USERS INTEREST PREDICTION MODEL
Based on 2nd Markov Model and Inter-transaction Association Rules

Yonggong Ren[1] and Alma Leora Culén[2]
[1] *School of Computer and Information Technology, Liaoning Normal University, Dalian, China*
[2] *Department of Informatics, University of Oslo, Oslo, Norway*

Keywords: Web Data Mining, Inter-transaction Association Rules, Dual-strategy Database, Dual Strategy Users Interest Prediction Model.

Abstract: The 2nd Markov Model and inter-transaction association rules are both known as key technologies for building user interest prediction models. The use of these technologies potentially improves the users surfing experience. The use of the 2nd Markov Model increases the accuracy of predictions, but it does not cover all the data. Therefore, in this paper we propose a dual strategy for a user interest prediction model that includes the entire data set and improves the accuracy of inter-transaction association rules. The foundation of our dual strategy is a new method of building a database based on the degree of user interest. Secondly, we integrate the 2nd Markov Model and inter-transaction association rules for predicting future browsing patterns of users. Experimental results show that this method provides more accurate prediction results than previous similar research.

1 INTRODUCTION

Web data mining may be applied towards helping users more efficiently acquire needed information from the huge quantity available on the Internet. Consequently, web data mining may be used as a means for improving the web surfing experience for users. In this paper we propose a model for predicting users' surfing patterns. Our model gives more accurate predictions than any other research results we have seen.

Much research has been done using the Markov Model to predict users' interests. Some examples relevant to our work in are: (Khalil et al., 2006), (Chimphlee et al., 2006), (Chimphlee et al., 2006), all three using both intra-transaction association rules and the Markov Model, and one that integrates clustering, intra-transaction association rules, and the Markov Model (Khalil et al., 2008). (Ren et. al, 2009) clustering algorithm could be used instead of the one proposed in (Khalil et al., 2008). However, we are aware of no research that shows how the construction of the database can affect the accuracy of prediction and, consequently, the efficiency of browsing. In this paper we propose a dual strategy, based on the degree of user interest, to build a database and then integrate inter-transaction association rules and the 2nd Markov Model. The model resulting from this includes a new database construction method and a new prediction method that we call the Dual-Strategy User Interest Prediction Model (DUIPM).

2 BACKGROUND

Complete web usage data mining results are specified as a set of web pages denoted by P, where $P = \{P_1, P_2 \dots P_n\}$, and n is the total number of pages visited during the mining process. The set of users is denoted by U, where $U = \{U_1, U_2, \dots, U_m\}$ and the users' surfing sessions set is denoted by S, where $S = \{S_1, S_2, \dots, S_m\}$. Each S_i is a set of web pages browsed by user U_i during a surfing session. $S_i = \{(P_1, t_1, \omega_1), (P_2, t_2, \omega_2), \dots, (P_n, t_n, \omega_n)\}$. S_i is clearly a subset of P.

There are two additional parameters associated with each S_i: the time the user spent in session and the weight associated with each visited page. Let t_k be the time which user U_i spent on the page P_k and ω_k the weight associated with page P_k. If $\omega_k = 1$, the user visited the page P_k; if $\omega_k = 0$, the user did not visit the page P_k.

Ren Y. and Culén A..
USERS INTEREST PREDICTION MODEL - Based on 2nd Markov Model and Inter-transaction Association Rules.
DOI: 10.5220/0003659502360241
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2011), pages 236-241
ISBN: 978-989-8425-79-9

Let U_i be a user in U. Let n be the frequency with which the user U_i visited some website, t the time U_i spent on one surfing session, N the total number of visits to that website by all users in U, and T the total time that all users spent on the website. If N and T are large, this may indicate that users like the website.

The *Dual-strategy User Interest Degree (DUID)* is then defined as follows:

$$DUID\ (n, U_i, t) = \frac{n \sum_{U_i \epsilon U} \sum_{t_i \epsilon T, \omega=1} t_i}{NT}.$$

Using all the users' surfing sessions to construct a prediction database, a dual strategy is decided upon as follows: if $DUID\ (n, U_i, t) < minsup\ p$, we use strategy one, else we use strategy two. Strategies one and two are described in section 3.1.

