International Conference on
Global Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Web Page Recommendation Using KNN Model and Genetic Algorithm

Phd Scholar Sumit Sharma, Associate Professor Dr. Pritaj Yadav
Department of Computer Science and Engineering,
RNTU, Bhopal, MP, India

Abstract- Recommender systems are very helpful on the internet that suggest things personalized just for you. They're great because they make it easier to find what you want without being overwhelmed by too much information. This paper is all about recommending web pages by looking at how people use websites and the content on those pages. They used a smart model called K-nearest neighbors to figure out which pages you might like based on what other people with similar interests have viewed. Then, they made these suggestions even better by using a clever algorithm called the elephant herd optimization. This work tested method on real website data to see if it actually works well. Since people's internet habits are always changing, it's important to have a recommendation system that can keep up. The results show that approach, called Elephant Herd-based Web page Recommendation (EHWPP), really makes things work better. In simpler terms, it's like having a smarter system that helps you find the web pages visitor want in a way that suits their interests.

Keywords- Data mining, Feature Extraction, Page Recommendation, Web mining.

I.INTRODUCTION

In contemporary times, the widespread utilization of websites and web pages has significantly augmented network traffic, imposing a substantial load on web servers. Despite the availability of large bandwidth connections for web clients, they still encounter high latencies during web navigation due to various overloaded elements, such as network congestion and prolonged message transference times. Addressing the reduction of users' perceived latency when browsing websites remains a pivotal area of research [1]. Over the past few years, numerous research efforts have been dedicated to mitigating web users' perceived latency, with prominent techniques including web caching, geographical replication, and pre-fetching.

Web caching has become a widely adopted approach, as it yields substantial latency savings. Major corporations often employ web replication through Content Delivery Networks to expedite website access times, albeit with some drawbacks, including cost implications that may render this solution unaffordable for many smaller companies.

The concept of web pre-fetching aims to proactively process object requests before users explicitly demand them, with the goal of reducing perceived latency. Leveraging previous web usage data and domain knowledge, predictive algorithms can calculate the next page to be accessed. While 1st-order Markov Models (Markov Chains) offer a straightforward way to capture sequential dependence [2], they fall short in considering the long-term memory aspects of web surfing behavior. This limitation

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

arises from their reliance on the assumption that the next state to be visited is solely a function of the current one. Higher-order Markov models [3], while more accurate in predicting navigational paths, introduce a trade-off between enhanced coverage and an exponential increase in state space complexity as the order increases [4].

The web log, serving as a registry of web pages accessed by diverse users at different times and locations, can be maintained at the server side, client side, or a proxy server, each with its own benefits and drawbacks in identifying users' relevant patterns and navigational sessions [5]. A client-side web log, for instance, captures only the web accesses by a specific client or user, proving beneficial in mining access sequences for that user.

II. RELATED WORK

Awad et al. [6] conducted a study in which they employed a hybrid approach, integrating Support Vector Machines with the All-kth Markov model, to ascertain calculations through the utilization of Dempster's rule. The augmentation of the discriminatory power was achieved by implementing attribute extraction of SVM. The study culminated in the incorporation of domain information, resulting in a reduction in the number of classifiers along with a concurrent decrease in calculation time.

In the research conducted by Mamoun A. Awad et al. [7], an exploration of the all-kth Markov model and the Markov model was undertaken for web prediction purposes. This exploration led to the development of an improved Markov model, specifically designed to address scalability issues in paths. A novel two-tier prediction framework was conceptualized to establish an EC classifier, constructed by examining training examples and the subsequently generated classifiers. The resulting model demonstrated enhanced efficiency, saving time and providing superior predictions. Comparative analysis, utilizing standard benchmark data and varying parameters of the Markov Model, underscored the significance of the proposed technique. The experiment validated the effectiveness of the modified Markov model.

Giorgos Kollias et al. [8] presented findings indicating that polynomials of stochastic matrices could be calculated using the product of Google matrices, which bear the same form as the original page rank formulation of Google. Each matrix was parameterized based on different damping factors, exhibiting desirable characteristics. The multidamping approach proved beneficial in identifying the highest page rank and provided approximate solutions. Furthermore, multidamping offered insights into the interpretation of current ranking in terms of user web-surfing habits. The multidamping strategy retained the original Monte Carlo-type framework, featuring distributed and parallel implementations. It introduced a novel link-based ranking system grounded in homogeneous products present in Google matrices. The algorithm also presented a method for calculating damping factors through analytical and numerically functional ranking.

