10.48047/jocaaa.2024.32.02.46

# Investigations and Analysis of Web Data Model and its Performance Evaluation

**Ashif Ali**
Ph.D. Research Scholar
Department of Computer Science
C. S. J. M. University, Kanpur
ashifali76@gmail.com

**Dr. Rashi Agarwal**
Faculty of Computer Science
Department of Computer Science
C. S. J. M. University, Kanpur

**Dr. Renu Jain**
Head of the Department
Department of Computer Science, C. S. J. M. University, Kanpur

***Corresponding Author:** Ashif Ali, Ph.D. Research Scholar**,** Department of Computer Science**,** C. S. J. M. University, Kanpur**,** ashifali76@gmail.com

## ABSTRACT

The web's growing size presents challenges for users who navigate the web to obtain relevant information. Web navigation prediction helps improve web cache performance, website design, and user preferences. However, modeling user behavior on the web faces challenges such as handling noisy navigations, unseen navigations, and deciding optimal thresholds for predictions. This research aims to conduct an empirical study of web navigation prediction models to find the best model and build an effective system. The All-Kth Modified Markov Model (KMMM) is found to be the best suited model for web navigation prediction. Other proposed models include backward browsing removal techniques, pruning models, novel hybrid models, and dynamic threshold-based models. These models improve website usability and user satisfaction, and help website owners understand future user preferences. This paper aims to improve web navigation prediction by proposing novel models for efficient prediction of user navigations. The models are tested on real web log datasets, such as University, e-commerce, product, and wikipages. The research involves capturing user navigation records, filtering out irrelevant data, identifying users using IP address, and identifying sessions. The models are then used to identify the next web page the user might visit. The paper addresses four major research gaps: no focussed empirical study, noisy data handling methods, low accuracy due to large sessions, and fixed thresholds for performance evaluation. The main contributions include evaluating models on common platforms, designing backward browsing elimination techniques, and integrating session cleaning models with dynamic thresholds.

**Keywords:** Navigation, Web User, Prediction Models, Markov, System Security, AI model, Web development.

## INTRODUCTION

This paper aims to address the challenges of web navigation prediction by proposing novel models for efficient prediction of user navigations on the web. The models are tested and

10.48047/jocaaa.2024.32.02.46

evaluated on real web log datasets, such as University, e-commerce, product, and wikipages. The web is a vast information repository, assessed by billions of users daily. Users often face difficulty in obtaining relevant information while browsing, leading to time loss [1]. To address this, an intelligent web recommender system is needed, which utilizes user browsing history in weblog files for recommendations. Web Navigation Prediction (WNP) is the process of discovering users' future navigation patterns based on past navigation behavior. The process involves capturing user navigation records, filtering out irrelevant data, identifying users using IP address, and identifying sessions using various techniques [2]. The user's navigation vectors are inputted into the pattern discovery phase, where WNP models are formed to identify the next web page the user might visit. Once the next possible web page is identified, model analysis of the pattern and recommendations are given to the appropriate users [3].

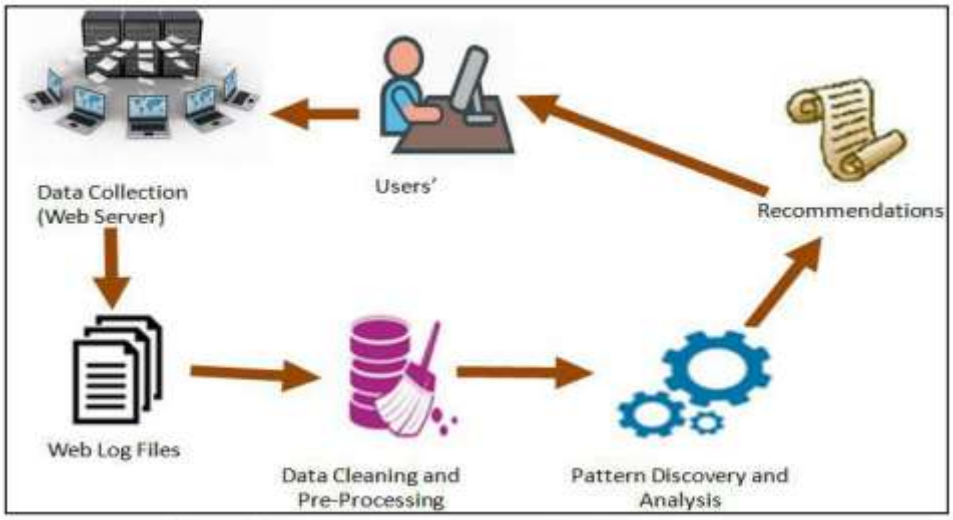


**Fig.-1 Framework for Web Navigation Prediction**

Web navigation systems aim to encourage website users to stay, peruse web content, and have a positive user experience. Web user navigation behaviour modelling can significantly benefit businesses by improving website structure, web cache performance, search engine recommendations, anomaly detection, location prediction, and personalizing browsing experiences [4]. For example, Amazon uses similar user navigation patterns for recommendations. Google Analytics mines navigation patterns to provide statistics like visit/session, page view, site referrer, conversion, bounce rate, and visitor visit, which can improve website design and impact conversions, sales, and bounce rates. Navigation patterns can also be used to predict users' future locations, such as nearby restaurants, shopping malls, movie theatres, and popular tourist locations [5]. However, modelling user navigation patterns presents challenges, such as noisy navigations, backlinks, and misleading navigations. Additionally, a prediction model may have new or unseen navigations that are not learned by the model, making it difficult to learn and improve performance [6].

