# Characterising Digital Information Disposal for Efficient Academic Outreach

Mukesh Kumar
*Department of Computer Science & Engineering,*
*Guru Jambheshwar University of Science & Technology - Hisar (Haryana)*

Prof. (Retd.) Dharminder Kumar
*Department of Computer Science & Engineering,*
*Guru Jambheshwar University of Science & Technology - Hisar (Haryana)*

**Abstract - This paper presents an in-depth study into user web navigation patterns, utilizing web server logs to analyse session data and predict page transitions using Markov models. Markov chains allow the system to dynamically model the URL access patterns that are observed in navigation logs based on the previous state. Machine learning and graph-based approaches were used to segment sessions, calculate transition probabilities, predict user behaviour, and construct Markov transition graphs. These findings aim to improve website design, enhance the overall user experience, and predict web navigation patterns.**

**Keywords: Graph-based Approaches, Markov Models, Machine Learning, Session Segmentation, User Experience, User Behaviour Prediction, User Web Navigation, Website Design, Website Optimization**

## I. INTRODUCTION

With the advent of internet and the digital information sharing it is of key interest to organisations and institutions to keep their information accessible to a broader population. This is particularly true in the case of educational institutions that cater to a larger audience and keep the information intact, factual, and easily accessible. Maintaining efficient navigation is among the critical aspects of the organization's interest. A recent study has defined information accessibility as a basic necessity. In this work, we study one such educational organization and evaluate the ease of access for the students to navigate and find the popular or important webpages, as well as webpages where people get stuck or the webpages that need redesign.

Page transitions and session patterns highlight critical interactions and enable the prediction of future user actions. In this study, we investigate these patterns by analysing session data drawn from web server logs and predicting page transitions using Markov models to analyse transition patterns, predict behaviour, and visualize page transition behaviour. Markov models provide a probabilistic framework that enables the prediction of subsequent actions based on past behaviour [1]. The analysis also includes filtering invalid session data, segmenting sessions, and generating a transition graph to uncover the structure of user behaviour. The insights derived from this analysis have the potential to assist in the optimization of website design, thereby promoting improved user experience.

In our work we particularly look at web usage mining with the help of web logs.

## II. PRELIMINARY CONCEPTS

In this section, we introduce the necessary concepts for our work.

*Introduction to web usage mining*

Data mining includes Web Usage Mining (WUM) [2] [3] as a subset, which contains extraction of information from available web pages over the Internet. A key data mining technique is WUM that helps in identifying and analysing the activity of page visitors. Along with visitor characterization, it also tracks the users over time. WUM has three sub categories in the context of our work.

**Web content data:** Web data contains of pages created using html, multimedia such as images, videos, and audio. The layouts of a html page are important in driving the user experience on the websites. While

technologies like XML, JSP etc. are popular recently focus has shifted towards python based web pages [4]. Each of the technologies can give various types of web content data.

*Web structure data:* When the content are rendered in a web page, there are usually have hyperlinks which forms a network of pages. The form of moving from one page to another is called as a navigation which leads into browsing web pages. Structure of a web site and its applications are immense in offering users with an easily navigable information [5].

*Web usage data:* While there is content and structure to web pages, the usage statistics are equally important. The usage data is stored in logs, and these are automatically collected by the web application [6]. The files are created as soon as users start interacting with the web site. Data which arises from the log files captures a lot of valuable information like:

1) *Association rules:* The relations between pages which are most well connected and frequently visited together by users are defined by this rule. Using this rule, we can identify the most important pages, and put them together in one location. The rule helps in constructing websites in such a way that the usage becomes easier. Also, logs are a general way of identifying information about the incoming requests [7]. However, the technique lacks the efficiency because of rules involved.

2) *Classification***:** In order to segregate the web pages into blocks containing very similar documents, we use classification. To extract we have to find the prime features that define / differentiate each of the classes. There are numerous kinds of algorithms available for doing classification, like; K-NN (K- nearest neighbours), Logistic Regression, Support vector Machines (SVM) among others. An application may include tracking of customer data as a frequent customer or non-frequent customer categories [8]. The techniques can also be extended to multiclass classification tasks, where the documents need not be belonging to a single class but into multiple classes.

3) *Clustering:* An extension to the earlier technique is called clustering where instead of identifying the class to which a document might belong, we club similar documents together [9]. The clustering algorithms, are of two types in the context of WUM. Firstly, *user-based* clustering works on the behaviour of users who are visiting a given website. Then, there is *usage-based* where instead of behaviour tagging, we identify the usage of a website. An application of clustering is to make sure similar information pages which are relevant stays together.

## III. METHODOLOGY

### *A. Log file structure*

Log files are a standard tool for computer systems developers and administrators. They record the ”what happened, when, by whom” of the system.

| Field name | Description | Significance |
|---|---|---|
| IP | Client who made the request | To identify the remote host |
| User Identifier | RFC 1413 identity | Usually contains value ’-’ |
| Username | To identify the requesting user | Helpful in request authentication |
| Date time | To log the event time | Helpful in identifying the instance |
| Request | Type of request | Signifies the call type |
| Status | Return code by server | Used for diagnostic purposes |
| Size | The return object size | Used for diagnostic purposes |

TABLE 1: Log-file and the field description as found in our dataset

The information available in the log files generated by an organization are as shown in Table II. Firstly, we see an *IP address (remote_host)* related to the request made, and this is useful in identifying the remote host. Then, we see a *user identifier*, which is usually a '-', and this is used in conjunction with an RFC 1413. A person identifier field is also seen, which is used only when a request authentication is necessary. *Time and Date* fields is pretty standard, and helps in estimating the exact time when a log was recorded. The *request type* shows what kind of connection is being made, the values in our site are get and post. When a response is made the success or failure of the response is indicated in the *status* field. Finally, a *size* field is useful for the client to know if a transmission has been successful, as well as to keep a tab on what was sent through.