Deepa and Raajan [9] developed a preprocessing methodology to transform log files into client sessions, facilitating mining and understanding of session ranges while identifying the least demanded pages by users. Recognizing the importance of information preprocessing before applying any mining algorithm, they established a record file applied to CTI record files. Client reviews were incorporated, and a filtering method was implemented to exclude the least demanding pages or resources.

Om Prakash et al. [10] proposed a web page prediction method based on Bhattacharya distance (WS-BD) and weighted support, with the primary objective of enhancing customer satisfaction. The initial step involved obtaining sequential patterns through weighted support, filtering patterns using the

PrefixSpan algorithm based on duration, recurrence, and frequency. The interesting sequential patterns were then clustered using a proposed dice similarity-based Bayesian fuzzy clustering. Ultimately, the predicted web page was determined using the Bhattacharya distance of a two-level match.

III. PROPOSED PAGE RECOMMENDATION MODEL

This section brief proposed Artificial Immune and Bootstrap Bagging based Webpage Recommendation (AIBBWR). Fig. 1 shows the flow of the model for feature collection and bootstrap bagging model training. In order to increase the recommendation hit artificial immune based testing is performed. So fig. 2 shows testing flow of the model. Some of nations used in the explanation of whole work is list in table 1.

1. Pre-Processing

Input raw dataset of a website need to be pre-process for the feature extraction either log or content. Each feature has its own value set, so steps should of wen and content is separate. This can be understand by let input raw looks like: 35.199.25.81 - - [05/Jul/2023:12:00:23 +0530] "GET /2020/08/15/list-of-journals-without-publication-fee HTTP/1.1" 301 278 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"

A web log shown above is taken from the dataset and it has information about visitor browser, timestamp, operating system, etc. out of these important identification is IP address and webpage link [13, 14]. So work will extract Ip, webpage link and timestamp from the log.

Raw WebData	Processed WebData (PRD)
35.199.25.81 - - [05/Jul/2023:12:00:23 +0530] "GET /2020/08/15/list-of-journals-without-publication-fee HTTP/1.1" 301 278 "-" "Mozilla/5.0 (Windows NT 6.1; WOW64; Trident/7.0; rv:11.0) like Gecko"	35.199.25.81
	05/Jul/2023:12:00:23 +0530
	"/2020/08/15/list-of-journals-without-publication-fee"

PWD←Pre_Processing(RWD)

Weblog Timestamp, IP and page link is use for the weblog feature extraction. Each sequential IP is used for the log generation where set of webpages were collect. In order to educe the size of data each page from the log is save in a file have unique pages.

So If IP 35.199.25.81 visit three pages "/2020/08/15/list-of-journals-without-publication-fee" then "/2020/08/15/list-of-journals-without-publication-fee" "/2018/04/20/fast-publication-journals-impact-factor" then "/2020/04/21/journals-for-publication/" and page ids are 1, 2, and 3. Then weblog is {1•2•3}. Similarly other set of logs were generated.

WL←weblog_Pre-procesing(PWD)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

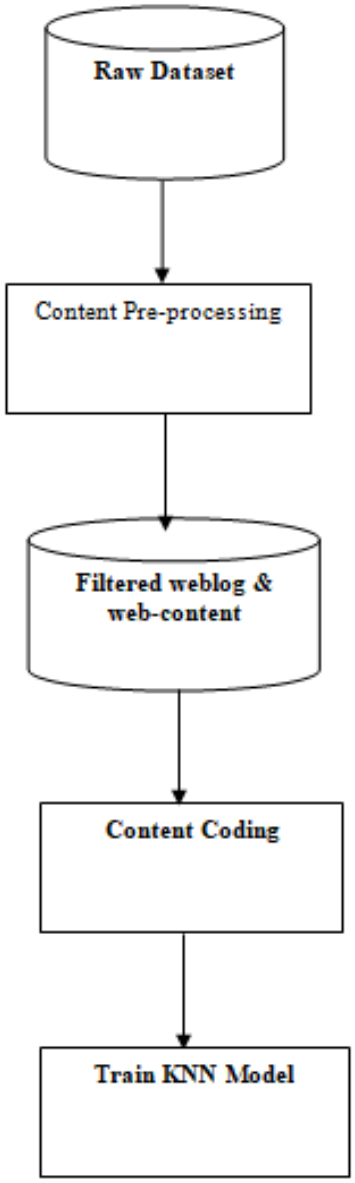

Fig. 1 Block diagram of EHWPP.