10.48047/jocaaa.2024.32.02.46

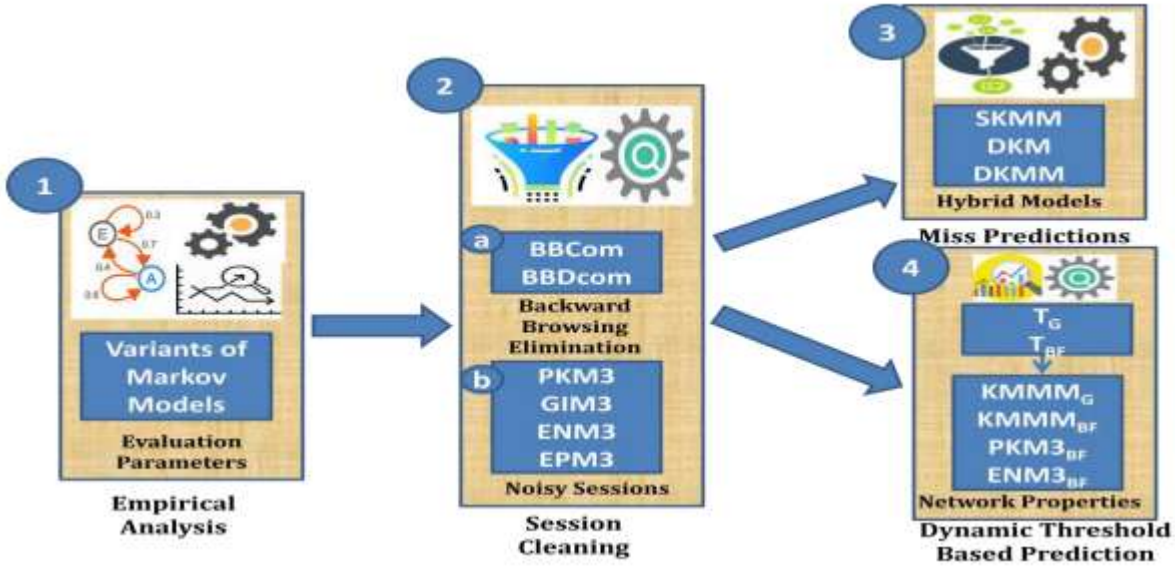


**Fig.-2 The four main research steps of this paper**

The web has evolved from static information representation to dynamic interaction, making it difficult to find relevant resources without capturing the meaning or semantics [7]. Web data modelling techniques capture and translate complex web applications into easily understood representations, ensuring high data quality and reducing development time, maintenance time, and redundancy. There are three basic levels of data models: Conceptual, Logical, and Physical [8]. A semantic web data model integrates vast amounts of online data and requires customer concern for describing inherent semantics. A requirement framework is needed to capture requirements and generate an equivalent conceptual model in an automated way [9].

The conceptual level design for web data models offer a better way to express problems and increase understand ability from the end user's perspective. However, a strategic approach for implementing the conceptual model into a physical level database schema is needed. Object Management Group's Model-Driven Architecture (MDA) provides an open, vendor-neutral approach to separate business and application logic from the underlying technology platform. Platform Independent Model (PIM) exhibits a specified degree of platform independence of data declaration, while Platform-Specific Model (PSM) is derived from PIM by transformation(s) using necessary rules for describing platform-specific details [10].

Performance evaluation is crucial for data model performance, with correctness, completeness, adaptability, and understands ability being major parameters. Techniques for semantic data model performance evaluation have been proposed, but they are mostly focused on evaluating semantic words or semantic queries. This paper proposes a conceptual level web data model to understand complex relationships within entities, aiming to express existing relations through abstractions and efficiently embed semantics within them [11].

## MAJOR OBJECTIVES AND FINDINGS

This paper aims to improve the prediction performance of web navigation models and recommend the most relevant information (webpage) to the user with minimum efforts. Four

10.48047/jocaaa.2024.32.02.46

major research gaps have been identified: no focussed empirical study that determines the best model for user navigation prediction based on parameters such as prediction accuracy, model accuracy, coverage, state-space complexity, and failure cases; existing methods for noisy data handling either generate longer sessions or require multiple scans or have high computation cost; previous studies addressed unseen data by integrating probabilistic models with machine learning classifiers that attained low accuracy due to a large number of sessions; and existing techniques use fixed thresholds for evaluating the performance of web navigation prediction which requires multiple iterations for fixation [12].