*B. Average statistics of users visits in year 2021, 2022 and 2023.*

| **Field** | **Statistic (Year 2021)** |
|---|---|
| Average unique IP addresses | 3925 (per day) |
| Average sessions | 4534 (per day) |
| Average unique pages | 156 (per day) |

| **Field** | **Statistic (Year 2022)** |
|---|---|
| Average unique IP addresses | 1608 (per day) |
| Average sessions | 1885 (per day) |
| Average unique pages | 129 (per day) |

| **Field** | **Statistic (Year 2023)** |
|---|---|
| Average unique IP addresses | 1445 (per day) |
| Average sessions | 1739 (per day) |
| Average unique pages | 143 (per day) |

TABLE 2 : Group the session data by the timestamp. Count the distinct average unique IP addresses, Sessions, unique Pages per day

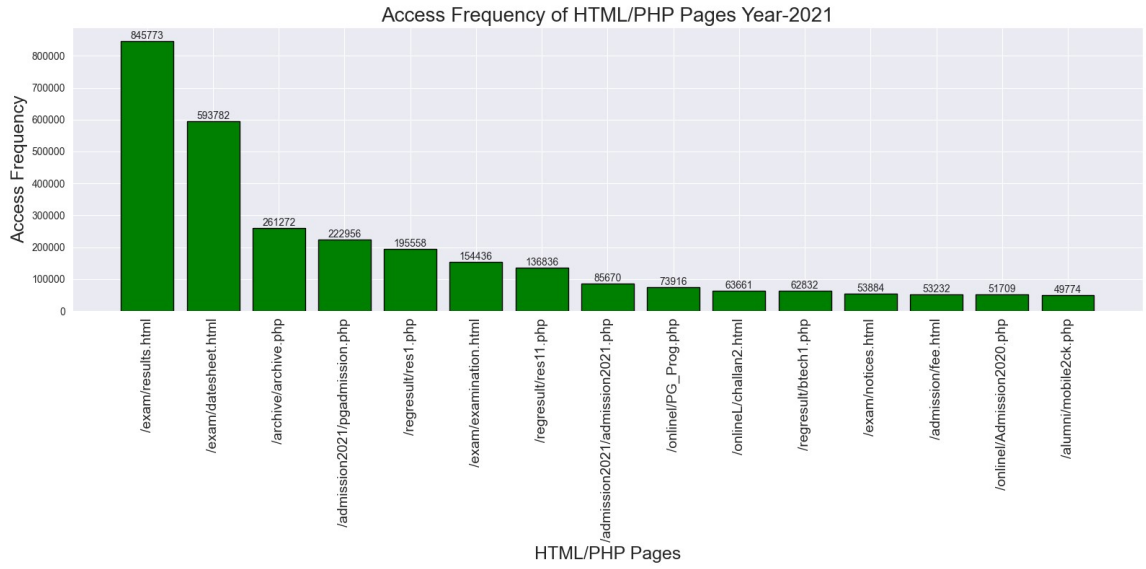


Figure 1 : The most accessed pages : results.html, datesheet.html, archive.php in user activity for 2021. Percentage of access frequency of top three webpages are 58% in top fifteen webpages, indicating their prominence in user activity for year 2021

