A product recommendation system for solving the cold start problem

Shrutika Chouhan[1], Rupendra Pratap Singh Hada[2]

SAGE University, Kailod kartal, Indore-Dewas Bypass, Indore (M.P.), India (452020)[1]
SGSITS Indore (M.P.), India (452020)[2]

shruti.lchouhan@gmail.com[1]

roodra248@gmail.com[2]

Abstract – **The recommendation of product or service based on application needs different product attributes and user requirements to be analyzed. But if both the kinds of information are unavailable then such kind of issue in the recommendation system is known as the clod start problem.[1] In this presented work, the effort is made in order to understand and resolve the issue of recommendation system. Thus, a four phase recommendation system is introduced in this work. In first phase, the web usage data is preprocessed which is used for further recommendations of products. In next phase, the frequent pattern based recommendation is performed. In third phase, the recommendation is made on the basis of user current search (user click streams) and similar user behaviors available in web access logs. Additionally, the final recommendation is made on the basis of filtering the results obtained in third phase and the cost, brand and social review of the product or service. The implementation of the system is performed on JAVA technology and the results in terms of accuracy, error rate is computed. The results show the prediction is accurate and required less computational resources. Thus, in recommendation system design the model is acceptable for use and future extension of work for solving the issue of cold start problem.**

Keywords: recommendation system, cost and brand filter, cold start problem, frequent pattern mining, k-nearest neighbor.

I. INTRODUCTION

Data mining is distributed in a number of different domains of engineering and science. In a number of places, the data mining techniques are used for different applications such as text data analysis, decision making, predictions and others. [4] According to the use and applications of data mining techniques that are termed according to the applications i.e. in text analysis it is termed as text mining, in web data analysis it is termed as web mining and others. In this presented work the web mining is the key area of investigation and application design. The web mining consumes the web data in form of web access logs, web page contents, web page links and others [12]. In this presented work the web usages mining and web content mining techniques are used for designing a web recommendation system [15].

The web recommendation system is a classical domain of research and study and day by day that is improved continuously. The need of recommendation system is to suggest the end client for a service or product to be use according to their behavior and the requirements. In this presented work the issue of clod start problem is considered for study and solution design. In clod start problem the web application or the recommendation system does not have a priori information about the user habit or web navigation behavior and based on their current web usages pattern the prediction about the suitable service and product is need to be performed. Because the recommendation engine having limited information about the end user, therefore likelihood of accurate prediction is fewer. In this context a new behavior analysis based recommendation model is proposed to work for improving the accuracy of prediction with limited information [28].

II. PREVIOUS WORK

This section provides the recently made efforts to design the techniques of improving the current recommendation system using data mining algorithm. The study of different articles and research paper is included in this section. Yongzheng Zhang et al. [28] presents a outline for foreseeing a client's buy practices on web-based business sites from the client's web-based life profile. They particularly go for comprehension if the client's profile data in an informal organization (for instance Facebook) can be used to anticipate what classes of items the client will buy (for precedent eBay Electronics). The paper gives a broad investigation on how client's Facebook profile data corresponds to buys on eBay, and examines the execution of various proficiencies and learning calculations on the undertaking of procurement conduct expectation. Since an ever increasing number of individuals utilize the small-scale blogging stage Twitter to pass on their necessities and wants, it has turned into an especially fascinating medium for the undertaking of distinguishing business exercises. Potential purchasers and merchants can be reached straightforwardly in this way opening up novel points of view and financial conceivable outcomes. By identifying business expectation in tweets, this work is viewed as an initial step to unite purchasers and merchants. Bernd Hollerit et al.[29] present a programmed strategy for distinguishing business expectation in tweets where sensible exactness 57% and review 77% scores have been accomplished. Furthermore, bits of knowledge into the nature and attributes of tweets showing business aim has been given along these lines adding to see how individuals express business exercises on Twitter. .
Mi Zhang et al. [30] proposed a semi-administered gathering learning calculation. The calculation develops unique (powerless) expectation models utilizing precedents with various settings and after that utilizes the bed raining technique to permit each (frail) forecast model to gain from the other forecast models. The technique has a few recognized focal points over the standard proposal strategies for tending to the chilly begin issue. To start with, it characterizes a fine-grained setting that is more exact for demonstrating the client thing inclination. Second, the strategy can normally bolster directed learning and semi-regulated realizing, which gives an adaptable method to join the unlabeled information. The proposed calculations are assessed on two certifiable datasets. The trial results additionally demonstrate that with this strategy the suggestion precision is essentially enhanced contrasted with the standard calculations and the cool begin issue is to a great extent mitigated. Yongfeng Zhang et al. [31] proposed the Explicit Factor Model (EFM) to create logical proposals, in the meantime keep a high expectation precision. Creators first concentrate unequivocal item includes (i.e. viewpoints) and client suppositions by expression level conclusion investigation on client audits, at that point create the two proposals and dis-suggestions as indicated by the particular item highlights to the client's advantages and the concealed highlights learned. In addition, intuitional include level clarifications concerning why a thing is or isn't prescribed are produced from the model. Disconnected exploratory outcomes on a few certifiable datasets show the upsides of our system over focused gauge calculations on both rating forecast and best K proposal assignments. Online examinations demonstrate that the point by point clarifications make the proposals and dis-suggestions more persuasive on client's acquiring conduct. Most existing exploration about online trust accepts static trust relations between clients. Little work exists examining trust development in an online world. Examining on the web trust development faces interesting difficulties since usually,

