ISSN (Online): 2230-7893
www.IJCEM.org

A Heuristic Approach for Web Log Mining using Bayesian Networks

N. B. Kadu[1], D. M. Thakore [2]

[1, 2] Bharti vidyapeeth C.O. E. Pune, India.
kamleshkadu@rediffmail.com, dmthakore@bvucoep.edu.in

Abstract

In the era of globalization and World Wide Web, the Web Applications are playing vital role in our daily life. When more users are using a web application more stress will be applying on the servers. So the whole system or web server may get slow down hereby the process may also slow down or even it may crash down. This makes the users to keep waiting for longer period for the response from the web server. A heuristic approach is presented to catalyze this slow down process by applying mining concept for the web log of the web server, which in deed consist of all the transactions carried out by the end user. By applying Bayesian Networks concept we stream the mined data so perfectly that in the future we can route the end users request to the web server based on the Bayesian results. This makes the system to run on long go and sustain in too busy scenario also.

Keywords: Bayesian Networks, Web Server, Web Log, Data Sets, Data Mining, Graph Model, Node.

I. INTRODUCTION

A national ID card as a general identification document could be used in many different situations, both in dealing with government agencies and private entities. In fact, one of the main goals of introducing a national ID card is the so called 'synergy effect' of replacing multiple identification documents with a single, standardized, and widely recognized document. A prominently discussed use of national ID cards is immigration and border control.

Another important use of national ID cards is to authenticate a person's entitlement to government services. Services such as welfare payments and health could be made dependent on the presentation of an ID card. In addition, voter registries could be based on national ID cards.

In line with the a fore mentioned synergy effect, a national ID card could also replace identification documents that are currently used in the private sector, but that are issued by the government for other purposes - e.g., driver's licenses, health cards, and social insurance numbers. It is likely that private entities would switch to asking customers for their national ID card for identification and age verification, rather than dealing with a multitude of different identification documents such as driver's licenses, health cards or birth certificates.

A citizen card helps in having all the utilities and services under one unique Id. This system not only helps us to know the information about the services or utilities but also it gives the information about the character of the citizen in credit rating while He/she performing the transactions like banking, Electricity, Banking, Insurance, Tax, Provident Fund, Telephone Municipality, Credit rating etc.

Credit rating tells the behavior of the person whether citizen's character is good or bad. The notion of the above stated idea is in itself an exciting prospect. The Ministry of Home Affairs has commissioned TCS, a MNC software consultancy based in India, to do a feasibility study for the National ID card scheme.

This complete system gets the request from the many real time departments like banking, Electricity, Banking, Insurance, Tax, Provident Fund, Telephone Municipality, Credit rating etc through the web portal and these all web portal may perform slow in the high peak hours or on high load in the day to day life.

But all of this is implicitly traced in the web log file of the web server so that mining on these log file and by applying Bayesian Network we can distinguish these incoming request based on the priority so that complete system work smoothly and even we will get better performance of complete system.

Moreover in this proposed system we designed and embedded complete system so that all the real time departments and Citizen identity card department also on the same web application and by applying Bayesain result we succeed to simulate the result on the same platform.

The rest of the paper is organized as follows: section 2 will introduce about related work done so far.

section 3 will give proposed work, section 4 will explain expected results. Section5 will conclude this paper.

II. RELATED WORK

As mentioned earlier some work has been done on this topic. We now review important literature on BN learning. A BN is a probabilistic graphical model that represents uncertain knowledge (Jensen, 1996). Spiegelhalter and Lauritzen (1990) and Buntine (1991) discuss parameter learning of a BN from complete data, whereas Binder, Koller, Russel and Kanazawa (1997) and Thiesson (1995) discuss parameter learning from incomplete data using gradi- ent method. Lauritzen (1995) has proposed an EM algorithm to learn Bayesian network parameters, whereas Bauer, Koller and Singer (1997) describe methods for accelerating convergence of the EM algorithm.

Learning using Gibbs sam- pling has been proposed by Thomas, Spiegelhalter and Gilks (1992) and Gilks, Richardson and Spiegelhalter (1996). The Bayesian score to learn the structure of a BN is discussed by Cooper and Herskovits (1992), Buntine (1991), and Heck- erman, Geiger and Chickering (1995). Learning the structure of a BN based on the Minimal Description Length (MDL) principle has been presented by Bouck- aert (1994), Lam and Bacchus (1994), and Suzuki (1993). Learning BN structure using greedy hill-climbing and other variants was introduced by Heckerman and Gieger (1995), whereas Chickering (1996) introduced a method based on search over equivalence network classes. Methods for approximating full Bayesian model averaging were presented by Buntine (1991), Heckerman and Gieger (1995), and Madigan and Raftery (1994).

