

Streaming Association Rule (SAR) Mining with a Weighted Order-dependent Representation of Web Navigation Patterns

Yong Seog Kim, Utah State University

Abstract

This paper presents a streaming association rule (SAR) model based on a weighted order-dependent representation of Web traffic data. We first note that most previous studies in association rules did not explicitly consider visitors' navigation order among Web pages in the process of transforming raw Web traffic data sets into a pre-processed data set for further analyses. As an alternative to a popular boolean representation (0 for visited and 1 for non-visited pages), we explicitly consider navigation order of each visitors to determine similarity among navigation records. We also note that most traditional association rule mining (ARM) models are not scalable because they require multiple scans of data to re-calibrate a predictive model when there are new updates in original databases. The proposed SAR model takes a "divide-and-conquer" approach and requires only single scan of data sets to avoid the curse of dimensionality. We also developed several heuristics to eliminate redundant and insignificant association rules while minimizing the loss of information. According to experiments on a real-world data set, prediction models based on a weighted order-dependent representation returned significantly higher accuracy than models based on a boolean representation. It is also shown that our SAR model returned a very comparable prediction accuracy while maintaining a small fraction of association rules in traditional ARM models.

Keywords: Web mining, association rule, streaming association rule, weighted order-dependent representation

1 Introduction

Web mining has gained much attentions of many practitioners and researchers, where the main goal of Web mining is to discover novel navigation patterns and hidden structures from Web traffic data (Spiliopoulou et al., 2000). The popular techniques traditionally used for Web mining are collaborative filtering (Resnick et al., 1994), clustering (Gündüz and Özsu, 2003), association rules (Agrawal et al., 1993; Agrawal and Srikant, 1995), and Markov models (Sarukkai, 2000). Although all of these techniques have been applied to real-world problems with some success, their storage space and runtime requirements are expensive (Gündüz and Özsu, 2003). Therefore, there is a strong need for developing prediction models that are highly scalable and can overcome the curse of dimensionality.

One of our major contributions is that we develop and present a streaming association rule (SAR) model that makes traditional association rule mining (ARM) models scalable. In order to make application of SAR model on large data sets practical, we take a “divide-and-conquer” approach to improve the storage efficiency of ARM algorithms. To do this, we first divide a large data set into several small chunks of records that can be fitted into a limited memory. Then, we extract association rules from each chunks for local decision making (e.g., developing daily or weekly marketing strategies). At the same time, we incrementally update the final set of association rules extracted from new chunks for global decision making (e.g., developing monthly or yearly strategic plans). Overall, our solution is designed to meet well known data mining desiderata outlined by Fayyad *et al.* (Fayyad et al., 1996) for large-scale prediction methods: iterative (read chunks of the data at a time), scalable (require an approximately constant amount of memory), single pass (read the data set only once), and any-time performance (provide a “best answer” at any moment).

Another major contribution is that we improve the quality of managerial implications (i.e., interpretation of outputs) and practical application (i.e., predicting users’ navigation patterns) of ARM models. Note that traditional ARM models represent each navigation sequences as a series of two integer values, “1” for visited Web pages and “0” for other pages (i.e., a boolean representation) by assuming that each items have the same significance. However, such a simple representation ignores important information about sequential vis-

iting order of Web pages and hence may lead to a prediction model that does not generalize well. Therefore, we consider our SAR model with a “(weighted) ordered” representation of navigation records. Note that in a weighted order-dependent representation, each navigation sequence consists of integer values between “P” (for the first visited page among “P” pages available on the Web site) and “1” (for the last visited page). We also develop several heuristics to eliminate redundant and insignificant association rules while minimizing the loss of information. Comparative experimental results on a real-world dataset show that prediction models based on a weighted order-dependent representation returned significantly higher accuracy than models based on a boolean representation. It is also shown that our SAR model returned a very comparable prediction accuracy while maintaining a small fraction of association rules in traditional ARM models.

The proposed model in this study is different from most quantitative association rule (QAR) (Srikant and Agrawal, 1996; Aumann and Lindell, 2003; Cheung et al., 2005) or weighted association rule (WAR) methods (Tao et al., 2003; Liu et al., 1999; Wang et al., 2000). First, the weight in this study reflects “ordinal” information (i.e., visited order), while the weight in previous studies represents “cardinal” information (e.g., how important each item in the whole sales). Note that definitions and properties based on quantitative attributes in previous studies do not automatically carry to attributes with ordinal properties. Second, several models in (Tao et al., 2003; Liu et al., 1999; Wang et al., 2000) used pre-defined “absolute” weights of each items determined by business end users, while our model use “relative” weights in the sense that each page will have different weights depending on visitors. Therefore, weights in our study do not reflect the value from the end users’ perspective but from visitors’ perspective. Other algorithm suffers the curse of dimensionality (Srikant and Agrawal, 1996; Cheung et al., 2005) and limited representation of association rules (Aumann and Lindell, 2003). Finally, our approach is different from a sequential association mining algorithm in (Agrawal and Srikant, 1995) because their main goal was to discover inter-transaction patterns in a one-time fashion, while our association mining is to discover associations of intra-transactions in an incremental fashion. Further, their model used a simple boolean representation and hence cannot consider weights.

The remainder of this article is organized as follows. Section 2 reviews association rule

and introduce three prediction models based on association rules. In Section 3 we introduce several heuristics to prune association rules and explain the overall prediction procedure. Experimental results are presented and analyzed in Section 4. Finally Section 5 addresses directions of future research and concludes the paper.

2 Association rule mining and prediction models

2.1 Preliminaries on association rule

We first introduce fundamental notations and structures of association rules. Let's first define $\mathfrak{S} = \{i_1, i_2, \dots, i_n\}$ as the set of items where each item represents a product available in a shopping mall or a Web page in a Web site. Then a transaction is defined as a subset of \mathfrak{S} , and a data set, D , consists of a set of transactions. We also use $|D|$ to denote the number of transactions in D and $\text{count}(A)$ to denote the number of transactions containing A that is a non-empty subset of \mathfrak{S} . Then, the support of an itemset A , $\text{support}(A)$, is defined as $\text{count}(A)/|D|$ and an itemset is said to be frequent (or large) if its support is larger than a user-specified value (also called minimum support). Finally, an association is represented in the form $[A \Rightarrow B]$ where $A \subset \mathfrak{S}$, $B \subset \mathfrak{S}$, and $A \cap B = \phi$. For convenience, we refer to A as the assumption (or antecedent) of the rule and B as the consequent of the rule. To measure the interestingness or importance of each association rules, support and confidence are often computed. For a given association rule (R_i), $A \Rightarrow B$, the support and confidence of R_1 is defined as the $\text{support}(A)$ and the $\text{support}(A \cup B)$, respectively. Note that the support of R_i measures the prior probability of the antecedent while the confidence of R_1 measures the conditional probability of the consequent (B) given the antecedent (A). Intuitively, the higher the support of the rule the more prevalent the rule is, and the higher the confidence of the rule the more reliable the rule is (Brijs et al., 1999). Therefore, the problem of mining association rules is to generate all rules that have support and confidence greater than the user-specified minimum support and minimum confidence.