KNN Model

Weblog is sequence of web pages visit by a user in fix range of timestamp. KNN finds the match with the last page visit of the user log in s single session [16]. So input are set of webpage or weblog WL and web content to train the KNN [17].

Consider the weblog 1→2→3→4 then number of common terms between {1,4}, {2,4}, {3,4} is [2, 0, 1] web content feature. Further lets other weblog is 1→5→6→8 then number of common terms between {1,8}, {5,8}, {6,8} is [2, 1, 2] web content feature. If new visitor weblog is 2→3→6 then its feature respect to 4 is [0, 1, 1] and feature respect to 8 is [1, 2, 2] then chance of opening page is 8 more as compared to 4.

KNN←Train_KNN(WL, WC)

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Prediction Model

Elephants, as social creatures, live in social structures of females and calves. An elephant clan is headed by a matriarch and composed of a number of elephants. Female members like to live with family members, while the male members tend to live elsewhere. They will gradually become independent of their families until they leave their families completely.

- Some clans with fixed numbers of elephants cover the elephant population.
- A fixed number of male elephants will leave their family group and live solitarily far away from the main elephant group in each generation.
- A matriarch leads the elephants in each clan.

Input is features collect form the RWD in form of W, WC. As user can click on any page in the webpage so dynamic situation need to resolve. For getting the randomness in the work elephant herd optimization was used [18]. This generate set of pages that may be click by the visitor.

Elephant Clan Population

Random set of webpages were developed in the work by use of KNN model. So each clan have set of webpage obtained from the KNN model where input is visitor previous set of pages VPP and its combinations. Herd population C is matrix of m number of elephants and each elephant have n number of possible pages.

$$C \leftarrow Generate\_Herd(KNN, VPP, m, n)$$

Fitness

In order to evaluate the fitness of each elephant in C fitness is estimate. This work uses WL and WC feature so number of similar terms between the VPP and elephant possible pages is fitness in the work.

$$AF_m = \sum_{i=1}^{m} \sum_{j=1}^{n} Web\_Term\_Count(A_{i,j}, VPP \quad)$$

AF_m is fitness of m^{th} elephant in C.

Clan Update

A best solution matriarch, M is derived based on the fitness values of each elephant in clan in the population [19]. A number of the statuses were randomly changed based on the best matriarch, M feature set. The cloning is done by placing the best elephant set page in other elephant of clan.

$$C \leftarrow Clan\_update(M_b, C)$$

Separating

Low fitness elephant were removed from the clan in form of male elephant. This is done after estimating the new clan fitness value.

Possible Page prediction

The best-fitting chromosome in the final C population is considered for inclusion in the page prediction subset after t iterations (fitness function, crossover). This collection of pages from the user's chromosomal element is a probable set of pages for the user's VPP viewed pages.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

IV. EXPERIMENT & RESULTS

Implementation of AIBBWR is done on MATLAB 2016 version. Machine used for the experimental work have I3 processor and 4GB RAM, operated on windows operating system. Comparison of model was done with model proposed in [18].

Web Dataset: To test model, real weblogs were taken from https://ijsret.com for month on July-2023. This website has 2610 pages list on Google search engine. Weblog of 3776 were taken for the analysis and testing.

Results

Table 1 Precision value based comparison of next webpage recommendation models.

Testing Weblogs	PASO [20]	AIBBWR [21]	EHWPP
1962	0.586	0.7387	0.7769
2616	0.5904	0.6066	0.6026
3271	0.6129	0.6412	0.6763
3925	0.6398	0.6994	0.7016
4579	0.6515	0.7428	0.771

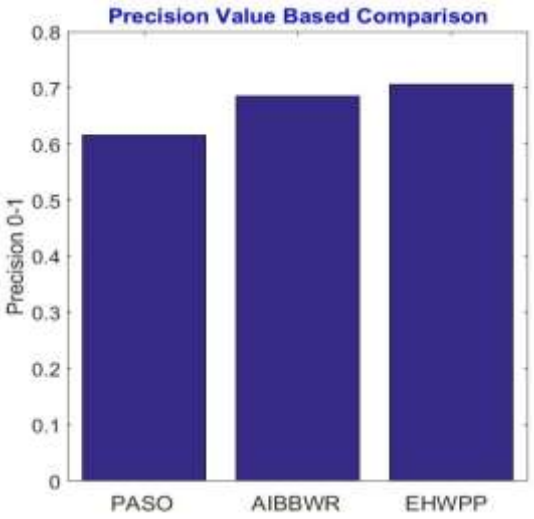

Fig. 2 Page prediction average precision value.