The main research question is "How can one analyze, model, and predict user future web navigation patterns efficiently?" The primary research question is split into four sub-research questions addressed in this paper. The research methodology adopted in this paper comprises of four steps: empirical analysis of web navigation prediction models over varied performance parameters; data cleaning; addressing unseen navigations; and proposing two dynamic threshold-based prediction models, KMMMG and KMMMBF [13].

The main contribution of this paper is providing a systematic methodology to improve and predict user future navigation behaviour on the Web. The major contributions of this paper include evaluating web navigation prediction models on common platforms and finding the best prediction model; designing backward browsing elimination techniques for noise elimination; and integrating session cleaning models with dynamic thresholds to boost the accuracy of the models [14-16].

## LITERATURE SURVEY

This Paper discusses various studies related to web navigation prediction, focusing on its importance, models, handling of noisy sessions, unseen navigations, and fixed thresholds. Web Navigation Prediction (WNP) is a system that predicts web pages a user might visit in the future based on their previously navigated web pages. These sessions are extracted from weblog servers and are varied in length, diversity, and size. The presence of longer or low-usage sessions can indicate unintuitive site links and make the model difficult to learn. Noisey sessions, such as backward navigation or misleading navigations, can make the prediction model difficult to learn [17-20]. Filtering approaches are required to remove noisy sessions. WNP can be applied in various applications, such as improving website structure, search engines, web caching systems, recommendation systems, anomaly detection, location prediction, and personalizing browsing experiences.

Web navigation prediction models have been developed to achieve various objectives, such as clustering similar users or sessions with common properties, classifying users based on traversed sessions, and using association rule mining techniques. However, these models struggle to scale well for large datasets and with large numbers of classes. Markov-based models, such as All-Kth Markov Model (KMM), Modified Markov Model (MMM), All-Kth Modified Markov Model (KMMM), Frequency-Pruned Markov Model (FPMM), Confidence-Pruned Markov Model (CPMM), and Error-Pruned Markov Model (EPMM), are used to discover future navigation patterns from existing navigation patterns [21-25].

Markov models are popular choices for modeling navigations, as they are stochastic processes and well suited for analysis and modeling of web access sequences of users on a website. However, other important metrics like model accuracy, coverage, and failure cases need to be analyzed to provide deeper insights about correct, incorrect, and miss predictions.

10.48047/jocaaa.2024.32.02.46

The main objective of this research is to discover future navigation patterns from current navigation patterns. Markov models and their variants have been popular choices in past studies, but more research is needed to compare these models empirically on a common platform [26].

## RESEARCH METHODOLOGY

The study evaluates the performance of dynamic threshold models with Top-k ranked fixed threshold models to demonstrate their effectiveness. The models were built using Netbeans and Java programming language, and the datasets were divided into training and testing segments. The performance of the proposed models was evaluated using real datasets, including the CTI Dataset, MSWEB dataset, and BMS WebView1 dataset. The CTI dataset consists of random surfing history of users, while the MSWEB dataset consists of records of 38000 users in February 1998. The BMS WebView1 dataset was part of the KDD Cup 2000 competition and contains 497 unique items and 59601 navigations [27].

The evaluation parameters used to measure the performance of the models were coverage and prediction accuracy. Coverage is the average number of predictions possible for the given test dataset, measured as the ratio of the number of predictions produced corresponding to each test session to the total test sessions available in the dataset. Prediction accuracy is the total correct predictions obtained by the model, measured as the sum of the total correct predictions relating to each test session to the total number of test sessions in the dataset [28].

The study compares the performance of proposed dynamic threshold-based WNP models KMMMG and KMMMBF over three real datasets: CTI, MSWEB, and BMS. The results show that longer sessions are rare, and the confidence of longer sessions is high for all datasets. The geometric threshold and branching factor threshold depend on support and confidence, with the threshold value reducing with longer sessions as longer training sessions are rare in the model [29]

The coverage metric is used to estimate the prediction ability of a model. In the CTI dataset, KMMM5 showed slight improvement over KMMM1. In the MSWEB dataset, KMMM1 had very little coverage, while KMMM5 showed significant improvement. The coverage of KMMMBF was more than KMMMG for all session length except 1-gram. In the BMS dataset, KMMMBF was higher than all other models. The study emphasizes the importance of the confidence parameter in evaluating the efficiency of WNP models [30].

## REQUIREMENTS & ANALYSIS

This paper proposes an improved prediction model for web navigation prediction, addressing unseen navigations. Markov-based models are popular for web navigation prediction, but their performance is primarily based on seen navigations. Unseen navigations are new sessions that have not been traversed by users, and the model does not learn these sessions during training. Conventional models like Support Vector Machine (SVM) and Neural Network (NN) have been integrated with Markov-based models to address unseen navigations.

10.48047/jocaaa.2024.32.02.46

However, these models are generally suitable for binary class problems with two classes, but their predictive power declines with increasing training data and classes. In this work, the authors combine Deep Neural Network (DNN) with enhanced Markov Models (KM2) and KM3 to address unseen, multi-class, and large data.

