Logic-Based Pattern Discovery

Alex Tze Hiang Sim, Maria Indrawan, Samar Zutshi, *Member*, *IEEE*, and Bala Srinivasan

Abstract—In the data mining field, association rules are discovered having domain knowledge specified as a minimum support threshold. The accuracy in setting up this threshold directly influences the number and the quality of association rules discovered. Often, the number of association rules, even though large in number, misses some interesting rules and the rules' quality necessitates further analysis. As a result, decision making using these rules could lead to risky actions. We propose a framework to discover domain knowledge report as coherent rules. Coherent rules are discovered based on the properties of propositional logic, and therefore, requires no background knowledge to generate them. From the coherent rules discovered, association rules can be derived objectively and directly without knowing the level of minimum support threshold required. We provide analysis of the rules compare to those discovered via the a priori.

Index Terms—Association rules, data mining, mining methods.

1 Introduction

The use of association rule mining technique is to describe the associations among items in a database. These associations represent the domain knowledge encapsulated in databases. Identifying domain knowledge is important because these knowledge rules usually are known only by the domain experts over years of experience. Thus, association rule mining is useful to identify domain knowledge hidden in large volume of data efficiently.

The discovery of association rules is typically based on the *support and confidence framework* where a minimum support (min_sup) must be supplied to start the discovery process [1]. A priori is a representational algorithm based on this framework and many other algorithms are a priori-like. Without this threshold specified, typically, no association rules can be discovered because the procedure to discover the rules will quickly exhaust the available resources.

Nonetheless, having to constrain the discovery of association rules with a preset threshold, in turn, requires in-depth domain knowledge before the discovery of rules can be automated. The use of min_sup generally assumes that:

- a domain expert can provide the threshold value accurately.
- the knowledge of interest must have occurred frequently at least equal to the threshold.
- a single threshold is enough to identify the knowledge sought by an analyst.

In practice, there are cases where these assumptions are not appropriate and rules reported lead to erroneous actions.

In this paper, we propose a novel framework to address the above issues by removing the need for a minimum support threshold. Associations are discovered based on logical implications. The principle of the approach considers that an association rule should only be reported when there is enough logical evidence in the data. To do this, we consider both presence and absence of items during the mining. An association such as $beer \Rightarrow nappies$ will only be reported if we can also find that there are fewer occurrences of $\neg beer \Rightarrow nappies$ and $beer \Rightarrow \neg nappies$ but more of $\neg beer \Rightarrow \neg nappies$. This approach will ensure that when a rule such as $beer \Rightarrow nappies$ is reported, it indeed has the strongest statistical value in the data as comparison was made on both presence and absence of items during the mining process. In addition, the inverse case of customer not buying beer and customer not buying nappies should have statistics that support the rule being discovered due to the logic properties of an equivalence.

By considering this new approach in finding data pattern, a solution toward fulfilling domain-driven data mining requirements [2] can be made. The proposed algorithm suggests a solution in two areas:

1. It eliminates the need to use different intelligence models and its combinations as suggested in [2] to determine appropriate threshold for the mining algorithms. The proposed algorithm discovers the natural threshold based on observation of data set. The different intelligence models can be used in conjunction with the proposed algorithm in determining the target item(s) to be considered during the mining process. Hence, assuming that there are different intelligence models and a way of synthesizing it, the proposed algorithm can incorporate it to determine the target item(s) as an expression of business problem that one wants to solve.

- *A.T.H. Sim is with the Department of Information Systems, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia (UTM), Johor, 81300 UTM Skudai, Malaysia. E-mail: alex@utm.my.*
- *M. Indrawan is with the Caulfield School of Information Technology, Faculty of Information Technology, Monash University, Melbourne, VIC 3145, Australia. E-mail: maria.indrawan@infotech.monash.edu.au.*
- *S. Zutshi is with the Information Systems and eBusiness Department, Faculty of Information Technology, Swinburne University, Lilydale, VIC 3140, Australia. E-mail: szutshi@groupwise.swin.edu.au.*
- *B. Srinivasan is with the Clayton School of Information Technology, Faculty of Information Technology, Clayton Campus, Monash University, VIC 3800, Australia. E-mail: bala.srinivasan@infotech.monash.edu.au.*

Manuscript received 1 Apr. 2009; revised 21 Sept. 2009; accepted 7 Nov. 2009; published online 9 Mar. 2010.
Recommended for acceptance by C. Zhang, P.S. Yu, and D. Bell.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDESI-2009-04-0296.
Digital Object Identifier no. 10.1109/TKDE.2010.49.

1041-4347/10/$26.00 © 2010 IEEE Published by the IEEE Computer Society

2. It provides a logical underpinning to the discovery process of patterns. Currently, the illustration of the mapping of constraints to the discovery process in this paper is based on support value. However, it may be replaced by another constraint. The challenge is in finding appropriate mapping of constraints expressed in different intelligence models into a proportional logic equivalence that is recognized by the proposed algorithm.

The description of our approach is presented in Sections 3-5. In Section 2, we highlight, in detail, the adverse effects of finding association rules passing a min_sup. In Section 3, we explain our general framework to discover association rules that can be mapped to different modes of logic implications in propositional logic. In Section 4, we focus on mining rules that map to a specific mode, logical equivalences, via our proposed framework. In Section 5, we illustrate the use of this specialized framework in finding association rules from a set of transactional records. A corresponding algorithm is provided. In Section 6, we discover rules on a well-known domain and compare the rules found to those discovered via a priori algorithm. We conclude our contributions in Section 7.

2 Issues Using a Minimum Support Threshold

Issues with discovering association rules reverberate around loss of rules and quality of rules discovered. Specifically, if rules are lost, it is misleading to report an incomplete set of rules and at the same time create a sense that all available rules have been found. This situation misleads a decision maker into thinking that only these rules are available which, in turn, will lead a decision maker to reason with incomplete information. For example, it is erroneous to assume that a subset of an incomplete set of rules has the strongest rules. Reasoning with incomplete information while not knowing it may lead to inappropriate conclusion or decisions.

Qualitywise, association rule mining is known to report on every detail of associations among items but unable to identify in specific the type of knowledge rules required. This is especially true when the association rules required mix between those infrequently and frequently observed rules. A large proportion of rules that fall in-between these frequencies of occurrences quickly collude the results discovered. In fact, rules discovered must not be too *rare*; otherwise, the mining process could take forever or its reported results are too large and difficult to process.

We analyze the literature concerning the loss of rules and discovering association rules involving rare items in the following two sections.

2.1 Loss of Association Rules Involving Frequently Observed Items

Some frequent association rules are lost due to the heuristics involved in setting a minimum support threshold. Use of a minimum support threshold to identify frequent patterns assumes that an ideal minimum support threshold exists for frequent patterns, and that a user can identify this threshold accurately. Assuming that an ideal minimum support exists, it is unclear how to find this threshold [3]. This is largely due to the fact that there is no universal standard to define the notion of being frequent enough and interesting.

The strength value of association rules has been occasionally debated in statistics. In one case, Babbie et al. [4, p. 258] discuss the dispute between the authors in [5] and [6]. The dispute centers on the scales used during data analysis. The difference in the scales used by the authors would lead to different levels of effective thresholds being set should the situation be applied to data mining. If the scale in [5] is adopted for mining, then the minimum support threshold set based on [6] would be lower. This case shows that one user's understanding of an ideal strength value may be different from another's.

