Applied Computing and Informatics (2015) **xxx**, xxx–xxx

Saudi Computer Society, King Saud University
Applied Computing and Informatics
(http://computer.org.sa)
www.ksu.edu.sa
www.sciencedirect.com

ORIGINAL ARTICLE

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

D.A. Adeniyi, Z. Wei, Y. Yongquan *

Department of Computer Sc. & Technology, College of Information Sc. & Engineering, Ocean University of China, Qingdao, Shandong, China

Received 3 August 2014; revised 29 September 2014; accepted 17 October 2014

KEYWORDS

Automated;
Data mining;
K-Nearest Neighbor;
On-line;
Real-Time

Abstract The major problem of many on-line web sites is the presentation of many choices to the client at a time; this usually results to strenuous and time consuming task in finding the right product or information on the site. In this work, we present a study of automatic web usage data mining and recommendation system based on current user behavior through his/her click stream data on the newly developed Really Simple Syndication (RSS) reader website, in order to provide relevant information to the individual without explicitly asking for it. The K-Nearest-Neighbor (KNN) 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. To achieve this, web users RSS address file was extracted, cleansed, formatted and grouped into meaningful session and data mart was developed. Our result shows that the K-Nearest Neighbor classifier is transparent, consistent, straightforward, simple

* Corresponding author.
E-mail address: i@yangyongquan.com (Y. Yongquan).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.aci.2014.10.001

to understand, high tendency to possess desirable qualities and easy to implement than most other machine learning techniques specifically when there is little or no prior knowledge about data distribution.

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 [13,31,2]. Web usage mining is the application of data mining technique to automatically discover and extract useful information from a particular web site [2,22,30].

The term web mining was believed to have first came to be in 1996 by Etzioni in his paper titled "The World Wide Web: Quagmire or Gold mine" and since then attention of researchers world over has been shifted to this important research area [26]. 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. According to Federico and Pier [9], over 400 papers have been published on web mining since the early paper published in 1990s.

The Really Simple Syndication (RSS) reader website was 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 study aimed at designing and developing an automatic, online, Real-Time web usage data mining and recommendation system based on data mart technology. The system is able to observe users/clients navigation behavior by acting upon the user's click stream data on the RSS reader web site, so as to recommend a unique set of objects that satisfies the need of an active user in a Real-Time, online basis. The user access and navigation pattern model are extracted from the historical access data recorded in the user's RSS address URL file, using appropriate data mining techniques.

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 [24]. For instance, if a user seems to be searching for politics news on china daily on his/her visit to the RSS reader site, more politics news headlines from other dailies such as CNN politics news will be recommended to the user with the required feed needed to be added to his/her profile in order to access such news headlines asides his/her originally requested news. This is aimed at assisting the user to get relevant information without explicitly asking for it, so as to ease and fasten navigation on the site without too many choices being presented to the user at a time, More so, the study will assist the web designer and administrator to re-arrange the content of the web site in order to improve the impressiveness of the web site by providing online Real-Time recommendation to the client.

To achieve this, the web users RSS address URL file was extracted, cleansed, formatted and grouped into meaningful session for data mining analysis, data mart was developed, this is as a result of the fact that the raw URL file extracted is not well structured to be used directly for data mining [19]. In designing the data mart, the process of URL data acquisition and model extraction was implemented using database management software specifically the Structured Query Language, MySQL 2008 [20]. The process of the development of the automatic Real-Time web usage mining and recommendation application was done by adopting the Java programming language with NetBeans as the editor and compiler [21]. The MATLAB software was used for interpretation and graphical presentation of the result obtained [17]. A thorough presentation of the experimental result was done in order to assist the site designer and administrator to improve the content and impressiveness of the said RSS reader site. Fig. 1 showing the architecture of the overall system can be seen in Supplementary material.

2. Related work

This section reviews some related works pertinent to this study, the review is specifically organized into subsections as follows:

2.1. Web data mining

Zdravko and Daniel [31], described web data mining as application of data mining techniques to discover patterns in web content, structure and usage. It is a branch of applied artificial intelligence that deals with storage, retrieval and analysis of web log files in order to discover users accessing and usage pattern of web pages [26].

2.2. Forms of data mining system

Two forms of data mining tasks were identified by researchers over the years, these includes; predictive and descriptive [13,1,7].

In predictive data mining task, inference is performed on current data in a database in order to predict future values of interest while in descriptive task, data in a database are classified by characterizing the general properties of the data, it finds pattern describing the data in the database so as to present the interpretation to the user [13,1,8].

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

2.2.1. Classification of data mining system

Data mining system can be classified using different criteria. Jiawei and Micheline [13], identified these criteria as kind of database mined, kind of knowledge mined, type of technique utilized and according to type of application adapted. Federico and Pier [9], stated further that in web usage data mining task, different techniques can be adopted, but the issue is how to determine which technique is most appropriate for the problem at hand. A multiple approach or an integrated technique that combines the benefits of a number of individual approaches can be adopted by a comprehensive data mining system [28]. [13,15,16], stated that there are different techniques for data classification which includes; decision tree classifier, Bayesian classifier, K-Nearest Neighbor classifier, and rule base classifier. In our work, the K-Nearest Neighbor classification method was adopted.