accessible information is from detached perception. In this paper, sociology speculations are leveraged to build up a system that empowers the investigation of online trust development. Specifically, Jiliang Tang et al.[32] proposed a system of advancement trust, eTrust, which misuses the elements of client inclinations with regards to online item audit. Creators present specialized insights about demonstrating trust development and perform trials to indicate how the abuse of trust advancement can help enhance the execution of online applications, for example, rating and trust expectation. In view of the above literature survey it can be seen that there are many demographic based recommendation systems, which detect the users' purchase intents from their micro-blogs, based on matching the users' demographic information extracted from their public profiles and the systems are based on previous purchase records. It has been seen that in this system design the information related to end-user is required, therefore if the recommendation system does not have such information then accurate prediction is a challenging task. So, in this project web access log is used and other potentially useful features are incorporated in the system such as frequently purchased products, clusters of likes of the user's friends.

III. PROPOSED WORK

This chapter provides the detailed discussion about the proposed approach for resolving the target clod start problem in recommendation system design.[1] Therefore the understanding of the problem domain and the detailed methodology of system design is presented in this chapter.

A. Problem Understanding

Web mining is a technique where the data mining algorithms employed on web data for discovering the application centric patterns. Web data can be available in form of direct or indirect. Here the direct means the available information visible to all openly such as web contents and the indirect means some of the web data is generated on servers and not directly open for publicly such as web access logs. Web mining techniques are used in various applications such as recommendation system design, review analysis and others. In this presented work the web recommendation system is the key area of study [13]. Traditionally, the recommendation systems are used with e-commerce web applications to guide or suggest the appropriate service and product to the end client. But now in these days in various applications such as banking, web browsing and other domains the recommendation engines are used. The recommendation system consumes the users' previous information or history of behavior or the frequent patterns and other relevant information that helps to understand the requirements of end user or frequency of buying the products or services. All these factors are used for predicting the products for the end client but when the information about the end client or product is not available then prediction is not an easy task. This kind of problem in recommendation system design is known as the clod start problem. In cold start problem there are not a single information is available that help to make a predictive data model. Therefore, some new strategy is required that help to understand the activity of user and the product to make suitable suggestions to the end user. This section provides the understanding about the core problem which is needed to deal with the solution and the next section provides the details about the proposed solution methodology and their functional aspects.

B. Proposed Methodology

The proposed recommendation engine design for solving the problem of cloud start problem is defined in four main modules. The modules and their outcomes are discussed in this section in detail.

a. Web Usage Data Analysis

The main aim of this phase of data analysis is to make the suitable quality of data, for utilizing with the recommendation model.