For data mining, different minimum support thresholds would result in inconsistent mining results, even when the mining process is performed on the same data set. That is, a lower minimum support threshold would result in more association rules being found, and a higher minimum support threshold would result in fewer association rules being found. Some users will find fewer association rules compared to others who use a lower minimum support threshold. For the latter, association rules associated with frequent items should be discovered but are lost. Hence, association rule mining technique discovers some domain knowledge (in form of association rules) subject to the accuracy in determining min_sup. We consider this situation as a case of losing association rules involving frequent items.

The problem of losing frequent association rules may thus have no solution to it apart from lifting the minimum support threshold.

2.2 Loss of Association Rules Involving Infrequently Observed Items

Some infrequent association rules are actionable. Typically, a data set contains items that appear frequently while other items rarely occur. For example, in a retail fruit business, fruits are frequently observed but occasionally bread is also observed. Some items are rare in nature or infrequently found in a data set. These items are called rare items [7], [8], [9]. If a single minimum support threshold is used and is set high, those association rules involving rare items will not be discovered. Use of a single and lower minimum support threshold, on the other hand, would result in too many uninteresting association rules. This is called the *rare item problem* defined by Mannila [10] according to Liu et al. [7].

The latter pointed out that in maintaining the use of a minimum support threshold to identify rare item sets, many users will typically group rare items into an arbitrary item so that this arbitrary item becomes frequent. Another practice is to split the data set into two or several blocks according to the frequencies of items, and mine each block using a different minimum support threshold. Although some association rules involving rare item sets can be discovered in this way, some association rules involving both frequent and rare items across different blocks will be lost.

Instead of preprocessing the transaction records, Liu et al. [7] proposed using multiple minimum thresholds called *minimum item supports* (*MIS*s). A minimum support is set on each item in a data set. Hence, we have a finer granularity of a minimum support threshold compared to the classic approach. Use of *MIS* results in association rules being found in which item sets occur infrequently and below a minimum support threshold. Nonetheless, a user needs to provide an *MIS* threshold for each item. This is arguably difficult to do, especially when the process of providing a minimum item support threshold is ad hoc and requires multiple revisions [11].

Lin et al. [12] devised an approach for finding the minimum support threshold for each item. This approach does not need a preset minimum support threshold but a parameter β that determines the actual discovery of frequent item sets [13]. Yun et al. [13] proposed a minimum support threshold called a *second support* to segregate item sets that occurred infrequently from coincidences, and a *minimum relative support*, which is the maximum of the proportion of the support of an item set against the support of each item within the item set. To search for infrequent and interesting association rules, a user is required to preset a minimum support threshold for a *second support* and a *minimum relative support*. All infrequent item sets that meet both thresholds are then articulated. Another research by Koh et al. [9] devised a calculation to determine a reasonable threshold for a minimum support called a *minimum absolute support* to identify an item set above coincidences. The *minimum absolute support* (*minabssup*) is the least number of collisions above a significance level (that is, 100 percent minus a confidence level p). For example, given that the number of transaction records $\mathrm{N} = 1{,}000$ and item sets x and y occur $\mathrm{a} = 500$ and $\mathrm{b} = 500$ times, respectively, and confidence level $\mathrm{p} = 0.0001$, then the *minabssup* value is 274 [9]. That is, there is 99.99 percent confidence that any co-occurrence below 274 is a product that occurs by chance.

The research by Liu et al. [7], Lin et al. [12], Yun et al. [13], and Koh et al. [9] has been important in establishing a *minimum item support* threshold with finer granularity although different criteria were injected for identifying *minimum item support* values. The common aim, however, was to offset heuristics when setting up a minimum support threshold. In all these approaches, we see that state-of-the-art association rule mining has drifted from the original idea of mining frequent patterns alone to considering other patterns as well. These include patterns above a parameter β that signifies how items should be valued [7], patterns that are not negatively correlated [12], patterns that are infrequent but with a relatively high occurrences [13], and patterns that are above coincidences [9]. Using a minimum support threshold alone cannot identify these patterns specifically.

2.3 Association Rules That Are Measured Using Other Measures of Interestingness

A number of researchers have pointed out that association rules are not necessarily interesting even though they may have at least a minimum support threshold and a minimum confidence threshold. Brin et al. [14] show that association rules discovered using a support and confidence framework may not be correlated in statistics. Webb [15, p. 31] and Han and Kamber [16, p. 260] demonstrate that some association rules are not interesting due to the high marginal probabilities in their consequence item sets. Brin et al. [17] argue that frequently co-occurred association rules (even with high confidence values) may not be truly related. These association rules show item sets co-occurring together, with no implications among them. Scheffer [18] highlights that in many cases, users who are interested in finding items that co-occur together are also interested in finding items which are connected in reality. Having a minimum support threshold does not guarantee the discovery of interesting association rules, as such rules may need to be further processed and quantified for interestingness.

The usage of *leverage* and *lift* are good alternatives in mining association rules without relying on pruning a minimum support threshold. The authors in [3] mined arbitrarily top k number of rules using *lift, leverage*, and *confidence* without using a preset minimum support threshold. The use of *leverage* and *lift* is also fundamental in designing a new measure of interestingness. Among such work, authors in [19] considered *lift* as also one of the 12 interesting criteria to generate interesting rules called an *informative rule set*. Authors in [20] devised a conditional-probability-like measure of interestingness based on *lift* (termed as *dependence*) called the Conditional Probability Increment Ratio ($CPIR$) and used this to discover required interesting rules accordingly.

Apart from the use of *lift*, *leverage*, and its derived measure of interestingness, measures of interestingness that consider a deviation from independence have also been used. These include *conviction*, proposed by Brin et al. [17], *collective strength* as devised by Aggarwal and Yu [21], and *Pearson's correlation coefficient* and *conviction* as reported by Blanchard et al. [22]. *Collective strength* has domain ranges from 0 to infinity. An implication strength of 1 indicates that the strength of the rule is exactly as expected under the assumption of statistical independence. Antonie and Zaïane [23] used *Pearson's correlation coefficient* to search for both positive and negative association rules that have a strong correlation. (By contrast, Lin et al. [12] remove the negative association rules.) The search algorithm [23] found the strongest correlated rules, followed by rules with moderate and small strength values. The ranges of threshold values on the *correlation coefficient* were predetermined. The measure of *conviction* [17] was an improvement over *lift*, which only considers the marginal probability of the consequence item set. A major drawback of *lift* or *leverage* is that they are not asymmetrical measures that measure the real implications of a rule. They do not distinguish the difference between the direction $X \Rightarrow Y$ and $Y \Rightarrow X$ in a rule's strength value. The measures of interestingness *lift* [14] and *leverage* [24] reflect more stringent co-occurrences because association rules found have item sets X and Y being statistically independent of one another. Without the combined use of *confidence*, however, association rules discovered will be nondirectional. In fact, symmetrical measures of interestingness [25] include *collective strength, Pearson's correlation coefficient, Chi-square, and Chi-squared-*based measures such as *Phi* and *Cramer's V*, where the source of a directional relation is unknown. Association rules identified by these interestingness measures in them, without the use of other measures of interestingness, are not implicational because they do not describe the asymmetry strength of a rule between two item sets. In addition, if the new measures of interestingness are used after a preset minimum support threshold, then some association rules could still be lost.