2.3. Overview of some related data mining techniques

Below is a brief overview of some of the data mining techniques according to different scholars in the field as it relates to our work.

Decision tree: The use of classification and regression tree (CART) was adopted by Amartya and Kundan [1] in their work. In constructing a decision tree, they applied both the gini index(g) and entropy value (e_i) as the splitting indexes, the model was experimented with a given set of values, different sets of results were obtained for both the outlook, humidity, windy, Temp, and Time for execution. The result of the experiment shows that the best splitting attribute in each case was found to be outlook with the same order of splitting attributes for both indices.

The decision tree technique has the restriction that the training tuples should reside in memory, so, in the case of very large data, decision tree construct therefore becomes inefficient due to swapping of the training tuples in and out of the main and cache memories. As a result of this a more scalable approach such as the KNN method, capable of handling training data that are too large to fit in memory is required.

The SOM model: Self Organizing Map (SOM) or Kohonen neural network model was explored by Xuejuu et al. [30], in their work, to model customers navigation behavior. The model was used to create clusters of queries based on user session as extracted from web log with each cluster representing a class of users with similar characteristics, in order to find the web links or product of interest to a current user on a Real-Time basis. The experimental result of the SOM model performance was compared with that of K-Means model, and the SOM model was found to outperform the K-Means model with value of correlation co-efficient of SOM model scoring twice that of K-means result.

Our work shares essentially the same goals as SOM, but differs in its construction. In SOM, the user profiles have been pre-determined offline by the offline

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

usage pattern discovery module, while in our work, user profiles are determined online, thereby making real time response and recommendation faster.

The path analysis model: Resul and Ibrahim [26], in their work used the path analysis method to investigate the URL information of access to the Firat University web server, web log file so as to discover user accessing pattern of the web pages, in order to improve the impressiveness of the web site. They explain further that, the application of path analysis method provides a count of number of time a link occur in the data set, together with the list of association rules which help to understand the path that users follow as they navigate through the Firat University web site.

The Path analysis model is based on information from the clients' previous navigation behavior, the method provides a count of number of time a link occur in the dataset. Though our work shares the same goal of recommendation but again differs in its approach, which is based on user's current navigation behaviors rather than previous navigation behavior as in path analysis method.

Bayesian classifier model: Decision rule and Bayesian network, support vector machine and classification tree techniques were used by Rivas et al. [27], to model accidents and incidents in two companies in order to identify the cause of accident. Data were collected through interview and modeled. The experimental result was compared with statistics techniques, which shows that the Bayesian network and the other methods applied are more superior than the statistics technique. Rivas et al. [27], stated further that the Bayesian/K2 network is of advantage as it allows what-if analysis on data, which make the data to be deeply explored.

In theory, Bayesian classifier is said to have minimum error rate in comparison with all other classifier but in practice this is not always the case, due to inaccuracy in assumptions made for its use, such as class conditional independency and the lack of available probability data which is usually not the case when using KNN method.

The K-Nearest Neighbor (KNN): Many researchers have attempted to use K-Nearest Neighbor classifier for pattern recognition and classification in which a specific test tuple is compared with a set of training tuples that are similar to it. [12], in their own work introduced the theory of fuzzy set into K-Nearest Neighbor technique to develop a fuzzy version of the algorithm. The result of comparing the fuzzy version with the Crisp version shows that the fuzzy algorithm dominates its counterpart in terms of low error rate. In the work of [11]. The K-Nearest Neighbor algorithm was used alongside with five other classification methods to combine mining of web server logs and web contents for classifying users' navigation pattern and predict users' future request. The result shows that the KNN outperformed three of the other algorithms, while two of them performed uniformly. It was also observed that KNN archives the highest F-Score and A(c) on the training set among the six algorithms. [25], as well adopted the KNN classifier to predict protein cellular localization site. The result of the test using stratified crossvalidation shows the KNN classifier to perform better than the other methods which includes binary decision tree classifier and the naïve Bayesian classifiers.

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

2.4. Justification for using KNN algorithm over other existing algorithm

The K-Nearest Neighbor (K-NN) algorithm is one of the simplest methods for solving classification problems; it often yields competitive results and has significant advantages over several other data mining methods. Our work is therefore based on the need to establish a flexible, transparent, consistent straightforward, simple to understand and easy to implement approach. This is achieved through the application of K-Nearest Neighbor technique, which we have tested and proved to be able to overcome some of the problems associated with other available algorithms. It is able to achieve these by the following:

- Overcoming scalability problem common to many existing data mining methods such as decision tree technique, through its capability in handling training data that are too large to fit in memory.
- The use of simple Euclidean distance to measure the similarities between training tuples and the test tuples in the absence of prior knowledge about distribution of data, therefore makes its implementation easy.
- Reducing error rate caused by inaccuracy in assumptions made for usage of other technique such as the Naïve Bayesian classification technique, such as class conditional independency and the lack of available probability data which is usually not the case when using KNN method.
- Providing a faster and more accurate recommendation to the client with desirable qualities as a result of straightforward application of similarity or distance for the purpose of classification.