Now, we provide an example to illustrate the limitation of traditional ARM models and to contrast association rules based on a boolean and an ordered representation method.

Data sets	Navigation sequence	$Data^{boolean}$			$Data^{order}$		
		p_1	p_2	p_3	p_1	p_2	p_3
D_1	$p_1 \rightarrow p_2$	1	1	0	3	2	0
	$p_1 \rightarrow p_3 \rightarrow p_2$	1	1	1	3	1	2
	$p_1 \rightarrow p_3 \rightarrow p_2$	1	1	1	3	1	2
D_2	$p_1 \rightarrow p_2$	1	1	0	3	2	0
	$p_1 \rightarrow p_2$	1	1	0	3	2	0
	$p_2 \rightarrow p_1$	1	1	0	2	3	0

Table 1: Boolean and order-dependent representation of navigation patterns

Suppose that we partition the whole data set into two blocks, D_1 and D_2 , and each block consists of three navigation records collected from a Web site that has three Web pages, p_1 , p_2 , and p_3 . Table 1 shows two representation formats for each records. Let’s take an example of $p_1 \rightarrow p_3 \rightarrow p_2$. In a boolean representation, it is represented as $p_1 = 1, p_3 = 1 \Rightarrow p_2 = 1$, meaning that visitors who visited p_1 and p_3 also visited p_2 . In an order-dependent representation, it is represented as $p_1 = 3, p_3 = 2 \Rightarrow p_2 = 1$, meaning that visitors who visited p_1 “first” and p_3 “second” also visited p_2 “third”. We denoted data sets obtained from a boolean and an order-dependent representation as $Data^{boolean}$ and $Data^{order}$, respectively in Table 1.

For these rules to be considered useful, their support and confidence should satisfy pre-defined minimum support and confidence (say, 50% for both criteria). If we represented a given example with a a boolean representation, its support and confidence in D_1 is 67%(=2/3) and 100%(=2/2) (that is, p_1 and p_3 appear in two transactions that always include p_2). With an ordered representation, its support and confidence is also 67%(=2/3) and 100%(=2/2) (that is, $p_1 = 3, p_3 = 2$ appear in two transactions that always include $p_2 = 1$). We computed support and confidence values for all patterns in D_1 , D_2 and a combined data ($D_1 + D_2$), and summarized results in Table 2.

Note that the first rule, $p_1 = 3 \Rightarrow p_2 = 2$, in a boolean representation satisfies minimum criteria in all three data sets (D_1 , D_2 , and $D_1 + D_2$), while in an order-dependent representation, it satisfies criteria in only two databases (D_2 and $D_1 + D_2$). Therefore, two representation methods would return different sets of association rules, which affects the

Data sets	$p_1 = 3 \Rightarrow p_2 = 2$				$p_1 = 3, p_3 = 2 \Rightarrow p_2 = 1$			
	$Data^{boolean}$		$Data^{order}$		$Data^{boolean}$		$Data^{order}$	
	sup	conf	sup	conf	sup	conf	sup	conf
D_1	3/3	3/3	3/3	1/3	2/3	2/2	2/3	2/2
D_2	3/3	3/3	2/3	2/2	0/3	NA	0/3	NA
$D_1 + D_2$	6/6	6/6	5/6	3/5	2/6	2/2	2/6	2/2

Table 2: Association rule candidates with support and confidence values.

prediction accuracy. In contrast, the second rule, $p_1 = 3, p_3 = 2 \Rightarrow p_2 = 1$, satisfies both minimum criteria in D_1 with both representation methods. However, it does not satisfy criteria in $D_1 + D_2$ for both representation methods mainly because low support in D_2 . This indicates that an interesting rule found in one database is not necessarily an useful rule in another database and a local rule from one chunk database can be lost when records in all chunks are considered to find a global set of rules.

2.2 Description of data and prediction models

The data set used in this study is the msnbc data set and its description in this section is a short summary of description available at kdd.ics.uci.edu/databases/msnbc/msnbc.html. This data set comes from Web server logs for msnbc.com and news-related portions of msn.com on September, 28, 1999. Each sequence of nearly one million visitors (989,818 records) in the data set corresponds to page views of a user. In particular, this data set records navigation patterns in visiting order and at the level of URL category to greatly reduce the dimensionality. Each category contains from 10 to 5,000 pages and 17 categories in this data set are “front page”, “news”, “tech”, “local”, “opinion”, “on-air”, “misc”, “weather”, “health”, “living”, “business”, “sports”, “summary”, “bbs” (bulletin board service), “travel”, “msn-news”, and “msn-sports”. We use two terms, pages and URL categories, interchangeably for notational convenience from now on.

Once we transformed the original sequences using into either a boolean or an order-dependent representation, we divided transformed records into 20 sub-data sets with approximately equal number of records. Finally, all records with only one visiting page were

removed from each sub-data sets (a total of 602,945 records) because these records cannot provide any useful information for prediction. As a result, each sub-data set contains about 20,000 records except the last sub-data with 15,639 records. The last data is used for only testing purpose while the first 19 data sets are used for mining association rules. From the first 19 data sets, we constructed the following three prediction models to predict next moves of navigators in both boolean and order representation. We summarized three prediction models as follows:

Model 1 (Simple model): This ARM model consists of association rules found from $Data_{t-1}$ to predict next moves of navigators in $Data_t$. This model is denoted as ARM^{simple} ($bool^{simple}$ or $order^{simple}$).

Model 2 (Model without rule merge): This ARM model consists of all association rules found from all available data sets ($Data_{t-i}$, where $i > 0$) to predict next moves of navigators in $Data_t$, and is denoted as ARM^{wrm} ($bool^{wrm}$ or $order^{wrm}$).