Based on a series of web pages that a user has visited, the Markov Model can predict the page the user will visit next. Let W be a set of pages visited during one surfing session, sorted by the length of visiting time. If some user has already visited i pages, the equation below may be used to calculate what the next page P_{i+1} would most likely be.

The equation $P_{i+1} = P(P_{i+1}\ |W) = P(P_{i+1} = P \mid P_i, P_{i-1}, \ldots, P_1)$ represents the *Markov Model (MM)* for predicting the next page to be visited with the highest probability.

As can be seen from the equation, as W and i increase, the probability $P\ (P_{i+1}|W)$ also grows, and thus the resulting prediction becomes more accurate. But with the larger W and i, the calculations become heavier and consequently the efficiency of the calculations becomes lower. Therefore we introduce a parameter k to control the quantity of data. Adding k leads to the equation known as k^{th}*-Markov Model*:

$$P_{i+1} = P(P_{i+1} = P \mid P_i, P_{i-1}, \ldots, P_{i-(i-k)}).$$

The number of sampled web pages is thus controlled by the parameter k; in fact, the prediction can be made using exactly k pages (see (Chen et al., 2004) and (Mobasher et al., 2002)).

Looking at association rules and their usage in data mining, we use the definition of inter-transaction association rules as given in (Tung et al., 1999). A transaction in our case is simply a set of pages visited by a single user, EP extended set of pages and α confidence.

An implication of the form $X \Rightarrow Y$ is called an *inter-transaction association rule* if

(1) $X \in EP, Y \in EP, X \cap Y \subseteq \phi$
(2) $\exists\ p_i(1) \in X, 1 \leq i \leq n, \exists p_i\ (j\ ) \in Y, 1 \leq i \leq n, j \neq 1$
(3) $\alpha = \sigma\ (X \cup Y)/\sigma\ (X) \geq \min\ conf.$

From the papers (Mobasher et al., 2002) and (Mobasher et al., 2001) we see that these rules give superior results in comparison with other methods involving association rules.

The traditional association rules may be considered as intra-transaction, i.e. they are limited to associations within the transaction. The difference between "intra" and "inter" is that while "intra" remains within the transaction, "inter" can find relationships across different items, thus breaking the barrier of a single transaction. In order to further clarify this distinction, we provide two examples: one of intra-transaction association rules and one of inter-transaction association rules.

Example 1. If user U_1 visits web page A and then page B, by intra-transaction association rules we may be able to conclude that a user U_2, after visiting page A, will visit page B with a probability of 0.5.

Example 2. If user U_1visits web pages A and B, and user U_2 visits web pages A, B and then C, inter-transaction rules may allow for the conclusion that user U_3 will, after visiting page A, visit page C with a probability of 0.2.

In (Deshpande et al., 2004) selective Markov Models are discussed. We will use the following definition: a Markov Model with pruning of redundant data and data whose frequency is less than some minimum confidence is called the *Frequency Pruned Markov Model (FPMM)*.

With these definitions in mind, we are in a position to discuss our dual-strategy user interest prediction method.

3 DUAL-STRATEGY USER INTEREST PREDICTION METHOD

When the value of k is small in a k^{th}-Markov Model, the model is called the low Markov Model. The higher the value of k is, the higher the accuracy, but also the complexity of computations. For this reason, the low model of sufficient accuracy, and high efficiency, if combined with a strategy for increasing accuracy, is to be preferred. One of the main issues with the low Markov Model is that of data coverage. Our approach to remedying this problem is to use inter-transaction association rules and recover some missing association rules covering the missing data. This increases the coverage and thus the accuracy, while it does not increase the complexity of computation associated with the use of higher order

Markov Models.

The *Dual-strategy User Interest Prediction Method (DUIPM)* is then the method that integrates the k^{th}-Markov Model and inter-transaction association rules for low values of k.

For addressing this trade-off between accuracy and complexity of computation, we use the frequency pruned Markov Model to pre-process the data. Referring to (Khalil et al., 2008), accuracy of the 1st to 4th FPMM is compared on four different databases: D1, D2, D3 and D4. Results are shown in Figure 1 and Table 1. As seen from Figure 1, the 2nd FPMM is more accurate than the 1st.