2.5. Significance of the study

Available published literature makes it clear that though web based recommendation systems are increasingly common, there still available many problem areas calling for solutions. The fact is that most existing works lack scalability and capability when dealing with on-line, Real-Time search driven web sites, more so, the recommendation quality and accuracy of some are doubtful, since they mostly relied on historical information based on clients' previous visit to the site, rather than his immediate requirement. Some recommendation systems as well, create a lot of bottleneck through system computing load when handling scaled web site at peak visiting time thereby slowing down the recommendation process.

To solve the above issues the following solutions were made through our system.

- Scalability problems common to many existing recommendation system were overcome through combine on-line pattern discovery and pattern matching for real time recommendation, in this regard our algorithm works better than decision tree algorithm.

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

– Our result indicates that the adoption of the K-NN model can lead to a more accurate recommendation that outperformed many other existing models. In most cases the precision rate or quality of recommendation by our system is equal to or better than 70%, meaning that over 70% of product recommended to a client will be in line with his immediate requirement, making support to the browsing process more genuine rather than a simple reminder of what the user was interested in on his previous visit to the site as seen in path analysis technique.
– Our recommendation engine collects the active users' click stream data, matches it to a particular user's group in order to generate a set of recommendation to the client at a faster rate, therefore overcoming the problem of bottleneck caused by system computing load when dealing with scaled web sites at a peak visiting time, as it is in many existing data mining methods.
– Our system provides a precise recommendation to the client based on his current navigation pattern, thereby overcoming time wastage in finding the right product or information caused by presentation of many irrelevant choices to the client at a time as it is in many existing systems.

Hence, the proposed approach is capable of addressing the issues and provides a straightforward, simple to understand and easy to implement web usage classification and recommendation model.

3. Methodology

This section presents detail description of the realization and implementation of web usage data mining system. The presentation of the application of the proposed methodology for the analysis of users' RSS address file of the RSS reader website was showcased. We have developed an online, Real-Time recommendation expert system that can assist the web designer and administrator to improve the content, presentation and impressiveness of their website by recommending a unique set of objects that satisfies the need of active user based on the user's current click stream.

3.1. Overview of steps in performing web usage data mining task

Data mining task can be categorized into different stages based on the objective of the individual analyzing the data [1,7].

The overview of the task for each steps is presented in detail in four subsections as follows:

3.1.1. Data acquisition, preprocessing and data mart development

Data acquisition: This refers to the collection of data for mining purpose, and this is usually the first task in web mining application [6]. The said data can be collected from three main source which includes (i) web server (ii) proxy server and (iii) web client [9]. In this study, the web server source was chosen for the fact that it is the richest and most common data source, more so, it can be used to collect large amount of information from the log files and databases they represent. The user profile information, the access and navigation pattern or model are extracted from the historical access data recorded in the RSS reader site, users' address database. The data are so voluminous as it contains so many detailed information such as date, time in which activities occur, saver's name, IP address, user name, password, dailies name, required feed, news headlines, and contents, as recorded in the database file. In fact, the original document is about 5285 pages.

Data pre-processing: In the original database file extracted, not all the information are valid for web usage data mining, we only need entries that contain relevant information. The original file is usually made up of text files that contains large volume of information concerning queries made to the web server in which in most instance contains irrelevant, incomplete and misleading information for mining purpose [30,11]. Resul and Ibrahim [26], described data preprocessing as the cleansing, formatting and grouping of web log files into meaningful session for the sole aim of utilizing it for web usage mining.

Data cleansing: Data cleansing is the stage in which irrelevant/noisy entries are eliminated from the log file [18]. For this work the following operations were carried out: (i) Removal of entries with "Error" or "Failure" status. (ii) Removal of requests executed by automated programs such as some access records that are automatically generated by the search engine agent from access log file and proxies. (iii) Identification and removal of request for picture files associated with request for a page and request include Java scripts (.js), and style sheet file (iv) Removal of entries with unsuccessful HTTP status code, etc.

Data mart development: Two crown corporation [29], explained that data mart is a logical subset of data warehouse. If the data warehouse DBMS can support more resources, that will be required of the data mining operation, otherwise a separate data mining database will be required. Since the raw log file is usually not a good starting point for data mining operation, the development of a data mart of log data is required for the data mining operation. In this work a separate data mart of users' RSS address URL was developed using relational database Management software MySQL [20,19].

3.1.2. Transaction identification

There is need for a mechanism to distinguish different users so as to analyze users access behavior [11]. Transaction identification is meant to create meaningful clusters of references for each user. Xuejuu et al. [30], stated that a user navigation behavior can be represented as a series of click operations by the user in time sequence, usually call click stream, which can further be divided into units of click descriptions usually referred to as session or visit.

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

Session identification: According to [30,11], a session can be described as a group of activities carried out by a user from the user's entrance into the web site up to the time the user left the site. It is a collection of user clicks to a single web server. Session identification is the process of partitioning the log entries into sessions after data cleansing operation [18,3]. In order to achieve this Xuejuu et al. [30], suggested the use of cookies to identify individual users, so as to get a series of clicks within a time interval for an identified user. One session can be made up of two clicks, if the time interval between them is less than a specific period.