In this phase the web usages data is consumed, therefore the web access log is accepted as input to the system. The web access log is generated at the web server end where the data or application is parked. The web server maintains the entry on a log file for each request and response to the application. The **Web access** log file contains a number of different attributes such as user IP address, time stamp of user request and response, protocol information, web methods, resource name and others.[25] All the data available in web access log is not utilized with the recommendation system design. Therefore, the **preprocessing** techniques are adopted to remove or clean the unwanted data from the web access log. After preprocessing of data, the target attributes such as user IP address, resource name, time stamp are preserved in web log and the remaining information is removed from the file. The remaining information is transformed into a relational data table for utilizing with the further processes [16].

a. Primary Recommendation

The given process model in figure 2.2 demonstrate the first stage recommendation based on the initial activity of user.

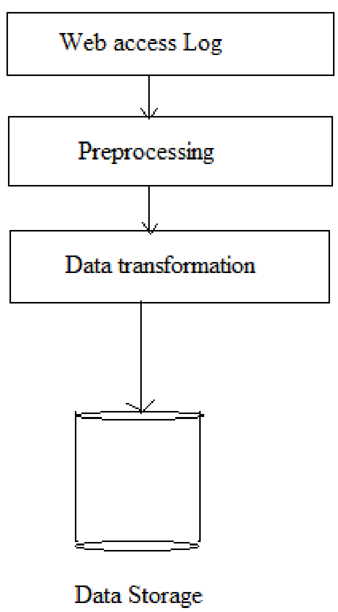

Figure 2.1 web usage analysis

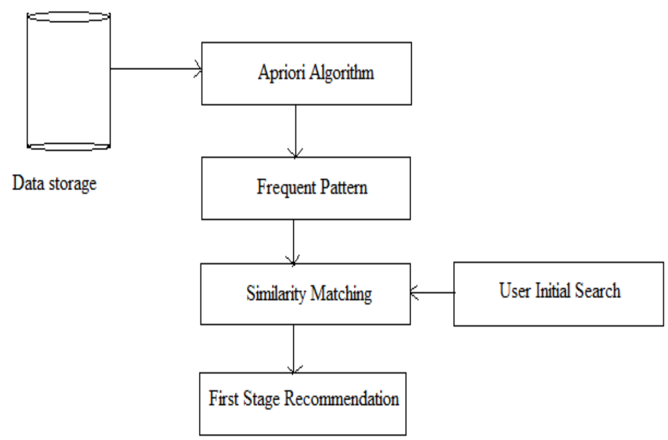

Figure 2.2 primary recommendation

The preprocessed information which is stored in database is utilized in this phase. First the data is applied on **apriori algorithm**. The apriori algorithm performs **frequent pattern** mining on the data. The frequent patterns are the subset of the entire set of data which is selected on the basis of frequently available attributes. After computing the frequent patterns, the user’s first activity or initial search product is considered. According to the user’s search product and the available frequent set the similar set of data pattern is concluded as the first stage prediction. The first stage recommendation is just based on the user search data and available frequent item which is considered for recommendation. After first stage of prediction the search space is reduced in next phase of recommendation.

b. Click Stream Based Recommendation

First stage of prediction provides a list of products that are relevant to the user search product and frequently purchased by the different users. Therefore, it is a significant large list of products. In this phase the user behavior and requirements are evaluated based on the user click stream observation.

In first stage of prediction user can select some products according to their interest. The interested products are recognized here as the user click events in the product suggested in initial prediction. The selected product by the user and web access data is used here with the KNN (k-nearest neighbor) algorithm for finding the similar kinds of pattern in web log history[30]. Here the user product selection behavior can be matched with the other user's product selection behavior. Therefore, the similar behavior based products are suggested in this phase as the second stage of recommendation.

a. Final Stage of Recommendation

The aim of this module is to precise the prediction by using collected information from the last two stages of predictions.