Model 3 (Model with rule merge): This ARM model is different from ARM^{wrm} in the sense that association rules from all data sets are combined and filtered before they are used for prediction. This model is denoted as ARM^{rm} ($bool^{rm}$ or $order^{rm}$).

3 Association rule pruning and prediction algorithm

3.1 Association rule pruning

The straight-forward application of the ARM algorithm is very likely to find association rules such as $p_5=0 \Rightarrow p_{16}=0$, meaning that users who did not visit p_5 did not visit p_{16} . We found in our preliminary experiments that 998 rules out of 1,000 rules specify relationship between non-visited pages when we applied a traditional ARM algorithm to several sub-data sets. However, these rules are not very useful because our main goal is to predict the next pages that visitors are likely to visit. Further, these rules do not provide any insights what causes visitors not to visit certain pages. Therefore, we developed several heuristics to reduce

potentially huge set of redundant and insignificant association rules. We summarized these heuristics as follows:

- Pruning Rule 1: This pruning rule is designed to eliminate association rules that include non-visited pages in their antecedents and consequences. In practice, we modified the original ARM algorithm (Agrawal and Srikant, 1995) in such a way that large itemsets consisting of p_i whose weight is equal to zero do not generate association rules.
- Pruning Rule 2: If a boolean rule R_i^b is a special case of another boolean rule R_j^b , R_i^b is considered redundant and hence pruned. For example, $P_1, P_2 \Rightarrow P_6$ contains $P_1 \Rightarrow P_6$, but does not contain $P_3 \Rightarrow P_6$.
- Pruning Rule 3: All ordered rules of which visited pages in their antecedents are not in the immediately visited sequence are pruned. For example, a rule, $p_1 = 4, p_3 = 2 \Rightarrow p_2 = 1$, is not considered useful in our study because p_3 is not visited immediately after p_1 (if p_3 were visited immediately after p_1 , the weight of p_3 should be 3 ($p_3 = 3$), not 2 ($p_3 = 2$). This pruning rule is applied mainly because we are only interested in sequentially ordered rules only.
- Pruning Rule 4: All ordered rules should satisfy the following relationship between antecedent and consequences: $\min(weights^{ant}) = \max(weights^{con}) + 1$ where $weights^{ant}$ and $weights^{con}$ represent weights of antecedents and consequences, respectively. According to this pruning rule, a rule $p_1 = 4, p_3 = 3 \Rightarrow p_2 = 1$, is pruned. This pruning rule is to focus on the prediction of the very next movement.
- Pruning Rule 5: If an ordered rule R_i^o is contained in another ordered rule R_j^o , R_i^o is redundant and pruned. For example, $P_1, P_2 \Rightarrow P_6$ contains $P_1 \Rightarrow P_6$, but does not contain $P_3 \Rightarrow P_6$. Note also that $P_1, P_2 \Rightarrow P_6$ in an ordered representation does not contain $P_2 \Rightarrow P_6$, while $P_1, P_2 \Rightarrow P_6$ in a boolean representation contains $P_2 \Rightarrow P_6$.

3.2 Association rule mining and prediction

Each association rule provides useful information about relationships among Web pages and helps site administrators re-organize Web contents and structures of Web site to promote

```

Given a set of rules from train data and a minimum similarity threshold,  $\kappa$ 
for each record  $record_i$  in the test data
  for each rule  $rule_j$  compute the similarity between  $record_i$  and  $rule_j$ 
  endfor find the rule,  $rule_{max}$ , with the maximum value of similarity,  $maxSimilarity$ 
  if ( $maxSimilarity > \kappa$ )
    predict the next page request based on  $rule_{max}$ 
  else
    predict the next page request based on popularity-voting
  endif
endfor

```

Figure 1: Summary of prediction algorithm

desired flows of surfers (Pitkow and Pirolli, 1999). Association rules can be also used to predict visitors' next page request. We outlined our prediction algorithm based on association rules in Figure 1. In general, we first computed the similarity between all available rules and the chosen navigation pattern for prediction. Then the best matching rule with the maximum value of similarity is chosen for prediction purpose. However, if no matching rules were discovered, the proposed system relies on the popularity-voting, returning most frequently visited pages among pages that the visitor did not visit yet. There will be no matching rules when the computed similarity is less than the pre-defined minimum similarity threshold, κ .

To compute similarity, we adopted a vector space model that originally used in the field of information retrieval (IR). Note that an association rule R_i can be represented as a vector: $R_i = (w_{i1}, w_{i2}, w_{i3}, \dots, w_{iP})$, where w_{ik} is the weight of p_k in an association rule R_i . Then, similarity between two association rules, R_i and R_j can be defined as the cosine value of the angle between them as follows:

$$similarity(R_i, R_j) = \cos \theta = \frac{R_i \bullet R_j}{|R_i||R_j|} = \frac{\sum_{k=1}^P w_{ik}w_{jk}}{\sqrt{\sum_{k=1}^P w_{ik}^2}\sqrt{\sum_{k=1}^P w_{jk}^2}}$$

4 Experimental results

4.1 Scalability of ARM^{simple} , ARM^{rm} , and ARM^{wrm} models

We used the number of association rules to compare three models in terms of scalability, assuming that the more rules, the more computationally expensive to implement models. Note also that we prefer to provide decision makers with a manageable number of association

