Automated web usage data mining and proposal system using K-Nearest Neighbour (KNN) classification method

Dr.Ayesha Taranum[*1],Shwetha K S[*2],Yathiraj G R[*3],T D Roopamala[*4]

[*1]*Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India-570002*

[*2]*Department of Computer Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, Karnataka, India-560078*

[*3]*Department of Computer Science and Engineering, Coorg Institute of Technology, Ponnampet, Karnataka, India-571216*

[*4]*Department of Computer Science and Engineering, Sri Jaya Chamarajendra college of Engineering, Mysuru, Karnataka, India-570006*

Abstract: The Internet has become a non-existent data offering. It provides many of the required user details. So much of the information provided offers a few options for the user to choose from. Choosing the right product or information has become a time-consuming and tedious task. To make the information selection process easier, a recommendation system has been developed. This may suggest an Internet site or an online site or product for the user to decide on. This program is trained to promote user experience. The Dailies of India represent the database as well from that user is suggested to their interested browsing prospect that meets the necessity of the explicit user at a specific time.

Keywords: Automated, Data Mining, K-Nearest Neighbour, Recommendation system, Web usage Data Mining

1 INTRODUCTION

Data mining extracting information from a large number of viewing data sets, which you can find unexpected relationships with pattern hidden in data, summarize data on new ways to do it understandable and useful for data users. The web data mining use data mining use Automatic process of detection and extraction useful information from a particular website. In over the years, there has been an explosive growth in the number of explorers in the area of web mining, especially for web use mines. Today a number of websites were developed intentionally to read dailies stories on the line across the Globe, however lack of ways to identify customer movement pattern as well cannot provide a real-time response the client needs, therefore, access to relevant information it takes time makes a profit of in line services will be reduced. The program is by be able to identify users / clients of roaming behaviour with works with detailed user streaming data, so that recommend a set of unique features that

satisfies the file need for active user Real Time, online base. How Neighbourhood Dividing K Nearest used online and in real time to exploit web usage data mining process to identify clients / visitors click on stream data that you are comparing to a specific user group and recommend red tail browsing option meets the specific user requirement provided time.

1.1 Web mining

Web mining is information integration collected traditional methods of data mining and strategies wit h information collected on World Wide Web. Web mining is used for understanding Customer conduct; check the performance of a specific website. Web Mining lets you look patterns data through

- Content mining
- Structure mining
- Usage mining

Content mining is employed to look at the info collected by search engines and web spiders. Structure mining is used to look at data associated with the structure of the particular website. Website and Usage mining are used to examine data associated with a specific user's browser also because the data gathered by forms the user may have submitted during web transaction.

1.2 Web usage mining

Web procedure mining is a grouping in web mining. This web mining permits for collecting Web admittance in sequence for Web pages. This data can present the paths primary to accessed Web pages. This in sequence is often composed robotically into admission logs via the Web server. CGI scripts present functional in sequence such as referrer logs, user contribution information and assessment logs. This statistics is essential to the generally use of data mining for firms and also their internet/ intranet applications and information access.

1.3 Data Classification

Characterization is an information mining (AI) procedure that is familiar with foresees bunch membership for information examples. For example, you might want to utilize characterization to estimate whether the climate on a specific day it will be "radiant", "blustery" or "shady". Generally utilized grouping procedures incorporate choice trees and neural networks.

1.3.1 K nearest Neighbours calculation

In design disclosure, the k Nearest Neighbors Calculation (or k NN in short) is a non parametric technique utilized for characterization. The info Comprises of k closest preparing models in the element space. In k NN arrangement, the yield is class participation. An article is arranged dependent on the most extreme vote of its neighbors, with the article being relegated to the class that is generally regular among its k closest neighbors k is a positive whole number, ordinarily little). In the event that k 1, the article is allotted to the class of that solitary closest neighbor. In k NN relapse, the yield is the property

estimation for the item. This worth is normal estimations of its k closest neighbors. K NN is a sort of occurrence based learning, or apathetic learning, in light of the fact that the capacity is only approximated locally and all calculation

is postponed until grouping is finished. The k NN calculation is one of the least difficult of all machine learning calculations. For grouping, it is significant to relegate weight to the commitment of the neighbors, with the goal that the closer neighbors can offer more to the normal than the far off ones. For instance, a typical weighting plot comprises in giving each neighbor a load of 1/d, where d addresses the distance to the neighbor. The neighbors are taken from an assortment of items for which the class (for k NN arrangement) or the Object property estimation (for k NN relapse) is known. This could be considered as the preparation set for the algorithm. A deformity of the k NN calculation is that it is delicate to nearby design of the information. The Calculation has nothing to do with and isn't to be mistaken for k methods, which is another mainstream AI method.