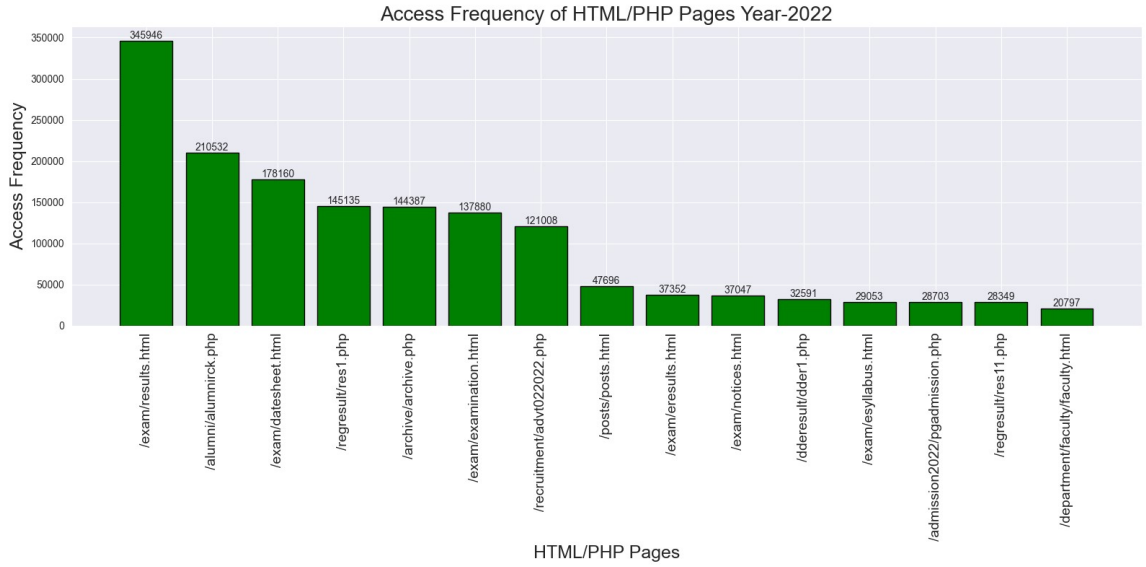


Figure 2 : The most accessed pages : results.html, alumnirck.php, datesheet.html in user activity for 2022. Percentage of access frequency of top three webpages are 47% in top fifteen webpages, indicating their prominence in user activity for year 2022.

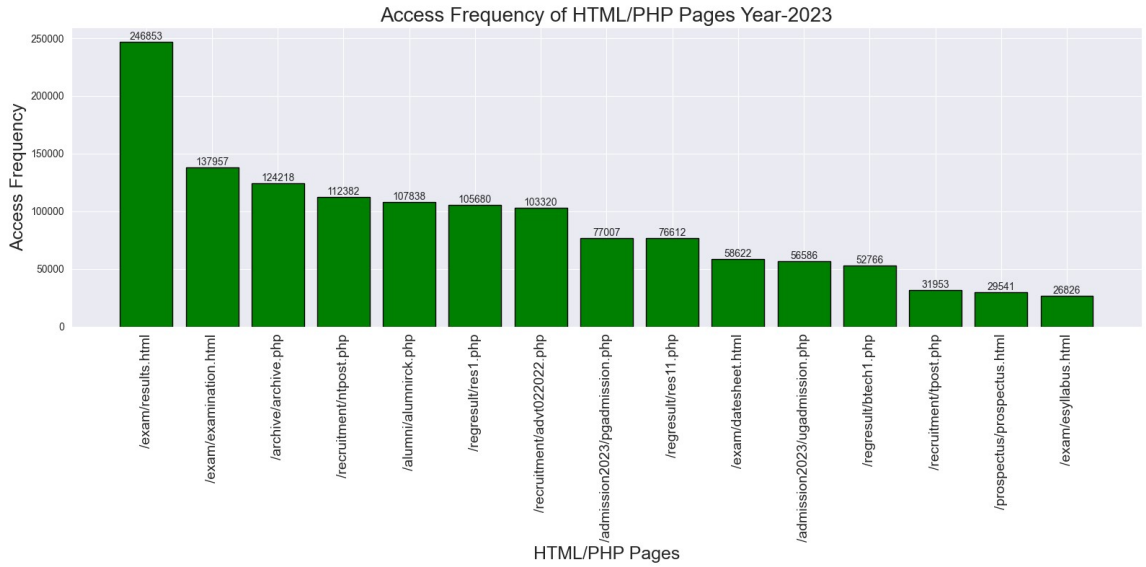


Figure 3 : The most accessed pages : results.html, examination.html, archive.php in user activity for 2023. Percentage of access frequency of top three webpages are 37% in top fifteen webpages, indicating their prominence in user activity for year 2023.

In above plots, we see distribution of popularity among pages. We can see a clear distinction between webpages rarely accessed and the most popular pages.

*C. Data Pre-processing and Cleaning*

For this work, data was collected from web server logs of about three years between 2021 to 2023. The data covers a wide range of events, annual and seasonal. 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.

- The dataset involves rows of user hits of a web server logs. Steps undertaken included:
- Removed invalid records (e.g., unknown URLs, corrupted remote hosts) and rows with missing critical values.
- Cleaned invalid page transition data by removing self-loops (e.g., source and target pages being the same).
- Filtered sessions based on time and known or unknown URL patterns.
- Kept transitions within 2021, 2022, and 2023 for the analysis.
- Extracted unique session-level and page-level information.
- Assigned unique user sessions using a 30-minute inactivity threshold.

*D. Clustering Method*

In the pre-processing phase, sessions are identified from web log file. Web log sessions are the time-stamped sequence of user IP (remote_host) click stream navigations. A user navigation action such as searching desired information or clicking on a web result is recorded in the weblog file. A session is a set of web pages (Pi) which is navigated by the user IP (remote_host). A session_id, S is represented with webpages as {P1; P2; :Pn} where n is the session length. Each session is assigned a unique session ID [10].

First, we first partition users (remote_host) into sessions with navigated webpages clusters and each session is of 30 minutes inactivity threshold. Consider a new session if the remote host changes or inactivity exceed 30 minutes. Then, for each session_id, we display the behaviours of the users within that session_id of clustered webpages. The focus of our paper is on the visualization and session aspects of such data.

The server-logs have been converted into a set of sequences, one sequence for each user (remote_host).