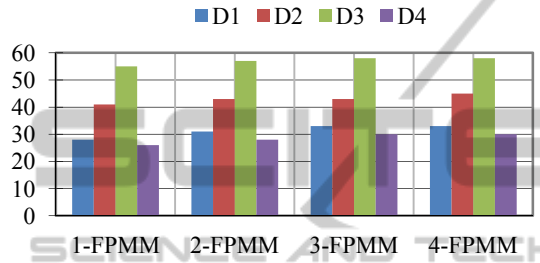

Figure 1: The contrast of the 1st, 2nd, 3rd and 4th FPMM based on accuracy (in percentages).

Table 1 shows that the 2nd FPMM covers the data much better than the 1st FPMM and is closer to the 3rd and 4th. Therefore, we choose the 2nd FPMM for our dual strategy.

Table 1: The data coverage of the 1st- 4th FPMM.

	1-FP	2-FP	3-FP	4-FP
D1	745	9162	14977	17034
D2	502	6032	18121	22954
D3	623	5290	11218	13697
D4	807	7961	19032	23541

3.1 Data Pre-processing

Data from the web log cannot be used directly; part of the data is redundant, and part is not relevant for the computations to follow. Thus, pre-processing of the data is a necessary step in order to increase the efficiency of the algorithms. Some examples of data that need to be eliminated include redundant data, error logs, and graphical, video and audio files.

We now use an example involving four users and their surfing sessions in order to show how we construct our Dual-Strategy Database. We show only the elimination of data by FPMM (redundant data and low frequency data). Table 2 shows the original database, which includes the surfing sessions from four users. The items in each session are all web pages that a specific user has visited. Table 3 shows explicitly which pages were visited by users, in which order, and including the time they spent on the session. Table 4 provides the frequency of every web page. From definition of FPMM, the items whose frequency is less than some minimum frequency value are pruned.

Table 2: The original database.

A,G,T,A,C,S,G,J,R,A,D,H,M,D,J F,D,H,N,I,J,E,A,C,D,H,M,I,J,G,M A,F,I,J,E,C,D,H,N,I,J,G,D,H,N,C,I,J,G,A,N F,L,S,D,H,N,J,Q,E,I,P,C,I,O,A,D,H,M
A,C,G,A,D,H,M,C,F,C,G,R,I,P,H,O,J A,I,J,B,A,E,C,T,D,H,M,I,Q,G
A,F,I,B,A,E,D,H,N,P,I,Q,F,J,D,H,N,G,C F,D,H,M,I,J,E,H,F,I,J,E,D,H,M,A,G,N F,D,H,N,J,A,D,A,E,D,J,R,H,N,G,C,F,G
A,C,D,E,G,C,A,F,N,H,M

Table 3: Surfing sessions for four users.

User	Session	Time
U_1	A,F,I,J,E,C,D,H,N,I,J,G,D,H,N,C,I,J,G,A,N	150s
	F,D,H,N,I,J,E,A,C,D,H,M,I,J,G,M	300s
	F,D,H,M,I,J,E,H,F,I,J,E,D,H,M,A,G,N	120s
	A,C,D,E,G,C,A,F,N,H,M	260s
U_2	A,C,G,A,D,H,M,C,F,C,G,R,I,P,H,O,J	20s
	A,G,T,A,C,S,G,J,R,A,D,H,M,D,J	10s
U_3	A,F,I,A,E,D,H,N,I,F,J,DH,N,G,C	40s
	A,I,J,A,E,C,D,H,M,I,G	50s
	F,D,H,N,J,A,D,A,E,D,J,H,N,G,C,F,G	30s
U_4	F,D,H,N,J,E,I,C,I,A,D,M	10s

Table 4: The frequency of each page.

Page	A	B	C	D	E	F	G	H	I
Freq.	18	2	13	18	9	11	13	18	14

J	L	M	N	O	P	Q	R	S	T
15	1	9	11	2	3	3	3	2	2

Assuming that the minimum confidence value is set to 4, web pages B, L, O, P, Q, R, S and T are eliminated from the database.