3.1.3. Pattern discovery

Pattern discovery is the key process of web mining which includes grouping of users based on similarities in their profile and search behavior. There are different web usage data mining techniques and algorithms that can be adopted for pattern discovery and recommendation, which includes, path analysis, clustering, and associate rule. In our work, we have experimented with the K-Nearest Neighbor classification technique as described in Section 3.2 in order to observe and analyze user behavior pattern and click stream from the pre-process to web log stage and to recommend a unique set of object that satisfies the need of an active user, based on the users' current click stream.

3.1.4. Pattern analysis

Pattern analysis is the final stage in web usage mining which is aimed at extracting interesting rules, pattern or statistics from the result of pattern discovery phase, by eliminating irrelevant rules or statistics. The pattern analysis stage provides the tool for the transformation of information into knowledge. We have incorporated an SQL language to develop a data mart using MySQL DBMS software specifically created for web usage mining purpose in order to store the result of our work [16]. The data mart is populated from raw users RSS address URL file of the RSS reader's site that contains some basic fields needed; our experiment result is presented in Section 4.

3.2. Our approach

The problem at hand is a classification problem, therefore the K-Nearest Neighbor method of data mining is ideal. The objective of the system is to create a mapping, a model or hypothesis between a given set of documents and class label. This mapping was later to be used to determine the class of a given Test(unknown or unlabeled) documents [31]. The K-Nearest Neighbor model is the simplest and most straightforward for class prediction, it is the most popular similarity or distance based text and web usage classification and recommendation model [31].

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

3.2.1. K-Nearest-Neighbor technique

According to Leif [14], a non-parametric method of pattern classification popularly known as K-Nearest Neighbor rule was believed to have been first introduced by Fix and Hodges in 1951, in an unpublished US Air Force School of Aviation Medicine report. The method however, did not gain popularity until the 1960s with the availability of more computing power, since then it has become widely used in pattern recognition and classification [13]. K-Nearest Neighbor could be described as learning by analogy, it is learnt by comparing a specific test tuple with a set of training tuples that are similar to it. It is classified based on the class of their closest neighbors, most often, more than one neighbor is taken into consideration hence, the name K-Nearest Neighbor (K-NN), the "K" indicates the number of neighbors taken into account in determining the class [13]. The K-NN algorithm has been adopted by statisticians as a machine learning approach for over 50 years now [31]. The K-NN is often referred to as "Lazy learner" in the sense that it simply stores the given training tuples and waits until it is given a test tuple, then performs generalization so as to classify the tuple based on similarities or distance to the stored training tuples. It is also called "instance based learner". The lazy learner or instance based learner does less work when presented with training tuples and more work during classification and prediction, therefore makes it computational expensive, unlike the eager learners that when given a training tuple construct a classification model before receiving the test tuple to classify, it is therefore very ready and eager to classify any unseen tuples. [13,31,1]. [13,14], stated that the K-NN error is bounded above twice the Baye's error rate.

3.3. The working of K-Nearest Neighbor classifier

The K-Nearest Neighbor classifier usually applies either the Euclidean distance or the cosine similarity between the training tuples and the test tuple but, for the purpose of this research work, the Euclidean distance approach will be applied in implementing the K-NN model for our recommendation system [13].

In our experiment, suppose our data tuples are restricted to a user or visitor/-client described by the attribute Daily Name, Daily Type and News category and that X is a client with Dayo as username and Dy123 as password.

The Euclidean distance between a training tuple and a test tuple can be derived as follows:

Let X_i be an input tuple with p features $(x_{i1}, x_{i2}, \ldots, x_{ip})$
Let n be the total number of input tuples $(i = 1, 2, \ldots, n)$
Let p be the total number of features $(j = 1, 2, \ldots, p)$
The Euclidean distance between Tuple X_i and $X_t (t = 1, 2, \ldots, n)$ can be defined as

$$d(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \ldots + (x_{ip} - x_{tp})^2} \quad (3.1)$$

In general term, The Euclidean distance between two Tuples for instance $X_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \ldots, x_{2n})$ will be,

$$\text{dist}(x_1, x_2) = \sqrt{\sum\nolimits_{i=1}^{n} (x_{1i} - x_{2i})^2} \tag{3.2}$$

Eq. (3.2) is applicable to numeric attribute, in which we take the difference between each corresponding values of attributes tuple x_1 and x_2, square the result and add them all together then get the square root of the accumulated result this gives us the distance between the two points x_1 and x_2 [13,14]. In order to prevent attributes with initially large ranges from outweighing attributes with initial smaller ranges, there is a need to normalize values of each attributes before applying Eq. (3.2).

The min–max normalization can be applied to transform for instance value V of a numeric attribute A to V^1 in the range [0,1] by using the expression

$$V^1 = \frac{V - \text{min}A}{\text{max}A - \text{min}A} \tag{3.3}$$

min$A$ and max$A$ are attribute A, minimum and maximum values [13].

In K-NN, classification, all neighboring points that are nearest to the test tuple are encapsulated and recommendation is made based on the closest distance to the test tuple, this can be defined as follows:

Let C be the predicted class

$$C_i = \{x \in C_p;\ d(x, x_i) \leq d(x, x_m), i\#m\} \tag{3.4}$$

The nearest tuple is determined by the closest distance to the test tuple. The K-NN rule is to assign to a test tuple the majority category label of its K-Nearest training tuple [14].

