipment fault was essential to manage fault [20]. An analysis was carried out on the efficacy of power tool for diagnosing fault and data mining (DM) for predicting fault. Therefore, a fault diagnosis expert system (FDES) was constructed for detecting equipment faults in a precise way. The accuracy was enhanced by optimizing and enhancing the ID3 algorithm. The experiments revealed the efficiency of enhanced algorithm to attain superior accuracy in comparison with others. Moreover, the Apriori association rule (AAR) algorithm was presented and proved robust for the power equipment elements, such as weights, in operation and to mine ARs among components.

V. A. Hameed, et.al (2023) established the Apriori algorithm and association rule mining (ARM) methods in Perl [21]. The fundamental objective was to analyse dataset, containing UK online gift store retailer sales transactions, so that frequent itemsets and association rules (ARs) were recognized. This algorithm assisted the traders in managing their inventory and maximizing the sales when the products were provided. The customers were bought these products in accordance with their earlier purchasing behavior. The outcomes indicated that the top ten frequent itemsets and their support values were highlighted with a list of ARs. Moreover, the purchasing behavior of candlelight vessels, bags, and tableware objects was found together. The established methods proved useful for small online gift store retailers for improving their sales efficacy and customer satisfaction based on ARs after recommending personalized products.

Y. Liu, et.al (2022) analysed that Apriori association rule (AAR) algorithm had some limitations of lower accuracy to recommend an algorithm in the process of present e-commerce systems [22]. For overcoming these issues, an association rule data mining (ARDM) algorithm was developed in which multi-item support tree (MST) was integrated with support. The recursive iteration was considered for producing conditional database and adjusting the least support in a dynamic way when the data was mined for attaining frequent item sets. This process resulted in making DM more effectual and maximizing the accuracy of recommendation. The experimental results revealed that the developed algorithm was feasible for changing the number of ARs, and enhanced the efficacy of DM and the accuracy of recommendation.

L. Fu, et.al (2022) designed a general modeling and analysis process for risk connections on the basis of association rule mining (ARM) and the weighted network (WN) theory [23]. The Apriori algorithm was implemented for mining the strong ARs among risks. Thereafter, the mining outcomes were put together with expert views to build the risk interaction network (RIN). The findings depicted that the designed network was applicable on both scale-free and small-word properties to illustrate that the DFPP accidents were not arbitrary events. However, they were occurred due to strong associations among security risks among multiple stakeholders. This approach was helpful for contractor in managing network risk and effective to manage security for DFPP.

3. Research Methodology

3.1 Data Cleaning and Pre-processing

The process of transforming the usage, content, and structural data from the many data sources into the understandable format is known as pre-processing. Because of the incompleteness of the available data, usage pre-processing is the most tedious work in the Web Mining process. Only the IP address, agent, and server side click stream can be utilized to identify users and server sessions unless a client-side tracking technique is employed. Text, images, scripts, and other materials, including multimedia, are transformed into formats that are helpful for the Web Usage Mining procedure as part of the content preprocessing process. This usually entails carrying out content mining tasks like clustering and categorization. In Web Usage Mining, pattern finding algorithms, filter the input to or output from the content of a website. Hypertext linkages between page views give a website its structure. The structure can be acquired and pre-processed similarly to a website's content. Once more, links and dynamic content present more issues than static page views. For every server session, a new site structure might need to be created.

Data collection: For this work, data was collected from web server logs of about three years between 2021 and 2023. Python frameworks are used to extract the necessary attributes from the log files. The two steps methodology adopted was first to eliminate the unnecessary log entries and then capture only a part of all the necessary fields usually available in a log file.

1. Imported a large web server log dataset.
2. Handled mixed data types, dropped irrelevant columns, and extracted target fields for analysis.
3. Handled missing or invalid data by filtering out records to maintain data integrity
4. Extract column names and count the total entries per column.
5. Ensure the 'time' column is properly formatted to datetime.
6. Removed unnecessary columns, retaining only 'time' and 'WebPage' for analysis.