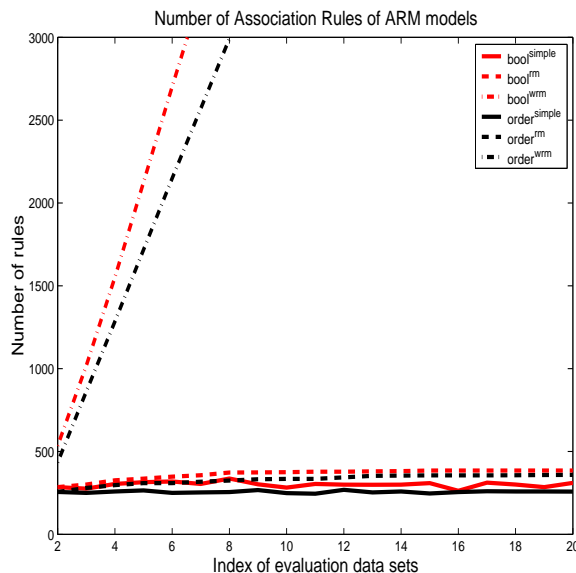


Figure 2: Relationship between models and size of association rule sets

rules rather than inundating them with a multitude of patterns discovered in the data. To report the number of rules for ARM^{simple} models, we extracted a set of association rules from each sub-data sets. To do this, we used 0.01 and 0.1 as minimum support and confidence to find association rules from $Data^{bool}$. The number of rules extracted from 19 training sets varied from 429 to 595 boolean association rules. On average, 536 rules were extracted from each sub-data and a total of 10,187 rules from all 19 sub-data sets were extracted. However, we used much low values (0.001 and 0.01) as minimum support and confidence level to find association rules from $Data^{order}$. It could be argued that the choice for these parameters are rather subjective. However, note that association rules from $Data^{order}$ typically have very low values of support and confidence because the weight of each items in an ordered association rule can vary from 1 to P , while the weight in a boolean association rule is either 0 or 1. Further, association rules with low minimum support and confidence can be more useful for prediction than popularity-voting, although they are not useful to draw managerial insights about navigation patterns. We found that, on average, 427 ordered rules were found in each sub-data and a total of 8,123 rules from all 19 sub-data sets were extracted. The Figure 2 graphically presented our findings.

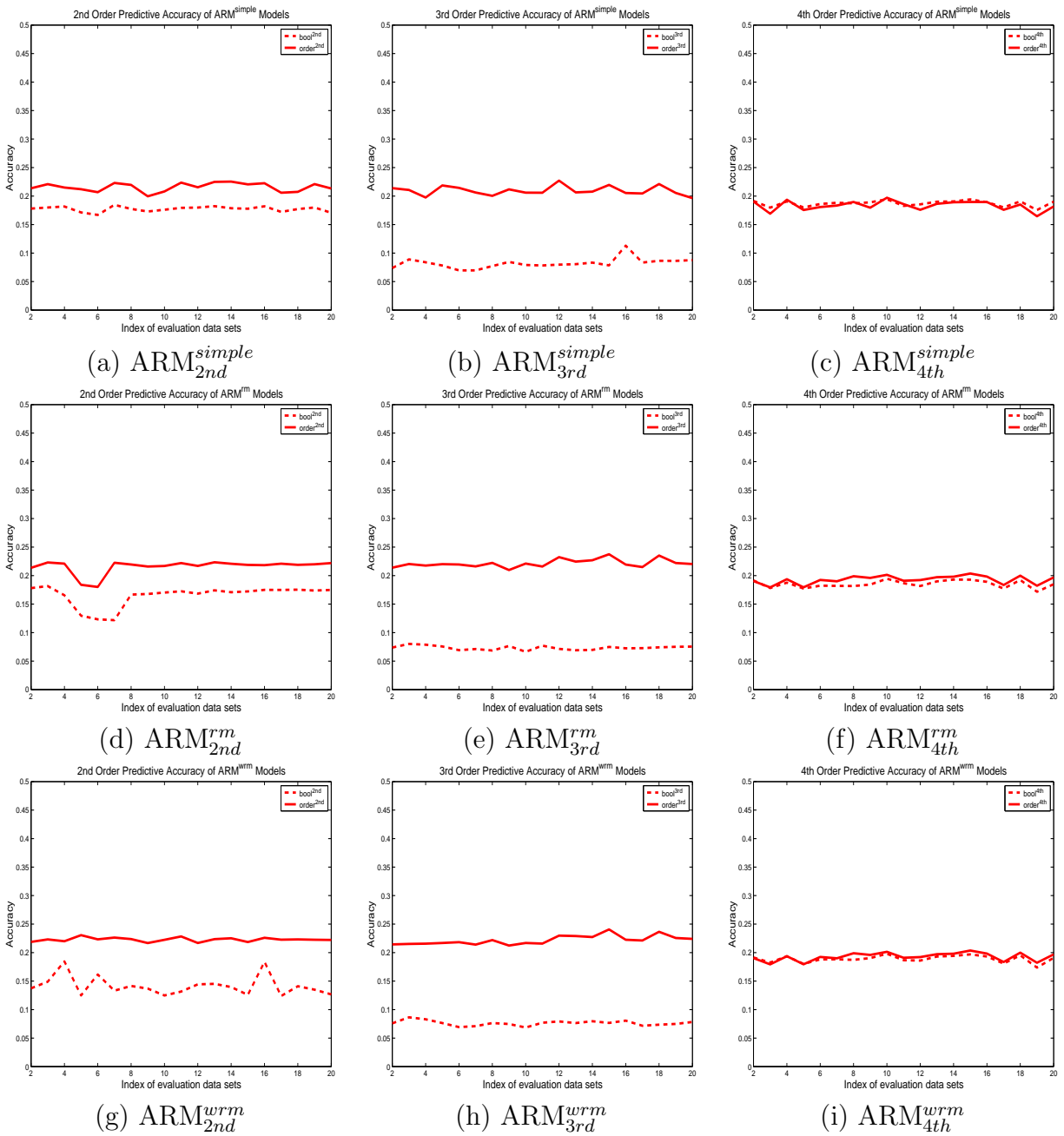
The number of rules in ARM^{wrm} models increase exponentially as more rules are extracted