(a) each session_id symbol represents possible categories of Web pages requested by the user.
(b) Random select 5 sessions IDs to ensure unbiased for analysis.
(c) Session_id : Unique identifier for user sessions.
(d) Sequence : List of source-target page transitions for the corresponding session.

| Session ID | Sequence |
|---|---|
| 1007324 | [['archive.php', 'datesheet.html'], ['datesheet.html', 'notices.html'], ['notices.html', 'formsstud.html'], ['formsstud.html', 'examination.html'], ['examination.html', 'archive.php']] |
| 1389099 | [['faculty.html', 'phm.html'], ['phm.html', 'seats.html'], ['seats.html', 'bio.html'], ['bio.html', 'adp.html'], ['adp.html', 'Admission2020.php'], ['Admission2020.php', 'archive.php']] |
| 81538 | [['btech.php', 'admission2021.php'], ['admission2021.php', 'btech.php'], ['btech.php', 'adp.html'], ['adp.html', 'admission2021.php'], ['admission2021.php', 'btech.php']] |
| 1479708 | [['examination.html', 'archive.php'], ['archive.php', 'Admission2020.php'], ['Admission2020.php', 'PHd2020.php'], ['PHd2020.php', 'calendar.html'], ['calendar.html', 'formsstud.html'], ['formsstud.html', 'eresults.html'], ['eresults.html', 'faculty.html'], ['faculty.html', 'hsb.html']] |
| 302884 | [['archive.php', 'vcmsg.html'], ['vcmsg.html', 'ec.php'], ['ec.php', 'planningboard.html'], ['planningboard.html', 'registrarmsg.html'], ['registrarmsg.html', 'daa.html'], ['daa.html', 'proctor.html'], ['proctor.html', 'hostels.html'], ['hostels.html', 'dsw.html'], ['dsw.html', 'uacolleges.html'], ['uacolleges.html', 'court.html'], ['court.html', 'coe.html'], ['coe.html', 'directors.html'], ['directors.html', 'officers.html'], ['officers.html', 'english.html'], ['english.html', 'faculty.html'], ['faculty.html', 'fee.html'], ['fee.html', 'deans.html']] |

Table 3 : shows a sample of such sequences (Year - 2021)

| Session ID | Sequence |
|---|---|
| 223408 | [['vipgh.html', 'hostels.html'], ['hostels.html', 'iqac.html'], ['iqac.html', 'security.html'], ['security.html', 'hostels.html'], ['hostels.html', 'security.html']] |
| 74133 | [['archive.php', 'uacolleges.html'], ['uacolleges.html', 'examination.html'], ['examination.html', 'results.html'], ['results.html', 'eresults.html'], ['eresults.html', 'esyllabus.html']] |
| 489830 | [['rti.html', 'gjucontacthsr.html'], ['gjucontacthsr.html', 'ucic.html'], ['ucic.html', 'advt022022.php'], ['advt022022.php', 'archive.php'], ['archive.php', 'phdadmission.php'], ['phdadmission.php', 'examination.html'], ['examination.html', 'students.htm'], ['students.htm', 'grievance.php']] |
| 34523 | [['coe.html', 'registrarmsg.html'], ['registrarmsg.html', 'seats.html'], ['seats.html', 'hsb.html'], ['hsb.html', 'haryana.html'], ['haryana.html', 'res1.php']] |
| 527563 | [['results.html', 'res1.php'], ['res1.php', 'results.html'], ['results.html', 'res1.php'], ['res1.php', 'results.html'], ['results.html', 'res1.php']] |

Table 4 : shows a sample of such sequences (Year - 2022)

| Session ID | Sequence |
|---|---|
| 411674 | [['schsylaf.php', 'hindi.html'], ['hindi.html', 'archive.php'], ['archive.php', 'cse.html'], ['cse.html', 'archive.php'], ['archive.php', 'consent-form.php'], ['consent-form.php', 'archive.php']] |
| 568793 | [['iqac.html', 'ugadmission.php'], ['ugadmission.php', 'bpharmacy.php'], ['bpharmacy.php', 'prospectus.html'], ['prospectus.html', 'phdadmission.php'], ['phdadmission.php', 'bed.php'], ['bed.php', 'btech.php'], ['btech.php', 'esyllabus.html']] |
| 214552 | [['examination.html', 'advt022022.php'], ['advt022022.php', 'archive.php'], ['archive.php', 'advt022022.php'], ['advt022022.php', 'examination.html'], ['examination.html', 'advt022022.php'], ['advt022022.php', 'examination.html'], ['examination.html', 'archive.php']] |
| 305636 | [['advt022022.php', 'examination.html'], ['examination.html', 'advt022022.php'], ['advt022022.php', 'examination.html'], ['examination.html', 'advt022022.php'], ['advt022022.php', 'archive.php'], ['archive.php', 'advt022022.php']] |
| 436363 | [['res.php', 'res1.php'], ['res1.php', 'mcom.php'], ['mcom.php', 'res1.php'], ['res1.php', 'mcom.php'], ['mcom.php', 'res1.php'], ['res1.php', 'res.php']] |

Table 5 : shows a sample of such sequences (Year - 2023)

*E. Markov Model Chain*

The most prominent model for describing human navigation on the Web is the Markov chain model, where Web pages are represented as states and hyperlinks as probabilities of navigating from one page to another In a Markov chain. The dynamics of a user's navigation session, in which visits a number of pages by following the links between them, can thus be represented as a sequence of states. One of the most influential assumptions in this field to date is the so-called Markovian property, which postulates that the next page that a user visits depends only on her current page, and not on any other page leading to the current one [11].