Variable
remote_host
ident
remote_user
time
request_method
url
path
params
query_str
fragment
protocol
status
size
Page_Path
WebPage

Table 1 : Weblog file variables

Variable	Frequency
time	5060640
WebPage	5060640

Table 2 : Dataset on 'Time' and 'WebPage'

3.2 Data Processing and Pattern Discovery

Methods and algorithms from an array of fields, including statistics, data mining, machine learning, and pattern recognition, are used in pattern discovery. This study does not, however, aim to outline every algorithm and technique that may be found in these fields. The types of mining operations that have been used on the Web domain are described in this section.

- Statistical Analysis: Web session data is analysed using statistical approaches to get fundamental information such as average time on page, page views, and navigation pathways. These measurements inform marketing decisions, identify unwanted access, and help tune system performance.
- Association Rules: Association Rule Mining identifies groups of frequently visited pages, even if they are not explicitly connected. It supports business and marketing initiatives, improves prefetching for lower latency, and aids in website restructuring.
- Clustering: Clustering groups related people or sites according to content or browsing habits. It is employed for related link recommendation, tailored content distribution, and market segmentation
- Classification: Classification uses supervised learning to place people into predetermined groups. Through the analysis of demographic and behavioural tendencies, it helps with user profiling for targeted marketing.
- Sequential Patterns: This method determines which pages users usually visit in order by recognizing time-ordered trends across sessions. It helps with trend forecasting and ad placement.
- Dependency Modeling: Dependency modeling uses models such as Bayesian networks or HMMs to identify patterns in user behavior. For improved site optimization, it forecasts future actions and simulates user journeys. The dataset is processed as follows -
  1. Transformed time fields into month-year formats.
  2. Extracted 'month_year' from timestamps and Grouped data by time and webpages, filtering events with high visit frequencies.
  3. Filtered frequencies above 5000 to focus on popular pages, followed by pivoting data into a time-series matrix.

3.3 Pattern analysis

The final phase in the entire Web Usage mining process, as shown-

- Extracted monthly patterns by converting 'time' column to month-year format.
- Aggregated webpage frequencies and created a pivot table representing monthly visitation trends across webpages, enabling temporal analysis.
- Filtered the pivot table to retain webpages with significant visitation trends, using thresholds to ensure meaningful patterns.
- Create a matrix-like structure: rows as month-year, columns as WebPages, and their frequency.
- Retain columns with a total sum of frequency >= 5000 and keep the top 20 columns.
- Visualize high-traffic web pages over time with a heatmap to identify trends.
- Generated a heatmap using Seaborn to visualize the temporal frequency distribution of popular web pages.
- Heatmaps illustrated high-traffic webpages over time using vibrant color palettes.

month_year	results.html	archive.php	examination.html
2021-01	200414	94996	32653
2021-02	55184	45247	9446
2021-03	54810	38732	13308
2021-04	40305	33218	15363
2021-05	28136	30575	10630
2021-06	35701	36289	11539
2021-07	43375	39436	21182
2021-08	70017	31974	13217
2021-09	103205	34219	7670
2021-10	87653	42200	5413
2021-11	62143	27604	7379
2021-12	65003	20813	7326

Table 3 : Time-series frequency matrix of webpages (month-year)

The purpose of table is to create a time-series frequency matrix of webpage visits with rows representing distinct time periods (month-year) and columns representing webpages. Each cell contains the frequency of visits to a specific webpage during a given time period. This matrix enables analysis of user behavior over time and serves as the input for visualization, association rule mining, and clustering tasks. There are 25 webpages columns and only 3 are shown for a sample.

Visualized webpage visit frequency trends using heatmaps, highlighting variation across time periods and pages.