In discovering knowledge rules as in the methods above, we are interested in proposing a framework that does not

TABLE 1
Truth Table for a Material Implication

p	q	$p \supset q$
T	T	T
T	F	F
F	T	T
F	F	T

TABLE 2
Truth Table for an Equivalence

p	q	$p \equiv q$
T	T	T
T	F	F
F	T	F
F	F	T

incur arbitrary threshold values, thus minimizing the caveats of threshold settings.

3 A Generalized Association Rule Mining Framework

We propose a novel association rule mining framework that can discover association rules without the need for a minimum support threshold. This enables the user, in theory, to discover knowledge from any transactional record without the background knowledge of an application domain usually necessary to establish a threshold prior to mining.

To introduce our framework, this section starts with the distinction between an association rule and the different modes of an implication as defined in propositional logic. The topic of implication from logic is raised because our proposed mining model is based on an association rule's ability to be mapped to a mode of implication. If an association can be mapped to an implication, then there is reason to report this relation as an association rule. Otherwise, without a priori such as the minimum support threshold, many association rules would be found, and we would need to report all of them. An implication having a rule where the left-hand side is connected to the right-hand side correlates two item sets together. This implication exists because it is true according to logical grounds, follows a specific truth table value, and does not need to be judged to be true by a user. The rule is reported as an interesting association rule if its corresponding implication is true.

3.1 An Implication

In an argument, the truth and falsity of an implication (also known as a *compound proposition*) ($\rightarrow$) necessarily rely on logic [26]. Each implication, having met specific logical principles, can be identified (for example, one may be a *material implication*, while the other may be an *equivalence*). Each has a set of different truth values. This will be explained later. We highlight here that an implication is formed using two *propositions* p and q. These propositions can be either true or false for the implication's interpretation. For example, "apples are observed in a customer market basket" is a true interpretation if this has been observed. From these propositions, we have four implications

1. $p \rightarrow q$,
2. $p \rightarrow \neg q$,
3. $\neg p \rightarrow q$, and
4. $\neg p \rightarrow \neg q$.

Each is formed using standard symbols "$\rightarrow$" and "$\neg$." The symbol "$\rightarrow$" implies that the relation is a mode of implication in logic, and "$\neg$" denotes a false proposition. We give an example for implications 1 and 2 below:

1. If "apples are observed in a customer market basket," then "bread is observed in a customer market basket" $p \rightarrow q$.
2. If "apples are observed in a customer market basket," then "bread is NOT observed in a customer market basket" $p \rightarrow \neg q$.

The truth and falsity of any implication is judged by "anding" ($\wedge$) the truth values held by propositions p and q. In a fruit retail business where no bread is sold, the implication that relates p and q will be false based on the operation between truth values; that is, $1 \wedge 0 = 0$. The second implication based on the operation will be true because $1 \wedge 1 = 1$. Hence, we say that the latter implication $p \rightarrow \neg q$ is true, but the first implication $p \rightarrow q$ is false. Each implication has its truth and falsity based on truth table values alone.

There are a number of modes of implication [27, pp. 31-58]. We highlight two modes of implication and their truth table values in the next two sections.

3.2 Material Implication

A material implication ($\supset$) meets the logical principle of a *contraposition*. A contrapositive (to a material implication) is written as $\neg q \rightarrow \neg p$. For example, suppose, if customers buy apples, that they then buy oranges is true as an implication. The contrapositive is that if customers do not buy oranges, then they also do not buy apples. If an implication has the truth values of its contrapositive, $\neg(p \wedge \neg q)$, it is a material implication [28, p. 36]. That is, $p \supset q \textit{ iff } \neg(p \wedge \neg q)$.

The truth table for a material implication is shown in Table 1.

3.3 An Equivalence

An *equivalence*($\equiv$) is another mode of implication. In particular, it is a special case of a material implication. For any implication to qualify as an equivalence, the following condition must be met [29, p. 20]: $p \equiv q \textit{ iff } \neg(p \textit{ xor } q)$ where truth table values can be constructed in Table 2.

An equivalence has an additional *necessary* condition [30]. Due to this condition, propositions are now deemed both necessary and sufficient relates with ("iff," in short). One of many ways to prove an equivalence is to show that the implications $p \rightarrow q$ and $\neg p \rightarrow \neg q$ hold true together. The latter is also named an *inverse* [28, p. 48]. Suppose, if customers buy apples, that they then buy oranges is a true implication. The inverse is that if customers do not buy apples, then they do not buy oranges.

We summarize in this section that a typical statement of the format "if ... then" is a *conditional* or a *rule*. If this conditional also meets specific logical principles with a truth table, they are an *implication*. Among many modes of implications, a *material implication* relates propositions together. An *equivalence* is a special case of the former, where propositions are necessarily related together all the time and are independent of user knowledge. In other words, equivalence is necessarily true all the time and judged purely based on logic. We are interested in finding association rules that map to this equivalence. By mapping to this equivalence, we can expect to find association rules that are necessarily related with true implication consistently based on logic. These are the association rules deemed interesting. In addition, the process of finding such association rules will be independent of user knowledge because the truth and falsity of any implication is based purely on logical grounds.

3.4 Mapping Association Rules to Equivalences

We have previously explained the distinctions between an association rule and an implication. We have also highlighted the motivation to map an association rule to an implication. This section explains how to map an association rule to an equivalence.

A complete mapping between the two are realized in three progressive steps. Each step depends on the success of a previous step. In the first step, item sets are mapped to propositions in an implication. Item sets can be either observed or not observed in an association rule. Similarly, a proposition can either be true or false in an implication. Analogously, the presence of an item set can be mapped to a true proposition because this item set can be observed in transactional records.

Having mapped the item sets, an association rule can now be mapped to an implication in a second step. An association rule has four different combinations of presence and absence of item sets. Similarly, there are four different implications depending on the truth value held by its propositions. Hence, an association rule can be mapped to an implication that has a truth value (either true or false).

Finally, in a more specific implication, the association rule has a set of four truth table values. Having mapped item sets and association rules, we can now map association rules into specific modes of implication that have predefined truth table values. We focus on equivalence. Based on a single transaction record in association rule mining, we show the mapping from association rules to equivalences below.

3.4.1 *Mapping Using a Single Transaction Record*

An item set has two states. In a single transaction record, an item can either be present or absent from the transaction record. It follows then that a proposition can either be true or false. If an item set is observed in a transaction record, it is analogous to having a true proposition. In the same way, item sets are mapped to propositions p and q as follows:

- Item set X is mapped to $p = \mathrm{T}$, if and only if X is observed.
- Absence of item set X, that is, $\neg X$ is mapped to $p = \mathrm{F}$, if and only if X is not observed.
- Item set Y is mapped to $q = \mathrm{T}$, if and only if Y is observed.
- Absence of item set Y, that is, $\neg Y$ is mapped to $q = \mathrm{F}$, if and only if Y is not observed.

TABLE 3
Mapping of Association Rules to Equivalences

Equivalences:	$p \equiv q$	$\neg p \equiv \neg q$
Association Rules:	$X \Rightarrow Y$	$\neg X \Rightarrow \neg Y$

True or False on Association Rules	**Required Conditions (to map associations to equivalences)**	
T	$X \Rightarrow Y$	$\neg X \Rightarrow \neg Y$
F	$X \Rightarrow \neg Y$	$\neg X \Rightarrow Y$
F	$\neg X \Rightarrow Y$	$X \Rightarrow \neg Y$
T	$\neg X \Rightarrow \neg Y$	$X \Rightarrow Y$

Each component of an association rule is now mapped to propositions. Using the same mapping concept, an association rule can be mapped to a true or false implication. An association rule consists of two item sets X and Y. Following the mappings above:

- Item sets X and Y are mapped to p and $q = \mathrm{T}$, if and only if X and Y are observed.