Learning the structure of BN from incomplete data was considered by Chick-ering and Heckerman (1997), Cheeseman and Stutz (1996), Friedman (1998), Meila and Jordan (1998), and Singh (1997). The relationship between causality and Bayesian networks has been discussed by Heckerman and Gieger (1995), Pearl (1993), and Spirtes, Glymour and Scheines (1993). Buntine (1991), Fried- man and Goldszmidt (1997), and Lam and Bacchus (1994) discuss how to se- quentially update the structure of a BN based on additional data. Applications of Bayesian network to clustering (AutoClass) and classi_cation has been pre- sented in (Cheeseman and Stutz, 1996; Ezawa and T, 1995; Friedman, Geiger and Goldszmidt, 1997; Singh and Provan, 1995). Zweig and Russel (1998) have used BNs for speech recognition, whereas Breese, Heckerman and Kadie (1998) have discussed collaborative fltering methods that use BN learning algorithms.

Applications to causal learning in social sciences has been presented by Spirtes et al. (1993) An important problem is how to learn the Bayesian network from data in distributed sites. The centralized solution to this problem is to download all datasets from distributed sites. Kenji (1997) has worked on the homogeneous distributed learning scenario. In this case, every distributed site has the same feature but different observations. In this paper, we address the heterogeneous case, where each site has data about only a subset of the features. To our knowledge, there is no significant work that addresses the heterogeneous case.

III. PROPOSED METHOD

In this approach, we used data from real world domain – a web server log data. This approach illustrates the ability of the proposed collective learning approach to learn the parameters of a Bayesian Network from the real world web log data.

Web Server log contains records of user interactions when request for the resources in the servers is received.

Web log mining can provide useful information about different user profiles. This in turn can be used to offer personalized services as well as to better design and organize the web resources based on usage history.

In our application, the raw web log file was obtained from the web server, in which we hosted Our Application Named **Unique Identity Number (UIN)** (Whose Description is given in Introduction) in LAN.

There are mainly three steps in our process.

1) First we preprocess the raw web log file to transform it to a session form, which is amendable to our application. This actually Involves identifying a sequence of logs as a single session. Based on the IP address (or Cookies if available) and time of access. Each session corresponds to the logs from a single user in a single web session. We consider each session as a data sample.

These things can be performed by the Using concepts of Java Sessions, RMI, and And Cookies.

2) In our second step we categorize the resource (Jsp, Video, Audio, text etc) requested from the server into different categories. For our Example, Based on the different resources on the UIN Web Server, We consider mainly eight categories. They are

U-Unique identity Number, T-Telephone, E-Electrical, G- Gas, R-RTO, B-Bank, I-Insurance, P-Pass Port.

These categories are our main features in our implementation, In general, we would have several tens (or perhaps a couple of hundred)of categories, Depending on the web server. This categorization has to be done

ISSN (Online): 2230-7893
www.IJCEM.org

carefully, and would have to be automated for a large web server.

Finally, Each feature value in a session is set to one or zero, depending on whether the user requested resources corresponding to the category. An 8-feature, Binary dataset was thus obtained, which was used to learn a Bayesian Network.

3) A central Bayesian Network was first obtained using the whole dataset. We then split these features into two sets corresponding to the respective scenario in the web server. and then by applying Bayesian Network theorem we will get the mining result.

Bayesian Networks

Bayesian network is a complete model for the variables and their relationships; it can be used to answer probabilistic queries about them. For example, the network can be used to find out updated knowledge of the state of a subset of variables when other variables (the *evidence* variables) are observed. This process of computing the *posterior* distribution of variables given evidence is called probabilistic inference. The posterior gives a universal sufficient statistic for detection applications, when one wants to choose values for the variable subset which minimize some expected loss function, for instance the probability of decision error. A Bayesian network can thus be considered a mechanism for automatically applying Bayesian theorem to complex problems.