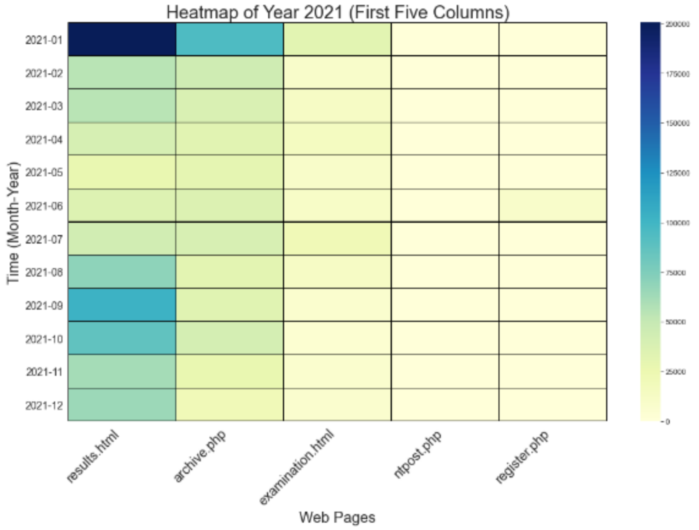

Fig 1 : Heatmap of first five pages

- Darker or warmer colors indicate pages with higher visit frequencies in specific months.
- Lighter or cooler colors represent pages with comparatively lower visit frequencies.
- Through this visualization, high-traffic periods and consistently visited webpages can be identified, helping to uncover navigation trends and user engagement insights.

3.4 Association Rule Mining:

For educational websites, the inspection of user navigational behavior is very important in the context of knowing resource access trends, learning styles, and content delivery optimization. Pattern analysis in this scenario means the discovery of common occurrences of page visits within user sessions. One example of a popular pattern analysis technique is the Apriori algorithm, a procedure used to reveal the frequent itemsets and association rules in large datasets like web server logs. The Apriori algorithm is based on the idea that if an itemset is frequent, then all of its subsets must also be frequent. Let D be a database of transactions (where each transaction is a user session containing multiple page accesses), and let $I = \{i_1, i_2, i_3, \dots., i_n\}$ be the set of all distinct web pages. Each transaction $T \subseteq I$ corresponds to the pages visited in a single session. The support of an itemset $X \subseteq I$, denoted as support(X) is defined as the proportion of transactions in which X appears:

$$support(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|} \quad (1)$$

The support threshold σ is pre-defined to filter out itemsets that occur infrequently. Itemsets satisfying $Support(X) \geq \sigma$ are called frequent itemsets. Further, the algorithm generates association rules of the form $X \rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The confidence of such a rule is given by:

$$confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \quad (2)$$

Confidence is a metric which gives the conditional probability of a user that visits pages in X also visiting pages in Y. The Apriori technique first identifies all the frequent 1-itemsets (individual pages that are frequently visited). Afterward, the frequent 1-itemsets are combined to make frequent 2-itemsets, and only those 2-itemsets whose support is greater than σ are kept. This iterative process is applied repeatedly by adding one page and extending the itemsets step by step, until there are no more such sets. This is shown mathematically as:

$$L_k = \{X \subseteq I || X| = k \text{ and } support(X) \geq \sigma\} \quad (3)$$

where L_k denotes the set of frequent itemsets of size k. For educational website hit data, each session is treated as a transaction composed of a set of visited web pages. The database D thus becomes a structured collection of sessions. Preprocessing, which is the preliminary step of data mining, refers to the cleaning (e.g., removal of robot clicks, non-educational page hits) and splitting of the sessions. The Apriori algorithm is then used to extract frequent patterns such as course page visits, homepage-to-registration page clicks, or resource download clusters after the preprocessing has been completed. Therefore, pattern recognition through the Apriori algorithm is able to convert collected educational web traffic into user common behaviour knowledge that is quantitative which not only can be a basis for further site structure, content arrangement and personalized learning experiences but also allows the recommendations to reach users effectively.