That is, an association rule $X \Rightarrow Y$ is mapped to an implication and is deemed interesting if and only if both item sets are observed from a single transaction record. Similarly, all four mappings from the association rule to its implications are given below:

- $X \Rightarrow Y$ is mapped to implication $p \rightarrow q$, if and only if both X and Y are observed.
- $X \Rightarrow \neg Y$ is mapped to implication $p \rightarrow \neg q$, if and only if X is observed and Y is not observed.
- $\neg X \Rightarrow Y$ is mapped to implication $\neg p \rightarrow q$, if and only if X is not observed and Y is observed.
- $\neg X \Rightarrow \neg Y$ is mapped to implication $\neg p \rightarrow \neg q$, if and only if both X and Y are not observed.

Having mapped association rules to implications, we use the same mapping concept to map association rules to equivalences based on specific truth table values. An equivalence has truth table values (T,F,F,T) (see Table 2) for implications $p \rightarrow q$, $p \rightarrow \neg q$, $\neg p \rightarrow q$, and $\neg p \rightarrow \neg q$, respectively. An association rule is mapped to an equivalence if each implication is either true or false.

For example,

- Association rules $X \Rightarrow Y$ is mapped to $p \equiv q$, if and only if
  - $X \Rightarrow Y$ is true;
  - $X \Rightarrow \neg Y$ is false;
  - $\neg X \Rightarrow Y$ is false; and
  - $\neg X \Rightarrow \neg Y$ is true.

We summarize all the mappings and the conditions required in Table 3.

In each mapping from an association rule to an equivalence, four conditions need to be checked. The four conditions to be passed on $X \Rightarrow Y$ and $\neg X \Rightarrow \neg Y$ are the same. This is highlighted in Table 3. (Note that the other four conditions on both $X \Rightarrow \neg Y$ and $\neg X \Rightarrow Y$ are also the same as shown.)

Generally, each condition testing requires at least one transaction record to conclude as either true or false.

However, because there are four conditions, mapping from an association rule to an equivalence cannot be carried out on a single transaction record. A mechanism is required to judge if the first association rule from each group can be mapped to a true equivalence having met all four conditions. This is because an association rule holds item sets that can be observed over a portion of transaction records. This leads us to perform mapping on multiple transaction records as described in the next section.

3.4.2 Mapping Using Multiple Transaction Records

Previously, item sets have been mapped to propositions p and q if each item set is observed or not observed in a single transaction. In data containing multiple transaction records, an item set X is observed over a portion of transaction records. This total number of observations is given by the cardinality of the transactions in database D that contain X, known as support:

$$\text{support}, \mathrm{S}(X) = |D_X|. \tag{1}$$

A support $\mathrm{S}(X)$ denotes the number of times X which is observed in the entire data. Similarly, support $\mathrm{S}(\neg X)$ denotes the number of times X which is not observed in the entire data. Based on this understanding, the interestingness of an item set is a relative comparison between the total number of observations of its presence and its absence:

- If $\mathrm{S}(X)$ has a greater value than $\mathrm{S}(\neg X)$, then item set X is mostly observed in the entire data, and it is interesting.
- Conversely, item set X is mostly not observed in the entire data, and it is not interesting; the absence of item set X is interesting.

Based on a relative comparison between the presence and absence of an item set, each item set can be mapped to propositions p and q within multiple transaction records:

- if $S(X) > S(\neg X)$, then
  - Item set X is mapped to $p = \mathrm{T}$.
  - Item set $\neg X$ is mapped to $p = \mathrm{F}$.
- if $S(\neg X) > S(X)$, then
  - Item set X is mapped to $p = \mathrm{F}$.
  - Item set $\neg X$ is mapped to $p = \mathrm{T}$.

An item set having mapped to a proposition is said to be interesting. The above mapping involves only a single item set.

To judge if a union of two item sets, such as (X, Y) (i.e., $X \cup Y$), is comparatively interesting, a multiple comparisons over three other possible combinations are necessary. This is to ensure that none of these combinations is more observed than the initial combination. For example, to judge if the union of item set (X, Y) is mostly observed in transactions, the number of transactions that contain $(X, \neg Y)$, $(\neg X, Y)$, or $(\neg X, \neg Y)$ must be lower than the portion of transactions that contain item set (X, Y). Otherwise, the item set (X, Y) cannot be judged as interesting. Extending this understanding, it can be seen that only one item set is deemed interesting and the others deemed not interesting for each combination of the presence and absence of item sets contained within item sets (X, Y).

TABLE 4
Association Rules and Supports

Association Rule	Support
$X \Rightarrow Y$	S(X, Y)
$X \Rightarrow \neg Y$	S(X, $\neg Y$)
$\neg X \Rightarrow Y$	S($\neg X$, Y)
$\neg X \Rightarrow \neg Y$	S($\neg X$, $\neg Y$)

Having discussed the interestingness of item sets, we extend the concept to an association rule. In addition to the previous discussion, an association rule always involves two item sets. These are (X, Y), $(X, \neg Y)$, $(\neg X, Y)$, or $(\neg X, \neg Y)$. An association rule $X \Rightarrow Y$ can be mapped to implication $p \rightarrow q$ if and only if item set (X, Y) is interesting. Otherwise, if item set (X, Y) is not the most observed set, then the association rule $X \Rightarrow Y$ cannot be judged as interesting. For each combination of items contained within item sets (X, Y), only one association rule is deemed interesting and the other association rules are deemed not interesting. The interesting one is the most observed association rule.

Following our discussion on judging the interestingness of the union of two item sets, an association rule is judged interesting by having the support value of two item sets. For example, $\mathrm{S}(X, Y)$ is higher than the support values that have one or none of the two item sets, such as $\mathrm{S}(X, \neg Y)$, $\mathrm{S}(\neg X, Y)$, and $\mathrm{S}(\neg X, \neg Y)$. Table 4 denotes the supports for each association rule.

We now can map association rules to implications as follows:

- $X \Rightarrow Y$ is mapped to an implication $p \rightarrow q$, if and only if
  - $\mathrm{S}(X, Y) > \mathrm{S}(X, \neg Y)$;
  - $\mathrm{S}(X, Y) > \mathrm{S}(\neg X, Y)$; and
  - $\mathrm{S}(X, Y) > \mathrm{S}(\neg X, \neg Y)$.
- $X \Rightarrow \neg Y$ is mapped to an implication $p \rightarrow \neg q$, if and only if
  - $\mathrm{S}(X, \neg Y) > \mathrm{S}(X, Y)$;
  - $\mathrm{S}(X, \neg Y) > \mathrm{S}(\neg X, Y)$; and
  - $\mathrm{S}(X, \neg Y) > \mathrm{S}(\neg X, \neg Y)$.
- $\neg X \Rightarrow Y$ is mapped to an implication $\neg p \rightarrow q$, if and only if
  - $\mathrm{S}(\neg X, Y) > \mathrm{S}(X, Y)$;
  - $\mathrm{S}(\neg X, Y) > \mathrm{S}(X, \neg Y)$; and
  - $\mathrm{S}(\neg X, Y) > \mathrm{S}(\neg X, \neg Y)$.
- $\neg X \Rightarrow \neg Y$ is mapped to an implication $\neg p \rightarrow \neg q$, if and only if
  - $\mathrm{S}(\neg X, \neg Y) > \mathrm{S}(X, Y)$;
  - $\mathrm{S}(\neg X, \neg Y) > \mathrm{S}(X, \neg Y)$; and
  - $\mathrm{S}(\neg X, \neg Y) > \mathrm{S}(\neg X, Y)$.