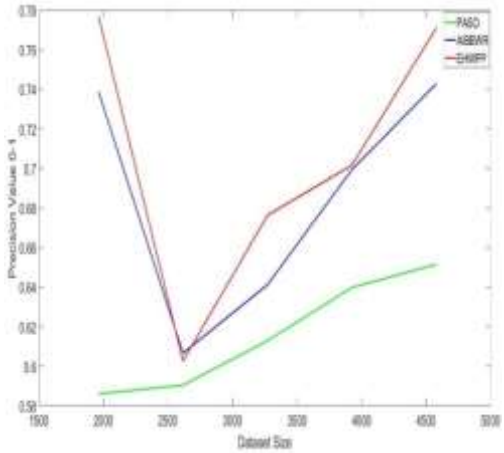

Fig. 3 Page prediction precision value as per different dataset size.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 1, fig. 2 and fig. 3 shows precision values of the webpage prediction models. It was found that use of elephant herd optimization algorithm has improved the performance. Further it was shown with increase of weblogs precision value was enhanced. EHWPP has increased the precision by 2.825% as compared to AIBBWR.

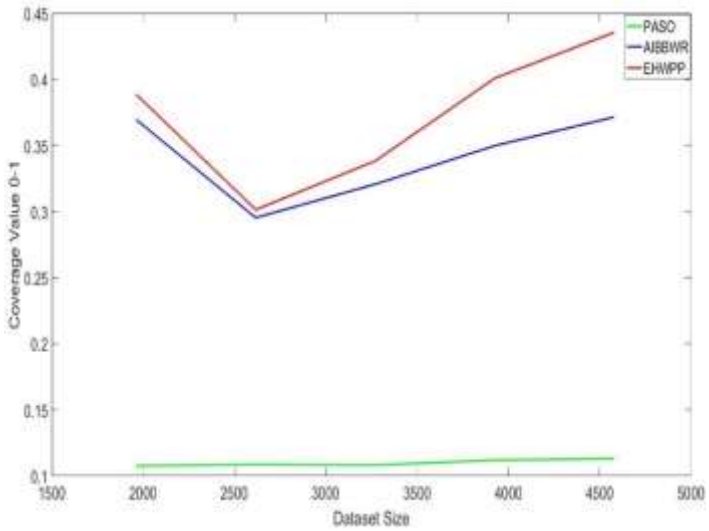

Fig. 4 Page prediction coverage value as per different dataset size.

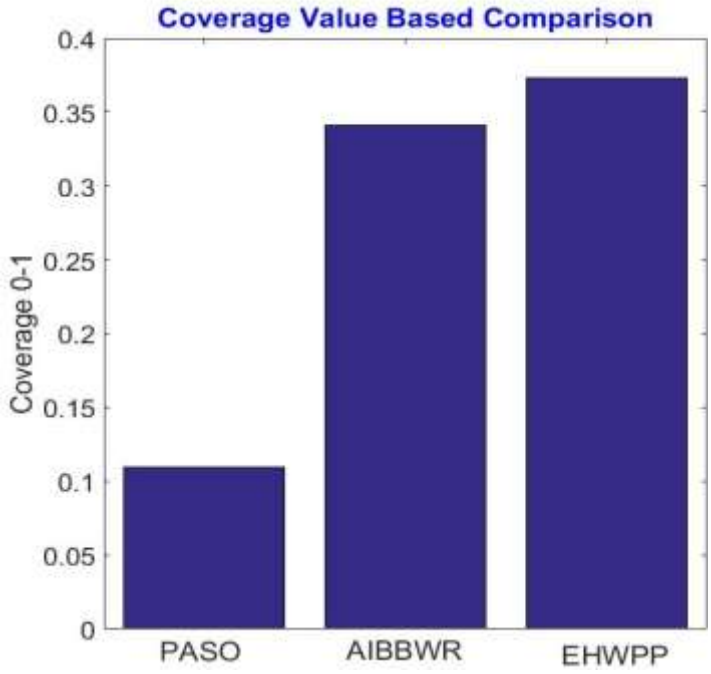

Fig. 5 Page prediction average coverage value

Fig. 4 and 5 shows that use of genetic algorithm for page set prediction has increases the performance of coverage. It was also found that out of Artificial Immune, Elephant herd algorithm and PSO Elephant perform well in all set of dataset. EHWPP model has improved the coverage parameter by 8.24% as compared to AIBBWR.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