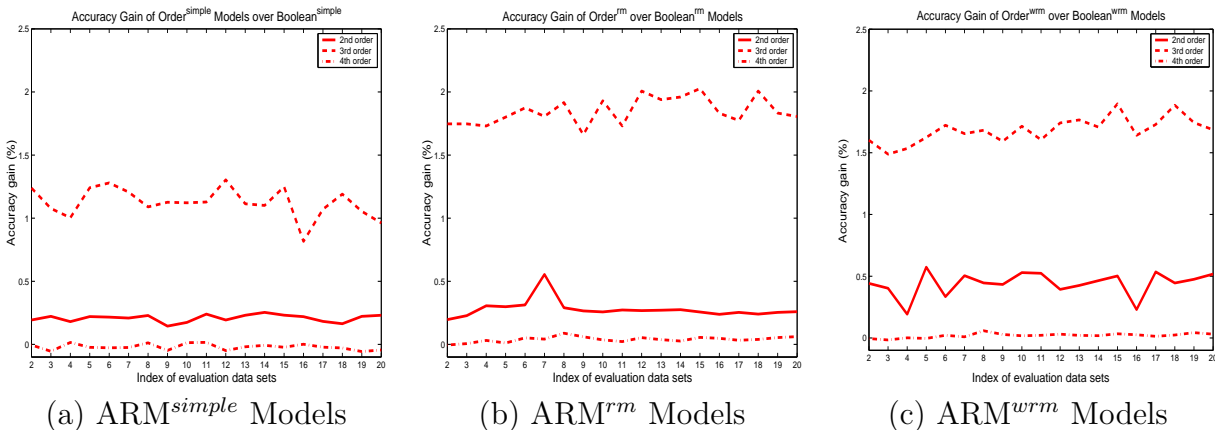
from successive data sets. For example, $order^{wrm}$ model had more than 3,000 rules when only first seven sub-data sets were used to extract association relationship. In fact, the number of rules in an ARM^{wrm} model at time t is equal to the sum of the number of rules that an ARM^{simple} model extracted from each $Data_{t-i}$ where $i > 0$. In contrast, $order^{rm}$ model maintained a constant number of rules mainly because new rules were added to final rule sets only when they are significantly different from association rules extracted earlier. Compared to ARM^{wrm} models, ARM^{rm} models maintained only a fraction of rules in ARM^{wrm} models by eliminating redundant rules. For example, $bool^{rm}$ started with 256 rules and extracted a total of 359 rules when all data sets were used to find rules. We also observed that as more data sets were used to extract rules, the fewer rules were added to the final set of rules. Note that ARM^{simple} models may maintain fewer rules on certain data sets than ARM^{rm} models because they kept different sets of rules for each data sets. However, ARM^{rm} models maintained only one set of 359 rules to predict records in all data sets. Therefore, we concluded that in terms of scalability and generalizability, ARM^{rm} is the best, followed by ARM^{simple} and ARM^{wrm} .

4.2 Prediction accuracy and accuracy gain

In this section, we compared prediction accuracy of ARM^{bool} and ARM^{order} models. Remember that for each boolean and ordered representation, we constructed three models— ARM^{simple} , ARM^{rm} , and ARM^{wrm} —and compared their accuracy in terms of predicting navigators' next page request given information of previously visited pages. Note that more scalable models are not necessarily more accurate models. The Figure 3 showed prediction accuracies of each models for prediction tasks of the secondly, thirdly, and fourthly visited pages in each of 19 evaluation data sets. We simply refer to the n -th (e.g., second) order prediction task as the prediction task to predict n -thly (e.g., secondly) visited page from now on.

Figure 3(a)-(c) contrasted the performance of $bool^{simple}$ and $order^{simple}$ models. We noted that $order^{simple}$ models showed the best performance (21.57%) in the second order prediction, but the worst accuracy (18.33%) in the fourth order prediction. Considering the prediction accuracy of a trivial model is about 6% ($= 1/17$, where 17 is the number of Web

Figure 3: Overall prediction accuracy of ARM^{bool} and ARM^{order} models

Figure 4: Accuracy gain of ARM^{order} over ARM^{bool} models

categories), the performance of $order^{simple}$ is very encouraging. In contrast, the prediction accuracy of $bool^{simple}$ dramatically varied in each prediction tasks, ranging between 8.22% in the third order prediction and 18.72% in the fourth order prediction task. According to t -test, the difference of prediction accuracy between $bool^{simple}$ and $order^{simple}$ models was significant in the second and third order prediction task ($\alpha = 0.01$). We did not expect a significant difference between two models in the second order prediction task because only information given to each models was the-first-visited-page index. However, all experimental results confirmed that even in the second order prediction task, the performance of ARM^{order} was significantly better than that of ARM^{bool}. The most significant performance difference (20.94% vs 8.22%) was found in the third order prediction task. Similar patterns were observed from the comparison of $bool^{rm}$ and $order^{rm}$ models (Figure 3(d)-(f)), and $bool^{wrm}$ and $order^{wrm}$ models (Figure 3(g)-(i)). In conclusion, ARM^{order} models showed significantly superior performance to ARM^{bool} in the second and third order prediction task ($\alpha = 0.01$), but there were no performance difference when prediction order is higher than fourth order. We attributed this finding to the fact that predictions of two models in higher order prediction tasks were made mainly by finding the most frequently visited pages since there were very few rules of higher orders and hence it was difficult to find a matching rule.

To quantify and visualize the improvement of prediction accuracy made by ARM^{order} models compared to the performance of ARM^{bool} models, we computed an accuracy gain

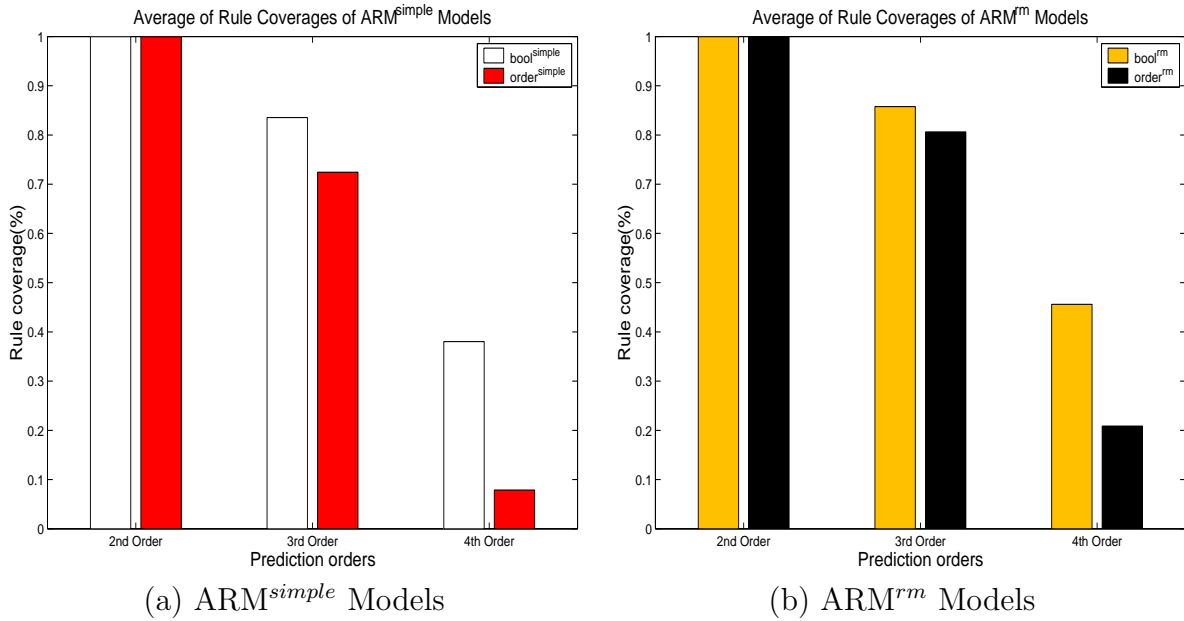
with the following formula: $gain^o = \frac{p^o - p^b}{p^b} \cdot 100(\%)$ where p^o and p^b represent the performance of an ARM^{order} model and an ARM^{bool} model, respectively. Figure 4 showed that $order^{rm}$ and $order^{wrm}$ showed the greatest improvement of prediction accuracy in the third and second order prediction task, respectively. However, none of three models showed significant improvement in the fourth order prediction.