The 0-order Markov model is the unconditional base-rate probability $p(x_k) = p_i(x_k)$, which is the page visit probability. The 1-order Markov model looks at page-to-page transition probabilities: $p(x_{i+1}|x_i) = Prob(X = x_{i+1}|X1 = x_i)$. The K-order Markov model considers the conditional probability that a user transitions to an $n^{th}$ page given his or her previous $k = n–1$ page visits:

$$p(xn|x_{n-1},\ldots,x_{n-k})$$

$$= Prob(Xn = xn|Xn–1 = x_{n-1},\ldots,X_{n-k} = x_{n-k}).$$

Our methodological approach relied on k-th Markov Models [kth] for page transition probabilities, segment sessions, and generate page transition graphs that depict user navigation flows. Furthermore, we utilized kth-order predictions for forecasting user behavior.

- Sessions were formed with a threshold of 30 minutes between actions.
- By leveraging session-based transition data, analyse user behaviour with a focus on calculation transition probabilities and providing insights on predicting future webpage visits. [12].
- Graph Construction: Directed transition graphs were created visualizing source, target, and predicted pages, similar to the dynamic clustering-based Markov model proposed by [13] for web usage mining.
- The k-th order Markov Model provided valuable insights into use navigation patterns. It helped identify key nodes and transition probabilities, which can be used in –

- Personalizing user experience
- Designing efficient navigation paths
- Optimizing content recommendation systems

- Transition probabilities were calculated using k=3.
- K=3 in the context of Markov model refers to third-order markov chain.
- It means that the prediction of the next page depends on the last three visited pages.
- E.g. if a sequence is [‘A’,’B’,’C’], then the prediction for ‘D’ is based on the probability of transitioning from [‘A’,’B’,’C’] to ‘D’.

*F. State transition matrix*

Having seen the sets of pages, and frequent visits history, we use k-th order Markov models to estimate probabilities of transition from session clusters [14].

(a) A user arrives at the Web site and is assigned to a particular session_id with some probability of webpages transitions.
(b) Probabilities in the display are encoded by intensity (higher probabilities are brighter) [15].
(c) Each row represents a ‘Source Page’.
(d) Each column represents a ‘Target Page’.
(e) Fill_value=0 ensures that any missing transition values are filled with ‘0’ (indicating no direct transitions between pages).
(f) Resulting matrix, displaying the page-to-page transition probabilities for the random session with source and target pages from the session data.
(g) Once we have the transition probabilities, and visualization we build Markov Model with the help of website logs.

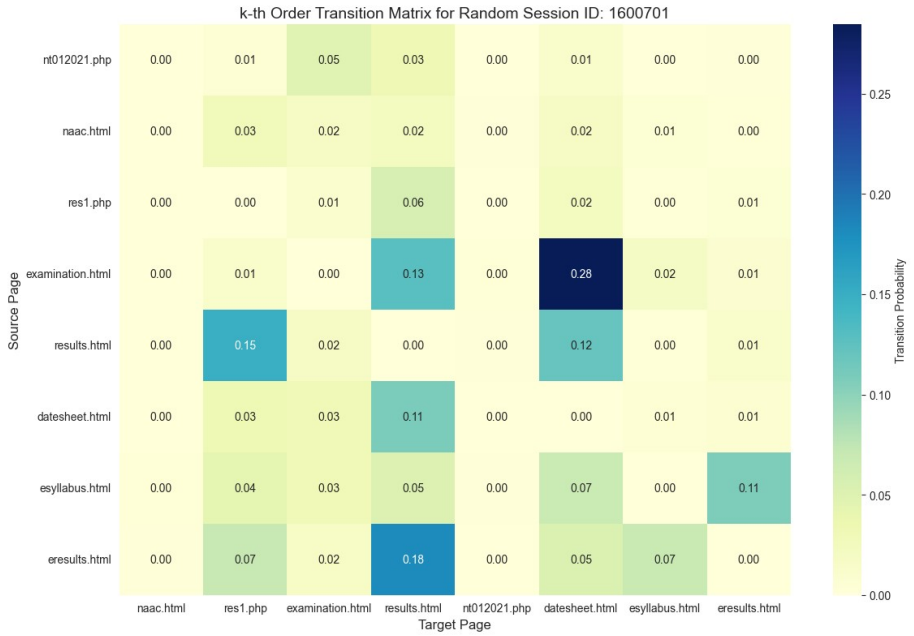


Figure 4 : kth order Markov Model Transition matrix of a random session
(Year - 2021)