We give the name *pseudoimplication* to association rules that are mapped to implications based on comparison between supports. By pseudoimplication, we mean that the implication approximates a real implication (according to propositional logic). It is not a real implication because there

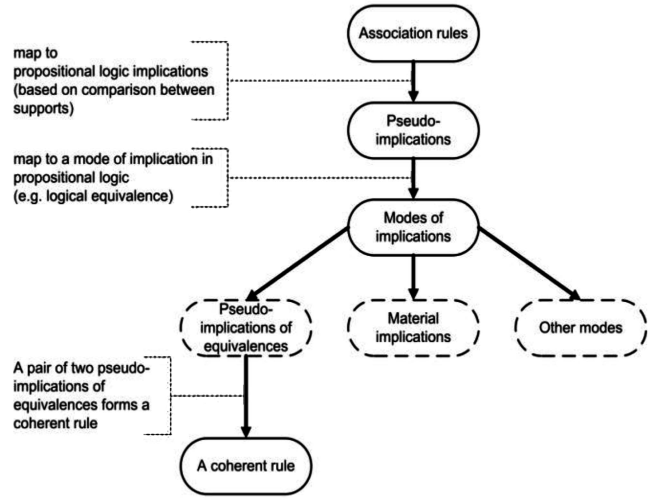


Fig. 1. A generalized framework of association rules that based on pseudoimplications.

are fundamental differences. Pseudoimplication is judged true or false based on a comparison of supports, which has a range of integer values. In contrast, an implication is based on binary values. The former depends on the frequencies of co-occurrences between item sets (supports) in a data set, whereas the latter does not and is based on truth values.

Using the concept of pseudoimplication, we now can further map association rules to specific modes of implications such as material implications and equivalences. Each follows the same truth values of the respective relations in logic. In a material implication relation, for example, association rules can be finally mapped to equivalences according to their truth table values. That is, the following conditions must be met:

$$\begin{aligned}&\mathrm{S}(X,Y) > \mathrm{S}(X,\neg Y),\\ &\mathrm{S}(X,Y) > \mathrm{S}(\neg X,Y),\\ &\mathrm{S}(\neg X,\neg Y) > \mathrm{S}(X,\neg Y), \text{ and}\\ &\mathrm{S}(\neg X,\neg Y) > \mathrm{S}(\neg X,Y),\end{aligned} \tag{2}$$

and are given below:

- $X \Rightarrow Y$ is mapped to $p \equiv q$ if and only if (2) is met.
- $X \Rightarrow \neg Y$ is mapped to $p \equiv \neg q$ if and only if (2) is met.
- $\neg X \Rightarrow Y$ is mapped to $\neg p \equiv q$ if and only if (2) is met.
- $\neg X \Rightarrow \neg Y$ is mapped to $\neg p \equiv \neg q$ if and only if (2) is met.

Each of the rules, having mapped to equivalence from propositional logic, is called a *pseudoimplication of equivalence*. We show the possible mappings in Fig. 1.

We have shown how pseudoimplications can be created by mapping association rules to two modes of implications: material implication and equivalence. We shall focus on pseudoimplications of equivalences because equivalences are special cases in material implications according to propositional logic. Being a special case, a rule's left-hand side and right-hand side are deemed more related than compared to material implication. In fact, both sides are so much related that they are equivalence. As a result, an equivalence relation is also bidirectional. That is, a rule's left- and right-hand sides are interchangeable. Having exchanged both sides, this new equivalence follows also the truth table values of equivalence. This characteristic is not observed in material implications and remains as a "more relaxed" mode of implication. In mining rules from data sets and without requiring background knowledge of a domain, we need a strong reason to identify the existence of rules. Therefore, pseudoimplications of equivalences are our primary focus in finding rules.

4 Coherent Rules Mining Framework

The pseudoimplications of equivalences can be further defined into a concept called *coherent rules* (see Fig. 1). We highlight that not all pseudoimplications of equivalences can be created using item sets X and Y. Nonetheless, if one pseudoimplication of equivalence can be created, then another pseudoimplication of equivalence also coexists. Two pseudoimplications of equivalences always exist as a pair because they are created based on the same conditions as shown in (2). Since they share the same conditions, two pseudoimplications of equivalences:

$$X \Rightarrow Y \quad \text{and} \quad \neg X \Rightarrow \neg Y \tag{3}$$

coexist having mapped to two logical equivalences $p \equiv q$ and $\neg p \equiv \neg q$. The result is a *coherent rule* that meets the same conditions of (2).

Coherent rules meet the necessary and sufficient conditions and have the truth table values of logical equivalence, as shown in Table 2. By definition, a coherent rule consists of a pair of pseudoimplications of equivalences that have higher support values compared to another two pseudoimplications of equivalences (see (2)). Each pseudoimplication of equivalence is an association rule with the additional property that it can be mapped to a logical equivalence. Association rules decoupled from coherent rules have the following strengths:

- Association rules decoupled from coherent rules can be reasoned as logical implications, whereas association rules cannot. This is because coherent rules inherit logic property such as contrapositives having truth table values of logical equivalence. As an example, an association rule $X \Rightarrow Y$ does not imply that $X \Rightarrow \neg Y$ is false or has a weaker strength value compare to the first. On the contrary, a coherent rule of' "$X \Rightarrow Y$ and $\neg X \Rightarrow \neg Y$" implies that the strength value of $X \Rightarrow \neg Y$ is weaker than $X \Rightarrow Y$.
- The concept of coherent rules is independent from any background knowledge such as a specific understanding of an application domain. Therefore, coherent rules do not require a user to preset an arbitrary minimum support threshold to define a frequent pattern. Coherent rules can be identified via truth table values. The discovery of coherent rules and their related association rules thus avoids many of the problematic issues discussed in Section 2.

4.1 The Differences in Setting Up Thresholds

There are fundamental differences between the thresholds used in the *coherent rules mining framework* and the threshold set by a user for the *support and confidence framework*.

- Unlike the threshold in the latter framework that needs to be determined by users as a hard rule and applied to all items, the former establishes natural setting of four thresholds by comparing the frequency of the presence and absence of individual items as required by logic. See (2). Due to this, no prior setting of an artificial threshold by users is needed, as the thresholds are found based on the observation of data itself.
- Each threshold used by the *coherent rules mining framework* has a support value between being observed in a single transaction record and being observed in all transaction records. This support range covers all possible association rules. Whereas in *support and confidence framework* or all other frameworks that require a preset of a minimum support threshold, the search space below a minimum support threshold has been excluded. Consequently, these frameworks cannot discover all coherent rules because some coherent rules would have support values lower than a minimum support threshold. To generate all possible association rules in order to postprocess them for coherent rules will be too costly.