The three proposed models are Shallow All-Kth Modified Markov (SKMM), Deep All-Kth Markov (DKM), and Deep All-Kth Modified Markov (DKMM). The Dempster rule of combinations is used to integrate these models. The paper is organized into preliminaries, proposed models, research methodology, results, and summarizing findings.

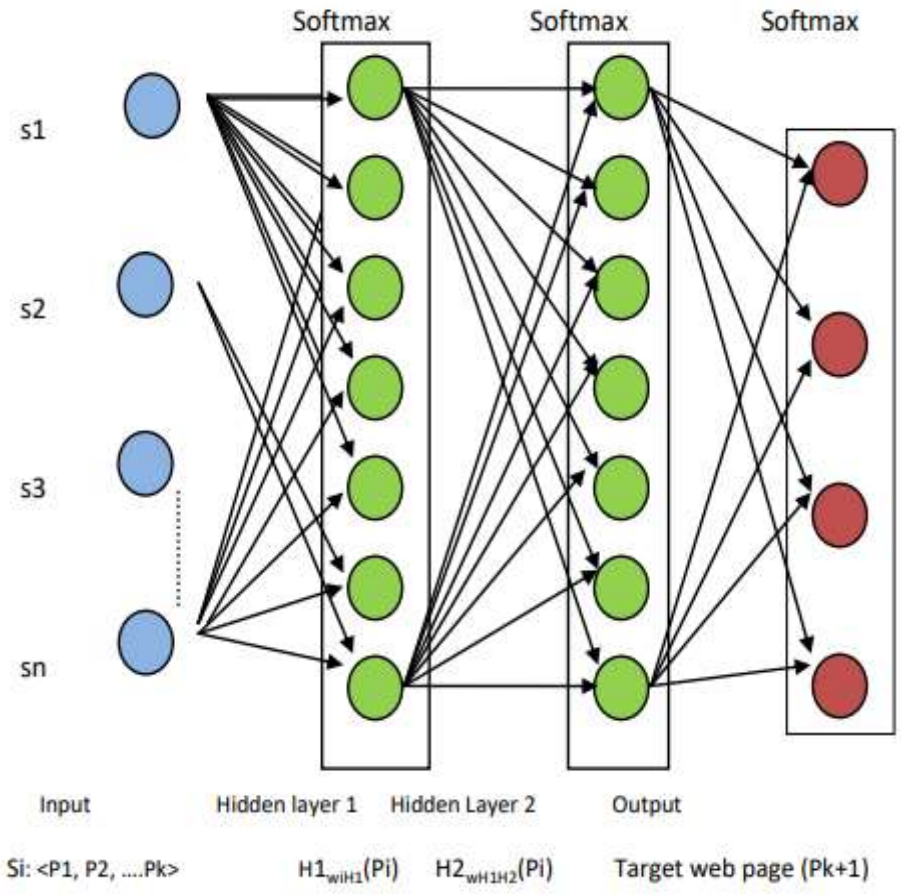


**Fig.3 Architecture of Deep Neural Network for Web Navigation Prediction**

This section discusses the basic architecture details of Neural Networks (NNs) used in this work, including Shallow Neural Network (SNN) and Deep Neural Network (DNN). SNN is a three-layer architecture with input, hidden, and output layers. It passes sessions successively in the forward direction, with the input layer forwarding the input session to the hidden layer and then forwarding data units to the output layer. The hidden layer processes the data, and the result is obtained from the output layer units.

