ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Distributed Learning Predictors through Web Log Mining using Bayesian Networks

Mr. Suresh
Research Scholar
Karunya University
Coimbatore, India
Suresh.kalaimani@gmail.com

Dr (Mrs) Sujni Paul*
Associate Professor
Karunya University
Coimbatore, India
Sujni_paul@yahoo.com

Mrs.Beulah Christalin Latha
Assistant Professor [S.G]
Karunya University
Coimbatore, India

Abstract: Distributed learning discusses the various strategies in which learners are separated but communicate between themselves through the learning coaches. This makes the learners improve the learning skills and decrease learning times. A large part of hidden information resides in a weblog server (i.e.), user IP address, location, time, Number of entries. Web log designer can analyse these information and rank the weblogs based on this information. In this paper weblog analysis of data is done through distributed system by using the Bayesian networks. The behaviour of a web site's users may change so that the designer can trace out the user behaviour from web log to make predictions, according to the frequent patterns accessed through the log files For the classification purpose Bayesian classifier is used which is based on the probability theory.

Keywords- weblogs; distributed node; distributed learner; Bayesian classifier; mining.

I. INTRODUCTION

Web mining is an application of data mining technique to extract knowledge from web data. Web data is of three type's web content, web usage, web structure. In this research paper a web log data which is a homogeneous dataset is taken and mined for tracking the user preferences [1], [2]. For instance if the website owner has a website, the owner has the authority to analyse the efficiency and growth of the website by the support of website designer. Here web designer is the key role to improve the count hit and access pattern of a website. Weblog files are viewed from which the information can be easily analysed for the user's requirement. In this paper a web site is

Considered and compared with three different countries people requirement through distributed environment. In web log server the information like number of user entry, location, IP address, most visited number of pages, time are considered [2]. Now web site designer can analyse three countries by web log data information to know which country people like this website and can analyse the reason for why the other country people are not keen on the website. So from these types of predictions the website performance can be improved. Then each site identifies the observations that are most likely local and non-local variables and transmits a subset of these observations to a central site. After taking these weblog data from three countries the overall performance of the site is evaluated by performing the mining process. The evaluation is done using a classification algorithm in data mining known as the bayes theorem and theoretical and practical implementations are carried out.

II. RELATED WORKS

Mining from homogeneous data constitutes an important class of Distributed Data Mining (DDM). Kargupta, Philip Chan and Hillol [KPH, 00] proposed the Collective Data Mining (CDM) framework for data mining from distributed heterogeneous data. For learning BN from distributed datasets, Kenji [Ken, 97] introduced an algorithm that can handle the homogeneous distributed learning scenario. In this paper the Learning Management System dataset is taken and this is viewed from the distributed nodes of different countries. An important problem is to learn the Bayesian network from data in distributed sites. The centralized solution to this problem is to download all datasets from distributed sites and based on the learner's interest the materials are sent. Kenji (1997) has worked on the homogeneous distributed learning scenario. In this case, every distributed site has the same feature but different observations.

III. DISTRIBUTED LEARNING

The distributed learning is the method of distributing the dataset available in a centralized repository to distributed resource nodes. Distributed learning is an instructional modelling technique that allows the instructor, student and its content to be located in non-centralized locations [10,11]. As the data is in different location the learning process occurs at different time and place.

A. *Features of distributed learning*

[a] Learners gain greater hold in checking how, when and where their learning occurs. They also increase their level of responsibility for their own learning methods and styles.

[b] Faculty gain greater ability to organise and design environments of distributed learning. In distributed learning environments, it is important to recognize that not only the learning environments can be distributed but rather the students learning experiences are shaped by these distributed networks, and the lecture circulate through these networks [6]. There is a very wide expanding range of information and communication technologies available for the development of distributed learning environments.

B. Emerging technologies

The development of high performance computing and communications is creating new learning schemes through world-wide web. For example, interpersonal interactions across network channels lead to the formation of virtual communities [5]. The advance in computer-supported collaborative learning through multimedia/hypermedia and experiential simulation offer the potential to create shared learning environments available at anyplace and any time.

