International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-4, Issue-2, February 2016

Web usage mining using Improved KNN Algorithm

Dr.P.Tamijeselvy, Sangavi. S, Suvetha. T, Umashankari. T

Abstract: **Internet has become an up scalable supply of knowledge. It provides numerous amount of user required information. Such vast handiness information provides several decisions for the user to pick out. Selecting an accurate product or information has become a time consuming and tedious task. To make information selection method easier, recommendation system has been developed. This may suggest a web page or web site or product for the user to decide on. This system is trained to suggest the resource for the user. Dailies of India represent dataset and from that user are suggested to their interested dailies and news. The classification is done using improved K-NN (K-Nearest Neighbour) classification algorithm and it has been trained to be used on-line and in real time to identify visitors click stream data, matching it to a specific user cluster and suggest a tailored browsing possibility that meet the requirement of the specific user at a specific time.**

Key-Words: **-Data mining, Improved KNN Classification, KNN classification, Clustering, Web server log files**

1 INTRODUCTION

Data mining is the extraction of knowledge from large amount of observational data sets, to discover unsuspected relationship and pattern hidden in data, summarize the data in novel ways to make it understandable and useful to the data users. Web usage mining is the application of data mining technique to automatically discover and extract useful information from a particular web site. In recent years, there has been an explosive growth in the number of researches in the area of web mining, specifically of web usage mining. Nowadays the number of websites were developed for the purpose of reading dailies news on-line across the Globe, but lack ways of identifying client navigation pattern and cannot provide satisfactory Real-Time response to the client needs, so, finding the appropriate news becomes time consuming which makes the benefit of on-line services to become limited. The system is able to observe users/clients navigation behavior by acting upon the user's click stream data, so as to recommend a unique set of objects that satisfies the need of an active user in a Real-Time, online basis. The K-Nearest Neighbor classification method was used online and in Real-Time to exploit web usage data mining technique to identify clients/visitors click stream data matching it to a particular user group and recommend a tailored browsing option that meet the need of the specific user at a given time.

1.1 Web mining

Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the world wide web. Web mining is used to understand customer behavior, evaluate the effectiveness of a particular website. Web mining allows you to look for patterns in data through

- Content mining
- Structure mining
- Usage mining

Content mining is used to examine the data collected by search engines and web spiders. Structure mining is used to examine data related to the structure of the particular website. Website and Usage mining are used to examine data related to a particular user's browser as well as the data gathered by forms the user may have submitted during web transaction.

1.2 Web usage mining

Web usage mining is a category in web mining. This web mining permits for collecting Web access information for Web pages. This data can offer the paths leading to accessed Web pages. This information is often collected automatically into access logs via the Web server. CGI scripts provide useful information such as referrer logs, user subscription information and survey logs. This data is vital to the overall use of data mining for firms and also their internet/ intranet applications and information access.

1.3 Data Classification

Classification is a data mining (machine learning) technique that is accustomed to predict cluster membership for data instances. For instance, you would like to use classification to forecast whether the weather on a particular day it will be "sunny", "rainy" or "cloudy". Widely used classification techniques include decision trees and neural networks.

1.3.1 K-Nearest Neighbours algorithm

In pattern discovery, the k-Nearest Neighbours algorithm (or k-NN in short) is a non-parametric method used for classification. The input

consists of k nearest training examples in the feature space.

In k-NN classification, the output is a class membership. An object is classified based on the utmost vote of its neighbours, with the object being assigned to the class that is most common among its *k* nearest neighbours (*k* is a positive integer, typically small). If *k* = 1, then the object is assigned to the class of that single nearest neighbour.

In k-NN regression, the output is the property value for the object. This value is average values of its *k* nearest neighbors. *k*-NN is a type of instance-based learning, or lazy learning, because the function is merely approximated locally and all computation is delayed until classification is done. The *k*-NN algorithm is one of the simplest of all machine learning algorithms.

For classification, it is valuable to assign weight to the contribution of the neighbors, so that the nearer neighbors can contribute more to the average than the distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/*d*, where *d* represents the distance to the neighbor.

The neighbors are taken from a collection of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This could be thought of as the training set for the algorithm.

A defect of the *k*-NN algorithm is that it is sensitive to local structure of the data. The algorithm has nothing to do with and is not to be confused with k-means, which is another popular machine learning technique.