The Figure 3 can be also used to compare ARM^{simple} , ARM^{rm} , and ARM^{wrm} models. For example, Figure 3(a), (d), and (g) showed the prediction accuracy of ARM^{simple} , ARM^{rm} , and ARM^{wrm} models in the second order prediction. We noted that $order_{2nd}^{wrm}$ showed significantly higher prediction accuracy (22.28%) than either $order_{2nd}^{simple}$ (21.57%) or $order_{2nd}^{rm}$ (21.57%) at $\alpha = 0.05$. However, there was no significant difference in terms of prediction accuracy between $order^{rm}$ and $order^{wrm}$ models in the third and fourth order prediction ($\alpha = 0.05$). This is encouraging because $order^{rm}$ models maintained comparable accuracy with a much smaller number of association rules compared to $order^{wrm}$. In contrast, $bool_{2nd}^{wrm}$ returned significantly lower accuracy (14.16%) than $bool_{2nd}^{simple}$ (17.73%) and $bool_{2nd}^{rm}$ (16.51%) models ($\alpha = 0.05$). However, $bool_{3rd}^{wrm}$ performed significantly better than $bool_{3rd}^{rm}$ model ($\alpha = 0.05$), but there was no significant difference between $bool_{4th}^{wrm}$ and $bool_{4th}^{rm}$ models ($\alpha = 0.05$).

4.3 Rule coverage and rule accuracy

Note that the overall accuracy reported in Section 4.2 may not truly reflect the usefulness of association rules because not all records are predicted based on association rules. Some records are predicted based on the popularity of Web pages when no matching rules are found. In fact, the overall accuracy, $accu^t$, can be computed as follows: $accu^t = accu^r \times portion^r + accu^p \times portion^p$, where $(accu^r, portion^r)$ and $(accu^p, portion^p)$ represent a pair of prediction accuracy and record portions predicted by association rules and popularity-voting, respectively. In particular, we used terms, “rule coverage” and “rule accuracy”, to indicate $accu^r$ and $portion^r$, respectively. We anticipated that rule coverage in higher order prediction would be lower than rule coverage in lower order prediction mainly because it is relatively difficult to find matching rules. However, as long as a matching rule is identified in higher order prediction, it is very likely to predict correctly and hence boost rule accuracy.

To validate our claim, we not only computed rule coverages of three models— ARM^{simple} ,

Figure 5: Rule coverages of ARM^{bool} and ARM^{order} models

ARM^{rm}, and ARM^{wrm}—in each one of 19 evaluation data sets, but also computed averages of rule coverages across all evaluation data sets. We first showed averages of rule coverages of ARM^{simple} and ARM^{rm} models in Figure 5(a) and (b). We did not show rule coverages of ARM^{wrm} models because they were identical to those of ARM^{rm} models. The outputs were consistent with our expectation. The rule coverages of both models in higher order prediction were much lower than those in lower order prediction. For example, an average of rule coverage of *bool^{simple}* models in the second order prediction was 99.96%, while the corresponding rule coverage in the third and fourth order prediction was only 83.54% and 38.01%, respectively. This means that more than 60% of records were predicted using popularity-voting in the fourth order prediction with *bool^{simple}* model. We also noted that the rule coverages of ARM^{bool} models were higher than those of ARM^{order} models. For example, in the third order prediction, the rule coverage of *bool^{simple}* model was 83.54%, which was higher than that of *order^{simple}* model. This gap became more significant in the fourth order prediction, 38.01% vs. 7.87%. Similar patterns were observed from comparison of *bool^{rm}* and *order^{rm}* models in the third and fourth order prediction task. This is reasonable because it is much difficult to find a matching ordered rule than a matching boolean rule for a given record.