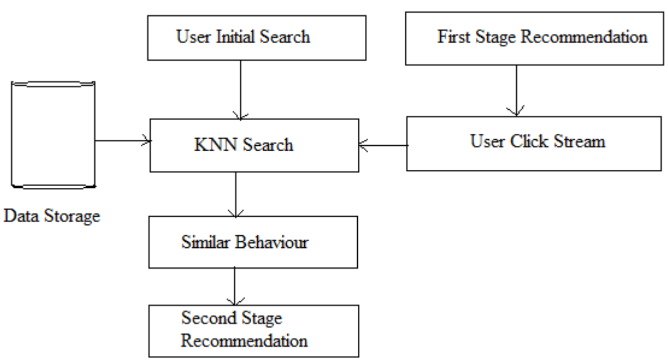

Figure 2.3 second stage of prediction

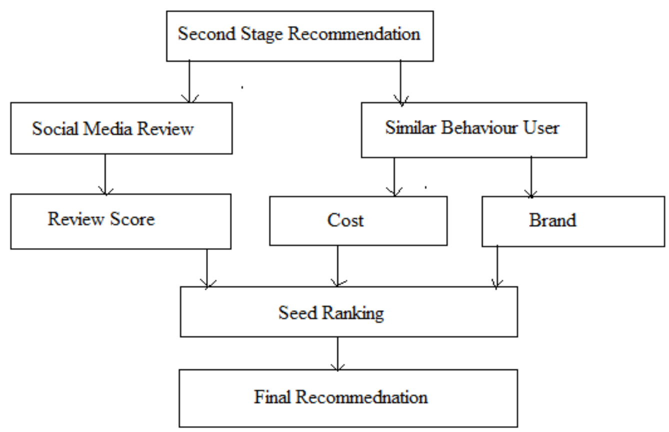

Figure 2.4 final stage of recommendation

The figure 2.4 shows the narrow down the search space for making precise prediction for the end user. Basically the second stage of prediction contains fewer amounts of data than the primary prediction. But the second stage of prediction also contains a significant amount of data to predict the appropriate product to end user. Therefore based on computed additional information the accurate prediction tried to make. The list of products are collected which is obtained in previous stage and three factors are concluded. First, the price range of a product which is purchased by the different users and having similar behavior of product selection, secondly, the brands of products, which is accepted by the similar behavior users. Finally the reviews of products are considered which is extracted from the social media[27]. All the computed factors are combined to rank the list of recommended products in second stage. Additionally, based on the ranking of these three factors products are recommended to user.

C. Proposed Algorithm

The above given four modules of web recommendation is explained in this section using the simple working steps.

Input: web access log W, Initial Search query Q

Output: recommended products RP

Process:

1. $R_n = readLogFile(W)$
2. $P_n = preProcessData(R_n)$
3. $F_n = Apriori.ComputeFrequentSet(P_n)$
4. $for(i = 1; i \leq n; i++)$
    a. $if\left(F_i.contains(Q)\right)$
        i. $RP.Add(F_i)$
    b. $end\ if$
5. End for
6. Return RP

Table 2.1 preprocessing of data first stage of prediction

Input: first stage of prediction RP, pre-processed data P_n

Output : second stage of recommendation SR

Process:

1. $for(i = 1; i \leq RP.length; i++)$
    a. $if(RP_i.Clicked)$
        i. $CL.Add(RP_i)$
    b. $end\ if$
2. End for
3. $S = KNN.Search(CL, P_n)$
4. $SR = createListProduct(S)$
5. Return SR

Table 2.2 second stage of prediction

Input : second stage of prediction SR

Output : Final recommendation R

Process:

1. $for(i = 1; i \leq SR.length; i++)$
    a. $C = GetCost(SR_i)$
    b. $B = getBrand(SR_I)$
    c. $RE = getSocialReview(SR_i)$
2. $end\ for$
3. $Rank = C * B * RE$
4. $R = Sort(SR, Rank)$
5. Return R

Table 2.3 Final Stage of prediction

IV. RESULT ANALYSIS

This chapter provides the performance evaluation of the proposed recommendation model. Therefore the computed performance parameters are reported in this section.

A. Accuracy

Accuracy is the measurement of correctness of prediction in data mining and machine learning. That is the ratio of correctly recognized sample and total samples to be recognized. Using this formula we compute the accuracy of model.

$$accuracy = \frac{correctly\ identifyied\ samples}{total\ samples} X100$$

The accuracy of the proposed recommendation model for cold start problem is described in figure 3.1 and table 3.1. In the diagram experiments conducted with the system is listed in X axis and corresponding obtained accuracy of experiments in form of percentage accuracy is included in Y axis. According to the experimental observations the accuracy of the proposed model is increasing with the amount of data available in different experiments.