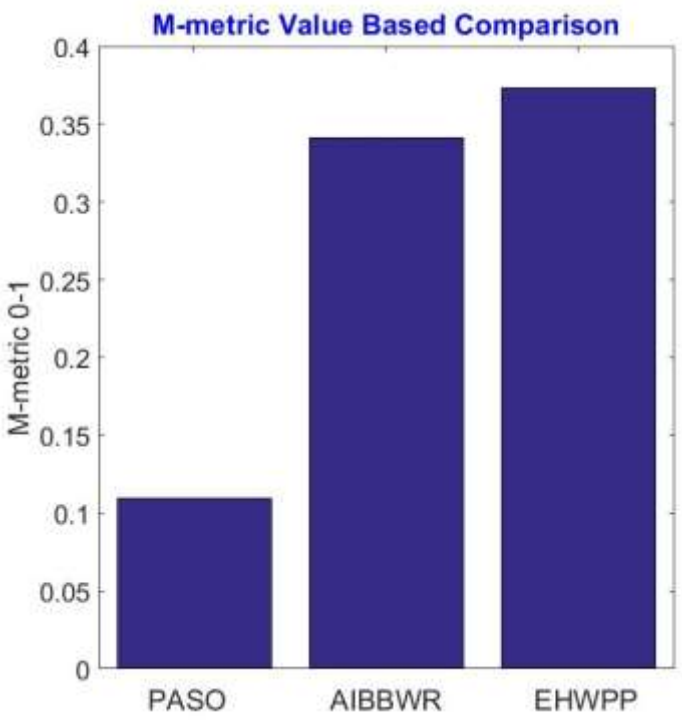

Fig. 6 Page prediction average M-metric value

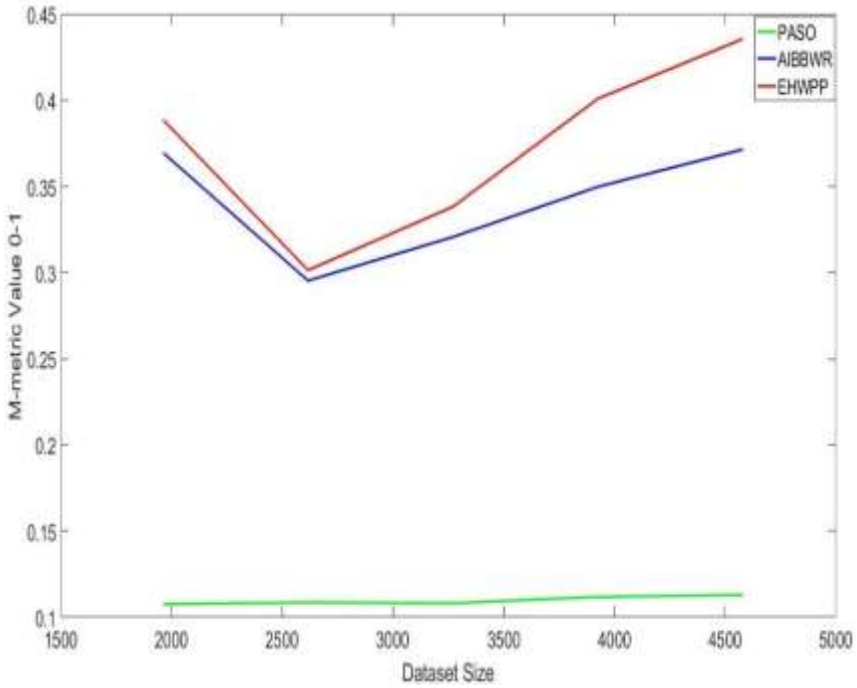

Fig. 7 Page prediction M-metric value as per different dataset size.