- Transformed the matrix to a binary format for Apriori algorithm analysis.
- The Apriori algorithm identifies frequent itemsets in transactional datasets, which in this case represent webpage visits over time.
- Identified frequent itemsets and generated association rules based on confidence and lift metrics.
- It works by iteratively building itemsets of increasing length (starting with single items) and calculating their support, which is
- The proportion of transactions that contain the itemset. Itemsets with support above a minimum threshold are retained as frequent.
- This method enables the discovery of strong patterns or associations in user navigation, such as which webpages are often visited together.
- Applied Apriori algorithm to the binary-encoded frequency matrix for identifying frequent itemsets.
- Generated association rules based on confidence and lift metrics, revealing strong correlations between webpage visitations.
- Filtered rules with lift >= 2, emphasizing significant relationships for further exploration.
-

antecedents	frozenset ({'results.html'})	frozenset ({'archive.php'})	frozenset ({'results.html'})	frozenset ({'examination.html'})
consequents	frozenset ({'archive.php'})	frozenset ({'results.html'})	frozenset ({'examination.html'})	frozenset ({'results.html'})
antecedent support	0.9189189	1	0.9189189	0.972973
consequent support	1	0.9189189	0.972973	0.9189189
support	0.9189189	0.9189189	0.9189189	0.9189189
confidence	1	0.9189189	1	0.9444444
lift	1	1	1.0277778	1.0277778
Representativity	1	1	1	1
Leverage	0	0	0.0248356	0.0248356
Conviction	inf	1	inf	1.4594595

Table 4 : Webpage Analysis using Apriori algorithms

Table 4 results of association rule suggests that when users visit the webpages: results.html (antecedents), they are likely to also visit: archive.php (consequents). This rule has a support value of 0.92, indicating the fraction of transactions containing both the antecedent and consequent. With a confidence of 1.00, it suggests that 100.0% of transactions containing the antecedents also include the consequents. The lift of 1.00 suggests a strong positive association between these pages, as it is greater than 1, indicating they are visited together more frequently than random chance.

Rule 1 : results.html -> archive.html

- Support : 0.92
- Confidence : 1
- Lift : 1
- Leverage : 0

Rule 2 : archive.html -> results.html

- Support : 0.92
- Confidence : .91
- Lift : 1
- Leverage : 0

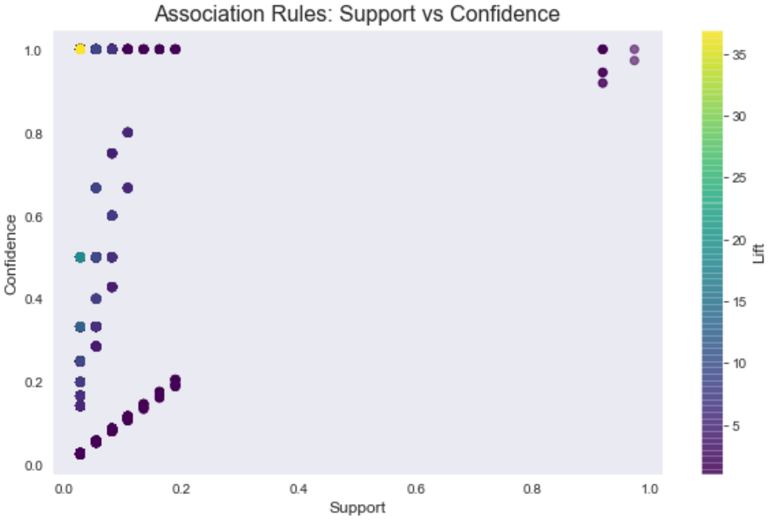

Fig 2 : Scatter plot visualizes the association rules

The above scatter plot visualizes the association rules with clustering results with support on the x-axis and confidence on the y-axis with each cluster represented using distinct colors. Each point represents an association rule, where:

- The x-axis and y-axis correspond to standardized feature values, representing webpage frequencies over time.
- The size and color of points indicate the lift value: higher lifts appear as bigger and darker points.
- Rules closer to the upper right corner demonstrate higher support (frequent occurrence in the dataset) and confidence (higher likelihood of consequents given antecedents), making them more significant.
- Each data point indicates a webpage's position within the feature space. The clear separation between clusters highlights distinct patterns or behaviors in webpage visitation.
- Clusters with overlapping points may suggest shared characteristics or transitional behaviors among webpages.
- This visualization aids in understanding how webpages group together based on user interaction patterns.