Experiment No	Accuracy %
1	92.8
2	91.3
3	93.7
4	94.8
5	95.2
6	94.1
7	95.6

Table 3.1 accuracy

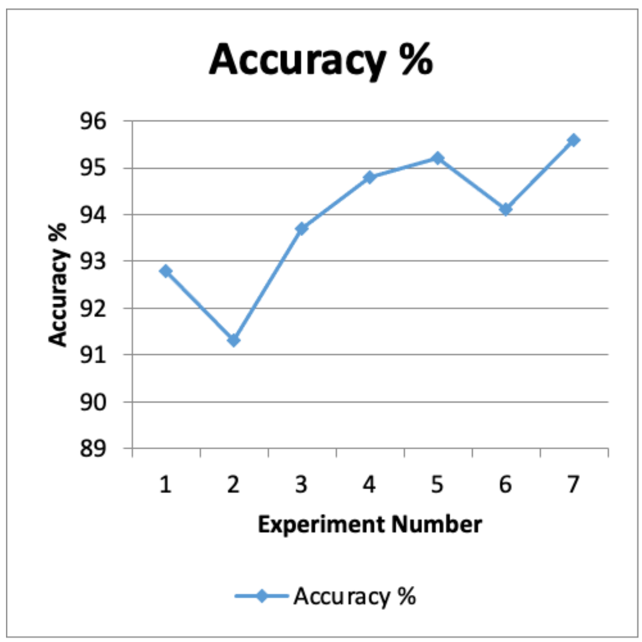

Figure 3.1 accuracy

B. Error Rate

Error rate is the measurement of incorrectly predicted outcomes among the total samples produced to be predicted. The inaccuracy of the method can be computed through the following formula.

$$error\ rate = \frac{misclassified\ samples}{total\ samples} X100$$

X axis includes the experiments conducted with the system and Y axis includes the respective error rate percentage of the predictive recommendation system.

Experiment No	Error rate %
1	7.2
2	8.7
3	6.3
4	5.2
5	4.8
6	5.9
7	4.4

Table 3.2 error rate

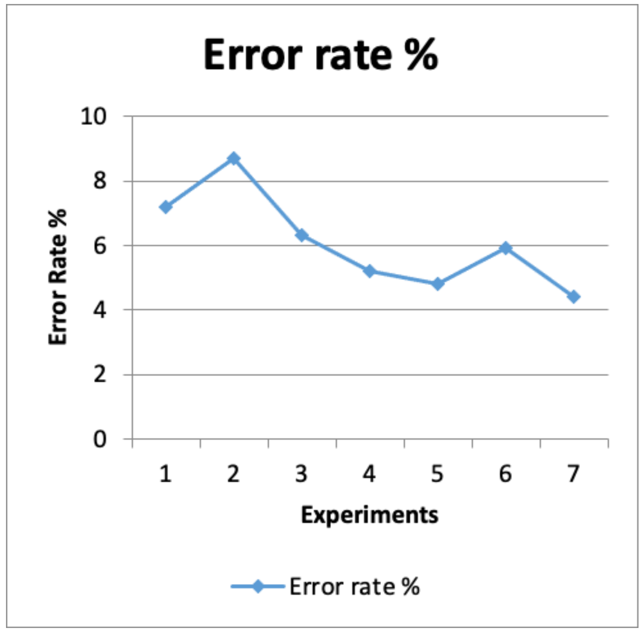

Figure 3.2 error rate

According to the line graphs the error rate of the model is remains consistent and not much varying in different experiments. Therefore, the proposed system is acceptable for solving the issue of cold start problem.