Fig. 6 and 7 shows m-metric values of the webpage prediction models. KNN model for prediction of page has increases the performance of web page prediction. Further it was shown with increase of weblogs m-metric value was enhanced. EHWPP has increased the precision by 5.81% as compared to AIBBWR.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

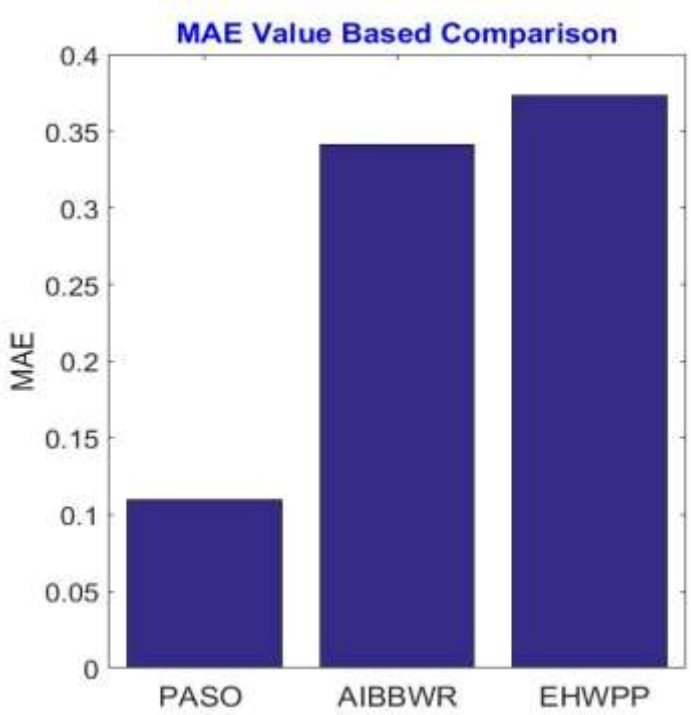

Fig. 8 Page prediction MAE value as per different dataset size

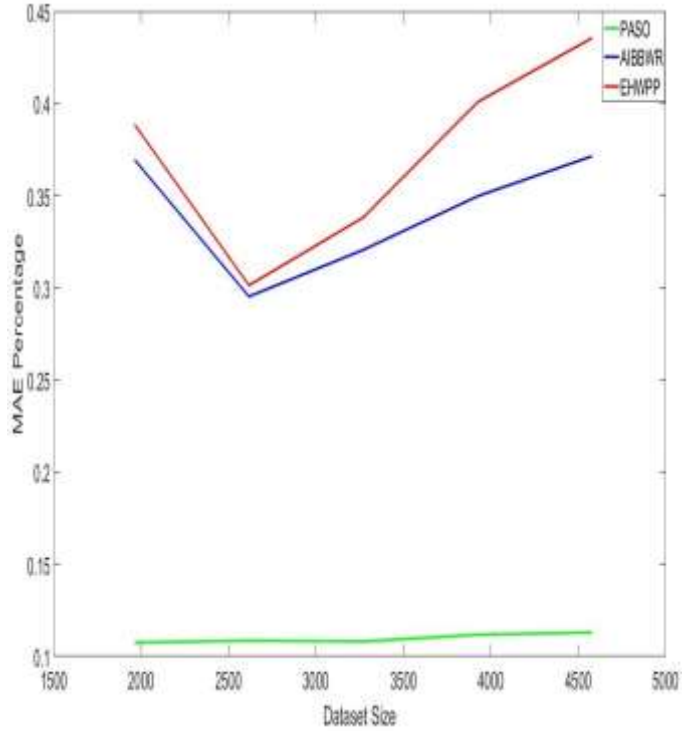

Fig. 9 Page prediction MAE value as per different dataset size.

Page prediction models errors was compared in fig. 8 and 9. It was found that EHWPP model has reduced the MAE by 18.24% as compared to AIBBWR. Table 7 shows that use of KNN model has reduces the RMAE by 10.06% compared to AIBBWR.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 2 RMAE value based comparison of next webpage recommendation models.