10.48047/jocaaa.2024.32.02.46

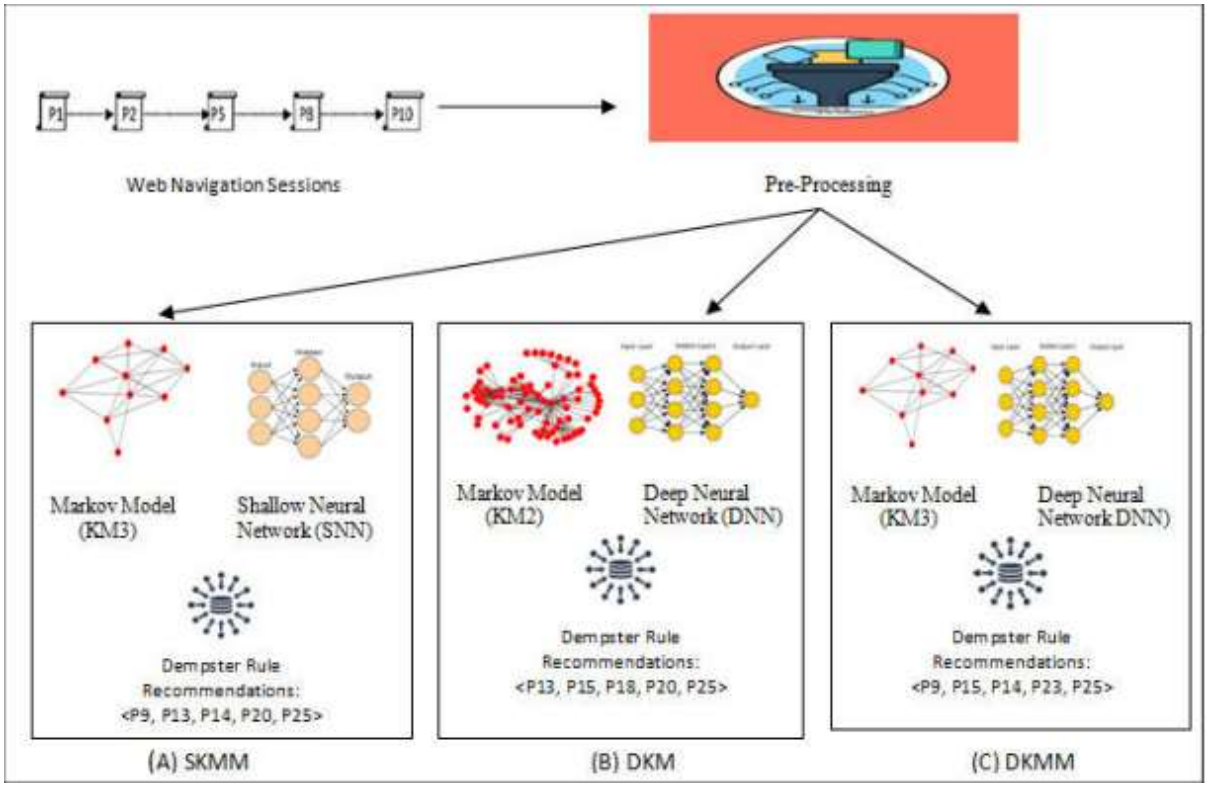


**Fig.-4 Proposed Hybrid models in web navigation prediction: (A) SKMM, (B) DKM, and (C) DKMM**

DNN, on the other hand, has more than one hidden layer. In WNP, the input session is the set of web pages, and the input is passed to each neuron (Perceptron) of the input layer. The Perceptron follows the modified Hebb rule, which considers connections with less network error in training.

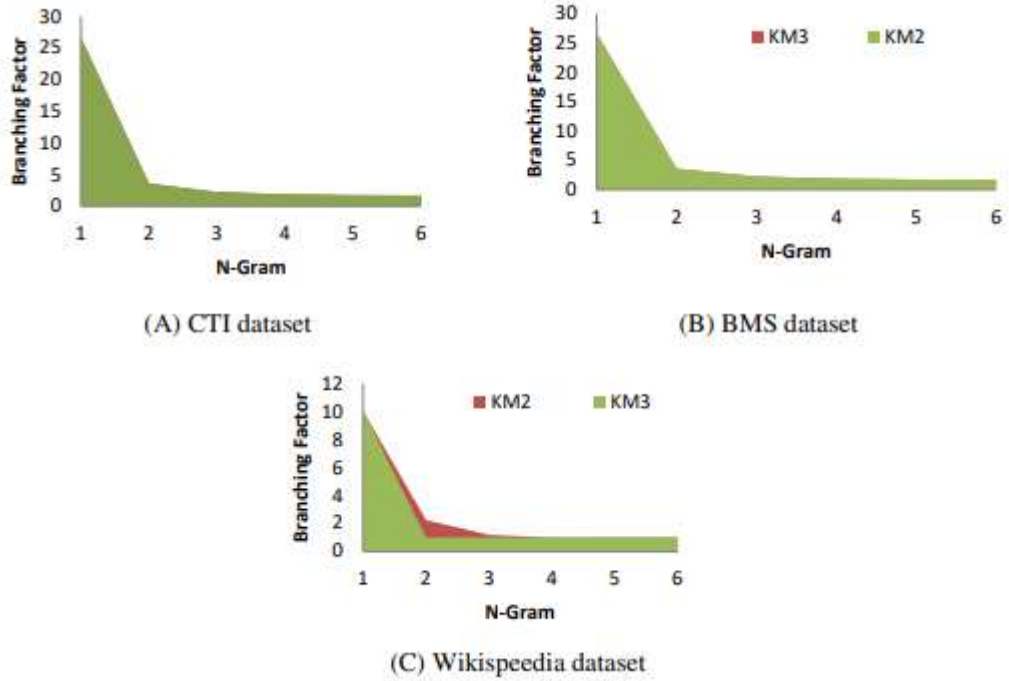


**Fig.5 Branching Factor of KM2 and KM3 over varied N-Grams**

10.48047/jocaaa.2024.32.02.46

In this work, two hidden layers are used to improve network predictions. The input layer does not perform computation, but passes the input to the subsequent hidden layer. The output of the input layer is equivalent to the input of the hidden layer 1 using weight *wij*