When a user U_i visits some web site for the first time, if the parameters from web log satisfy $DUID\ (n, U_i, t) < minsup\ p$, we use database strategy 1 to create the database for predicting the users interest. However, if the parameters from web log satisfy $DUID\ (n, U_i, t) > minsup\ p$, we use database strategy 2 to create the database. Thus, this process of building the database is named Dual-

strategy Database. The result of these two strategies is shown in Tables 5 and 6. Note that Table 5 has the first four rows unchanged, as strategy 2 applies to them. Strategy 1 applies to the remaining rows and the data in these rows is pruned. Strategy 2 applies to user U_1 only, and the result is shown in Table 6, including time user U_1 has spent on each session.

Table 5: The database after strategy 1 is applied.

A,G,T,A,C,S,G,J,A,D,H,M,D,J F,D,H,N,I,J,E,A,C,D,H,M,I,J,G,M A,F,I,J,E,C,D,H,N,I,J,G,D,H,N,C,I,J,G,A,N F,S,D,H,N,J,E,I,C,I,A,D,H,M
A,C,G,A,D,H,M,C,F,C,G,I,H,J A,I,J,A,E,C,D,H,M,I,G
A,F,I,A,E,D,H,N,I,F,J,D,H,N,G,C F,D,H,M,I,J,E,H,F,I,J,E,D,H,M,A,G,N F,D,H,N,J,A,D,A,E,D,J,H,N,G,C,F,G
A,C,D,E,G,C,A,F,N,H,M

Table 6: The database after strategy 2 is applied.

User	Long User Sessions	Time
U_1	A,F,I,J,E,C,D,H,N,I,J,G,D,H,N,C,I,J,G,A,N	150s
	F,D,H,N,I,J,E,A,C,D,H,M,I,J,G,M	300s
	F,D,H,M,I,J,E,H,F,I,J,E,D,H,M,A,G,N	120s
	A,C,D,E,G,C,A,F,N,H,M	260s

3.2 Prediction Strategy 1

The 2nd FPMM is the first component in our Dual-Strategy to predict user's interest. The second component is Dual-Strategy Database construction. Algorithm 1 then describes the first strategy for predicting user's interest, based on the 2nd FPMM and the Dual-Strategy Database construction applying the strategy 1.

```
Algorithm 1:
Input:  The original database.
Output: The results of the 2nd FPMM
For each U_i in U, for each t_i in T,
for ω = 1 and for each S_i in S
//Prepare for constructing a data
  base by strategy 1
 Create (Sum);
//Prepare for constructing a data
  base by strategy 2
 Create (Sum');
//Eliminate redundant data with FPMM
   S_i' = FPMM(S_i);
 Add (S_i') to Sum;
 If DUID (n, U_i, t) > minsup p,
     Add (S_i') to Sum';
 If DUID (n, U_i, t) < minsup p,
     Find the two most frequently
     appearing consecutive items
     Predict using (Sum);
 Else
     Predict using (Sum');
 End
```

3.3 Prediction Strategy 2

Data coverage is, as mentioned above, one of the main shortcomings of the 2nd FPMM, causing the inaccuracy in predicting user's interest. This issue is remedied by integrating inter-transaction association rules. A better prediction of the next web pages that the user may visit is obtained by using combination of inter-transaction rules and the 2nd FPMM.

```
Algorithm 2:
Input:  Dual-strategy Database
     (processed by Algorithm 1).
Output: The web pages most likely to
     be visited next.
When (database≠ null)
Set U as a vector;
//Using 2nd FPMM, find the two most
 frequent consecutive items
U = Find_2ndMM(2, line);
 //Find their next occurrence
Find_continious(U);
 //Get probability P_i for every next
   item
P_i=Get_probability(Find_continuous(U
 ));
  If(P_1 = P2 = ...  = P_n) {
//Extract items from the original
database up to the first occurrence
of most frequent continuous items
and items between the two
occurrences of the most frequent
continuous items in order to find
inter-transaction associations.
   New_database(Begin_of_transaction
After_continious,Before_continious);
 Find_iter_transaction_association_
rules (New_database);
Make_prediction(A,B,C ⇒ D);}
If(P_1 > P_2 > ... > P_n) {
//Record the page with the largest
 probability(P_max)
Remember(Page(P_1));
   New_database(Begin_of_transaction
After_continious,Before_continious);
Find_iter_transaction_association_
rules (New_database);
//Recommend the page which has P_max
//Recommend the results of rules
Make_prediction(Page(P_1));
Make_prediction(A,B,C ⇒ D);}
End
```