Testing Weblogs	PASO [20]	AIBBWR [21]	EHWPP
1962	0.3847	0.0396	0.0593
2616	0.4013	0.0553	0.059
3271	0.3787	0.0537	0.0561
3925	0.3597	0.049	0.0498
4579	0.3395	0.0454	0.046

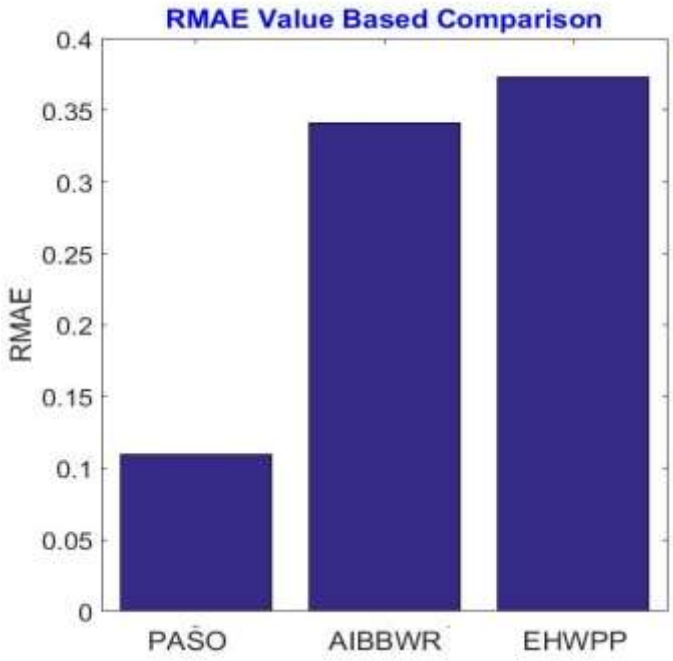

Fig. 10 Page prediction MAE value as per different dataset size.

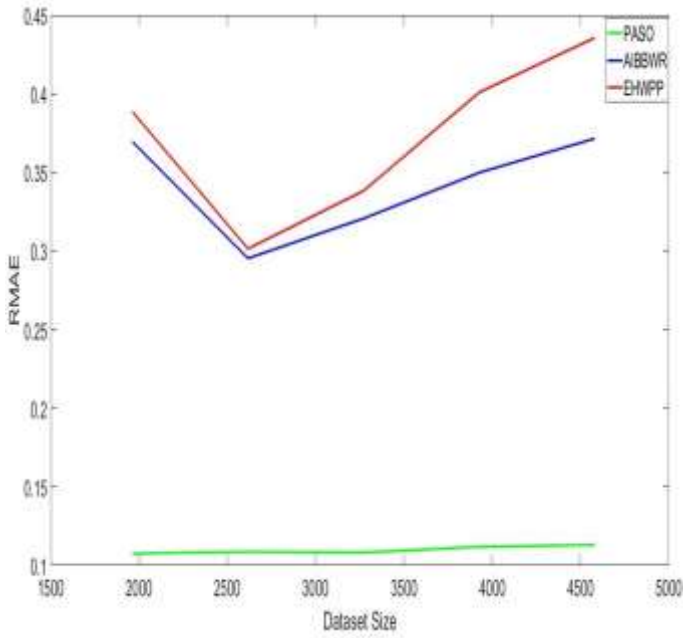

Fig. 11 Page prediction MAE value as per different dataset size.

Page prediction models errors was compared in fig. 10 and 11. It was found that EHWPP model has reduced the MAE by 18.24% as compared to AIBBWR. Table 2 shows that use of KNN model has reduces the RMAE by 10.06% compared to AIBBWR.

V. CONCLUSION

Prediction of web pages as per visitor might find interesting is beneficial for both the visitor and the website. However, accurately identifying the right set of pages is a challenging task. To tackle this challenge, the paper introduces a novel model called Artificial Immune and Bootstrap Bagging based Webpage Recommendation (AIBBWR). This model leverages both the website's weblog data and the content features of its pages to understand and learn from the behavior of visitors. EHWPP incorporates advanced techniques to enhance its recommendation performance. By applying this algorithm to the extracted features from the web data, the paper aims to make the webpage recommendations more accurate and tailored to individual preferences. The proposed model's effectiveness was evaluated through experiments conducted on real website data. The results reveal that EHWPP model has reduced the MAE by 18.24% as compared to AIBBWR. Use of KNN model has reduces the RMAE by 10.06% compared to AIBBWR. This suggests that AIBBWR outperforms existing methods in providing more accurate and relevant suggestions to users. Looking ahead, the paper suggests that future research could explore incorporating additional data, such as users' past site visits, to further enhance the recommendation process. This indicates an ongoing commitment to refining and evolving recommendation systems to better serve the diverse needs and preferences of web users.