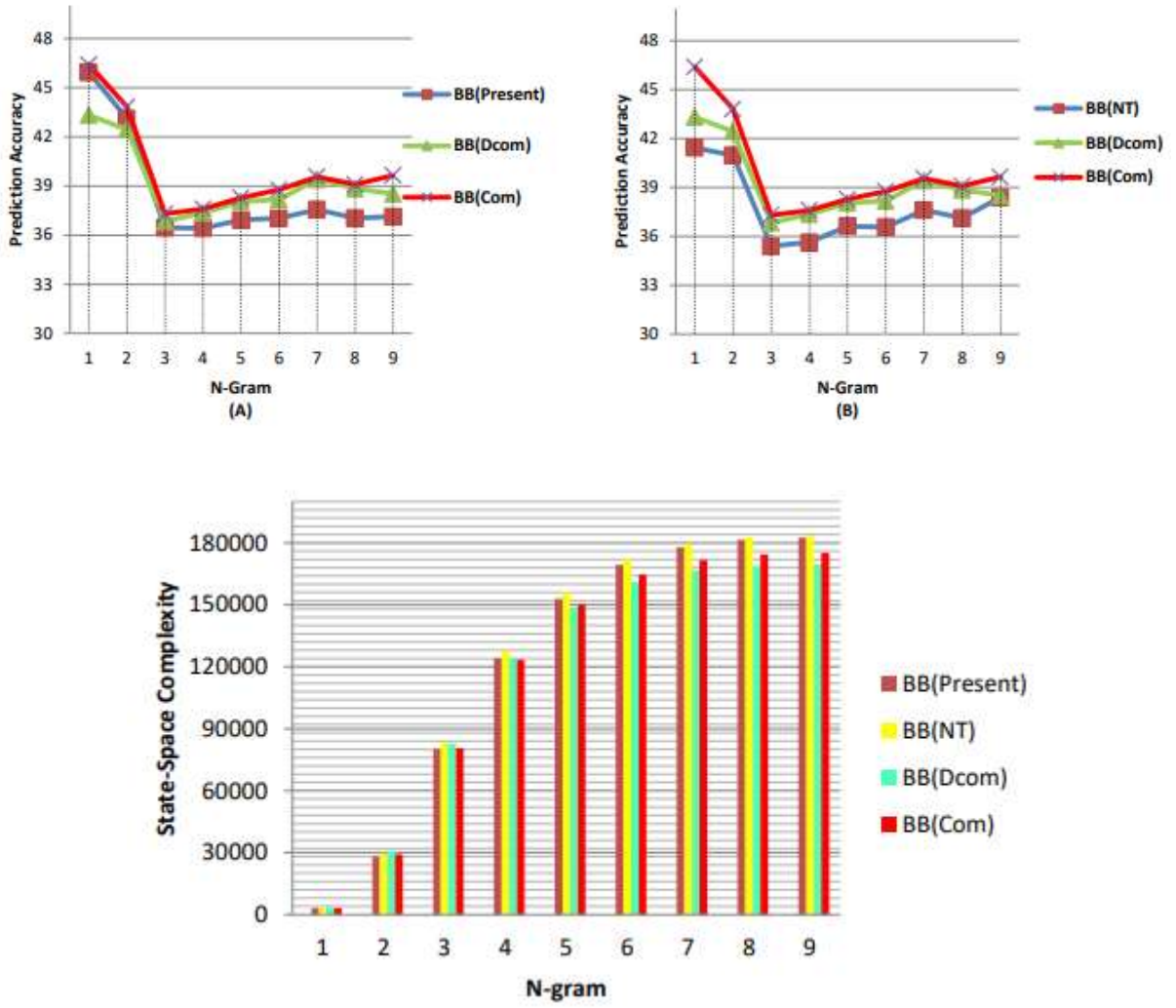


**Fig. 6 State-Space Complexity of backward browsing elimination techniques**

This paper focuses on identifying web sessions with noise and proposing new techniques to eliminate it. Two research dimensions were examined: backward browsing elimination and pruning sessions with negligible contribution to web navigation prediction. Backward browsing caused redundant web pages, making the model difficult to learn. Two techniques were proposed: BBCcom and BBCom, which produced shorter forward navigations and improved accuracy. The authors also proposed pruning models PKM3, GIM3, ENM3, and EPM3, which achieved low state-space complexity and comparable accuracy. PKM3 performed best for websites with high inter-linkages and simple navigations, with improvements of up to 5.45% over KMMM in the BMS dataset.