Applied filtered rules with lift >= 2 on the dataset to identify strong associations between webpage visits. A lift value greater than 2 indicates that the antecedent and consequent webpages are visited together at least twice as frequently as expected if the pages were independently visited. This helps in uncovering meaningful patterns, such as popularly co-visited webpages or navigation behaviors, which can guide user experience improvements and targeted recommendations.

antecedents	frozenset ({'esyllabus.html'})	frozenset ({'ntpost.php'})	frozenset ({'posts.html'})	frozenset ({'register.php'})
consequents	frozenset ({'ntpost.php'})	frozenset ({'esyllabus.html'})	frozenset ({'register.php'})	frozenset ({'posts.html'})
antecedent support	0.108108108	0.189189189	0.054054054	0.108108108
consequent support	0.189189189	0.108108108	0.108108108	0.054054054
support	0.054054054	0.054054054	0.027027027	0.027027027
confidence	0.5	0.285714286	0.5	0.25
lift	2.642857143	2.642857143	4.625	4.625

representativity	1	1	1	1
leverage	0.033601169	0.033601169	0.021183346	0.021183346
Conviction	1.621621622	1.248648649	1.783783784	1.261261261

Table 5 : Webpage Analysis using Filtered Apriori algorithms with strong association

Generated and filtered rules (lift >= 2) using metrics using scatter plots to show relationships between support, confidence, and lift of association rules.

Rule 1 : esyllabus.html -> ntpost.html

- Support : 0.05
- Confidence : 0.5
- Lift : 2.64
- Leverage : 0

Rule 2 : ntpost.html -> esyllabus.html

- Support : 0.5
- Confidence : 0.2
- Lift : 2.64
- Leverage : 0

Created a graph representation of filtered association rules, with nodes representing webpages and edges showing relationships based on lift values.

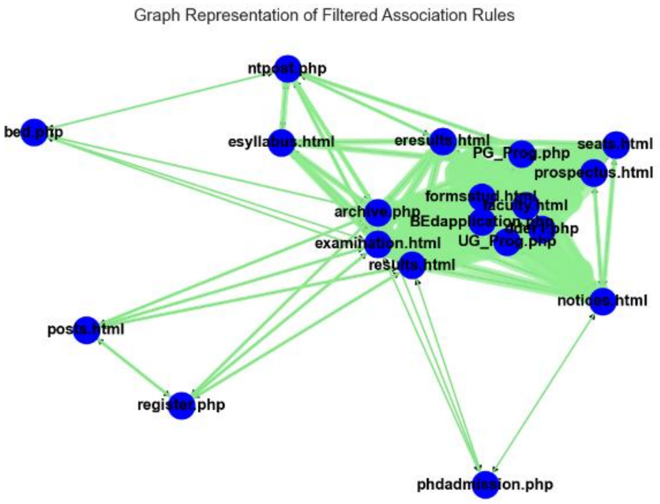

Fig 3: Visualizes the filtered association rules

This graph represents filtered association rules with a lift greater than or equal to 2, highlighting strong positive associations between webpages. Nodes are webpages, and directed edges point from antecedents to consequents, weighted by the lift value. Thicker edges indicate stronger correlations. This visualization reveals interconnected page visitation patterns, allowing insights into user navigation trends such as frequent page groups, potential optimization areas, or related content suggestions.

3.5 Modeling and Clustering

Utilized KMeans clustering to group webpages into three clusters based on visitation patterns, revealing unique user behaviour segments.