1.3.2 Decision Tree Classification

In constructing a decision tree, apply both the gini index(g) and entropy value (e_i) as the splitting indexes, the model is experimented with a provided set of values, and different sets of results were obtained for both.

The decision tree technique has the restriction that all the training tuples is ought be in main memory, so, in the case of terribly large data; this may lead to inefficiency of decision tree construct due to swapping of the training tuples in and out of the main memory and cache memory. As a result of this a more scalable method like the KNN method are capable of handling training data that are too large to fit in memory is needed.

1.3.3 Bayesian Classifier Model

Decision rule and Bayesian network[9], are classification tree and support vector machine techniques that were used to model accidents and incidents in two firms in order to spot the cause of accident. Data is collected through interview and is then modelled. The experimental result was then compared with statistics techniques, which revelled that the Bayesian network and the other methodologies applied are more superior to the statistics technique. In theory, Bayesian classifier is claimed to poses minimum error rate as compared with all other classifier techniques but in practice this is not the case, owed to incorrectness in assumptions made for its use, such as class conditional independency and the insufficiency of available probability of data which is usually not the case when using K-NN method.

2 RELATED WORKS

M.F. Federico, L.L.Pier [1] researched the area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers. In this paper they present a survey of the recent developments in this area that is receiving increasing attention from the Data Mining community. Web usage mining is used to discover interesting user navigation patterns and can be applied to many real world problems, such as improving Web sites/pages, making additional topic or product recommendations, user/customer behavior studies, etc. This article provides a survey and analysis of current Web usage mining systems and technologies. B.Lalithadevi, A. Mary Ida,W.Ancy Breen[3] researched on Web usage mining system which performs five major tasks: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. Each task is explained in detail and its related technologies are introduced. A list of major research systems and projects concerning Web usage mining is also presented, and a summary of Web usage mining is given in the last section**.** Quig Yang, Hairing Henery Zhang[4] researched on Caching is a well-known strategy for improving the performance of Web-based systems. The heart of a caching system is its page replacement policy, which selects the pages to be replaced in a cache when a request arrives. In this paper, they present a Web-log mining method for caching Web objects and use this algorithm to enhance the performance of Web caching systems. In our approach, we develop an n-gram-based prediction algorithm that can predict future Web requests. The prediction model is then used to extend the well-known GDSF caching policy. Web page mining is the application of data

mining technique to automatically discover and extract useful information from a particular website. It is computationally expensive to find the k nearest neighbours when the dataset is very large. KNN can have poor run-time performance when the training set is large. It is very sensitive to irrelevant or redundant features because all features contribute to the similarity and thus to the classification.

The fact is that most existing works lack scalability and capability when dealing with on-line, Real-Time search driven web sites. So system is proposed to recommend a web page or web site or product for the user to choose. This system will be trained to recommend the resource for the user. The classification is done by improved KNN (K-Nearest Neighbour) classification method has been trained to be used on-line and in Real-Time to identify clients/visitors click stream data, matching it to a particular user group and recommend a tailored browsing option that meet the need of the specific user at a particular time.

3 PROBLEM FORMULATION

The major problem of many online websites is that the presentation of many choices to the client at a time. This typically leads to strenuous and time overwhelming task in finding the right product or information on the site. Web usage mining is the application of data mining technique used to involuntarily discover and extract useful information from a particular search.

3.1 Proposed scheme

Though web based recommendation systems are common, there is still several downside areas calling for solutions. The reality is that the most of the existing works lack scalability and capability when working with on-line, and Real-Time driven web searches. So system is proposed to suggest a web page or web site or product for the user to choose. This system is trained to recommend the appropriate resource for the user. Here reader web site is chosen to recommend the user based on their needs. Dailies of India forms the dataset and from which user are recommended to their interested daily's and news. This classification is done using improved KNN (K-Nearest Neighbour) classification method that can be trained to be used on-line and in Real-Time to recognize visitors click stream data, matching it to a specific user group and recommend a customized browsing option that will meet the needs of the specific user at a particular time.