1.3.2 Decision Tree Classification

In developing a choice tree, apply both the gini index(g) and entropy esteem (e I) as the parting files, the model is explored different avenues regarding a gave set of qualities, and various arrangements of results were acquired for both The choice tree sleuth unique has the limitation that all the preparation tuples is should be in principle memory, thus, on account of horrendously huge information; this may prompt shortcoming of choice tree develop due to Trading of the preparation tuples all through the principle memory and ca che memory. Accordingly a more adaptable technique like the KNN strategy are fit for dealing with preparing information that are excessively huge to fit in memory is required.

1.3.3 Bayesian Classifier Model

Choice principle and Bayesian network [9], are classification tree and backing vector machine strategies that were utilized to show mishaps and episodes in two firms to detect the reason for mishap. Information is gathered through meet and is at that point demonstrated. The trial result was at that point contrasted and statistics procedures, which delighted that the Bayesian organization and the other techniques applied, are better than the measurements strategy. In principle, Bayesian classifier is asserted to acts least blunder rate like looked at

With any remaining classifier techniques yet practically speaking this isn't the situation, owed to inaccuracy in presumptions made for its utilization, like class

Restrictive independency and the inadequacy of accessible likelihood of information which is generally not the situation when utilizing K NN strategy.

2 RELATED WORKS

M.F. Federico, L.L.Pier [1] investigated the region of Web Mining which manages the extraction of intriguing information

from logging data created by Web workers. In this paper they present a review of the new improvements here that is accepting expanding consideration from the Data Mining people group. Web use mining is utilized to find intriguing client route designs and can be applied to numerous true issues, for example, improving Web locales/pages, making extra point or on the other hand item proposals, client/client conduct

Contemplates, and so forth this article gives a review and examination of current Web utilization mining frameworks and innovations. B.Lalithadevi, A. Mary Ida,W.Ancy Breen[3] explored on Web utilization mining framework which performs five significant assignments: I) information gathering, ii) information planning, iii) route design revelation, iv) design investigation and representation, and v) design applications. Each errand is clarified in detail and its related advancements are presented. A rundown of major research frameworks and ventures concerning Web utilization mining is additionally introduced, and an outline of Web use mining is given in the last area Quig Yang, Hairing Henery Zhang[4] explored on Caching is a notable technique for improving the performance of Web based frameworks. The core of a reserving framework is its page substitution strategy, which chooses the pages to be supplanted in a reserve when a solicitation shows up. In this paper, they present a Web log mining strategy for storing Web items and utilize this algorithm to upgrade the presentation of Web storing frameworks. In our methodology, we build up a n gram based forecast calculation that can anticipate future Web demands. The forecast model is at that point used to broaden the notable GDSF reserving strategy. Site page mining is the utilization of information mining strategy to naturally find and remove helpful data from a specific site. It is computationally costly to track down the k closest neighbors when the dataset is extremely huge. KNN can have helpless run time performance when the preparation set is huge. It is exceptionally touchy to unessential or excess highlights since all highlights add to the comparability and in this way to the grouping. The truth of the matter is that most existing works need scalability

What's more, capacity when managing on line, Real Time search driven sites. So framework is proposed to suggest a site page or site or item for the client to pick. This framework will be prepared to suggest the asset for the client. The classification is finished by improved KNN (K Nearest Neighbor) characterization technique has been prepared to be utilized on line and in Real Time to distinguish customers/guests click stream information, coordinating with it to a specific client bunch and suggest a custom fitted perusing choice that address the issue of the specific client at a specific time.