Coherent rules have more stringent constraints compared to typical association rules with a minimum confidence value. Although there are frameworks that find association rules based on measure of interests other than a minimum support (for example, a minimum confidence value), they need to generate both positive and negative association rules before matching and comparing the association rules using (2). This postprocessing requires a large amount of space and is typically costly. *Coherent rules mining framework* remains a unique framework to generate coherent rules from data directly.

4.2 Quality of Logic-Based Association Rules

Coherent rules are defined based on logic. This improves the quality of association rules discovered because there are no missing association rules due to threshold setting. A user can discover all association rules that are logically correct without having to know the domain knowledge. This is fundamental to various application domains. For example, one can discover the relations in a retail business without having to study the possible relations among items. Any association rule that is not captured by coherent rules can be denied its importance. These rules are either in contradiction with others (among the positive and negative association rules) or less stringent compared to the definition of logical equivalences.

As an example, consider that a nonlogic-based association rule is found within 100 transaction records between item i_1 and item i_2 with confidence at 75 percent and support at 30 percent. This association rule is not important if the absence of the same item i_1 (i.e., $\neg i_1$) is found associated with item i_2 with a higher confidence at 85 percent and a higher support at 51 percent. Without the further analysis, the first discovery misleads decision makers to conclude that item i_1 is associated with item i_2, whereas the relation having item $\neg i_1$ is, in fact, stronger. Coherent rules avoid this problem all together based on logic.

In the next section, we explain how to identify coherent rules from data sets.

TABLE 5
Artificial Transaction Records

id	Content of T_{id}	id	Content of T_{id}
1	i_2	14	$i_2, i_3, i_4, i_5, i_6, i_7$
2	i_6	15	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$
3	i_2	16	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$
4	i_3	17	$i_1, i_2, i_3, i_4, i_5, i_7$
5	i_3	18	i_1, i_2, i_3, i_4, i_6
6	i_3, i_4	19	i_1, i_2, i_3, i_4
7	i_3, i_4	20	i_1, i_2, i_3, i_5, i_6
8	i_3, i_4, i_5, i_7	21	i_1, i_2, i_3, i_5, i_6
9	$i_1, i_3, i_4, i_5, i_6, i_7$	22	i_1, i_2, i_5
10	$i_1, i_3, i_4, i_5, i_6, i_7$	23	i_2, i_5
11	i_1, i_3, i_4, i_5, i_7	24	i_1, i_2, i_4, i_7
12	$i_1, i_2, i_3, i_4, i_5, i_6, i_7$	25	i_2
13	i_2, i_3, i_4, i_5, i_7		

5 Finding Coherent Rules in Transaction Records

In this section, the concept of coherent rules is utilized to find coherent rules from transaction records. Each coherent rule is decoupled into two pseudoimplications of equivalences. Each pseudoimplication of equivalence is an association rule, which can be further mapped to a logical equivalence. To explain this more formally, we adopt the following notation.

Assume that $I = \{i_1, i_2, \ldots, i_n\}$, a set of items. Let T be a table of transaction records (relational table) such that $T = \{t_1, t_2, \ldots, t_m\}$. A task-relevant transaction record t_i holds a subset of items such that $t_i \subseteq I$. Assume that we can predetermine the total number of items contained in two independent supersets $A = \{a_1, \ldots, a_u\}$ and $C = \{c_1, \ldots, c_v\}$ such that $A \subset I, C \subset I$ and $A \cap C = \emptyset$. We are interested in coherent rules between the two item sets X and Y that meet the conditions in (2), where $X \subseteq A$, $Y \subseteq C$, $X \neq \emptyset$, and $Y \neq \emptyset$.

As an example, consider the set of transaction records listed in Table 5. Following our notation, the unique items contained in these transaction records are given by I and $I = \{i_1, i_2, i_3, i_4, i_5, i_6, i_7\}$. We want to find coherent rules, where consequence item set Y holds only a single item, say i_7. As a result, A holds arbitrarily any item other than item i_7 such that $A = \{i_1, i_2, i_3, i_4, i_5, i_6\}$ and $C = \{i_7\}$. Let P be a power set function. The power sets on item set A are given by I_A such that $I_A = P(A)$. Let I_C be the power sets on item set C such that $I_C = P(C)$. We are interested in coherent rules between the two item sets X and Y that meet the conditions in (2), where $X \in I_A$, $Y \in I_C$, $X \neq \emptyset$, and $Y \neq \emptyset$.

The power sets of I_A and I_C are given as

$$\begin{aligned} I_A = \{&\{null\}, \{i_1\}, \{i_2\}, \ldots, \{i_6\}, \\ &\{i_1, i_2\}, \{i_1, i_3\}, \ldots, \{i_1, i_2, i_3, i_4, i_5, i_6\}\}, \end{aligned} \tag{4}$$

and

$$I_C = \{\{null\}, \{i_7\}\}. \tag{5}$$

TABLE 6
Contingency Table for Antecedent X and Consequence Y

Frequency of co-occurrences		Consequence, Y	
		$Y = \{i_7\}$	$\neg Y = \neg\{i_7\}$
Antecedent, X	$X = \{i_1\}$	8	5
	$\neg X = \neg\{i_1\}$	3	9

A coherent rule is between two item sets X and Y such that $X \in I_A$, $Y \in I_C$, $X \neq \emptyset$, and $Y \neq \emptyset$.

For example, Table 6 contains the contingency table for i_1 and i_7, as follows:

From Table 6, the higher values of support are for $\{i_1\} \cup \{i_7\}$ and $\neg\{i_1\} \cup \neg\{i_7\}$; consequently:

$$\begin{aligned} \{i_1\} &\Rightarrow \{i_7\}, \\ \neg\{i_1\} &\Rightarrow \neg\{i_7\}. \end{aligned} \tag{6}$$

The above two rules can be mapped to logical equivalences since $Q1 > Q2$, $Q1 > Q3$, $Q4 > Q2$, and $Q4 > Q3$ in relate to (2), where:

1. $Q1 = \mathrm{S}(\{i_1\}, \{i_7\})$;
2. $Q2 = \mathrm{S}(\{i_1\}, \neg\{i_7\})$;
3. $Q3 = \mathrm{S}(\neg\{i_1\}, \{i_7\})$; and
4. $Q4 = \mathrm{S}(\neg\{i_1\}, \neg\{i_7\})$.

The pair of rules in (6) is a coherent rule that consists of two pseudoimplications of equivalences.

Drawn from the same data set, we can also give an example that is not a coherent rule based on a contingency table shown in Table 7.

From Table 7, the support of $\{i_2\} \cup \{i_7\}$ and $\neg\{i_2\} \cup \neg\{i_7\}$ is not the highest to support the rules:

$$\begin{aligned} \{i_2\} &\Rightarrow \{i_7\}, \\ \neg\{i_2\} &\Rightarrow \neg\{i_7\}. \end{aligned} \tag{7}$$

Instead, the support of $\{i_2\} \cup \neg\{i_7\}$ is the highest. Equation (7) cannot be mapped to logical equivalences since $Q1 \ngtr Q2$, $Q1 \ngtr Q3$, $Q4 \ngtr Q2$, and $Q4 > Q3$, where:

1. $Q1 = \mathrm{S}(\{i_2\}, \{i_7\})$;
2. $Q2 = \mathrm{S}(\{i_2\}, \neg\{i_7\})$;
3. $Q3 = \mathrm{S}(\neg\{i_2\}, \{i_7\})$; and
4. $Q4 = \mathrm{S}(\neg\{i_2\}, \neg\{i_7\})$.