3.2 System Overview

The dataset employed in this system is the user access database for a specific period of time, that was extracted, pre-processed and grouped into meaningful sessions and data mart was developed. The Improved K-Nearest Neighbour classification technique was used to investigate the uniform resource locator information of the users' address database. The user reader site data is stored in the data mart created. The overview of the system is shown in the Figure 1

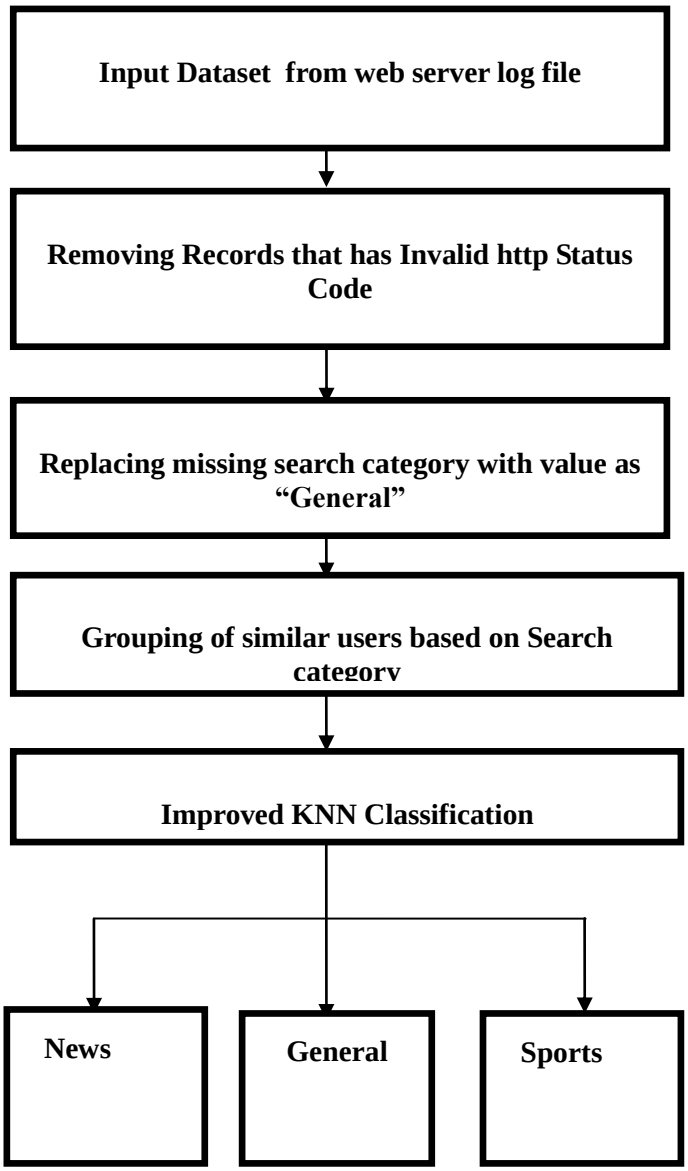

Figure 1: System Overview

3.2.1 Data Preprocessing

Data Acquisition refers to the collection of data for mining purpose, and this is usually the first task in data mining application. For web usage mining data is collected from web server, proxy server and web client. The web server source was chosen for the fact that it is the richest and most common data source, more so, it is possible to collect large amount of

information from the log files and databases they represent.

The original database file extracted, not all the data are applicable for web usage data mining, we solely want the entries that contain relevant information. The original file is usually made up of text files that have huge volume of data concerning queries made to the web server in which most of the instances contains irrelevant, incomplete and misleading data for mining purpose. Data cleansing is the stage in which inappropriate/noisy entries are eliminated from the log file.

In data preprocessing step data having inappropriate http status code are removed and missing search class is replaced with General.

3.2.2 Grouping Similar Users

In identification of similar users, search category is considered as classifier attribute. When count value of particular class is greater than one and it matches to any of the defined category, then it is grouped under one category. If there is only one user for particular search then that record will be eliminated.

3.2.3 Improved KNN Algorithm

Curse of dimensionality that is caused due to Euclidean distance is eliminated by finding weight of attributes in improved KNN methods. The traditional KNN text classification has three defects. First is that it has great computational complexity. When using traditional KNN classification, to find the K nearest neighbour samples for the given test sample, it is mandatory be calculate the similarities between all the training samples, as the dimensions of the text vector is generally high, so it has great calculation complexity in this process which makes the efficiency of text classification to be very low. Usually there are 3 methods to reduce the complexity of KNN algorithm: minimising the dimensionality of vector text [4]; using data set of small size; using improved algorithm which can accelerate to spot out the K nearest neighbour samples.