IV. DATA MINING IN DISTRIBUTED LEARNING

In this paper a distributed homogeneous dataset is taken from the Learning Management system where each site has observations corresponding to a subset of the attributes. It has the various attributes [11] . A naive bayes approach is used to learn a Bayesian Network (BN) from a distributed homogeneous dataset to aggregate all data to a single site and learn a BN from the merged data. Based on the type of the learning material of the learners the materials will be distributed.

Knowledge discovery (or data-mining) is involved by extracting knowledge from distributed databases and this knowledge base discovery uses the machine learning techniques. Distributed learning involves the major parameters such as the communication cost, computation to perform the mining process, knowledge integration. The communication is the time involved for a learner to view the course contents or the question bank from the distributed nodes. The computation is the time taken to perform the naive bayes classification for the prediction purpose. The obtained knowledge is then integrated so that the results of the predictions could be interpreted.

V. BAYESCLASSIFIER FOR DISTRIBUTED LEARNING

Simple Bayesian classifier is simple probabilistic classifier based on applying the bayes theorem also known as optimal; it has the independent attribute in a class [18]. The Bayesian classifier is a simple classification method, which classifies an instance *j* by determining the probability of it belonging to class *C*i.
These probabilities are calculated as:

P (Ci|A1= V1 &......& An= Vnj)

Where 'n' is Number of attributes, 'A' is the set of nodes, 'j' is the number of instance, 'i' is the number of distinct classes.

Where an example is represented as attribute-value pairs of the form Ai=Vi. If there are N independence attributes, then the probability is proportional to: P □ P A □C

When this independence assumption is made then the classifier is called Naive Bayes. The naive Bayes classifier is the simplest method; it assumes that all attributes of the examples are independent of each other given the context of the class. This is also called naive Bayes assumption. It is used to classify the dataset from two unknown classes.

Bayesian Theorem:

It consist of X number of data sample whose class label is unknown. H means hypothesis it's used to test in some way which is either proves or disproves the hypothesis. Concept of hypothesis is a very important part of the scientific method. Hypothesis means guess or possibility.

P (H/X) probability of that the hypothesis H holds given the given observed data sample X.

P (H/X) is the posterior probability H conditioned on X. Bayes theorem is useful in that provides a way of calculating the posterior probability P(H/X), from P(H),P(X) and P(X/H)

P (H/X) =P(X/H) P (H) / P(X)

Example

In this example consider that 5% of the people use a particular website in a day. But according to the web server log file data an 85% of the people used in a day. But also the mining test indicates the 15% of the people visit the web site (the false positives). Suppose in a web log mining used in Bayesian theorem, what's probability they have it.

P (W) = using website
P (M) = mining from weblog files
P (W) = 0.05%, P (W/M) = 0 .85%, p(W/M') = 0.15%, P(W/M) = ?

[a] Using website and mining from weblog files W ∩ M
[b] Using website and No mining W ∩ M'
[c] Not using website and mining from weblog files W' ∩ M
[d] Not using website and No mining W' ∩ M'

P (W/M) = p (W ∩ M) \ P (M).
= (0.05) (0.85)/0.185
= 0.0425/0.185
= 0.2297
P (M) = (0.05) (0.85) + (0.95) (0.15)
= 0.0425 + 0.1425
= 0.185

Bayes theorem

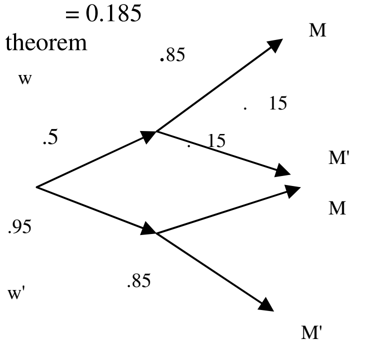


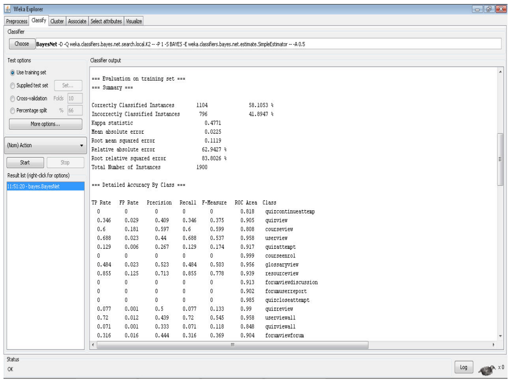
Figure 1. Tree structure of Bayes Theorem