1. KMeans clustering is applied to group webpages based on their visitation frequency patterns.
2. Analysed cluster distributions and to understand shared traits between grouped 1. The matrix is standardized to normalize the data, ensuring features have a mean of 0 and standard deviation of 1.
3. KMeans algorithm partitions the data into 3 clusters (as specified) by minimizing the variance within each cluster
4. Each webpage is assigned to a cluster, enabling the identification of patterns and shared characteristics between webpages based on user interactions.
5. Standardized the matrix and applied KMeans clustering to categorize web pages into 3 clusters.

```
Clustered Matrix Shape: (37, 21)
Cluster Centers:

[[ 0.68610093  0.81060546  0.12154693 -0.40065302 -0.28683987 -0.03623601
1.18623231 -0.23483143  0.35396828 -0.31846902 -
0.28886691  0.93294586
-0.34275642  0.51851852 -0.16666667  0.41878817 -0.16666667 -0.16666667
-0.16666667 -0.16666667]

[ 4.59081552  4.47099923  4.12727642 -0.40065302 -0.28683987  5.90848123
1.94751922 -0.23483143  1.79084429  1.85271167 -0.28886691  2.22232949
-0.34275642 -0.16666667  6.         -0.22010078  6.         6.
6.         6.        ]

[-0.39873051 -0.43579438 -0.19337773  0.14839001  0.10623699 -0.20675397
-0.46754111  0.0869746  -0.18431699  0.03753739  0.10698775 -0.39329045
0.12694682 -0.16666667 -0.16666667 -0.13144418 -0.16666667 -0.16666667
-0.16666667 -0.16666667]]


Cluster Assignments Count:

2   27
0    9
1    1
```

Table 5 : Cluster Matrix Shape

```
Cluster Summary (Mean Values):

        results.html   archive.php     examination.html   ntpost.php
Cluster
0    62828.444444    36894.333333    819.777778        0.00000
1   200414.000000  94996.000000   2653.000000        0.00000
2    24603.592593    17110.148148   11260.518519  4186.77777

        register.php   PG_Prog.php    notices.html     posts.html
```

```
Cluster
0   0.000000     1318.333333   5134.444444    0.000000
1   0.000000     47279.000000  7498.000000    0.000000
2   3007.074074      0.000000      0.000000  1704.518519

      prospectus.html   eresults.html   bed.php    seats.html
Cluster
0     2160.444444       0.000000        0.000000      2879.0
1     6128.000000    7613.000000        0.000000      5678.0
2      674.111111    1248.296296     1190.444444         0.0

esyllabus.html  nt012021.php  formsstud.html phdadmission.php
Cluster
0   0.000000       2526.555556        0.0       1517.333333
1   0.000000       0.000000         19906.0       0.000000
2   1115.407407  0.000000             0.0       210.555556

   BEdapplication.php  UG_Prog.php  faculty.html  dder1.php
Cluster
0   0.0                  0.0            0.0        0.0
1   12241.0            9978.0         8857.0    8738.0
2   0.0                  0.0            0.0        0.0

Number of Webpages per Cluster:
2   27
0    9
```

Table 5 : Cluster Summary

- Cluster 0: Represents webpages with moderate and consistent visitation trends.
- Cluster 1: Represents webpages with high visits, possibly reflecting high interest or popular resources.
- Cluster 2: Represents low-traffic webpages, indicating niche or infrequent access patterns.
- High-clustered webpages highlight core areas of interest for users.
- Low-clustered webpages could denote niche content or under-utilized resources that may need strategic focus or evaluation.
- Use high-cluster webpages to guide prioritization when updating website content.
- Evaluate low-cluster webpages to optimize navigation or determine the need for removal if relevance is low.
  The clustering in the graph visualization illustrates how webpages are grouped based on their visitation patterns, with each cluster representing a unique behaviour profile:
  - The cluster plots show the segmentation of webpages based on their visitation patterns. Each cluster is represented by a distinct color, highlighting groups of webpages with similar trends.
  - Webpages in the same cluster share comparable behavior over time, suggesting common user interests or functions.
  - Clusters with tight, dense points indicate webpages with highly consistent visitation patterns.
  - Overlapping areas between clusters may signify transitional behavior or shared features among those webpages.
  - The distribution and size of clusters provide insights into user navigation, helping to identify popular or niche segments of the website.