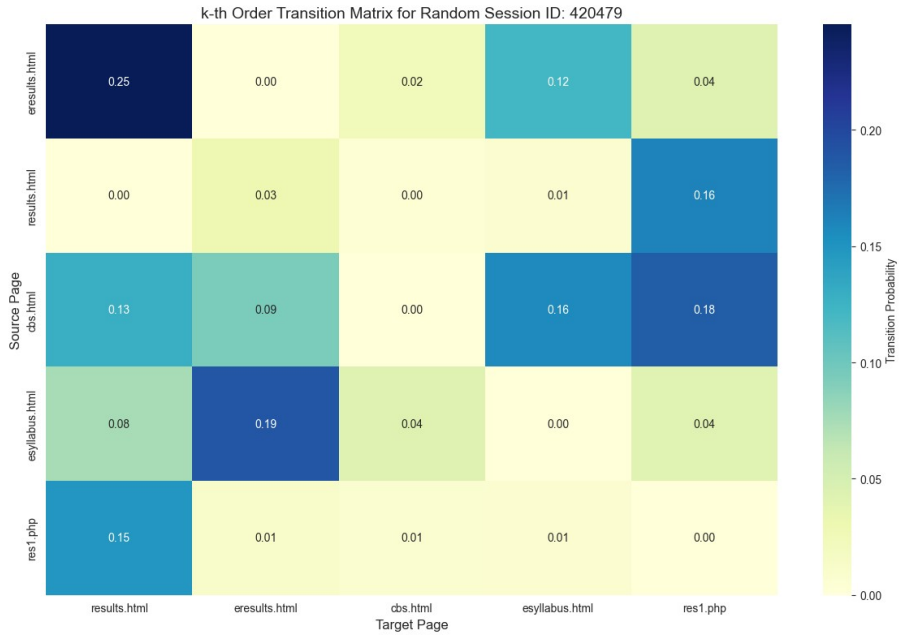


Figure 5 : kth order Markov Model Transition matrix of a random session (Year - 2022)

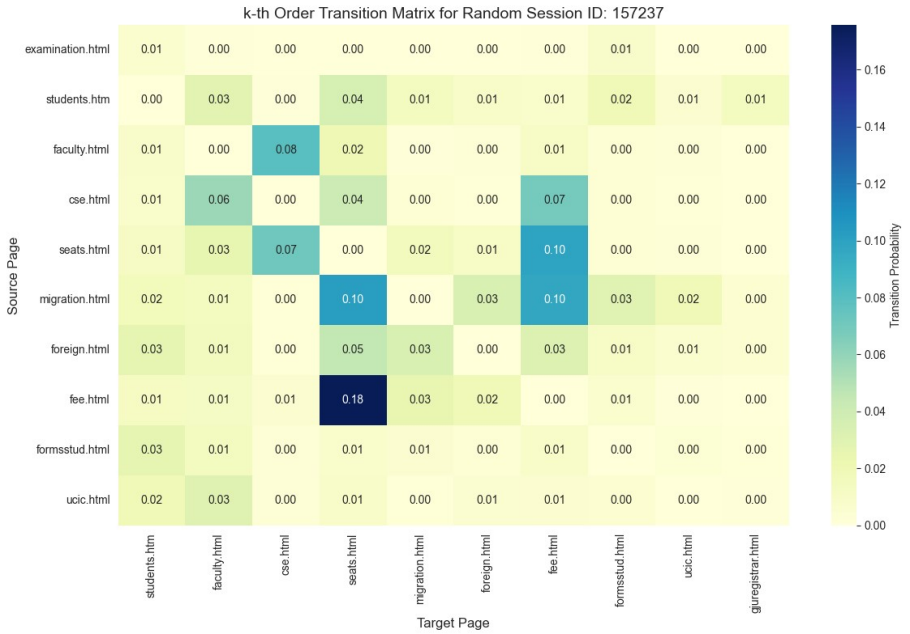


Figure 6 : kth order Markov Model Transition matrix of a random session (Year - 2023)

*G. Random Session Graph forming*

The graph represents the session-based page transitions captured using a k-th order Markov Model. The visualization helps to analyse user navigation patters and understand the probability distribution of webpage transitions [16].

This plot visualizes the page transition behaviour for a specific session.

Key aspects shown include:

- Nodes represent pages visited during the session. Node size
- Directed edges represent transitions between pages, with weights indicating the transition probabilities. Edge weights indicate the transition probabilities.
- Identify the most frequently visited and highly connected pages.
- Detect user preferences in navigating the website.

Purpose: Understanding user navigation patterns to optimize website structure.

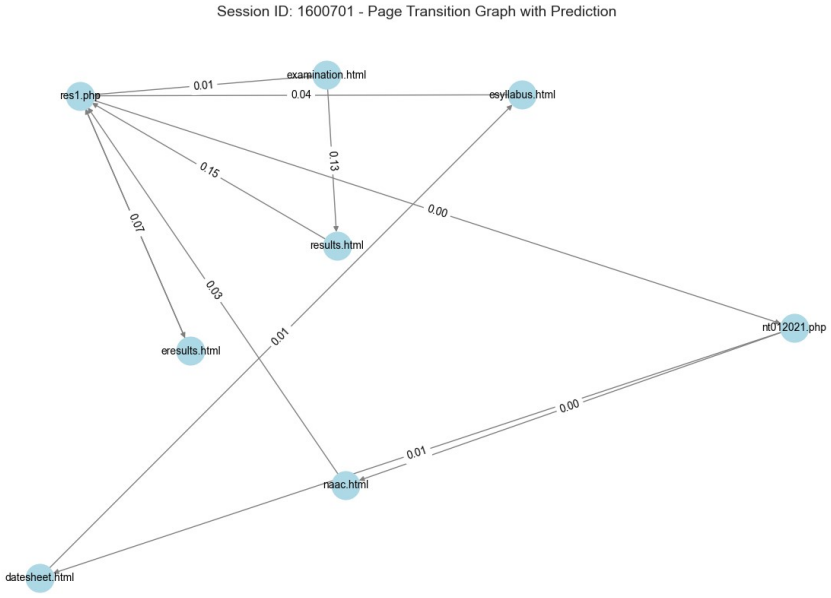


Figure 7 : DAG of a random session (Year -2021)