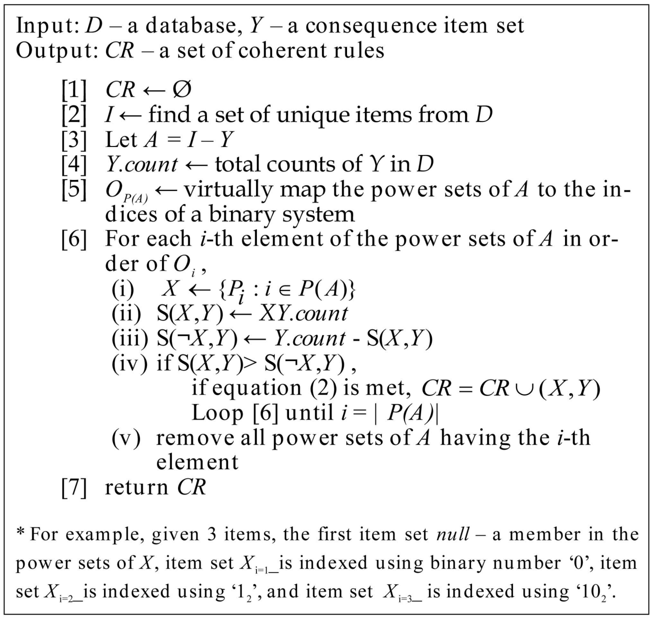
Hence, the pair of rules in (7) is not a coherent rule.

In this particular example, we have chosen i_7 as a target item. In real-life situations and business problems, the target item can either be determined by the user or by using a DDM intelligence model (if available).

The interestingness of the coherent rule shown in (6), however, cannot be reasoned because we do not have any background knowledge on the nature of items in this artificial data set.

TABLE 7
Contingency Table for Antecedent X and Consequence Y

Frequency of co-occurrences		Consequence, Y	
		$Y = \{i_7\}$	$\neg Y = \neg\{i_7\}$
Antecedent, X	$X = \{i_2\}$	7	9
	$\neg X = \neg\{i_2\}$	4	5



```
Input: $D$ – a database, $Y$ – a consequence item set
Output: $CR$ – a set of coherent rules

    [1] $CR \leftarrow \emptyset$
    [2] $I \leftarrow$ find a set of unique items from $D$
    [3] Let $A = I - Y$
    [4] $Y.count \leftarrow$ total counts of $Y$ in $D$
    [5] $O_{P(A)} \leftarrow$ virtually map the power sets of $A$ to the in-
        dices of a binary system
    [6] For each $i$-th element of the power sets of $A$ in or-
        der of $O_i$ ,
          (i)   $X \leftarrow \{P_i : i \in P(A)\}$
          (ii)  $\mathrm{S}(X,Y) \leftarrow XY.count$
          (iii) $\mathrm{S}(\neg X,Y) \leftarrow Y.count$ - $\mathrm{S}(X,Y)$
          (iv)  if $\mathrm{S}(X,Y) > \mathrm{S}(\neg X,Y)$ ,
                    if equation (2) is met, $CR = CR \cup (X,Y)$
                    Loop [6] until $i = |\ P(A)|$
          (v)   remove all power sets of $A$ having the $i$-th
                element
    [7] return $CR$

* For example, given 3 items, the first item set null – a member in the
power sets of $X$, item set $X_{i=1}$ is indexed using binary number '0', item
set $X_{i=2}$ is indexed using '$1_2$', and item set $X_{i=3}$ is indexed using '$10_2$'.
```

Fig. 2. A simple search for coherent rules algorithm (*ChSearch*).

The discovery of coherent rules is useful in application domains where the domain knowledge is not known to a user or is difficult to grasp. In the retail domain, the reasons for associations between items are not obvious. Customers have various reasons to buy different items together. Using mapping to logical equivalences, we can discover coherent rules and its association rules without the need to survey on customers. As a result, we know that some items are associated together based on logical grounds.

5.1 Search Algorithm

We propose to search for coherent rules by exploiting the antimonotone property found on the condition $\mathrm{S}(X, Y) > \mathrm{S}(\neg X, Y)$ targeting at a preselected consequence item set Y. We write a basic algorithm to generate the coherent rules in Fig. 2.

5.1.1 Distinct Features of ChSearch

We list some features of *ChSearch* compared to a priori. Unlike a priori, *ChSearch*:

- does not require a preset minimum support threshold. *ChSearch* does not require a preset a minimum support threshold to find association rules. Coherent rules are found based on mapping to logical equivalences. From the coherent rules, we can decouple the pair for two pseudoimplications of equivalences. The latter can be used as association rules with the property that each rule can be further mapped to a logical equivalence.
- does not need to generate frequent item sets. *ChSearch* does not need to generate frequent item sets. Nor does it need to generate the association rules within each item set. Instead, *ChSearch* finds coherent rules directly. Coherent rules are found within the small number of candidate coherent rules allowed through its constraints.

TABLE 8
Total Frequency of Class Attributes

#	Class Attributes	Frequency of Occurrence	%	Type of Association
1	*Reptile*	5/ 101	4. 95	Infreq.
2	*Mammal*	41/ 101	40.59	Freq.
3	*Invertebrate*	10/ 101	9.90	Freq.
4	*Insect*	8/ 101	7.92	Freq.
5	*Fish*	13/ 101	12.87	Freq.
6	*Bird*	20/ 101	19.80	Freq.
7	*Amphibian*	4/ 101	3.96	Infreq.

- identifies negative association rules. *ChSearch*, by default, also identifies negative association rules. Given a set of transaction records that does not indicate item absence, a priori cannot identify negative association rules. *ChSearch* finds the negative pseudoimplications of equivalences and uses them to complement both the positive and negative rules found. (Although a priori finds negative association rules if a transaction database is transformed into binary attributes, the rules found typically contradict one another. See a further discussion on this in Section 6.2.3.)

6 Experiments

6.1 Settings

We use the Zoo data set [31] to perform a series of experiments before test it on transaction records for Market Basket Analysis. This is because most transaction records have anonymous item names. Consequently, we could not further analyze for their relationships among items.

The Zoo data set contains a collection of animal characteristics and their classes in a Zoo. This data set was chosen because the characteristics of animals are well understood. The small size of this data set facilitates the interpretation of findings. As a result, the interestingness of found rules can be compared and contrasted based on general understanding.

The Zoo data set has seven classes of animals. It has a spectrum of frequency of occurrences in the transaction records. We show the frequencies of each class in Table 8. If the frequency of occurrence is below 5 percent, we consider the class of animal to be rare. Otherwise, the class is identified as frequent (that is, incurring at least 5 percent of support values).

Table 8 shows that five-class attributes have a frequency of occurrence of at least more than 5 percent. These are *mammal*, *invertebrate*, *insect*, *fish*, and *bird*. The class *mammal* is most frequently observed, followed by *bird* and *fish*. On the other hand, the classes *amphibian* and class *reptile* are the least observed at 3.96 and 4.95 percent, respectively.

6.2 Quality of Coherent Rules

We compare the coherent rules found using *ChSearch* with association rules found using a priori written by Borgelt [32]. Two thresholds are set for the a priori:

- The minimum support threshold is set at 5 percent [33], [34], [35].
- The minimum confidence threshold is set at 50 percent [1], [14].