Table I. Sample Dataset

Course	Time	IPAddress	Fullname	Action
09EC214(IIITA)	2009-09-13-05.28	192.168.89.19	07fd068RohanSoans	quizattemp
09EC214(IIITA)	2009-09-13-05.28	192.168.89.19	07fd068RohanSoans	quizview
09EC214(IIITA)	2009-09-13-05.28	192.168.89.19	07fd068RohanSoans	courseview
09EC214(IIITA)	2009-09-13-05.28	192.168.89.19	07fd068RohanSoans	userview
09EC214(IIITA)	2009-09-01-05.52	192.168.4.20	07fd068RohanSoans	quizattempt
09EC214(IIITA)	2009-09-01-05.52	192.168.4.20	07fd068RohanSoans	quizview
09EC214(IIITA)	2009-09-01-05.52	192.168.4.20	07fd068RohanSoans	courseview
09EC214(IIITA)	2009-08-08-18.21	192.168.89.19	07fd068RohanSoans	courseenrol
09EC214(IIITA)	2009-12-06-12.50	192.168.89.22	07fd069AndrewsTito	glossaryview
09EC214(IIITA)	2009-12-06-12.50	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.50	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.50	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.49	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.49	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.49	192.168.89.22	07fd069AndrewsTito	resourceview
09EC214(IIITA)	2009-12-06-12.49	192.168.89.22	07fd069AndrewsTito	courseview
09EC214(IIITA)	2009-08-11-11.17	192.168.89.122	07FD074PraisenEliasPraisenElias	glossaryview
09EC214(IIITA)	2009-08-11-11.17	192.168.89.122	07FD074PraisenEliasPraisenElias	courseenrol
09EC214(IIITA)	2009-08-11-11.17	192.168.89.122	07FD074PraisenEliasPraisenElias	courseview

VI. RESULTS AND DISCUSSION

The results of this research paper are simulated in WEKA tool. WEKA is open source java code created by researchers at the University of Waikato in New Zealand. It provides many different machine learning algorithms. Weka has a comprehensive set of classification tools. Many of these algorithms are very new and reflect an area of active development. This uses the real time data set from the Learning management system as shown in Table 1. The major attributes considered are course, time, ipaddress, fullname and action.



BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Folds 10

% 66

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        1104               58.1053 %
Incorrectly Classified Instances       796               41.8947 %
Kappa statistic                          0.4771
Mean absolute error                      0.0225
Root mean squared error                  0.1119
Relative absolute error                 62.9427 %
Root relative squared error             83.8026 %
Total Number of Instances             1900