- Session ID: 1600701
- Total Pages (Nodes): 8
- Total Transitions (Edges): 11
- Most Connected Page: ('res1.php', 7)
- Strongest Transition (Source -> Target, Probability): (('results.html', 'res1.php'), 0.1491034644194756)

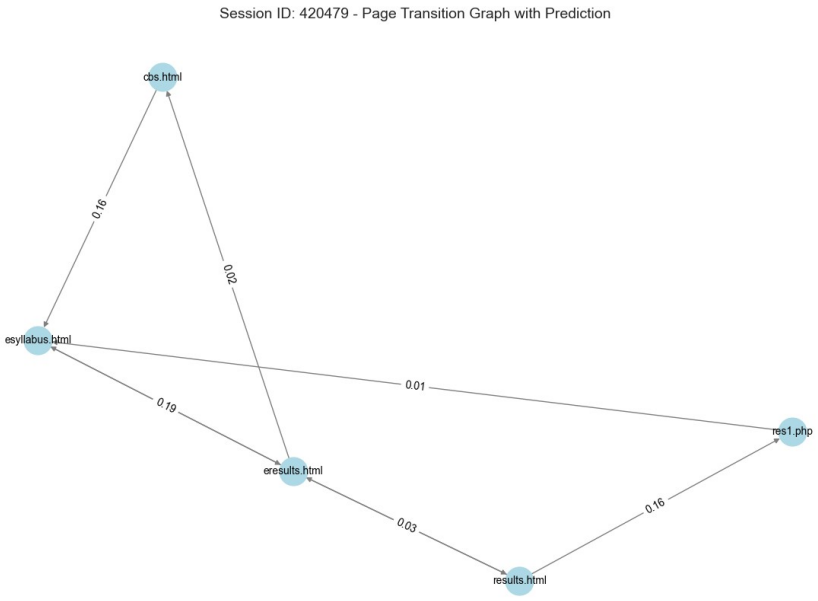


Figure 8 : DAG of a random session (Year -2022)

- Session ID: 420479
- Total Pages (Nodes): 5
- Total Transitions (Edges): 8
- Most Connected Page: ('eresults.html', 5)
- Strongest Transition (Source -> Target, Probability): (('eresults.html', 'results.html'), 0.2453825857519788)

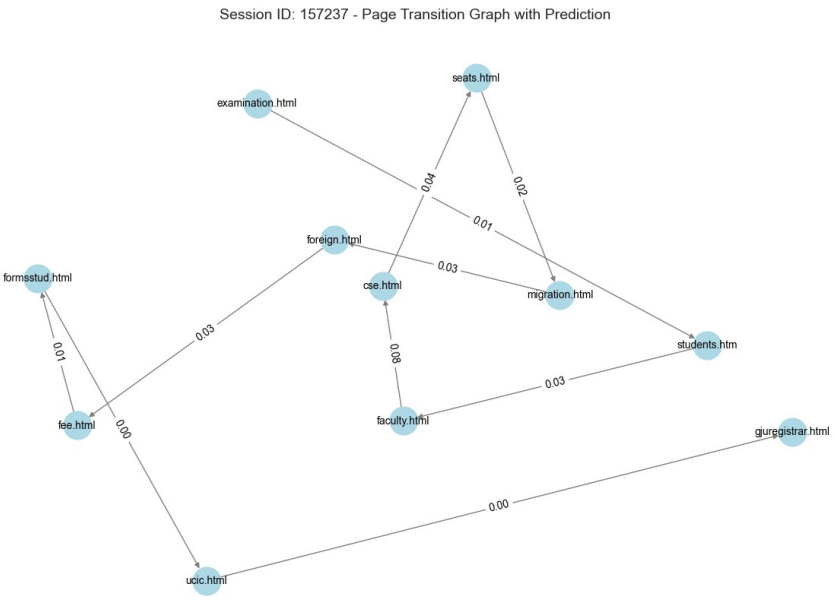


Figure 9 : DAG of a random session (Year -2023)

- Session ID: 157237
- Total Pages (Nodes): 11
- Total Transitions (Edges): 10
- Most Connected Page: ('students.htm', 2)
- Strongest Transition (Source -> Target, Probability): (('faculty.html', 'cse.html'), 0.0794824399260628)

To identify and characterize the different types of users in the user base, we use graph clustering, which works on top of graphs instead of an inside graph. This algorithm identifies similar graphs and clusters them into groups based on their structure. As we intend to design the user story based on the structure of the graph, the method fits our case perfectly. In the next subsection, we describe the graph clustering approach used in our work.

## IV. RESULTS AND DISCUSSION

The k-th order Markov Model provides valuable insights into user navigation patterns.

**Markov Chain Modelling -** Transition probabilities between pages were computed, and k-th order Markov models predicted the next page visited during the session.

**Session Segmentation -** User interactions were grouped into sessions based on temporal thresholds.

**Graph Analysis -** Directed graphs were constructed to track page-to page transitions, highlighted important nodes and edge.