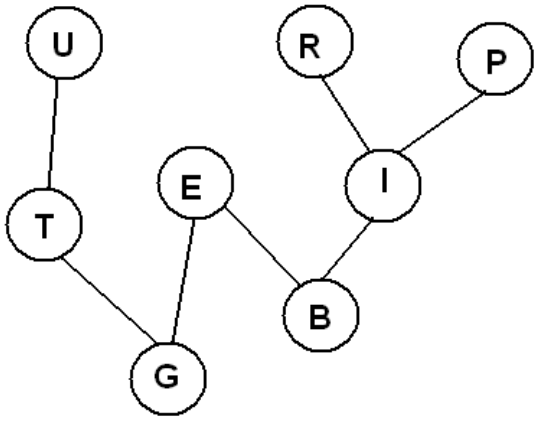
The most common exact inference methods are: variable elimination, which eliminates (by integration or summation) the non-observed non-query variables one by one by distributing the sum over the product; clique tree propagation, which caches the computation so that many variables can be queried at one time and new evidence can be propagated quickly; and recursive conditioning and AND/OR search, which allow for a space-time tradeoff and match the efficiency of variable elimination when enough space is used. All of these methods have complexity that is exponential in the network's tree width. The most common approximate inference algorithms are importance sampling, stochastic MCMC simulation, mini-bucket elimination, loopy belief propagation, generalized belief propagation, and variational methods.
A Bayesian network (BN) is a probabilistic graph model. It can be defined as a pair (G, p), where G = (V; E) is a directed acyclic graph (DAG) (Jensen, 1996; Heckerman, 1998). Here, V is the vertex set which represents variables in the problem and E is the edge set which denotes probabilistic relationships among the variables. For a variable X 2 V, a parent of X is a node from which there is a directed link to X. Let pa(X) denote the set of parents of X, then the conditional independence property can be represented as follows:

$$P(X \mid \mathcal{V} \setminus X) = P(X \mid pa(X)).$$

This property can simplify the computations in a Bayesian network model. For example, the joint distribution of the set of all variables in V can be written as a product of conditional probabilities as follows:

$$P(\mathcal{V}) = \prod_{X \in \mathcal{V}} P(X \mid pa(X)).$$

Then by ASIA model we will get a graph for all the departments that we taken in our scenario which helps to route the end users request through the web server seamlessly. A graph for our proposed system is shown below.



A proposed graph system for Bayesian System

IV Result Analysis

In our proposed system the end result will yields the lesser time for getting the response from the web server for the request raised by the end-user. This exponentially decrease in response time is due to the Bayesian Networks as this theorem will create a graph for the so called departments hits at the web server log data.

So the System will implement this predefined generated graph to re-route the request of the end user to the

appropriate departments or server directly without consuming much more time.

V Conclusion

In this paper problem of enforcing delay in response from the web server on high pay load scenario is successfully handled by using Bayesian networks on collected web log mined data from the real web server.

Here Bayesian network creates a graph based on the priority of the hits collected by the different server or different department on a centralized system. This graph implicitly provides a route or artificial intelligence to the system to re-route the request in busy load timings directly for desired pages .This drastically decreases the response time for the end users based on his/ her request history. This will be useful approach for minimizing the server crash problems or server delay problems in the dominated web word.

References

[1] Abe, N., Takeuchi, J. and Warmuth, M. (1991), Polynomial learnability of probabilistic concepts with respect to the Kullback-Leibler divergence, in `Proceedings of the 1991Workshop on Computational Learning Theory', pp. 277-289.

[2] Bauer, E., Koller, D. and Singer, Y. (1997), Update rules for parameter estimation in Bayesian networks, in D. Geiger and P. Shanoy, eds, `Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann, pp. 3-13.

[3] Beinlich, I., Suermondt, H., Chavez, R. and Cooper, G. (1989), The alarm monitoring system:
A case study with two probabilistic inference techniques for belief networks, in `Proceedings of the Second European Conference on Artificial Intelligence in Medical Care', Springer- Verlag, pp. 247-256.

[4] Binder, J., Koller, D., Russel, S. and Kanazawa, K. (1997), `Adaptive probabilistic networks with hidden variables', Machine Learning 29, 213-244.

[5] Bouckaert, R. R. (1994), Properties of Bayesian network learning algorithms, in R. L. de Man- taras and D. Poole, eds, `Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence', Morgan Kaufmann, pp. 102-109.

[6] Chen, R., Sivakumar, K. and Kargupta, H. (2001a), An approach to online Bayesian learning from multiple data streams, in H. Hargupta, K. Sivakumar and R. Wirth, eds, `Proceedings of the Workshop on Ubiquitous Data Mining: Technology for Mobile and Distributed KDD (In the 5th European Conference, PKDD 2001)', Freiburg, Germany, pp. 31-45.

[7] Zweig, G. and Russel, S. J. (1998), Speech recognition with dynamic Bayesian networks, in `Proceedings of the Fifteenth National Conference on Artificial Intelligence'.

[8] Suzuki, J. (1993), A construction of Bayesian networks from databases based on an MDL scheme, in D. Heckerman and A. Mamdani, eds, `Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence', Morgam Kaufmann, pp. 266-273.