REFERENCES

1. Kamilov, M., Hudayberdiev, M. and Khamroev, A. (2019). Algorithm for the Development of a Training Set that Best Describes the Objects of Recognition. In: Procedia Computer Science. vol.150, pp.116-122.
2. Miniukovich A., Figl K. Web Design Prototypicality: Commercial Banks Harvard Dataverse (2023),
3. Khamroev, A. (2017). An algorithm for constructing feature relations between the classes in the training set. In: Procedia Computer Science. vol.103, pp.244-247.
4. C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus, S. García. "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com", Expert Systems with Applications, Volume 39, Issue 12, 2012, Pages 11243-11249.
5. P. Verma and N. Kesswani, "Web Usage mining framework for Data Cleaning and IP address Identification," 2014.
6. 28. M. Awad, L. Khan, and B. Thuraisingham, "Predicting WWW surfing using multiple evidence combination," VLDB J., vol. 17, no. 3, pp. 401– 417, May 2008.
7. 29. Mamoun A. Awad and Issa Khalil. "Prediction of User's Web-Browsing Behavior: Application of Markov Model ".IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012 1131.
8. 30. Giorgos Kollias, Efstratios Gallopoulos, and Ananth Grama. "Surfing the Network for Ranking by Multidamping". IEEE Transactions On Knowledge And Data Engineering 2013.
9. 31. A. Deepa, and P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.
10. 32. Om Prakash, and A. Jaya. "WS-BD-Based Two-Level Match: Interesting Sequential Patterns and Bayesian Fuzzy Clustering for Predicting the Web Pages from Weblogs". The Computer Journal, Volume 63, Issue 1, 2020.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

11. Devasirvatham Weslin, Thiyagarajan Joshva Devadas. "Extricating web pages from deep web using deaima architecture" Theoretical Computer Science, Volume 931, 2022, Pages 93-103,
12. Aliaksei Miniukovich, Kathrin Figl. "The effect of prototypicality on webpage aesthetics, usability, and trustworthiness". International Journal of Human-Computer Studies, 2023.
13. Alshdaifat, E.; Alshdaifat, D.; Alsarhan, A.; Hussein, F.; El-Salhi, S.M.F.S. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. Data 2021, 6, 11.
14. Jayanti Mehra, Dr. R S Thakur. "An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1227-1232.
15. Pino-Mejías, R., Cubiles-de-la-Vega, MD., López-Coello, M., Silva-Ramírez, EL., Jiménez-Gamero, MD. (2004). Bagging Classification Models with Reduced Bootstrap. In: Fred, A., Caelli, T.M., Duin, R.P.W., Campilho, A.C., de Ridder, D. (eds) Structural, Syntactic, and Statistical Pattern Recognition. SSPR /SPR 2004. Lecture Notes in Computer Science, vol 3138. Springer, Berlin, Heidelberg.
16. Patra, S., & Ganguly, B. (2019). Improvising Singular Value Decomposition by KNN for Use in Movie Recommender Systems. Journal of Operations and Strategic Planning, 2(1), 22-34.
17. Nguyen, L.V.; Vo, Q.-T.; Nguyen, T.-H. Adaptive KNN-Based Extended Collaborative Filtering Recommendation Services. Big Data Cogn. Comput. 2023
18. G. Kanagaraj & G. Subashini (2023) Uniform distribution elephant herding optimization (UDEHO) based virtual machine consolidation for energy-efficient cloud data centres, Automatika, 64:3, 529-539.
19. Malihe Jafari, Eysa Salajegheh, Javad Salajegheh. "Elephant clan optimization: A nature-inspired metaheuristic algorithm for the optimal design of structures", Applied Soft Computing, Volume 113, Part A, 2021.
20. R. Manikandan. "A novel approach on Particle Agent Swarm Optimization (PASO) in semantic mining for web page recommender system of multimedia data: a health care perspective". Springer Science+Business Media, LLC, part of Springer Nature 10 January 2019.
21. Sumit Sharma, Dr. Pritaj Yadav Associate Professor. "Artificial Immune Based Web Page Recommendation Using Bootstrap Bagging Trained Model". Journal of Namibian Studies, 35 S1 2023.