3 PROBLEM FORMULATIONS

The serious issue of numerous online sites is that the introduction of numerous decisions to the customer at a time. This ordinarily prompts demanding and time overwhelming task in tracking down the correct item or data on the site. Web use mining is the utilization of information mining strategy used to automatically find

and concentrate valuable data from a specific pursuit.

3.1 Proposed Method

Though electronic suggestion frameworks are normal, there is as yet a few disadvantage territories calling for arrangements. Actually the majority of the

Existing works need adaptability and ability when working with on line, and Real Time driven web look. S o framework is proposed to recommend a web Page or site or item for the client to pick. This framework is prepared to suggest the suitable asset for the client. Here per user site is picked to suggest the client dependent on their necessities. Dailies of India shape the dataset and from which client are prescribed to their intrigued day by days and news. This characterization is finished utilizing improved KNN (K Closest Neighbor) characterization technique that can be prepared to be utilized on line and in Real Time to perceive visitors click stream information, coordinating with it to a Explicit client bunch and suggest are did perusing alternative that will address the issues of the explicit client at a specific time.

3.2 System Overview

The dataset utilized in this framework is the client access database for a particular timeframe, that was separated, pre handled and gathered into significant meetings and information shop was created. The Improved K Nearest Neighbour order method was utilized to examine the uniform asset finder information of the clients' location data set. The client per user site information is put away in the information store made. The outline of the framework is

Appeared in the Figure 1

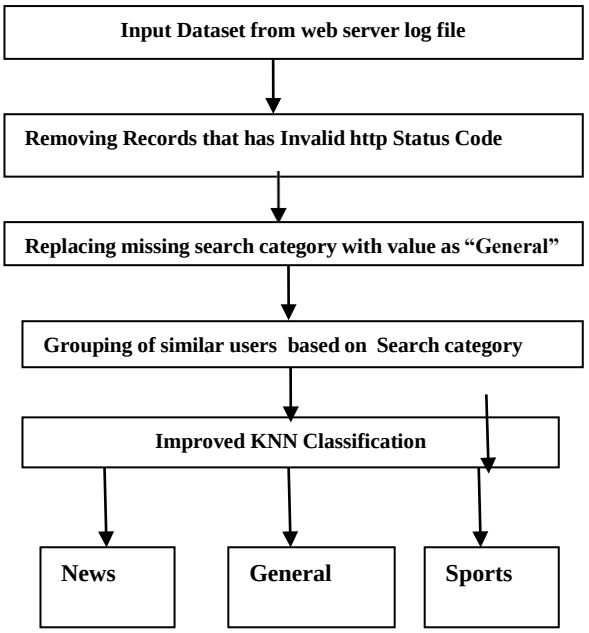

Figure 1: System Overview

3.2.1 Data Pre-processing

Information Acquisition alludes to the assortment of information for mining reason, and this is normally the primary assignment in information mining application. For web use mining information is gathered from web server, intermediary worker and web Customer. The web worker source was picked for the reality that it is the most extravagant and most normal information source, all the more thus; it is feasible to gather huge measure of data from the log documents and data sets they address. The original data set document extricated, not all the information is pertinent for web use

information mining, and we exclusively need the passages that contain significant data. The first record is generally comprised of text records that have immense volume of information concerning questions made to the web worker in which a large portion of the occurrences contains unimportant, deficient and misdirecting information for mining reason. Information purifying is the stage wherein wrong/loud passages are

Wiped out from the log record. In information pre processing step information having improper http status code is taken out and missing pursuit class is supplanted with General.

3.2.2 Grouping Similar Users

In distinguishing proof of comparative clients, search classification is considered as classifier trait. When tally value of specific class is more noteworthy than one and it matches to any of the characterized class, at that point it is assembled under one class. In the event that there is just a single client for specific inquiry then that record will be Wipe-out.