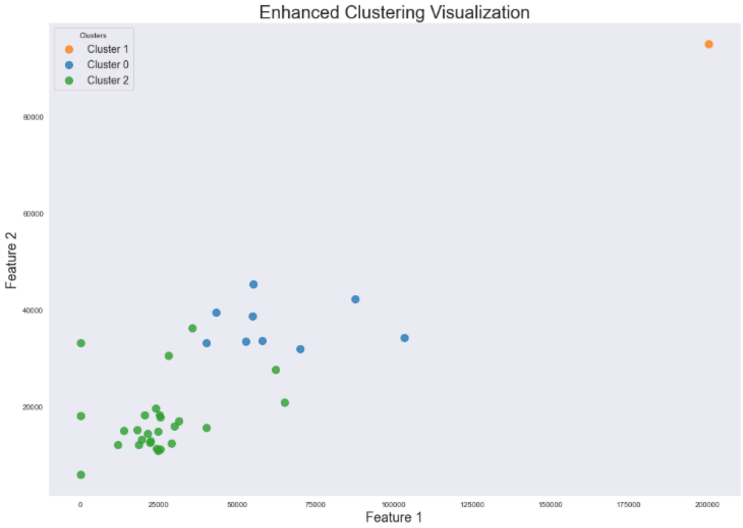

Fig 4 : Clustering Webpage distribution

- Cluster 0: Contains webpages with moderate and consistent traffic patterns, indicating general user interest.
- Cluster 1: Represents highly visited webpages, likely core resources or highly accessed content.
- Cluster 2: Includes infrequently visited webpages, often niche or specialized content.
- Nodes within the same cluster are highly interconnected, showing strong shared visitation trends.
- Larger node sizes or thicker edges indicate higher traffic or stronger correlations between pages.
- These clusters help identify user preferences and guide optimization of website content.
- For instance, 'results.html' (Cluster 1) is strongly linked with 'archive.php', highlighting a navigation pattern commonly shared by users seeking results-related information.

Conclusion

1. This approach helped to uncover meaningful patterns in user interaction with webpages.
2. Key association rules indicate strong navigation patterns linking specific resource pages, aiding in understanding user interaction paths.
3. Insights from the association rules and clusters pave the way for better personalization and design of user experiences.
4. Heatmaps and clustering highlight highly visited and niche webpage groups, providing actionable insights for website optimization.
5. Graph-based visualization underscores strong correlations between pages, indicating potential for targeted content recommendations.
6. Overall, this methodology emphasizes a comprehensive framework for analysing web traffic data, unveiling user behavioural trends, and supporting strategic decision-making.

Future Work

The analysis can be extended to real-time monitoring, recommendation systems, and predictive modeling for website optimization.

References

[1] M. Dehghani and Z. Yazdanparast, “Discovering the symptom patterns of COVID-19 from recovered and deceased patients using Apriori association rule mining”, Informatics in Medicine Unlocked, vol. 42, pp. 16-25, 7 September 2023

[2] P. Gao, “Research on Analysis of Students Using Mobile Phones in Ideological and Political Classrooms by Apriori Algorithm of Association Rules”, Procedia Computer Science, vol. 208, pp. 12-17, 2 November 2022

[3] X. Wang, C. Song and X. Lv, “Evaluation of Flotation Working Condition Recognition Based on An Improved Apriori Algorithm”, IFAC-PapersOnLine, vol. 51, pp. 129-134, 2018

[4] E. Çakır, R. Fışkın and C. Sevgili, “Investigation of tugboat accidents severity: An application of association rule mining algorithms”, Reliability Engineering & System Safety, vol. 209, pp. 1-12, 19 January 2021

[5] V. Robu and V. D. dos Santos, "Mining Frequent Patterns in Data Using Apriori and Eclat: A Comparison of the Algorithm Performance and Association Rule Generation," 2019 6th International Conference on Systems and Informatics (ICSAI), Shanghai, China, 2019, pp. 1478-1481