3.3.1. Computing distance for categorical attribute

A categorical attribute is a nonnumeric attribute such as color and object name. To calculate the distance, we simply compare the corresponding values of the attributes in tuple x_1 with that of x_2, if the values are the same, then the difference is taken to be zero(0), otherwise the difference is taken to be one(1). For instance, if two users, x_1 and x_2 click stream on the RSS reader site is both sport news category, then the difference is zero(0), but if tuple x_1 is sport and tuple x_2 is politics, then the difference is taken to be one(1) [13].

3.3.2. Missing values

If the value of a given attribute A is missing in tuple x_1 or x_2 or both, for categorical value, if either or both values are missing we take the difference to be one(1), in numeric attribute if x_1 and x_2 values are missing we also take the difference to be one(1), if only one value is missing and the other is present and normalized we can consider the difference to be $|1 - V^1|$ or $|0 - V^1|$ whichever is greater is chosen [13].

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

3.3.3. Determining the value of K, the number of Neighbor

In reality, the value of K is usually odd numbers, ie. $K = 1$, $K = 3$, $K = 5$, etc. this is obvious in order to avoid ties [14]. $K = 1$ rule is mostly referred to as the nearest neighbor classification rule. The value of K (Number of neighbor) can be determined by using a test set to determine the classification error rate, by experimenting with different values of K, starting with $K = 1$, then the K value with minimum error rate is selected [13]. Jiawei and Micheline [13], stated further that the larger the training tuple, the larger the value of K. Zdravko and Daniel [31], in their work, experimented with different values of up to $K = 19$, with and without distance weighting on a set of document collections, the experiment was run with a complete set of 671 attributes and concluded that a small set of relevant attributes works better than all attributes, that the experiment works perfect with $K = 1$, and $K = 3$ and with little improvement in $K = 5$. So, if K approaches infinity, the error rate approaches that of Baye's error rate [13]. Zdravko and Daniel [31], further stated that 1-NN makes a better prediction using single instance however large the training set is, but under the assumption that there is no noise and all attributes are equally important for classification.

In our work, we adopted 5 as the maximum value of K. We simply applied the distance weighted K-NN approach, in which we experimented for different values of K on our sample data, starting from $K = 1$, up to $K = 9$. We discovered that the experiment works better with $K = 1$, $K = 3$ and with little accuracy at $K = 5$, so, we selected $K = 5$, which gives us the minimum error rate. The algorithm for the K-Nearest Neighbor classifier model is shown in Fig. 2, in Supplementary material.

3.4. Application of K-Nearest Neighbor classification technique to predict user's class label in the RSS reader's web site

Example 1. Let us consider the RSS reader sites' client click stream as a vector with three(3) attributes: Daily name News category and Added required feed type, with users represented by $X_1, X_2, X_3, X_4, \ldots, X_{11}$ as the class labels as shown in Table 1. Assuming the class of user X_3 is unknown.

To determine the class of user X_3, we have to compute the Euclidean distance between the vector X_3 and all other vectors, by applying Eq. (3.2).

The Euclidean distance between two tuples for instance training tuple X_1 and test tuple X_3 ie.

$X_1 = (x_{11}, x_{12}, x_{13})$ and $X_3 = (x_{31}, x_{32}, x_{33})$ each with the following attributes as in Table 1.

X_1 = (CNN news, World, www.*world) and X_3 = (Punch ng, politics, www.*politics) will be:

$$\mathrm{dist}(x_1, x_3) = \sqrt{\sum_{i=3}^{n} (x_{1i} - x_{3i})^2}$$

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

Table 1 The RSS reader's data mart class labels training tuple.

Users	Daily's name	News category	Added required feed type	Class
X_1	CNN news	World	www.*world	World
X_2	China daily	Business	www.*business	Business
X_4	CNN news	Politics	www.*politics	Politics
X_5	Punch ng	Entertainment	www.*entertainment	Entertainment
X_6	Thisday news	Politics	www.*politics	Politics
X_7	Vanguard news	Sports	www.*sports	Sports
X_8	Complete football	Sport	www.*sports	Sports
X_9	Vanguard news	Politics	www.*politics	Politics
X_{10}	China daily	Politics	www.*politics	Politics
X_{11}	Thisday news	World	www.*world	World
X_3	Punch ng	Politics	www.*politics	?

Remember, for a categorical attribute as in Table 1, the difference (x_{11}, x_{31}) can be computed by simply compare the corresponding value of the attributes in tuple x_1 with that of x_3 as explained previously. If the values are the same then the difference is taken to be zero(0), otherwise, the difference is taken to be one(1). So, for ($x_{1,1}$ and $x_{3,1}$) ie. (CNN news and Punch ng), the difference is 1, for (x_{12} and x_{32}) ie. (World and Politics) the difference is 1, likewise for (x_{13} and x_{33}) ie., (www.*world and www.*politics) the difference is 1 as well, therefore,

$$\mathrm{dist}(x_1, x_3) = \sqrt{(x_{1,1} - x_{3,1})^2 + (x_{1,2} - x_{3,2})^2 + (x_{1,3} - x_{3,3})^2} \text{ this gives :}$$

$$\mathrm{dist}(x_1, x_3) = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} = 1.73205081$$

Repeating the same process in our example for all other tuple $x_2, x_4, \ldots, x_{11}$, the result of these calculation produced a stream of data as shown in Table 2, which shows the users sorted by their Euclidean distance to the user x_3 to be classified.