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0         0         0          0        0          0.818    quizcontinueattemp
  0.346     0.029     0.409      0.346    0.375      0.905    quizview
  0.6       0.181     0.597      0.6      0.599      0.808    courseview
  0.688     0.023     0.44       0.688    0.537      0.958    userview
  0.129     0.006     0.267      0.129    0.174      0.917    quizattempt
  0         0         0          0        0          0.999    courseenrol
  0.484     0.023     0.523      0.484    0.503      0.956    glossaryview
  0.855     0.125     0.713      0.855    0.778      0.939    resourceview
  0         0         0          0        0          0.913    forumviewdiscussion
  0         0         0          0        0          0.902    forumuserreport
  0         0         0          0        0          0.985    quizcloseattempt
  0.077     0.001     0.5        0.077    0.133      0.99     quizreview
  0.72      0.012     0.439      0.72     0.545      0.958    userviewall
  0.071     0.001     0.333      0.071    0.118      0.848    quizviewall
  0.316     0.016     0.444      0.316    0.369      0.904    forumviewforum
```

Figure 2. WEKA User interface

Figure 2 shows about the user interface of the classification algorithm. Bayes results of classification are obtained by this user interface with the major 5 parameters discussed in the table.

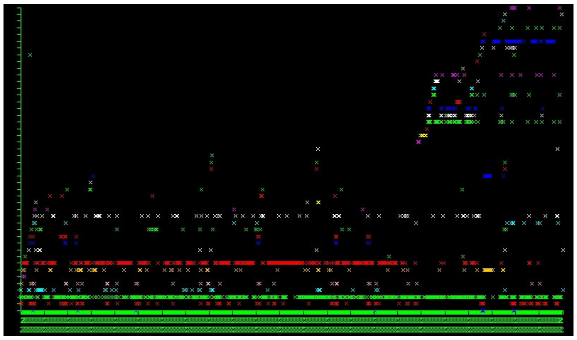


Figure 3. Time Vs.Action

Three distributed locations are considered in this work. From the above graph it is interpreted that the action course view is used by 1896 learners (Green)from all the three countries. In the same way the action resource view is used by 1805 (red) learners. This action is done at the time of 10.37 on 04/08/2009. The final maximum number of learners using the action view forum on

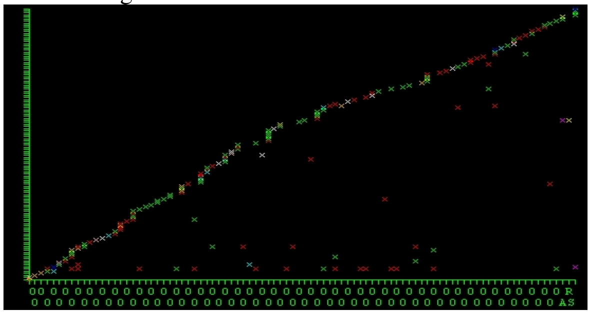


Figure 4. IP address Vs. full name

From the above graph the maximum number of IP address used is 192.168.89.182 and person full name while using the ipaddress was jebastin paul Christopher. This result distribution forms a rough linear curve.

Weka classifier graph Visualizer:

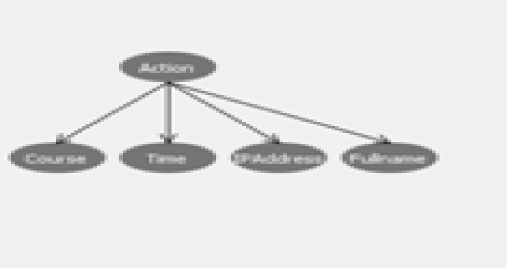


Figure 5. Tree form of the attributes

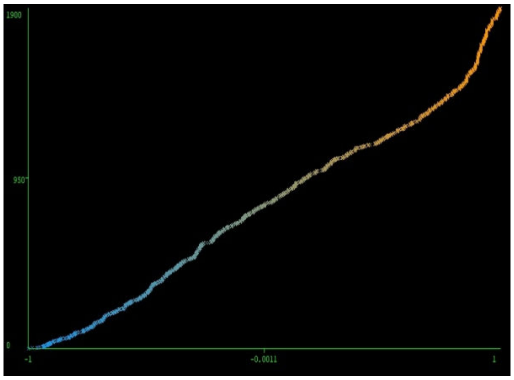


Figure. 6 Prediction margin curve

The prediction margin is the difference between the predicted probability for the actual class and the highest probability predicted for the other classes. The graph shows the overall distribution. From this information web site owner can easily predict that people location and their details.

VII.CONCLUSIONS AND FUTURE SCOPE

In this paper a weblog mining strategy is adopted through distributed environment. By using the bayes classification theorem in weka it is easy to predict the learner's behaviour from different countries. Based on these predictions the learner can improve their website efficiency and make popularity from day to day by using this method. Weblog data is obtained from the Learning Management System. This paper is focussed on a homogeneous data set in distributed environment. In future heterogeneous dataset can be considered.

VIII. REFERENCES

[1] Bouckaert, R. R. (2000), "Properties of bayesian network learning algorithms", in R. L. de Mantaras and D. Poole, eds, `Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence'

[2] Dasgupta, S. (1998), "The sample complexity of learning structure Bayesian networks", Machine learning