C. Time Consumption

The amount of time required to process the data for performing the prediction using the developed model is termed here as the time consumption. The time requirements of the proposed recommendation system are computed using the following formula:

$$time\ consumed = algorithm\ end\ time - start\ time$$

The time consumption of the proposed recommendation model for predicting appropriate product to end client is demonstrated in figure 3.3 and table 3.3. In this diagram the X axis contains the experiments performed with the data model and the Y axis contains the time consumed in experiments. During the experiments the time is measured in terms of milliseconds. The results show the time consumed in different experiments are not varying in large amount and remains consistent in different experiments. Thus, the system is acceptable for real world use of application.

Experiment No	Time consumed (MS)
1	188
2	198
3	193
4	201
5	210
6	183
7	191

Table 3.3 Time Consumption

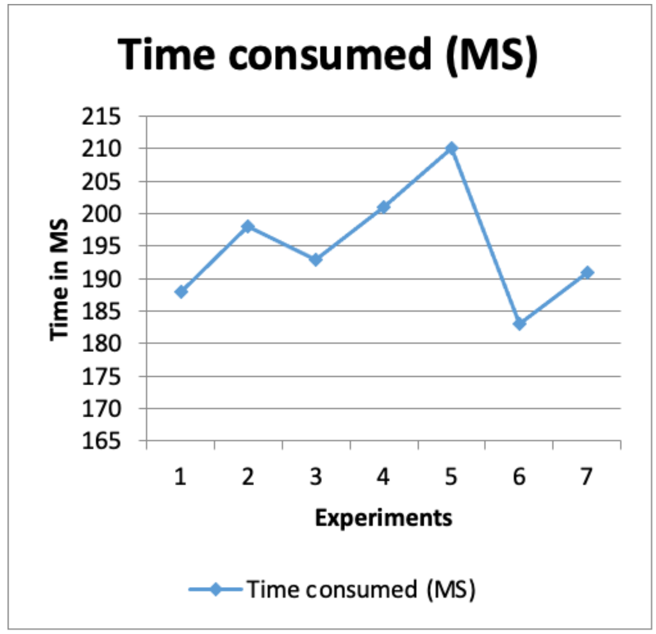

Figure 3.3 Time consumption

D. Memory Usages

The memory usages are measured in form of process takes the amount of main memory space for performing the computations. The java technology computes the memory usages in the following form.

$$memory\ usage = total\ allocated\ memory - free\ memory$$

The memory usages of the system are computed here in terms of megabytes (MB). The computed memory usage is described in figure 3.4 and table 3.4. In this diagram the X axis contains the experiments performed and Y axis shows the consumed main memory of the system. The memory usages of the system is depending on the amount of data to be process. Therefore, it is acceptable for predicting the products to the end user with low memory resource consumption.

Experiments	Memory usages (MB)
1	110
2	118
3	121
4	106
5	119
6	122
7	113

Table 3.4 memory usage

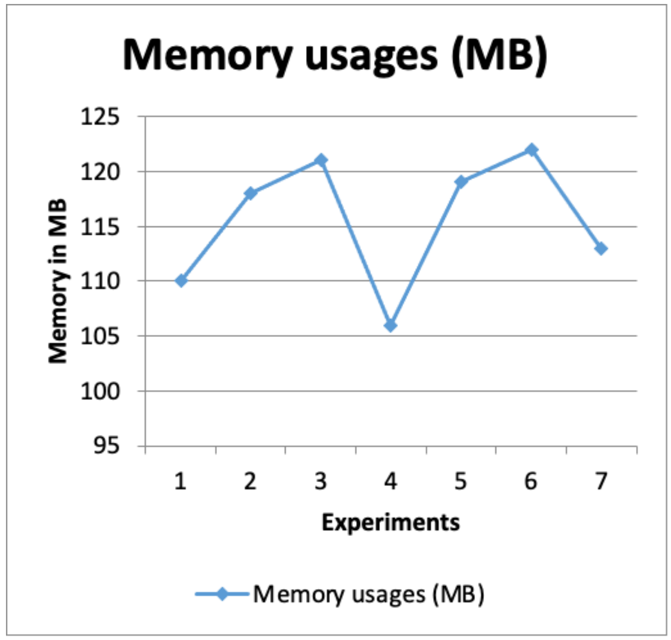

Figure 3.4 memory usage

V. CONCLUSION

This chapter provides the conclusion of the work performed for designing the solution for the clod start problem in recommendation system design. In addition of that the future work is also suggested in this chapter.