4 EXPERIMENTAL RESULTS

4.1 Experimental Data

After data pre-processing, the strategy to use is determined by the DUID. Sessions by users U_2, U_3 and U_4 satisfy the equation $DUID\,(n, U_i, t) < minsup\,p$, thus the strategy 1 is used, while U_1 does not satisfy the equation and therefore the strategy 2 is used (as shown in Tables 5 and 6, respectively). The two consecutive web pages which appear most frequently in the resulting database are I and J. Thus, the probabilities that I and J are followed by E and G, respectively, are:

$$P_{i+1} = argmax\{P(E|I,J)\} = argmax\{E = 0.57\},$$
$$P_{i+1} = argmax\{P(G|I,J)\} = argmax\{G = 0.43\}.$$

The probability of E being the next page is larger than that of G. There still may be some prediction results that algorithm 1 could not uncover, so we use algorithm 2 to find them. Using algorithm 2, a new association rules based database is created and used to predict the next pages.

Table 7: The predictions for user U_1 after visiting pages I and J.

Part of session	2nd FPMM	Prediction
A,F	I,J	E
C,D,H,N	I,J	G
D,H,N,C	I,J	G
F,D,H,N	I,J	E
A,C,D,H,M	I,J	G
F,D,H,M	I,J	E
H,F	I,J	E

The association rule from this database is thus $D \Rightarrow E, F, J$. The prediction result for user U_1 can be E (most likely), then G (see Table 7). If web page D appears in this session, by association rules, the prediction can be E and J.

When a new user visits this web site, the prediction can be made by building database using strategy 1. The results are as follows:

$$P_{i+1} = argmax\{P(M|H,D) = 0.5\},$$
$$P_{i+1} = argmax\{P(N|H,D) = 0.5\}.$$

Thus, the probabilities of both M and N are 0.5.

The inter-transaction association rule in this case is $J, I \Rightarrow M, E$. As before, if a new user visits pages J and I, then we can predict M and E as the next most likely pages to be visited. If the user does not visit pages J and I, pages M and N are most likely web pages to be visited next (see Table 8).

Table 8: The predictions for user U_1 after visiting pages D and H.

Part of session	2nd FPMM	Prediction
A,F,I,J,E,C	D,H	N
I,J,G	D,H	N
F	D,H	N
I,J,E,A,C	D,H	M
F	D,H	M
I,J,E,H,F,I,J,E	D,H	M
A,C,G,A	D,H	M
A,G,A,C,G,J,A	D,H	M
A,F,I,A,E	D,H	N
I,F,J	D,H	N
A,I,J,A,E,C	D,H	M
F	D,H	N
F	D,H	N
J,E,I,C,I,A	D,H	M

4.2 The Performance of the Model

In Figures 2 and 3 we compare different algorithms: the 2nd Markov Model (2nd MM), the 2nd Frequency Pruned Markov Model (2nd FPMM), and our Dual User Interest Prediction Model (DUIPM).

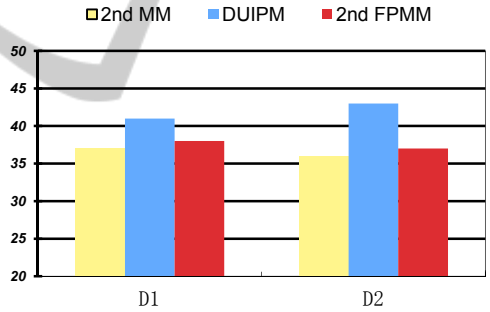

Figure 2: Accuracy based comparison between the 2nd MM, DUIPM and 2nd FPMM.

