Ijesm
Consulting, help, relaxation
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES AND MANAGEMENT
Vol. 1(1) Sept-Dec (2011)

DYNAMIC PREDICTION OF DATA MINING FOR E-COMMERCE THROUGH MARKOV CHAIN

Sanjay S Bhadoria *, Rohit Chandrawanshi[1]

[*,1] *Patel College of Science & Technology, Bhopal, M.P. – India*

**Corresponding Author*

E Mail: sanjay.bhadoria@gmail.com

ABSTRACT

Data mining has matured as a field of basic and applied research in computer science in general and e-commerce in particular. We limit our discussion to data mining in the context of e-commerce. We also mention a few directions for further work in this domain, based on the survey. In this paper, we introduce a dynamic approach that uses knowledge discovered in previous episodes. The proposed approach is shown to be effective for solving problems related to the efficiency of handling database updates, accuracy of data mining results, gaining more knowledge and interpretation of the results, and performance using Markov chain, named for Andrey Markov, is a mathematical system that transits from one state to another (out of a finite or countable number of possible states) in a chainlike manner

Key Words: E- Commerce, Markov chain.

1. Data Gathering, Cleaning, Preparation

. Data miners in industry estimate that 50% to 80% of efforts are in data gathering and cleaning.

. Complications:

. Many log files distributed in many servers.

. Identifying users, sessions.

. Crawlers & robots.

. Caching.

. Users with multiple computers.

2. Data Sets

We identify .users. using cookies, .visits. by looking for 30 minute gaps in user activity.

. Main data set after preliminary cleaning:

. 1.5 month time frame

. 2,000 users

. 10,000 visits

. 15,000 requests

. 5,000 unique URLs

3. Content Categorization

Summarize URL data using a small number of categories. Example:

Category Explanation

A Homepages

B Software & Accessories

C Shopping

D Special Offers

E General Product Information

F Specific Product Information

G Configure

H Shopping Cart, Quote Sheet

I Checkout

Ijesm Consulting, help, relaxation

J Order Status

Z None of the above

4. Modeling Sequence Data Pattern-based Approach

. Detect common patterns or sequences in training data. A classification scheme can then be based on presence or absence of certain patterns. (Association rules)

. For instance, if sequence .GHI. indicates a likely buy, we may want to look for this sequence in new data.

Pros:

. Good software is readily available for detecting associations and patterns in sequence data.

. Patterns themselves may be of interest to end-user.

Cons:

. Difficult to include additional variables.

. Little probabilistic rigor or meaning.

5. Modeling Sequence Data[Feature Vector Approach]

. Idea: Represent sequences as fixed-dimension vectors of features, then use a standard classification or clustering algorithm

. For instance, a possible feature representation of example visits:

.ABDDFG. → (1,1,0,2,0,1,1,0,0,0,0)

.GGGGGHH. → (0,0,0,0,0,0,5,2,0,0,0)

.HGHIIIIII. → (0,0,0,0,0,0,1,2,6,0,0)

Pros:

. Easy to implement.

. Straightforward to include additional variables such as demographics, time of day, etc.

Cons:

. Ignores sequence information.

. How to make it dynamic?

6. Modeling Sequence Data [Pair wise Distances Approach]

.First define a distance measure between sequences, then classify or cluster based on these distances

Pros:

. Flexible approach.

Cons:

. Obvious distance metrics and algorithms make predicting new sequences computationally demanding.

. Difficult to implement dynamically.

7. Markov Mixture Models

. Assume buy visits are generated by a Markov chain and non-buy visits are generated by a second Markov chain.

. Use Bayes. rule to determine the chain most likely to be the generator of a new sequence.

. Recently used for web log clustering and visualization

Ijesm Consulting, help, relaxation

. Similar to hidden Markov models used in speech processing (see [Rabiner 1989]) and gene sequencing (see [Durbin et.al. 1998]).

8. What is a Markov Chain?

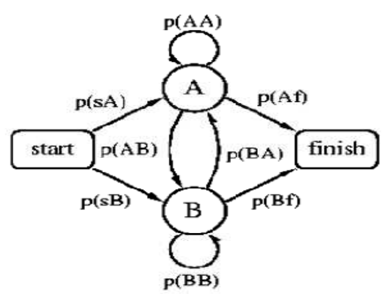
Markov chain is a discrete (discrete-time) random process with the Markov property. Often, the term "Markov chain" is used to mean a Markov process which has a discrete (finite or countable) state-space. Usually a Markov chain would be defined for a discrete set of times (i.e. a discrete-time Markov chain)[1] although some authors use the same terminology where "time" can take continuous values.[2][3] Also see continuous-time Markov process. The use of the term in Markov chain Monte Carlo methodology covers cases where the process is in discrete-time (discrete algorithm steps) with a continuous state space. The following concentrates on the discrete-time discrete-state-space case.

A "discrete-time" random process means a system which is in a certain state at each "step", with the state changing randomly between steps. The steps are often thought of as time, but they can equally well refer to physical distance or any other discrete measurement; formally, the steps are just the integers or natural numbers, and the random process is a mapping of these to states. The Markov property states that the conditional probability distribution for the system at the next step (and in fact at all future steps) *given* its current state depends only on the current state of the system, and not additionally on the state of the system at previous steps.

. Probabilistic models of a system which is assumed to transition between a discrete set of states.