3.2.3 Improved KNN Algorithm

Revile of dimensionality that is caused because of Euclidean distance is wiped out by finding weight of properties in improved KNN techniques. The conventional KNN text arrangement has three deformities. First is that it has incredible computational intricacy. When utilizing traditional KNN order, to discover the K closest neighbor tests for the given test test, it is compulsory be compute the similitude between all the preparation tests, as the measurements of the content vector is for the most part high, so it has extraordinary calculation on intricacy in this cycle which makes the productivity of text order to be low. For the most part there are 3 strategies to diminish the intricacy of KNN calculation: limiting the dimensionality of vector text [4]; utilizing informational collection of little size; utilizing devil meandered calculation which can speed up to spot out the K closest neighbor tests. Second is relying upon preparing set KNN calculation doesn't utilize extra information to point up the arrangement rules, yet the classifiers are created by the self training tests, this makes the calculation to rely upon preparing set exorbitantly, for model, it is important to re figure when there is a little change on preparing set. Last is there is no weight distinction between tests. The conventional KNN calculation treats all preparation tests similarly, and there is no variety between the examples, accordingly it don't match the real wonder when the examples
have lopsided circulation.

4 PROBLEM SOLUTIONS

The accompanying advances are carried out for the Propos d System

4.1 Data Pre processing

Information: Dataset is gathered from Web server log document

Yield: Pre handled Data set.

Steps

- Upload the dataset into worker.
- Check the missing qualities.

- If the missing worth happened for Search Class at that point supplant the worth as "general".

- Eliminate the records that have http status code more prominent than 400 and under 200

4.2 Grouping Similar Users

Info: Pre handled informational index

Output: bunched result informational index Steps

- The pre prepared dataset is taken as information.
- Search class groups will be shaped in light of check.
- If the check esteem is one, at that point the record will be killed

4.3 Steps for Improving KNN

- Determine Para meter K, where K is the number of closest neighbors
- Calculate the distance between the inquiry what not the preparation models
- Sort the distance and decide closest neighbor dependent on the kth least distance
- Gather the class Y of the closest neighbours
- Use straightforward greater part of the class of closest neighbors as the Prediction estimation of the inquiry distance [6] is determined as in condition (1)

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(\omega_i^2)(a_i(x)-a_i(y))^2} \quad \text{—— (1)}$$

where

ω-weight of the attribute

5 EXPERIMENTATION RESULTS

Information: Input Dataset from Web worker log records Output: Grouping of comparable clients dependent on search Precision considers all recovered reports, however it tends to be assessed at a given cut-off position, bookkeeping just the top most outcomes returned by the framework. This is called Precision at n. Exactness is the opportunity that (Randomly Selected) the recovered archive is important. Review

Performance Measures	KNN	Improved KNN
Precision	**50%**	**75%**
Recall	**17%**	**25%**
Accuracy	**90%**	**96%**

in data recovery is the small amount of archives that are pertinent to the inquiry that are effectively recovered.

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Number of relevant item in collection}}$$

Review is the opportunity that a (Randomly Selected) important record is recovered in a pursuit. Exactness is definitely not a dependable measurement for assessing the genuine presentation of

the classifier when the quantity of tests in various classes differs incredibly in light of the fact that it might yield misdirecting results.

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Number of item retrieved}}$$

$$\text{Accuracy} = \frac{TP+TN}{TN+FP+FN+TP}$$

Where

TN is the number of accurate negative cases

FP is the number of false positive cases

FN is the number of false negative cases

TP is the number of accurate positive cases

Genuine Positive (TP): These allude to the positive tuples that were effectively marked by the classifier.

Genuine Negative (TN): These are the negative tuples that were accurately named by the classifier

False Positive (FP): These are the negative tuples that were mislabelled as sure by the classifier

False Negative (FN): These allude to the positive tuples that were mislabelled as negative.

Table 1 portrays the exhibition investigation table of KNN and Improved KNN dependent on Precision, Recall and Accuracy.

Table 1: presentation examination of KNN and better KNN

Figure 2 depicts the presentation examination chart for comparing KNN and enhanced KNN based on accuracy, recollect and correctness.