The 1-NN approach simply picks the user with minimum distance to x_3, (the first one from the top of the list) and uses it's class label "politics" to predict the class of x_3, therefore recommends similar news headlines of "Politics" as in user x_9 class.

Table 2 Data showing users sorted by distance to user x_3.

User	Class	Distance to user x_3
x_9	Politics	1.000000000
x_{10}	Politics	1.000000000
x_6	Politics	1.000000000
x_4	Politics	1.141421356
x_5	Entertainment	1.141421356
x_2	Business	1.732050810
x_7	Sports	1.732050810
x_8	Sports	1.732050810
x_1	World	1.732050810
x_{11}	World	1.732050810

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

In example 1, 3-NN will as well classify "Politics" because it is the majority label in the top three classes. However, distance weighted K-NN can be helpful in determining the class a given test tuple belong, whenever there seems to be a ties [31]. For instance, the distance weighted 5-NN will simply add the distance for class politics as in our example in Table 2 and compare it with that of Entertainment whichever is greater is selected. i.e. 1.000000000 + 1.000000000 + 1.000000000 + 1.141421356 = 4.141421356 while that of entertainment is 1.141421356 thus, weight of "politics" > "Entertainment" Then the 5-NN will as well classify user x_3 as "Politics" because it has higher weight than entertainment. The distance weighted K-NN allows the algorithm to use more or even all instances instead of one instance as in 1-NN.

4. System evaluation and analysis of result

This section evaluates our system by applying the result of the experiment conducted. The result was presented and analyzed in order to evaluate the quality of our recommendation system based on K-Nearest Neighbor classification model. In the previous section we established that a class with minimum distance to the test tuple will be predicted for 1-NN or in case ties exist, the weighted distance predict a class with greater weighted distance as in 5-NN in example 1 and recommendation will be made based on this, for user with unknown class.

Software was developed with Java NetBeans programming language and MySQL DBMS was used in creating the data mart in order to implement our model using K-NN method. The sample interface from the automated on-line Real-Time recommendation system developed for the purpose, indicating the active user's click stream, a dialog box presenting his requested news headlines and a message box presenting Real-Time recommendation to the user based on his current request is shown in Fig. 3 in Supplementary material and the source code in Java NetBeans programming language for the system is also available as part of Supplementary material.

In this work, the number of class C of user X that can be recommended by the recommendation model is set at 5, "5" indicates different news categories headlines and user classes that could be presented to the active user, based on information from the user's click stream. However, this number could be increased or decreased depending on the available options at a given time.

In this study however, the computation of Euclidean distance that produced the set of values from which the closest distance $C_i = \{x \in C_p;\ d(x,x_i) \leqslant d(x,x_m),\ i \# m\}$, was not repeatedly shown, because of size, since the calculation follows the same procedures. Table 2 shows the sorted result according to distance to the test tuple.

Godswill [10], stated that in real life analysis, a model performance quality can only be measured by ability to predict accurately, the new data set rather than the training data set in which the model was trained. They explained further that the

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

predictive ability of a model will be questionable and cannot be used for prediction, if the model performs well in the training set but performs poorly in the test validation data set or new data set.

4.1. Presentation of result

Example 2: Using the data in Table 1, this time around assuming the class of user X_7 is unknown. We can determine the class of user x_7 based on his current click stream information by computing the Euclidean distance between the user x_7 and all other users as we did in example 1

X_1 = (CNN news, World, www.*world)
X_7 = (Vanguard news, Sports, www.*sports)
$\text{dist}(x_1,x_7) = \sqrt{\sum_{i=3}^{n}(x_{1i} - x_{7i})^2}$
Being categorical attributes,
Differences $(x_{1,1} - x_{7,1}) = 1$, $(x_{1,2} - x_{7,2}) = 1$, $(x_{1,3} - x_{7,3}) = 1$
Therefore applying Eq. (3.2) we have
$\text{dist}(x_1,x_7) = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3} = 1.73205081$
Repeating the whole process for all the available users produced a stream of data as in Table 3.

4.2. Analysis of the results

The MATLAB code [17,23], that was used for graphical analysis of the experimental result from Tables 2 and 3 as shown in Figs. 4,6 is available on request.

In order to model the users click stream in the RSS readers web site, The K-Nearest Neighbor classification technique of data mining was applied on the extracted users RSS address database. The data set was produced by computing the Euclidean distance between the test tuple and the training tuples as shown in example 1 and example 2 and data set presented in Tables 2 and 3 respectively.

Table 3 Data showing users sorted by distance to user x_7.