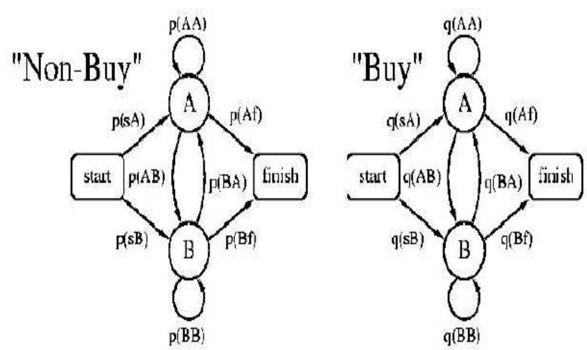
. Future behavior depends only on current state, not on past.

. In our case, states in chain correspond to clicks in a visit.



MMM for Dynamic Prediction Illustration

. Assume visit data are generated by one of two Markov chains:



. Is a given sequence generated by buy model or non-buy model?

MMM for Dynamic Prediction Computation

. **Fitting model parameters**: Maximum likelihood fit is simply transition matrices observed in training data.

. **Classifying a new sequence**: To classify .AAB, use Bayes. rule:

$$\Pr\{Buy \mid \text{"}AAB\text{"}\} = \frac{\Pr\{\text{"}AAB\text{"} \mid Buy\}\Pr\{Buy\}}{\Pr\{\text{"}AAB\text{"} \mid Buy\}\Pr\{Buy\} + \Pr\{\text{"}AAB\text{"} \mid Non-Buy\}\Pr\{Non-Buy\}}$$

where $\Pr\{Buy\}$ and $\Pr\{Non\text{-}Buy\}$ estimated apriori.

$$\Pr\{\text{"}AAB\text{"} \mid Non-Buy\} = p(sA)p(AA)p(AB)$$

$$\Pr\{\text{"}AAB\text{"} \mid Buy\} = q(sA)q(AA)q(AB)$$

9. Summary:

We have solved Clickstream analysis problem is real, interesting, and useful. We have developed Dynamic prediction methods through Markov Chain**.**

Markov chains are used in Finance and Economics to model a variety of different phenomena, including asset prices and market crashes. The first financial model to use a Markov chain was from Prasad *et al.* in 1974.[10] Another was the regime-switching model of James D. Hamilton (1989), in which a Markov chain is used to model switches between periods of high volatility and low volatility of asset returns.[11] A more recent example is the Markov Switching Multifractal asset pricing model, which builds upon the convenience of earlier regime-switching models.[12] It uses an arbitrarily large Markov chain to drive the level of volatility of asset returns. Dynamic macroeconomics heavily uses Markov chains. An example is using Markov chains to exogenously model prices of equity (stock) in a general equilibrium setting. Leontief's Input-output model is a Markov chain.

10. References :

[1] N R Srinivasa Raghavans¯*adhan*¯*a* Vol. 30, Parts 2 & 3, April/June 2005, pp. 275–289. © Printed in India.

[2] Quanten, S., De Valck, E., Mairesse, O. et al., Individual and Time- Varying Model Between Sleep and Thermoregulation. J Sleep Res 15:243-244, 2006

[3] Dynamic Data Analysis and Data Mining for Prediction of Clinical Stability Kristien Van Loona, Fabian Guiza b , Geert Meyfroidt c, Jean-Marie Aerts a, Jan Ramon b, Hendrik Blockeel b, Maurice Bruynooghe b, Greet Van Den Berghe c and Daniel Berckmans a,1 a 1Division Measure, Model & Manage Bioresponses, Katholieke Universiteit Leuven, Kasteelpark Arenberg 30, B-3001 Leuven, Belgium b Department of Computer Sciences, Celestijnenlaan 200a, B-3001 Leuven, Belgium. C Department of Intensive Care Medicine, University Hospital Gasthuisberg, Herestraat 49, B-3000 Leuven, Belgium

[4] Box, G. E., Jenkins, G. M., and Reinsel, G. C., Time series analysis: forecasting and control. New Jersey. Prentice-Hall International, 1994

Ijesm Consulting, help, relaxation

[5] Rangayyan, R.M., Biomedical Signal Analysis: A Case Study Approach. New York. Wiley Interscience, 2002

[6] Seeger, M., Gaussian Processes for Machine Learning. *Int J Neural Syst* 14(2): 69-106, 2004

[7] Minka, T. P., A Family of Algorithms for Approximate Bayesian Inference. PhD Thesis, Masachusetts Institute of Technology, 2001

[8] Applications of Data Mining to Electronic Commerce International Journal of Engineering and Information Technology Vol 2, No. 2 2010 waves publishers ISSN 0975-5292 (Print) IJEIT 2010, 2(2), 140-145 ISSN 0976-0253 (Online)

[9] Hand, D. J., Construction and assessment of classification rules. Chichester, England: John Wiley & Sons; 1997.

[10] Sakai, S., Kobayashi, K., Toyabe, S. I., et al., Comparison of the Levels of Accuracy of an Artificial Neural Network Model and a Logistic Regression Model for the Diagnosis of Acute Appendicitis. *J Med Syst* 31: 357-364, 2007

[11] Erol, F. S., Uysal, H., Ergun, U., et al., Prediction of Minor Head Injured Patients Using Logistic Regression and Mlp Neural Network. *J Med Syst* 29:205-15, 2005

Ijesm Consulting, help, relaxation