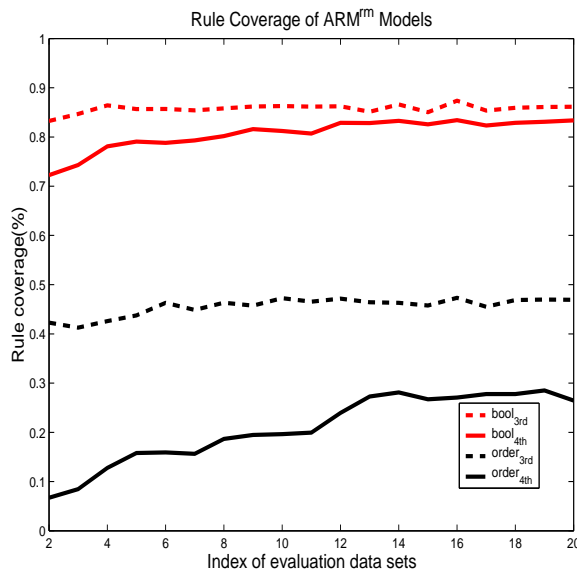
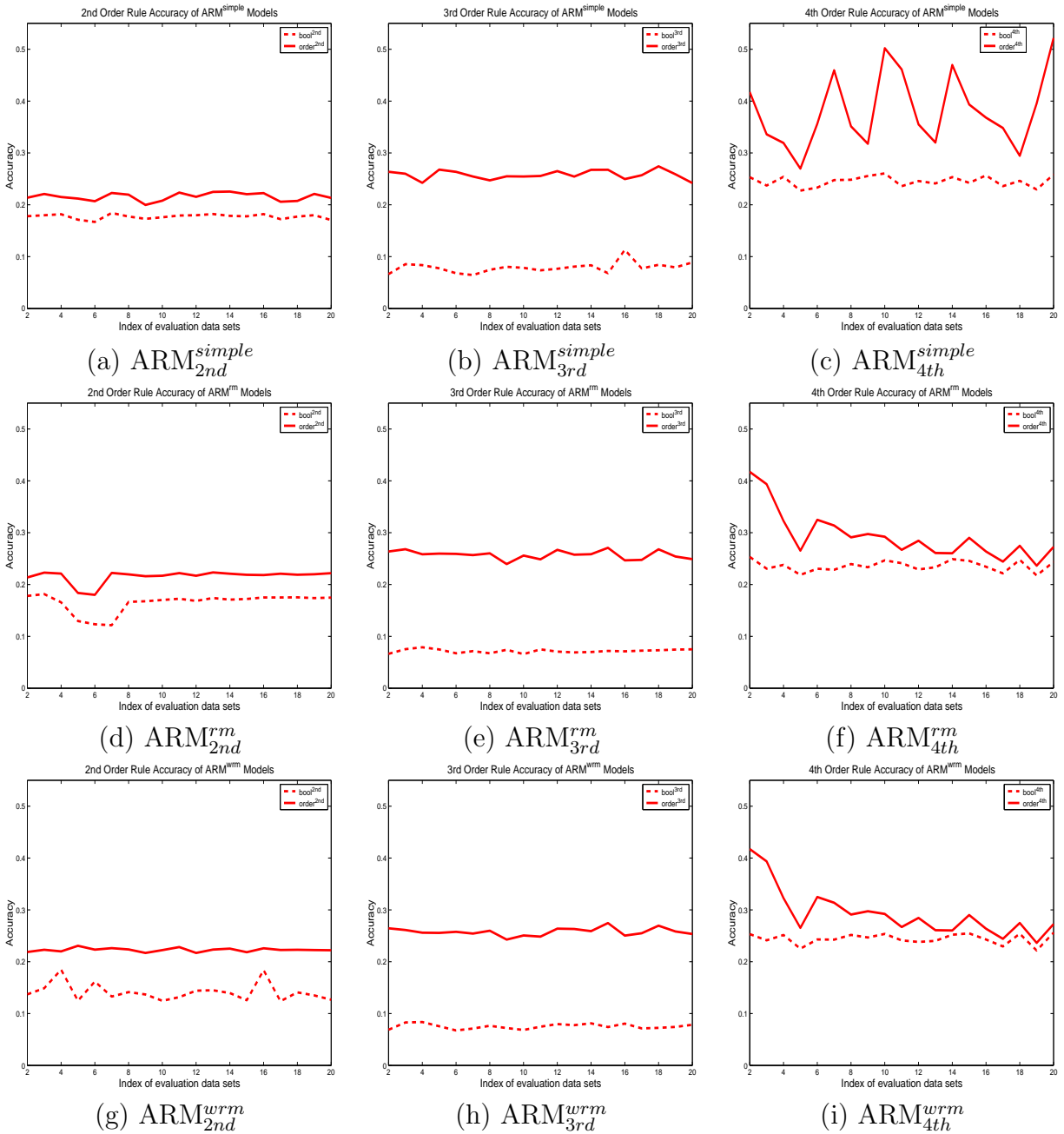


Figure 6: Rule coverages of ARM^{rm} models

In comparison of ARM^{simple} and ARM^{rm} models, both $bool^{rm}$ and $order^{rm}$ models showed higher rule coverage than $bool^{simple}$ and $order^{simple}$ models. This was expected because association rules in ARM^{rm} models are cumulative sets of association rules constructed from multiple data sets. In particular, experimental results confirmed that averages of rule coverages of ARM^{rm} and ARM^{wrm} models were identical, which proved the effectiveness of our heuristics.

Figure 6 showed rule coverages of $bool^{rm}$ and $order^{rm}$ models in each evaluation data sets in the third and fourth order prediction. We did not show coverages in the second order prediction because they were over 99% and there were no interesting trends across evaluation data sets. Note that rule coverages of both $bool^{rm}$ and $order^{rm}$ models steadily increased up to a turning point as new rules from next data sets were extracted, and remained at the same level. For the same order prediction task, $bool^{rm}$ models reached turning points earlier than $order^{rm}$ models. For example, rule coverage of $bool_{3rd}^{rm}$ and $order_{3rd}^{rm}$ model became constant starting from evaluation data D_4 and D_{12} , respectively. We attributed this finding to the fact that $bool^{rm}$ models have very limited space of candidate association rules due to their boolean restriction. We also noted that the rule coverage of the chosen model in lower order prediction reached a turning point earlier than that of the same model in higher order

Figure 7: Rule accuracy of ARM^{bool} and ARM^{order}

prediction. This is mainly because it is relatively difficult to find association rules of higher order.

Finally, we presented rule accuracies of three models in Figure 7. In general, trends of rule accuracies were similar to trends of overall accuracies in the sense that ARM^{order} models

returned significantly higher accuracy ($\alpha = 0.01$) than corresponding ARM^{bool} models for a given order prediction task. We also noted that rule accuracies of ARM^{order} models in the third and fourth order prediction were higher than overall accuracies of the same models shown in Figure 3. For example, the rule accuracy of $order_{3rd}^{simple}$ was 25.79%, which is a significant improvement from its overall accuracy of 20.94%. This is consistent with our notion that the rule accuracy will be higher than the overall accuracy because the rule accuracy was based on prediction made only by the best matching rule.

Several patterns in the fourth order prediction were noteworthy. First, the difference of rule accuracies of ARM^{order} and ARM^{bool} models (refer to Figure 7(c), (f), and (i)) is greater than the difference of overall accuracies of corresponding models (refer to Figure 3(c), (f), and (i)). This is mainly because the rule accuracy of ARM^{order}_{4th} model significantly improved, while the rule accuracy of ARM^{bool}_{4th} model only marginally improved. Second, an average of rule accuracies of $order_{4th}^{simple}$ model across evaluation data sets was the highest, returning an impressive rule accuracy of 38.19%. The trend of rule accuracies of $order_{4th}^{simple}$ models took saw-tooth shapes because the rule accuracy of an $order_{4th}^{simple}$ model on $Data_t$ is heavily dependent on the correlation between $Data_{t-1}$ and $Data_t$. Third, $order_{4th}^{rm}$ and $order_{4th}^{wrm}$ models returned the same level of rule accuracies across all evaluation data. Their rule accuracies over first few data sets were very high (41.73%, 39.36%, 32.27%, and so on), but steadily decreased as more rules with marginal rule accuracies were extracted from new data sets. The average of rule accuracy of both $order_{4th}^{rm}$ and $order_{4th}^{wrm}$ models over all evaluation data was 29.33%.