10.48047/jocaaa.2024.32.02.46

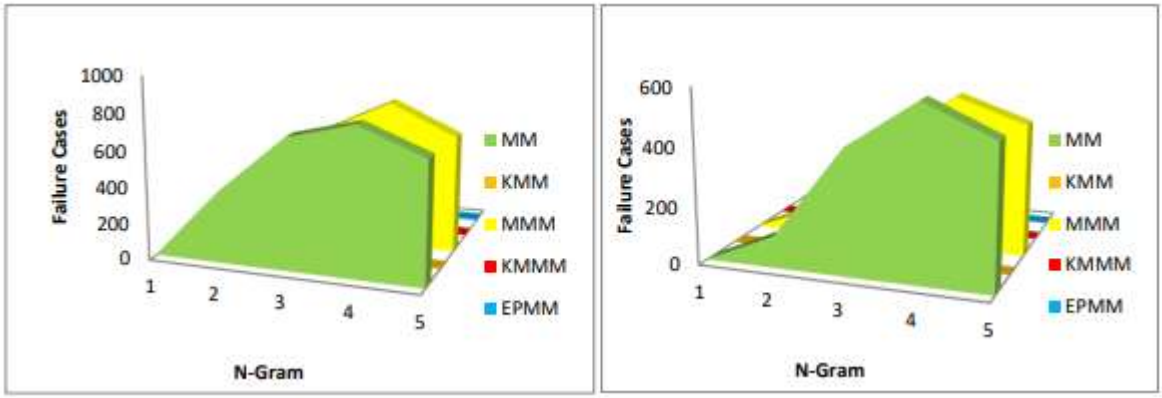


**Fig.7 Failure Cases required for building the Markov Models (A) CTI Dataset, and (B) MSWEB Dataset**

This paper compares the performance of various Markov-based models for Web Navigation Prediction (WNP). The models include traditional, modified, All-Kth, confidence-pruned, frequency-pruned, error-pruned, and all-kth modified Markov models. Results show that longer sessions affect model accuracy, coverage, state-space complexity, and failure cases. Nested lower-order models improve performance. KMMM performs best but has high state-space complexity. Markov-based models do not predict states unavailable in the training model, lowering their accuracy.

## CONCLUSION

This paper aims to address the challenges of web navigation prediction by proposing novel models for efficient prediction of user navigations on the web. The models are tested and evaluated on real web log datasets, such as University, e-commerce, product, and wikipages. Web navigation prediction (WNP) is the process of discovering users' future navigation patterns based on past navigation behaviour. The four main research steps of this paper include capturing user navigation records, filtering out irrelevant data, identifying users using IP address, and identifying sessions using various techniques. The user's navigation vectors are inputted into the pattern discovery phase, where WNP models are formed to identify the next web page the user might visit. Once the next possible web page is identified, model analysis of the pattern and recommendations are given to the appropriate users. Web navigation systems aim to encourage website users to stay, peruse web content, and have a positive user experience. Web user navigation behaviour modelling can significantly benefit businesses by improving website structure, web cache performance, search engine recommendations, anomaly detection, location prediction, and personalizing browsing experiences. Four major research gaps have been identified: no focussed empirical study that determines the best model for user navigation prediction based on parameters such as prediction accuracy, model accuracy, coverage, state-space complexity, and failure cases; existing methods for noisy data handling either generate longer sessions or require multiple scans or have high computation cost; previous studies addressed unseen data by integrating probabilistic models with machine learning classifiers that attained low accuracy due to a large number of sessions; and existing techniques use fixed thresholds for evaluating the performance of web navigation prediction which requires multiple iterations for fixation. The main contribution of this paper is providing a systematic methodology to improve and predict user future navigation behaviour on the Web. The major contributions of this paper include evaluating web navigation prediction models on common platforms and finding the best prediction model; designing backward browsing elimination techniques for noise elimination;

10.48047/jocaaa.2024.32.02.46

and integrating session cleaning models with dynamic thresholds to boost the accuracy of the models.

## References


1. Sunil K., Gupta S. and Gupta A., "A survey on Markov model," MIT International Journal of Computer Science and Information Technology, vol. 4, no. 1, pp. 29–33, 2014.
2. Chen M.S., Park J.S. and Yu P.S., "Efficient data mining for path traversal Patterns," IEEE Transactions on Knowledge and Data Engineering, vol. 10, no. 2, pp. 209–221, 1998.
3. Vishwakarma S., Lade S., Suman M. and Patel D., "Web user prediction by: Integrating Markov model with different features," International Journal of Engineering Research and Science and Technology, vol. 2, no.4, pp.74–83, 2013.
4. Straub DW, Watson RT. Research Commentary: Transformational Issues in Researching IS and Net-Enabled Organizations. Info Syst Res 2001;12(4):337-345. [doi: 10.1287/isre.12.4.337.9706]
5. Bayir M.A., Toroslu I.H., Demirbas M. and Cosar A., "Discovering better navigation sequences for the session construction problem," Data and Knowledge Engineering, vol. 73, no. 1, pp.58–72, 2012.
6. Fink D, Nyaga C. Evaluating web site quality: the value of a multi paradigm approach. Benchmarking 2009;16(2):259-273. [doi: 10.1108/14635770910948259]
7. Mamoun A. A., Latifur K., Bhavani T., "Predicting WWW surfing using multiple evidence combination," The VLDB Journal-The International Journal on Very Large Data Bases, vol. 17, no.3, pp. 401-417, 2008.
8. Markaki OI, Charilas DE, Askounis D. Application of Fuzzy Analytic Hierarchy Process to Evaluate the Quality of E-Government Web Sites. In: Proceedings of the 2010 Developments in E-systems Engineering. 2010 Presented at: DeSE'10; September 6-8, 2010; London, UK p. 219-224. [doi: 10.1109/dese.2010.42]
9. R.B Wagh, J.B. Patil, "Enhanced web personalization for improved browsing Experience," Advances in Computational Sciences and Technology, vol. 10, no. 6, pp. 1953- 1968, 2017.
10. Aranyi G, van Schaik P. Testing a model of user-experience with news websites. J Assoc Soc Inf Sci Technol 2016;67(7):1555-1575. [doi: 10.1002/asi.23462]