Second is depending on training set KNN algorithm does not use additional data to point up the classification rules, but the classifiers are generated by the self training samples, this makes the algorithm to depend on training set excessively, for example, it is necessary to re-calculate when there is a small change on training set.

Last is there is no weight difference between samples. The traditional KNN algorithm treats all training samples equally, and there is no variation between the samples, therefore it don't match the actual phenomenon when the samples have uneven distribution.

4 PROBLEM SOLUTION

The following steps are implemented for the Proposed System

4.1 Data Preprocessing

Input **:** Dataset is collected from webserver log file
Output **:** Pre-processed Data set.

Steps

- Upload the dataset into server.
- Check the missing values.
- If the missing value occurred for Search class then replace the value as "general".
- Eliminate the records that have http status code greater than 400 and less than 200

4.2 Grouping Similar Users

Input **:** Pre-processed data set
Output **:** clustered result data set

Steps

- The pre-processed dataset is taken as input.
- Search category clusters will be formed based on count.
- If the count value is one, then the record will be eliminated

4.3 Steps for Improving KNN

- Determine Parameter K, where K is the number of nearest neighbours
- Calculate the distance between the query and all the training examples
- Sort the distance and determine nearest neighbour based on the k-th minimum distance
- Gather the category Y of the nearest neighbours
- Use simple majority of the category of nearest neighbours as the Prediction value of the query distance [6] is calculated as in equation (1)

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

where

ϙ-weight of the attribute

5 EXPERIMENTATION RESULTS

Input :Input Dataset from Webserver log

files

Output :Grouping of similar users based on search

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

Precision takes all retrieved documents into account, but it can be evaluated at a given cut-off rank, accounting only the top most results returned by the system. This is called Precision at n.

Precision is the chance that (Randomly Selected) the retrieved document is relevant.

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

Recall in information retrieval is the fraction of documents that are relevant to the query that are successfully retrieved. Recall is the chance that a (Randomly Selected) relevant document is retrieved in a search.

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

Accuracy is not a reliable metric for evaluating the real performance of the classifier when the number of samples in different classes varies greatly because it may yield misleading results.

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

Where

TN is the number of true negative cases
FP is the number of false positive cases
FN is the number of false negative cases
TP is the number of true positive cases

True Positive(TP):These refer to the positive tuples that were correctly labelled by the classifier.

True Negative(TN):These are the negative tuples that were correctly labelled by the classifier

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

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

Table 1 depicts the performance analysis table of KNN and Improved KNN based on Precision, Recall and Accuracy.

Table 1: Performance analysis of KNN and Improved KNN

Figure 2 depicts the performance analysis graph for comparing KNN and Improved KNN based on Precision, Recall and Accuracy.

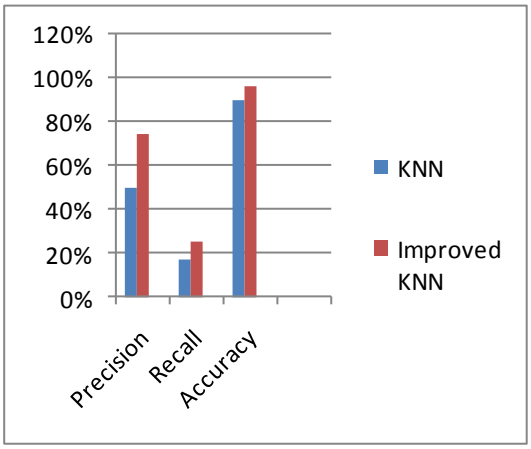

Figure 2: Performance analysis graph for comparing KNN and Improved KNN

6 CONCLUSION

The proposed system provides a basis for automatic Real-Time recommendation system. The system performs classification of users on the simulated active sessions extracted from testing sessions by collecting active users' click stream[6] and matches this with similar class in the data mart, so as to generate a set of recommendations to the client in a Real-Time basis using improved k- NN classification.

The System can be Future enhanced by

- Efficient features can be extraction method can be used to improve classification method.
- System can be enhanced to have different types of log records.
- Optimized classification techniques can be used to enhance the classification accuracy.

References

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

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.

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

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.

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.

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

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.

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

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.

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

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

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)

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)

Dr.P.Tamijeselvy [1], Sangavi. S [2], Suvetha. T [3], Umashankari. T [4]

Associate Professor [1], UG Scholar [2,3,4]
Department of Computer Science and Engineering
Sri Krishna College of Technology, Coimbatore, India.