A. Conclusion

The cold start is a critical problem in recommendation system. According to the problem statement the recommendation system not contain any information about the product and the end client behavior using which the prediction can be performed [1]. In this context need some additional efforts to extract the information about product and the user behavior for making accurate recommendations. In this context a data model is proposed that works in four main phases. First of all, the web usage data is analyzed and the preprocessing is performed [15]. This data help to understand the other user's behavior and finding the similar behavior users as the current user is. In next, the apriori algorithm is applied for frequent pattern mining. That results the frequently purchased product according to the current user search. That contains a significant amount of products. Thus, to refine the results the similar user behaviors of another user is obtained using the KNN search and user click stream analysis. Using the user likely behavior, the recommendations are made in second time. Finally, on more precise recommendation is developed. in this prediction the product information and their reviews are analyzed additionally the product brands and cost filter is applied for reducing the amount of data in second recommendation. That step based prediction help to make more accurate prediction or recommendation without any prior knowledge about the user behavior and product.

The implementation of the proposed solution is conducted using JAVA technology and the use of JAVA IDE. Additionally, for performance visualization, the MySQL Server is used as backend. After implementation of the required system the experiments are performed and based on experiments the results are summarized in table 4.1.

S. No.	Parameters	Observations
1	Accuracy	The accuracy is higher of the presented model and it is found between 91-96%.
2	Error rate	The error rate of the system is acceptable due to less amount of error probability, it is varying between 4-8%
3	Memory	The memory usages of the system depend on the amount of data to be process and it is varying between 110-139 MB
4	Time	Time remains consistent in different number of experiments and varying between 110-130 MS

Table 4.1 performance summary

The results described in table 4.1 demonstrate the performance of the proposed technique is acceptable for utilizing with the different applications of recommendation system design. The future extension of the work is described in next section.

B. Future Work

The proposed work is to solve the problem of cold start issue in recommendation system design. Therefore, a model is implemented and evaluated based on the experimental analysis the proposed technique found acceptable. In near future the following work is proposed for work.

1. Current system considers limited features for final recommendations of product thus in near future more attributes are explored for optimizing the performance of final prediction
2. The proposed system implemented using the ranking technique therefore in near future some probability based model is considered for enhancing the system performance

REFERENCES

[1] Zhao, W.X., Li, S., He, Y., Chang, E.Y., Wen, J.R. and Li, X., 2016.

Connecting social media to e-commerce: Cold-start product recommendation using micro blogging information. IEEE Transactions on Knowledge and Data Engineering, 28(5), pp.1147-1159.

[2] “An Introduction to Data Mining: Discovering hidden value in your data warehouse”, available online at:

http://www.thearling.com/text/dmwhite/dmwhite.htm

[3] What Is Data Mining?, available online at:

https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#DMCON002

[4] Smita and Priti Sharma, “Use of Data Mining in Various Field: A Survey Paper”, IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 3, (May-Jun. 2014), pp. 18-21

[5] Dhanalakshmi. D and Dr. J. Komala Lakshmi, “ASurvey on Data Mining Research Trends”, A Survey on Data Mining Research Trends, Volume 3, Issue 10 October, 2014 Page No. 8911-8919

[6] Samiddha Mukherjee Ravi Shaw and Nilanjan Haldar, “A Survey of Data Mining Applications and Techniques”, International Journal of Computer Science and Information Technologies, Vol. 6 (5), 2015, pp. 4663-4666

[7] Berson, Alex, and Stephen J. Smith, Building data mining applications for CRM, McGraw-Hill, Inc., 2002.

[8] Pal, Jiban K. ”Usefulness and applications of data mining in extracting information from different perspectives”, (2011).

[9] Sumathi, Sai, and S. N. Sivanandam. Introduction to data mining and its applications, Volume 29. Springer, 2006.

[10] Srivastava, T., Prasanna Desikan, and Vipin Kumar, ”Web mining–concepts, applications and research directions.” In Foundations and advances in data mining, pp. 275-307, Springer, Berlin, Heidelberg, 2005.