10.48047/jocaaa.2024.32.02.46

11. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess B, editors. Analyzing Qualitative Data. Abingdon-on-Thames: Routledge; 2002:187-208.
12. Akgül Y. Quality evaluation of E-government websites of Turkey. In: Proceedings of the 2016 11th Iberian Conference on Information Systems and Technologies. 2016 Presented at: CISTI'16; June 15-18 2016; Las Palmas, Spain p. 1-7. [doi: 10.1109/cisti.2016.7521567]
13. äws products, "Amazon Personalize, Real-time personalization and recommendation, based on the same technology used at Amazon.com", 2019 [Online]. Available: https://aws.amazon.com/personalize/.
14. Arrue M, Vigo M, Abascal J. Quantitative metrics for web accessibility evaluation. 2005 Presented at: Proceedings of the ICWE 2005 Workshop on Web Metrics and Measurement; 2005; Sydney.
15. European Commission (2016), "*Europa web guide*", The EU Internet Handbook, available at: http://ec.europa.eu/ipg/ (accessed 17 June 2018).
16. Bañón-Gomis A, Tomás-Miquel JV, Expósito-Langa M. Improving user experience: a methodology proposal for web usability measurement. In: Strategies in E-Business: Positioning and Social Networking in Online Markets. New York City: Springer US; 2014:123-145.
17. Dimopoulos C., Makris C., Panagis Y., Theodoridis E. and Tsakalidis A., "A web page usage prediction scheme using sequence indexing and clustering techniques," Data and Knowledge Engineering, vol. 69, pp. 371-382, 2010.
18. Wang WT, Wang B, Wei YT. Examining the Impacts of Website Complexities on User Satisfaction Based on the Task-technology Fit Model: An Experimental Research Using an Eyetracking Device. In: Proceedings of the 18th Pacific Asia Conference on Information Systems. 2014 Presented at: PACIS'14; June 18-22, 2014; Jeju Island, South Korea.
19. ISO (n.d), "*Terms & definitions*", Online Browsing Platform (OBP), available at: https://www.iso.org/obp/ (accessed 18 February 2021).
20. Jainari, M.H., Baharum, A., Deris, F.D., Mat Noor, N.A., Ismail, R. and Mat Zain, N.H. (2022), "A standard content for university websites using heuristic evaluation", in Arai, K. (Ed.), Intelligent Computing. SAI 2022. Lecture Notes in Networks and Systems, Springer, Cham, Vol. 506, doi: 10.1007/978-3-031-10461-9_19.

10.48047/jocaaa.2024.32.02.46

21. Król, K. and Zdonek, D. (2020), "*Aggregated indices in website quality assessment*", Future Internet, Vol. 12 No. 4, p. 72.
22. Law, R. (2019), "*Evaluation of hotel websites: progress and future developments*", International Journal of Hospitality Management, Vol. 76, pp. 2-9.
23. Morales-Vargas, A., Pedraza-Jiménez, R. and Codina, L. (2020), "*Website quality: an analysis of scientific production*", Profesional de la Información, Vol. 29 No. 5, p. e290508.
24. Thielsch, M.T. and Hirschfeld, G. (2019), "*Facets of website content*", Human–Computer Interaction, Vol. 34 No. 4, pp. 279-327.
25. Whitenton, K. (2021), "*Triangulation: get better research results by using multiple UX methods*", Nielsen Norman Group, available at: https://www.nngroup.com/articles/triangulation-better-research-results-using-multiple-ux-methods/ (accessed 4 March 2021).
26. Allison R, Hayes C, Young V, McNulty CAM. Evaluation of an educational health website on infections and antibiotics: a mixed-methods, user-centred approach in England. JMIR Formative Research. 2019 doi: 10.2196/14504. (forthcoming) [DOI] [PMC free article] [PubMed] [Google Scholar]
27. B. Naveen, J. K. Grandhi, K. Lasya, E. M. Reddy, N. Srinivasu and S. Bulla, "Efficient Automation of Web Application Development and Deployment Using Jenkins: A Comprehensive CI/CD Pipeline for Enhanced Productivity and Quality," *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India, 2023, pp. 751-756, doi: 10.1109/ICSSAS57918.2023.10331631.
28. "IEEE Recommended Practice for the Internet - Web Site Engineering, Web Site Management and Web Site Life Cycle," in *IEEE Std 2001-2002 (Revision of IEEE Std 2001-1999)* , vol., no., pp.1-114, 3 March 2003, doi: 10.1109/IEEESTD.2003.94235.
29. Suresh S., Prakash K. and Jose B., "Semantically enriched web usage mining for personalization," International Journal of Computer, Control, Quantum and Information Engineering, vol. 8, no. 1, pp.249–257, 2014.
30. Cooley, R., P. Tan, J. Srivastava., "Discovery of interesting usage patterns from Web data", 2000.