It helps identify key nodes and transition probabilities, which can be used in:

- Personalising user experience
- Designing efficient navigation paths
- Optimising content recommendation systems

Key results from the study include:

- Transitions probabilities were calculated using k=3
- Each user's session contained predicted next pages based on historical data

**Year - 2021**

**Graph Analysis:** {'Pages and Transitions': {'Total Pages': 186, 'Total Transitions': 1041374}, 'Most Connected Page': 'results.html', 'Strongest Transition': ('pay.php', 'payment.php')}

**Summary Statistics:** {'Unique URLs': 21855, 'Unique Sessions': 1645095, 'Unique Remote Hosts': 421804, 'Unique Pages': 186}

**Year - 2022**

**Graph Analysis:** {'Pages and Transitions': {'Total Pages': 190, 'Total Transitions': 662710}, 'Most Connected Page': 'examination.html', 'Strongest Transition': ('datesheet.php', 'datesheet.html')}

**Summary Statistics:** {'Unique URLs': 4167, 'Unique Sessions': 687197, 'Unique Remote Hosts': 236983, 'Unique Pages': 190}

**Year - 2023**

**Graph Analysis:** {'Pages and Transitions': {'Total Pages': 201, 'Total Transitions': 611674}, 'Most Connected Page': 'examination.html', 'Strongest Transition': ('PG_physical_seats.php', 'PHd2020.php')}

**Summary Statistics:** {'Unique URLs': 16214, 'Unique Sessions': 631194, 'Unique Remote Hosts': 174995, 'Unique Pages': 202}

## CONCLUSION

The application of a k-th order Markov model offers an efficient method for understanding and predicting user behaviour on websites. This approach, combined with future improvements in real-time modelling holds significant potential for enhancing web traffic analysis.

## REFERENCES


[1] J. Borges, R. Frias, and M. Levene, "Evaluating Variable Length Markov Chain Models for Analysis of User Web Navigation Sessions,", *IEEE Transactions on Knowledge and Data Engineering,* vol: 19, Issue: 4, Page(s): 441 – 452, 2006, DOI: 10.1109/TKDE.2007.1012.

[2] N. Sain and S. Tamrakar, "A Survey of Web Usage Mining based on Fuzzy Clustering and HMM,", *International Journal of Computer Science and Information Technologies*, vol. 3(4), 2012.

[3] N. Sain and S. Tamrakar, "Web Usage Mining & Pre-Fetching Based on Hidden Markov Model & Fuzzy Clustering,", *International Journal of Computer Science and Information Technologies*, vol. 3(4), 2012.

[4] S. A. Ehikioya and J. Zeng, "Mining web content usage patterns of electronic commerce transactions for enhanced customer services," *Eng. Reports*, vol. 3, no. 11, pp. 1–36, 2021, wileyonlinelibrary.com/journal/eng2 doi: 10.1002/eng2.12411.

[5] S. Sharma and A. Bhagat, "Data preprocessing algorithm for Web Structure Mining,", *IEEE Xplore Proc. 5th Int. Conf. Eco-Friendly Comput. Commun. Syst. ICECCS 2016,* pp. 94–98, 2017, doi: 10.1109/Eco-friendly.2016.7893249.

[6] P. Svec, L. Benko, M. Kadlecik, J. Kratochvil, and M. Munk, "Web Usage Mining: Data Pre-processing Impact on Found Knowledge in Predictive Modelling," *Procedia Comput. Sci.*, vol. 171, pp. 168–178, 2020, www.sciencedirect.com doi: 10.1016/j.procs.2020.04.018.

[7] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci*., vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[8] D. H. Kraft, M. J. Martín-Bautista, J. Chen, and D. Sánchez, "Rules and fuzzy rules in text: Concept, extraction and usage," *Int. J. Approx. Reason*., vol. 34, no. 2–3, pp. 145–161, 2003, doi: 10.1016/j.ijar.2003.07.005.

[9] L. Chen, S. S. Bhowmick, and W. Nejdl, "COWES: Web user clustering based on evolutionary web sessions," *Data Knowl. Eng*., vol. 68, no. 10, pp. 867–885, 2009, www.sciencedirect.com doi: 10.1016/j.datak.2009.05.002.

[10] P. Sengottuvelan, R. Lokeshkumar, and T. Gopalakrishnan, "An improved session identification approach in web log mining for web personalization," *J. Internet Technol*., vol. 18, no. 4, pp. 723–730, 2017, doi: 10.6138/JIT.2017.18.4.20150113.

[11] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier, Eds., "Detecting Memory and Structure in Human Navigation Patterns Using Markov Chain Models of Varying Order," *PloS One,* Vol 9, 2014, https://journals.plos.org/plosone/.

[12] H. Jindal and N. Sardana, "PKM3: an optimal Markov model for predicting future navigation sequences of the web surfers," *Pattern Anal. Appl*., vol. 24, no. 1, pp. 263–281, 2021, doi: 10.1007/s10044-020-00892-7.

[13] J. Borges and M. L. Birkbeck, "A Dynamic Clustering-Based Markov Model for Web Usage Mining,", 2004, https://arxiv.org/abs/cs/0406032.

[14] H. Jindal and N. Sardana, "Web navigation prediction based on dynamic threshold heuristics," *Journal of King Saud University – Computer and Information Sciences,* Published by Elsevier, vol. 34, no. 6, pp. 2820–2830, 2022, doi: 10.1016/j.jksuci.2020.03.004.

[15] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Visualization of Navigation Patterns on a Web Site Using Model-Based Clustering," *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000.

[16] N. Wermuth, D. Cox, "Graphical Markov Models," *International Ency. of the Social and Behavioral Sc.,* 2nd ed., 10, Elesevier, Oxford, *341–350*, 2015, https://doi.org/10.48550/arXiv.1407.7783.