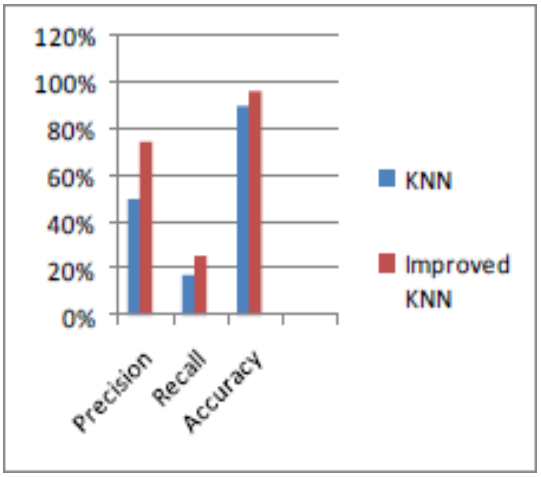

Figure 2: Performance analysis graph for comparing KNN and Improved KNN

6 CONCLUSIONS

The proposed framework gives a premise to programmed Real-Time suggestion framework. The framework performs characterization of clients on the re-enacted dynamic meetings separated from testing meetings by gathering dynamic clients' snap stream [6] and matches this with comparable class in the information store, to create a bunch of proposals to the customer in a Real-Time premise utilizing improved k-NN arrangement.

The System can be Future upgraded by

- Efficient highlights can be extraction technique can be utilized to improve order strategy.
- System can be improved to have various kinds of log records.

• Optimized grouping strategies can be utilized to improve the order precision.

References

[1] M.F. Federico, L.L. Pier, Mining interesting knowledge from weblog: a survey, J. Data Knowledge Eng. 53(2005) (2005) 225–241.

[2]S.Kaviarasan,K.Hemapriya,K.Gopinath, Semantic Web Usage Mining Techniques for Predicting Users' Navigation Requests, International Journal of Innovative Research in Computer Science and Communication Engineering,Vol. 3,Issue 5,[ISSN:2320-9801],2015.

[3]B.Lalithadevi, A. Mary Ida,W.Ancy Breen, A New Approach for Improving World Wide Web Techniques in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 3,Issue 1,[ISSN:2277 128X],2013

[4] Paul, N. Kenta, Better Prediction of Protein Cellular Localization Sites with the K-Nearest Neighbor Classifier, ISMB-97, Proceeding of America Association for Artificial Intelligence, USA, 1997, pp. 147–152.

[5] Qing Yang, Haining Henery Zhang, Web log mining for predictive web caching, IEEE Transactions on Knowledge and Data Engineering, Vol. 15 NO. 4. [104 1-4347/03/$17.00],2003.

[6] Richard Jensen, Chris Cornelia, A fuzzy K-Nearest Neighbour Classification. Springer Berlin Heidelberg, Vol-5306,[ISSN:0302-9743], 2008.

[7] Resul, T. Ibrahim, Creating meaningful data from web log for improving the impressiveness of a web site by using path analysis method, Journal of expert system with applications 36 (2008) (2008)6635–6644, http://dx.doi.org/10.1016/j.eswa.2008.08.067.

[8] Srivastava, R. Cooley, B. Mobasher, Data preparation for mining World WideWeb browsing patterns, J Knowledge Inform. Syst. 1 (1) (1999) 1–27.

[9]S. Amartya, K.D. Kundan, Application of Data mining Techniques in Bioinformatics, B.Tech Computer Science Engineering thesis, National Institute of Technology, (Deemed University), Rourkela, 2007.

[10]L. Shu-Hsien, C. Pei-Hui, H. Pei-Yuan, Data mining techniques and applications-A decade review from 2000 to 2011, Journal of expert system with applications 39 (2012)(2012) 11303–11311,http://dx.doi.org

[11]Shihua Cai,Liangxiao Jiang,Dianhong Wang.Survey of Improving K-NN for Classification. International Journal of Advanced Research in Computer Science and Software Engineering.

[12]T. Rivas, M. Paz, J.E. Martins, J.M. Matias, J.F. Gracia, J. Taboadas, Explaining and predicting workplace accidents using data-mining Techniques, Journal of Reliable Engineering and System safety 96 (7) (2011)

[13]Z. Shu-Hsien, C. Pei-Hui, H. Pei-Yuan, Data mining techniques and applications-A decade review from 2000 to 2011, Journal of expert system with applications 39 (2012) (2012)