The 2nd MM and the 2nd FPMM use databases created by strategy 1, while the use of strategy 2 is specific for DUIPM. The 2nd FPMM prunes redundant data, so its prediction results are somewhat more accurate than the ones using only the 2nd MM. The DUIPM is clearly the most accurate method, due to the integration of inter-transaction association rules in strategy 2. Figure 2 shows the accuracy based comparison.

On the other hand, DUIPM is much more complex than the 2nd FPMM and the 2nd MM. It uses two kinds of database construction methods and integrates inter-transaction association rules, so its performance is the lowest. 2nd FPMM has a pruning strategy, so its performance is higher than the 2nd-MM's, as shown in Figure 3.

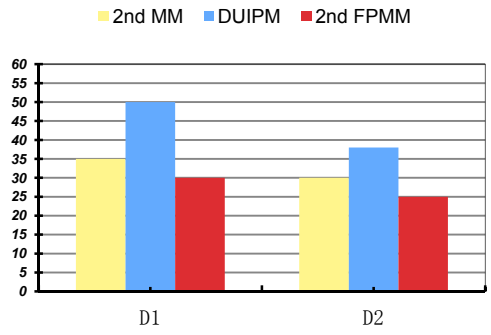

Figure 3: The performance based comparison of the three models.

In summary, the performance of DUIPM is the weakest, but computational power is increasing all the time and it may be justified to sacrifice some performance in favour of better prediction accuracy.

5 CONCLUSIONS

Based on the two methods for database building, integration of the 2nd FPMM and inter-transaction association rules, we propose a different model for users' interest prediction. The database building method and the use of inter-transaction association rules offsets the main disadvantage of the 2nd Markov Model - that it cannot cover enough data to make accurate predictions. As future work, we would concentrate on improving the performance of the algorithm and providing additional experimental results.

ACKNOWLEDGEMENTS

This research was supported by Science and Technology Plan Projects of Liaoning Province Grant No. 2008216014), Science Foundation for the Excellent Youth Scholars of Dalian (Grant No. 2008J23JH026), Liaoning Educational Committee (Grant No. L2010229) and SRF for ROCS, SEM.

REFERENCES

Chen, J. et al., 2004. Discovering Web usage patterns by mining cross-transaction association rules. In *Proceedings of the International Conference on Machine Learning and Cybernetics - Volume 5*, pp. 2655-2660

Chen, J. and Liu, W., 2006. Research for Web Usage Mining Model. In *Proceedings of International Conference on Intelligent Agents Web Technologies and International Commerce*, pp. 8-8.

Chimphlee, S. et al., 2006. Using Association Rules and Markov Model for Predit Next Access on Web Usage Mining. In *T. Sobh & K. Elleithy, eds. Advances in Systems, Computing Sciences and Software Engineering*. Dordrecht: Springer Netherlands, pp. 371-376.

Chimphlee, S. et al., 2006. Rough Sets Clustering and Markov model for Web Access Prediction. In *Proceedings of the Postgraduate Annual Research Seminar,* pp.470-476.

Deshpande, M. & Karypis, G., 2004. Selective Markov models for predicting Web page accesses. In *ACM Transactions on Internet Technology, Volume 4*, pp.163–184.

Khalil, F., Li, J. & Wang, H., 2008. Integrating recommendation models for improved web page prediction accuracy. In *Proceedings of the thirty-first Australasian conference on Computer science - Volume 74*, pp. 91–100.

Khalil, F., Li, J. & Wang, H., 2006. A framework of combining Markov model with association rules for predicting web page accesses. In *Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61*, pp. 177–184.

Mobasher, B. et al., 2002. Using sequential and non-sequential patterns in predictive Web usage mining tasks. In *Proceedings of the IEEE International Conference on Data Mining*, pp.669- 672.

Mobasher, B. et al., 2001. Effective personalization based on association rule discovery from web usage data. In *Proceedings of the 3rd international workshop on Web information and data management*, pp. 9–15.

Ren, Y., & Culén, A. L., 2009. Clustering Based on Data Attribute Partition and Its Visualization. In *Proceedings of the Second International Conferences on Advances in Computer-Human Interactions*, pp. 13–18.

Tung, A. K. H. et al., 1999. Breaking the barrier of transactions: mining inter-transaction association rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 297–301.