We also noted that rule accuracies of ARM^{simple}, ARM^{rm}, and ARM^{wrm} in the second order prediction (Figure 7(a), (d), and (g)) were almost identical to those of corresponding models in Figure 3(a), (d), and (g). This is mainly because rule coverages in the second order prediction were almost 100% and association rules were used to predict almost all records in the evaluation data sets. However, averages of rule accuracies in the third and fourth order prediction showed slightly different patterns from prediction accuracies. We noted that none of $order_{3rd}^{simple}$, $order_{3rd}^{rm}$, and $order_{3rd}^{wrm}$ models (25.73, 25.78, 29.33) performed better than other models, while $order_{4th}^{simple}$ performed significantly better than $order_{4th}^{rm}$ and $order_{4th}^{wrm}$ ($\alpha = 0.05$). In comparison of ARM^{bool} models, $bool_{4th}^{simple}$ model was the most accurate

(24.54%), followed by $bool_{4th}^{wrm}$ (24.42%) and $bool_{4th}^{rm}$ (23.59%).

5 Conclusions and future research

We summarized our contributions as follows:

- This paper introduces a novel way to incorporate sequential navigation order to find an interesting associations among Web pages. By weighting pages based on their visited orders, association rules can accurately represent visitors' interests and hence predict the next moves of visitors more accurately than boolean association models.
- This paper also introduces a simple but effective way to make traditional ARM models scalable by incrementally updating association rules without repeatedly scanning the updated databases. The proposed model also provides optimal solutions at anytime.
- The study also provides several heuristics that eliminate redundant rules and generate only rules that satisfy certain constraints. These heuristics significantly reduced the number of association rules, making tradition ARM models scalable.
- Three models— ARM^{simple} , ARM^{rm} and ARM^{wrm} models—are compared. Both ARM^{rm} and ARM^{wrm} models performed better than ARM^{simple} mainly because they utilized cumulative information from multiple data sets. Between ARM^{rm} and ARM^{wrm} models, ARM^{rm} models showed comparable prediction accuracy while maintaining significantly smaller number of rules than ARM^{wrm} models.

In future work, a more accurate prediction model will be studied. For example, we may be able to combine ARM^{bool} and ARM^{order} models for prediction purpose. In the current system, when there are no matching rules that satisfy minimum similarity, both models make a prediction based on popularity. However, in the new system, when no matching ordered rules found, the system searches for matching boolean rules that satisfy the requirement of minimum similarity. If a matching boolean rule were identified, the new system would use it for prediction. However, the new system will make a prediction based on popularity when no matching rules based on either an ordered or a boolean representation. This new system is

based on the idea that prediction accuracy of ARM^{bool} is higher than that of popularity-based prediction, but lower than that of ARM^{order} models. Another direction of future research is to study the effect of the minimum similarity (κ) on the prediction accuracy. Note that models with low values of κ increase rule coverage but are likely to use a rule for prediction although the chosen rule is not predictive. In contrast, models with high values of κ may make all predictions based on only popularity of Web pages because no rules satisfy the requirement of minimum similarity. Therefore, there must exist an optimal value of κ that maximizes prediction accuracy.

References

- Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Int'l Conf. on Management of Data*, pages 207–216, Washington, D.C.
- Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In Yu, P. S. and Chen, A. S. P., editors, *Proc. of 11th Int'l Conf. on Data Engineering*, pages 3–14, Taipei, Taiwan. IEEE Computer Society Press.
- Aumann, Y. and Lindell, Y. (2003). A statistical theory for quantitative association rules. *Journal of Intelligent Information Systems*, 20(3):255–283.
- Brijs, T., Swinnen, G., Vanhoof, K., and Wets, G. (1999). Using association rules for product assortment decisions: A case study. In *Proc. of 5th Int'l Conf. on Knowledge Discovery & Data Mining (KDD-99)*, pages 254–260, New York, NY. ACM Press.
- Cheung, D. W., Lian, W., Yiu, S. M., and Zhou, B. (2005). Density-based mining of quantitative association rules.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in knowledge discovery and data mining*. MIT Press.
- Gündüz, S. and Özsu, M. T. (2003). A web page prediction model based on click-stream

- tree representation of user behavior. In *Proc. of 9th Int'l Conf. on Knowledge discovery & data mining*, pages 535–540, New York, NY, USA. ACM Press.
- Liu, B., Hsu, W., and Ma, Y. (1999). Mining association rules with multiple minimum supports. In *Proc. of 5th Int'l Conf. on Knowledge discovery & data mining*, pages 337–341, New York, NY. ACM Press.
- Pitkow, J. and Pirolli, P. (1999). Mining longest repeating subsequences to predict World Wide Web surfing. In *Proc. of 2nd USENIX Symposium on Internet Technologies & Systems*, pages 11–14.
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. In *Proc. of 1994 ACM Conf. on Computer Supported Cooperative Work*, pages 175–186, New York, NY. ACM Press.
- Sarukkai, R. R. (2000). Link prediction and path analysis using markov chains. In *Proc. of 9th Int'l World Wide Web Conference*, pages 377–386, Amsterdam, The Netherlands. North-Holland Publishing Co.
- Spiliopoulou, M., Pohle, C., and Faulstich, L. (2000). Improving the effectiveness of a Web site with Web usage mining. In Masand, B. and Spiliopoulou, M., editors, *Advances in Web Usage Analysis and User Profiling*, pages 142–162, Berlin. Springer-Verlag.
- Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In *Proc. of 1996 ACM SIGMOD Int'l Conf. on Management of Data*, pages 1–12, New York, NY. ACM Press.
- Tao, F., Murtagh, F., and Farid, M. (2003). Weighted association rule mining using weighted support and significance framework. In *Proc. of 9th Int'l Conf. on Knowledge Discovery & Data Mining (KDD-03)*, pages 661–666.
- Wang, W., Yang, J., and Yu, P. S. (2000). Efficient mining of weighted association rules (WAR). In *Proc. of 6th Int'l Conf. on Knowledge Discovery & Data Mining (KDD-00)*, pages 270–274, New York, NY. ACM Press.