[6] K. Li, L. Liu and J. P. S. Catalão, “Impact factors analysis on the probability characterized effects of time of use demand response tariffs using association rule mining method”, Energy Conversion and Management, vol. 197, pp. 17-25, 1 October 2019

[7] Y. Li, Z. Zou and Y. He, “Study on the evolution of airport asphalt pavement integrated distress based on association rule mining”, Construction and Building Materials, vol. 369, pp. 789-795, 2 February 2023

[8] G. B. Baró, J. F. Martínez-Trinidad and M. S. Lazo Cortés, “A PSO-based algorithm for mining association rules using a guided exploration strategy”, Pattern Recognition Letters, vol. 138, pp. 1-15, 4 July 2020

[9] C. Wang, Y. Liu and Y. Wei, “Association rule mining based parameter adaptive strategy for differential evolution algorithms”, Expert Systems with Applications, vol. 123, pp. 54-69, 1 June 2019

[10] L. Zheng, "Research on E-Commerce Potential Client Mining Applied to Apriori Association Rule Algorithm," 2020 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Vientiane, Laos, 2020, pp. 667-670

[11] M. John and H. Shaiba, “Apriori-Based Algorithm for Dubai Road Accident Analysis”, Procedia Computer Science, vol. 163, pp. 218-227, 2019

[12] X. Xie, G. Fu and S. Jiang, “Risk prediction and factors risk analysis based on IFOA-GRNN and apriori algorithms: Application of artificial intelligence in accident prevention”, Process Safety and Environmental Protection, vol. 122, pp. 169-184, February 2019

[13] J. A. Delgado-Osuna, C. García-Martínez and S. Ventura, “Heuristics for interesting class association rule mining a colorectal cancer database”, Information Processing & Management, vol. 57, pp. 15-21, 25 January 2020

[14] H.-B. Wang and Y.-J. Gao, “Research on parallelization of Apriori algorithm in association rule mining”, Procedia Computer Science, vol. 183, pp. 641-647, 19 April 2021

[15] M. M. Hassan, A. Karim and A. S. M. Farhan Al Haque, “An Apriori Algorithm-Based Association Rule Analysis to detect Human Suicidal Behaviour”, Procedia Computer Science, vol. 219, pp. 1279-1288, 22 March 2023

[16] S. Guney, S. Peker and C. Turhan, "A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining," in IEEE Access, vol. 8, pp. 84326-84335, 2020

[17] X. Zhang and J. Zhang, “Analysis and research on library user behavior based on apriori algorithm”, Measurement: Sensors, vol. 27, pp. 458-463, 22 May 2023

[18] A. Ziakopoulos, E. Michelaraki and G. Yannis, “Association Rule Mining for Island and Mainland Road Crash Injuries in Greece”, Transportation Research Procedia, vol. 72, pp. 163-170, 13 December 2023

[19] X. Ren, "Application of Apriori Association Rules Algorithm to Data Mining Technology to Mining E-commerce Potential Customers," 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 1193-1196

[20] S. Haosong, G. Dongying, Z. Hang, Z. Helin and L. Jinze, "Power Equipment Fault Diagnosis and Prediction Based on Improved ID3 Algorithm and Apriori Association Rule Algorithm," 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 2022, pp. 607-610

[21] V. A. Hameed, M. E. Rana and L. H. Enn, "Apriori Algorithm based Association Rule Mining to Enhance Small-Scale Retailer Sales," 2023 IEEE 6th International Conference on Big Data and Artificial Intelligence (BDAI), Jiaxing, China, 2023, pp. 187-191

[22] Y. Liu and L. Wang, "An Optimized Association Rule Data Mining Algorithm," 2022 European Conference on Communication Systems (ECCS), Vienna, Austria, 2022, pp. 10-15

[23] L. Fu, X. Wang and M. Li, “Interactions among safety risks in metro deep foundation pit projects: An association rule mining-based modeling framework”, Reliability Engineering & System Safety, vol. 221, pp. 23-30, 6 February 2022