User	Class	Distance to user X_7
x_8	Sports	1.0000000000
x_9	Politics	1.1414213560
x_1	World	1.7320508100
x_2	Business	1.7320508100
x_3	Politics	1.7320508100
x_4	Politics	1.7320508100
x_5	Entertainment	1.7320508100
x_6	Politics	1.7320508100
x_{10}	Politics	1.7320508100
x_{11}	World	1.7320508100

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

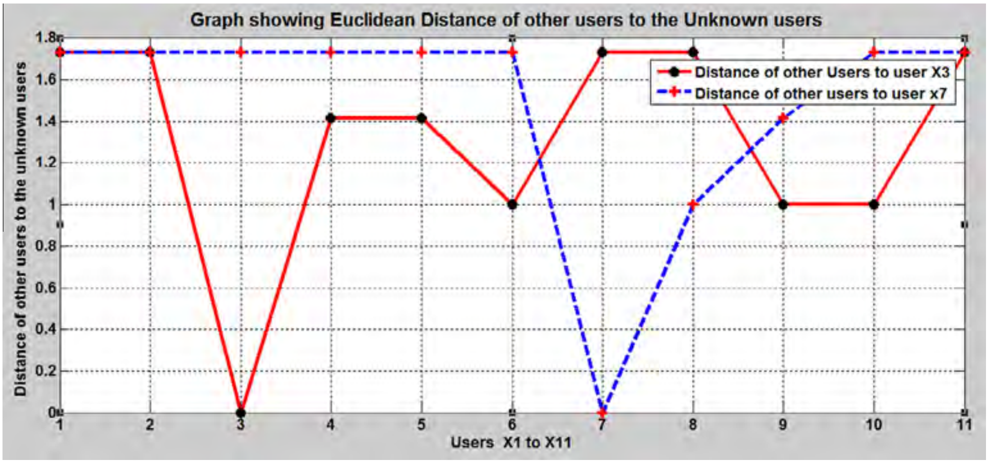

Figure 4 Graph showing Euclidean distance from the other User/Neighbor to user X_3 and X_7.

The K-Nearest Neighbor classifier predicts the class label with class C_i for which $C_i = \{x \in C_p; d(x,x_i) \leqslant d(x,x_m), i \#m\}$ for the unknown user class ie., the 1-NN classification simply picks the user with minimum distance to users X_3 and X_7 as the case may be (ie. The first user from the top of the list), in Table 2 for user X_3 and Table 3 for user X_7 respectively and use their class labels to predict the class of X_3 and X_7 respectively, therefore, recommend similar news headline of politics for user X_3 as in user X_9 class from Table 2 and Sports for user X_7 as in user X_8 class from Table 3 as shown in Figs. 4,6 respectively. Figs. 5 and 6 can be found in Supplementary material.

5. Summary of findings

Different criteria can be used to determine the quality and efficiency of a particular web site, which includes the following: contents, presentation, ease of usage, ease of accessing required information, waiting time of users, to mention just a few. In this study a novel approach is presented to classify users based on their current click stream, matching it to a particular user group popularly referred to as Nearest neighbor and recommend a tailored browsing option that satisfies the needs of the active user at a particular time, by applying the web usage data mining technique to extract knowledge required for providing Real-Time recommendation services on the web site.

We have conducted experiments on our designed experimental system. The data set used in the system is the RSS user access database for a two months period, which was extracted, pre-processed and grouped into meaningful sessions and data mart was developed. The K-Nearest Neighbor classification technique was used to investigate the URL information of the RSS users' address database of the RSS reader site as stored in the data mart created. Evaluating

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

sample testing session, the results are presented and analyzed. The results of our experiment indicate that the adoption of K-Nearest Neighbor model can lead to more accurate recommendation that outperformed other classification algorithms. In most cases the precision rate or quality of recommendation is equal to or better than 70%, this means that over 70% of news recommended to a client will be in line with his immediate requirement, making support to the browsing process more genuine, rather than a simple reminder of what the user was interested in on his previous visit to the site as seen in path analysis technique.

The findings of the experimental study can now be used by the designer and administrator of the web site to plan the upgrade and improvement of the web site, in order to ease navigation on the site without too many choices at a time as well as meeting their needed information without expecting them to ask for it explicitly, therefore improving the impressiveness of the web site.

6. Conclusion

Our work 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 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.

The result of our experiment shows that an automatic Real-Time recommendation engine powered by K-NN classification model implemented with Euclidean distance method is capable of producing useful and a quite good and accurate classifications and recommendations to the client at any time based on his immediate requirement rather than information based on his previous visit to the site.

7. Recommendation for future work

Our designed system is a proof-of-concept, prototype of idea for using web usage data mining with K-NN technique, and there are some aspects in which it can be improved in any future work. The study could be taken much further by investigating the users RSS address URL of the RSS reader in a continuous basis.

More research also need to be carried out on many other data mining techniques, comparing the result with this model, so as to determine the most effective model in handling a problem of this nature in the nearest future.

8. Uncited references

[4,5].

Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities (No. 201413065) and The National Key Technology R&D Program (No. 2012BAH117F03)

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aci.2014.10.001.