[3] G.F cooper and E. Herskovits. "Bayesian method for induction of probabilistic networks from data", machine learning, 2002

[4] R. Chen and K. Sivakumar. "A new algorithm for learning parameters of a bayesian network from distributed data" , 2002.

[5] R. Chen_ K. Sivakumar† H. Kargupta., "Learning Bayesian network structure from distributed data" , 1998

[6] Chen_ K. Sivakumar H. Kargupta., "Learning Bayesian networks from multiple streams", 1988

[7] Remco R. Bouckaert, "*Bayesian network classifiers*" in Weka 2004

[8] H. Kargupta, B. Park, D. Hershberger, and E. Johnson. "Collective data mining: A new perspective toward distributed data mining" , In H. Kargupta and P. Chan, editors, Advances in distributed and Parallel Knowledge Discovery, pages 133–184. AAAI/ MIT Press, Menlo Park, California, USA, 2000.

[9] Un Yong Nahm and Raymond J. Mooney. " A mutually beneficial integration of data mining and information extraction" , In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 627–632, Austin, TX, July 2000

[10] B. Park and H. Kargupta. "Distributed data mining: algorithms, systems, and applications", In Nong Ye, editor, Data Mining Handbook, pages 341– 358. IEA, 200

[11] Michael Wurst and Martin Scholz. "Distributed subgroup discovery", In Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-06), 2006

[12] Margaret H Dunham. "Data mining introductory and advanced topics [M*]*" , Beijing: Tsinghua University Press, 2003, p195-220

[13] G. R. Grimmett and D. R. Stirzaker. "Probability and random processes" Oxford Science Publications, 2ndedition, 2002

[14] B. Liu, W. Hsu, and Y. Ma. "Mining association rules with multiple minimum supports. Proc. of 1999 Int. Conf. on Knowledge Discovery and Data Mining" , August 1999

[15] I-Hsien Ting, Chris Kimble, Daniel Kudenko (2006)"UBB Mining: Finding unexpected browsing behaviour in clickstream data to improve a web site's design"

[16] Nasraoui O., "world wide web personalization" , Invited chapter in "Encyclopedia of data mining and data warehousing", J. Wang, Ed, Idea Group, 2005

[17] Mobasher, B., Cooley, R. and Srivastava, J. (2000) "Automatic personalization based on web usage mining" communications of the ACM, Vol. 43, No.8, pp. 142–151

[18] Eamonn J. Keogh, Michael J. Pazzani "Learning augmented bayesian classifiers:A comparison of distribution-based and classification-based approaches", 2002

AUTHORS

1. Dr (Mrs).Sujni Paul obtained her Bachelors degree in Physics from Manonmanium Sundaranar University and Masters Degree in Computer Applications from Bharathiar University during 2000. She completed her Ph.D in the area Data Mining from Karunya University. She is currently working as Associate Professor in Department of Computer Applications Karunya University,

Coimbatore, India for the past 8.5 yrs to till date. She is guiding 4 Ph.D research scholars and 1 M.Phil scholar. She is working in the area of parallel and distributed data mining..

2. Mr.Suresh ,obtained MCA degree in sri ram engineering college now currently doing M.phil in the Department of Computer Applications Karunya University, Coimbatore, India. My area of research is Data Mining. Now I am working in the of weblog mining through parallel and distributed environment.

3. Mrs.Beulah Christalin Latha is working as Assistant Professor[S.G] in the Department of Computer Applications Karunya University, Coimbatore, India for the past 8.5 yrs to till date. She is pursuing her Ph.D in the area of E-Learning.