[11] S. Vidya and K. Banumathy, “Web Mining- Concepts and Application”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4), 2015, pp. 3266-3268

[12] “Web Mining”, available online at: https://www.techopedia.com/definition/15634/web-mining

[13] Kosala, Raymond, and Hendrik Blockeel. ”Web mining research: Asurvey.”ACM Sigkdd Explorations Newsletter 2, no. 1 (2000): 1-15.

[14] Pal, Sankar K., Varun Talwar, and Pabitra Mitra. ”Web mining in soft computing Framework: relevance, state of the art and future directions.” IEEE Transactions on Neural Networks 13, no. 5 (2002): 1163-1177.

[15] Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. ”Web usage mining: Discovery and applications of usage patterns from web data.” ACM SIGKDD Explorations Newsletter 1, no. 2 (2000), pp. 12-23.

[16] P. Nithya, and P. Sumathi, “A Survey on Web Usage Mining: Theory and

Applications", International Journal of Computer Technology & Applications, Volume 3, Issue pp. 1625-1629

[17] Aparna Ranade, Abhijit R. Joshi, Ph. D," Techniques for Understanding User Usage Behavior on the Internet" International Journal of Computer Applications (IJCA), Volume 92 – No.7, April 2014

[18] Bhalla, Karan, and Deepak Prasad, Data preparation and pattern discovery for web usage mining, Dissertation 2007.

[19] C. Shahabi and F. b. Kashani, "Efficient and Anonymous Web Usage Mining for Web Personalization", Department of Computer Science Integrated Media Systems Center, University of Southern California, Los Angeles, California 90089-2561, USA.

[20] "Chapter 3: Process of Web Usage Mining", available online at: http://shodhganga.inflibnet.ac.in/bitstream/10603/44185/11/11_chapter%203.pdf

[21] Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: ASurvey", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013

[22] Madria, Sanjay Kumar, Sourav S. Bhowmick, W-K. Ng, and Ee-Peng Lim. "Research issues in web data mining." In International Conference on Data Warehousing and Knowledge Discovery, pp. 303-312. Springer, Berlin, Heidelberg,1999.

[23] Singh B, Singh HK. Web Data Mining research: A survey. In: 2010 IEEE International Conference on Computational Intelligence and Computing Research. IEEE; 2010

[24] Taghipour N, Kardan A, Ghidary SS. Usage-based web recommendations. In: Proceedings of the 2007ACM Conference on Recommender Systems - RecSys '07. New York, New York, USA: ACM.

[25] Srivastava, Jaideep, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. "Web usage mining: Discovery and applications of usage patterns from web data." Acm Sigkdd Explorations Newsletter 1, no. 2 (2000): 12-23.

[26] "SocialMedia", availableonlineat: http://www.investopedia.com/terms/s/social-media.asp, Investopedia

[27] Daniel Nations, “What Is Social Media? Explaining the Big Trend”, available

onlineat: https://www.lifewire.com/what-is-social-media-explaining-the-big-trend-3486616

[28] Zhang, Yongzheng, and Marco Pennacchiotti. ”Predicting purchase behaviors from social media.” In Proceedings of the 22nd international conference on World Wide

Web, pp. 1521-1532. ACM, 2013.

[29] Hollerit, Bernd, Mark Kröll, and Markus Strohmaier. ”Towards linking buyers and sellers: detecting commercial intent on twitter.” In Proceedings of the 22nd International Conference on World Wide Web, pp. 629-632. ACM, 2013.

[30] Zhang, Mi, JieTang, XuchenZhang, andXiangyangXue.”Addressing cold start recommender systems: A semi-supervised co-training algorithm”, In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 73-82. ACM, 2014.

[31] Zhang, Yongfeng, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. ”Explicit factor models for explainable recommendation based on phrase-level sentiment analysis.” In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 83-92. ACM, 2014.

[32] Tang, Jiliang, Huiji Gao, Huan Liu, and Atish Das Sarma,”eTrust: Understanding trust evolution in an online world”, In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 253-261.ACM, 2012.