References

[1] 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.
[2] F. Bounch, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, S. Ruggier, Web log data warehourseing and mining for intelligent web caching, J. Data Knowledge Eng. 36 (2001) 165–189, PH:S0169-023x(01)00038-6.
[3] L.D. Catledge, J. Pitkow, Characterizing browsing strategies in the world wide web, J. Comput. Networks ISDN Syst. 27 (6) (1995) 1065–1073, doi: 101016/0169-7552(95)00043-7.
[4] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining World Wide Web browsing patterns, J. Knowledge Inform. Syst. 1 (1) (1999) 1–27.
[5] R. Cooley, P.N. Tan, J. Srivastava J, Discovery of Interesting Usage Patterns from Web Data, International Workshop on Web Usage Analysis and User Profiling, ISBN 3-540-67818-2, p. 163–182.
[6] A. Dario, B. Eleno, B. Giulia, C. Tania, C. Silvia, M. Naeem, Analysis of diabetic patients through their examination history, J. Expert Syst. Appl. 40 (2013) 4672–4678, http://dx.doi.org/10.1016/j.eswa.2013.02.006.
[7] H. David, M. Heikki, S. Padhraic, Principles of Data Mining, The MIT press, Cambridge. Massachusetts, London, England, 2001, p. 2–20.
[8] F.N. David, Data mining of social networks represented as graphs, J. Comput. Sci. Rev. 7 (2013) (2012) 1–34, http://dx.doi.org/10.1016/j.cosrev.2012.12.001.
[9] M.F. Federico, L.L. Pier, Mining interesting knowledge from weblog: a survey, J. Data Knowledge Eng. 53 (2005) (2005) 225–241, http://dx.doi.org/10.1016/j.datak.2004.08.001.
[10] C.N. Godswill, A Comprehensive Analysis of Predictive Data Mining Techniques, M.Sc. Thesis, The University of Tennessee, Knoxville, 2006.
[11] L. Habin, K. Vlado, Combining mining of web server logs and web content for classifying users' navigation pattern and predicting users future request, J. Data Knowledge Eng. 61 (2007) (2006) 304–330, http://dx.doi.org/10.1016/j.datak.2006.06.001.
[12] M.K. James, R.G. Michael, A.G. James, A fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on System Man and Cybernetics, vol. SMC-15 No4.[0018-9472/85/0700-0580$01.00], 1985.
[13] H. Jiawei, K. Micheline, Data mining concept and Techniques, second ed., Morgan Kaufmann Publishers, Elsevier inc., USA San Francisco, CA 94111, 2006, p. 285–350.
[14] E.P. Leif, K-Nearest Neighbor. Scholarpedia 4(2):1883. Downloaded 27-04-2014,@ www.google.com, 2009.
[15] C. Luca, G. Paolo, Improving classification models with taxonomy information, J. Data Knowledge Eng. 86 (2013) (2013) 85–101, http://dx.doi.org/10.1016/j. datak.2013.01.005.
[16] T. Luigi, S. Giacomo, Mining frequent item sets in data streams within a time horizon, J. Data Knowledge Eng. 89 (2014) 21–37, http://dx.doi.org/10.1016/j.datak.2013.10.002.
[17] Mathworks Incorporation., MATLAB R2011b (7.13.0.564), Licence Number: 161052, USA, Mathworks Incorporation, 1984–2011.

Please cite this article in press as: D.A. Adeniyi et al., Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, Applied Computing and Informatics (2015), http://dx.doi.org/10.1016/j.aci.2014.10.001

[18] M. Michal, K. Jozef, S. Peter, Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor, J. Proc. Comput. Sci. 1 (2012) (2012) 2273–2280, http://dx.doi.org/10.1016/j.procs.2010.04.255.
[19] K. Mi-Yeon, H.L. Dong, Data-mining based SQL injection attack detection using internal query trees, J. Expert Syst. Appl. 41 (2014) (2014) 5416–5430, dx.doi.org/10.1016/j.eswa.2014.02.041.
[20] MySQL Corporation, MySQL Database Management System Software. USA MySQL/Oracle Corporation, 2008.
[21] NetBeans IDE 7.3, NetBeans java compiler. USA, Java/Oracle corporation, 2008.
[22] A. Niyat, K. Amit, K. Harsh, A. Veishai, Analysis the effect of data mining techniques on database, Journal of advances in Engineering & software 47 (2012) (2012) 164–169, http://dx.doi.org/10.1016/j.advengsoft.2011.12.013.
[23] I.O. Ogbonaya, Introduction to Matlab/Simulink, for engineers and scientist, 2nd edition., John Jacob's Classic Publishers Ltd, Enugu, Nigeria, 2008.
[24] C. Padraig, J.D. Sarah, K-Nearest Neighbor Classifier. Technical Report UCD-CSI-2007-4, University College Dublin, 2007.
[25] H. 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.
[26] D. 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.
[27] 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) 739–747, http://dx.doi.org/10.1016/j.ress.2011.03.006.
[28] 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/10.1016/j.eswa.2012.02.063.
[29] Two Crown Corporation, Introduction to Data Mining and Knowledge Discovery, third ed., Two crown corporation, 10500 Falls Road, Potamac, MD 20854, USA, 1999, pp. 5–40.
[30] Z. Xuejuu, E. John, H. Jenny, Personalised online sales using web usage data mining, J. Comput. Ind. 58 (2007) (2007) 772–782, http://dx.doi.org/10.1016/j.compind.2007.02.004.
[31] M. Zdravko, T.L. Daniel, Data mining the Web, Uncovering patterns in Web content, structure, and usage, John Wiley & sons Inc., New Jersey, USA, 2007, p. 115–132.