Based on this setting, all association rules not supported by at least 5 percent of transaction records are considered infrequent; otherwise, they are considered frequent. We group association rules into two categories:

- Infrequent association rules that contain less observed classes (for example, *reptile* and *amphibian*).
- Frequent association rules that have frequently observed classes (for example, *mammal*).

In reporting the results, the attribute values of the data are represented as attribute names followed by its possible values. For example, reptile(1) refers to $\text{reptile} = \text{``yes''}$, legs(4) refers to $\text{legs} = 4$.

6.2.1 Total Number of Rules Discovered

The algorithm *ChSearch* discovers all 265 coherent rules for *mammal*. In comparison, the algorithm a priori finds 20,853 association rules (based on a priori implementation release 4.27 by [32]) without the constraints mentioned in Section 6.2. Out of these, 20,588 or 98.7 percent of the association rules can be argued redundant according to logic. If the constraints are in-place, a priori finds 387 (frequent and positive) association rules of all sizes, where one-third of these rules are redundant—nonlogic based. We detailed the comparison in the next section.

6.2.2 Infrequent Rules

Rules involve classes *reptile* and *amphibian* especially cases of *reptile(1)* and *amphibian(1)* are considered infrequent rules. Theoretically, a priori could not find association rules involving them. Two coherent rules are found by our algorithm:

Coherent rule 1.

$$\{eggs(1), toothed(1), breathes(1), tail(1)\} \Rightarrow \{reptile(1)\}, \tag{8}$$

and

$$\neg\{eggs(1), toothed(1), breathes(1), tail(1)\} \Rightarrow \neg\{reptile(1)\}. \tag{9}$$

Coherent rule 2.

$$\{aquatic(1), leg(4)\} \Rightarrow \{amphibian(1)\}, \tag{10}$$

and

$$\neg\{aquatic(1), leg(4)\} \Rightarrow \neg\{amphibian(1)\}. \tag{11}$$

If each of the coherent rules is decoupled into two association rules, then a total of four association rules are found by our approach. Further examination of the support and confidence values (see (12)-(15)) suggests that these association rules will not be found by a priori due to support lower than the threshold, regardless the true knowledge that the rule has reported—an animal with eggs, toothed, and breathes using lungs with tail is usually a reptile:

$$\begin{gathered}\{eggs(1), toothed(1), breathes(1), tail(1)\} \Rightarrow \{reptile(1)\},\\ support = 3.0\%, confidence = 75.0\%,\end{gathered} \tag{12}$$

TABLE 9
Shortest Frequent Rules Found for Class Mammal

#	*ChSearch*	*Apriori*
1	$milk(1) \Rightarrow mam.\ (1)$, $milk(0) \Rightarrow mam.(0)$	$milk(1) \Rightarrow mam.(1)$ [S=40.6%, C=100%]
2	$hair(1) \Rightarrow mam.(1)$, $hair(0) \Rightarrow mam.(0)$	$hair(1) \Rightarrow mam.(1)$ [S=38.6%, C=90.7%]
3	$leg(4) \Rightarrow mam.(1)$, $\neg leg(4) \Rightarrow mam.(0)$	$leg(4) \Rightarrow mam.(1)$ [S=30.7%, C=81.6%]
4	$catsize(1) \Rightarrow mam.(1)$, $catsize(0) \Rightarrow mam.(0)$	$catsize(1) \Rightarrow mam.(1)$ [S=31.7%, C=72.7%]
5	$toothed(1) \Rightarrow mam.(1)$, $toothed(0) \Rightarrow mam.(0)$	$toothed(1) \Rightarrow mam.(1)$ [S=39.6%, C=65.6%]
6	Nil	$domestic(1) \Rightarrow mam.(1)$ [S=7.9%, C=61.5%]
7	Nil	$breathes(1) \Rightarrow mam.(1)$ [S=40.6%, C=51.2%]

$$\neg\{eggs(1), toothed(1), breathes(1), tail(1)\} \Rightarrow \neg\{reptile(1)\},\quad support = 94.1\% \text{ and } confidence\ 97.9\%, \tag{13}$$

$$\{aquatic(1), leg(4)\} \Rightarrow \{amphibian(1)\},\quad support = 4.0\%, confidence = 57.1\%, \tag{14}$$

$$\neg\{aquatic(1), leg(4)\} \Rightarrow \neg\{amphibian(1)\},\quad support = 93.1\%, confidence = 100.0\%. \tag{15}$$

In the next section, we highlight that our approach can discover rules containing infrequent items that may not be discovered based on support and confidence framework.

6.2.3 Frequent Rules

We repeat the experiment on the most observed class *mammal*. We list and compare the shortest rules found by *ChSearch* and a priori algorithms in Table 9. Rules found by only one approach are highlighted.

From the experimental result, the algorithm *ChSerach* finds five coherent rules for *mammal* (or 10 association rules, because each coherent rule consists of a pair of association rules) but a priori finds seven association rules. Of these seven shortest association rules, association rules 6 and 7 in Table 9 are not reported by *ChSearch*.

We detail the contingency tables (Tables 10 and 11, respectively) for these two association rules and explain why these rules are not interesting and should not be reported.

TABLE 10
Contingency Table for Domestic(D) and Mammal(M)

Frequency of co-occurrences		Consequence, $\mathbb{Y}$		
		Y='M'	$\neg Y$=¬'M'	Total
Antece-dent, $\mathbb{X}$	X='D'	8	5	13
	$\neg X$=¬'D'	33	55	88
	Total	41	60	101

TABLE 11
Contingency Table for Breathes(B) and Mammal(M)

Frequency of co-occurrences		Consequence, $\mathbb{Y}$		
		Y='M'	$\neg Y$=¬'M'	Total
Antece-dent, $\mathbb{X}$	X='B'	41	39	80
	$\neg X$=¬'B'	0	21	21
	Total	41	60	101

Association rule $domestic(1) \Rightarrow mam.(1)$ is reported by a priori based on support value of 7.9 percent and confidence of 61.5 percent. We argue that this rule should not be reported because it is contradicted by other associations rules with similar attribute such as

1. $domestic(0) \Rightarrow mam.(1)$, [S = 80.5%, C = 37.5%],
2. $domestic(0) \Rightarrow mam.(0)$, [S = 54.5%, C = 62.5%].

The above two association rules have support values higher than 7.9 percent. Reporting association rules with weaker support while stronger association rules exist is misleading. In addition, the rules in 1 and 2 are contradictory if they are reasoned as logical implications. That is, if "*not domestic*" (*domestic*(0)) is associated with *mammal*, then it cannot be associated with "*not mammal*" logically. Hence, these rules are not reported by *ChSearch* as implicational association rules.

A priori, however, reports that *mammal* is a domestic animal and identifies this relationship as one of the strongest association rules with a confidence of 61.5 percent. In fact, a detail analysis on Zoo data set reveals that there are 41 sets of *mammal* (such as buffalo, bear, and elephant) but only eight of them are domestic animals. Many more mammals (that is, 33 or 80.5 percent) are not domestic type. The latter is never highlighted; instead, a weaker rule is reported. If this weaker rule is used in a business application, it could cause a wrong decision. *ChSearch* avoids reporting this misleading association rule.

In another case, based on Table 9, the association rule:

$$breathes(1) \Rightarrow mammal(1)\quad support = 40.6\%, confidence = 51.2\% \tag{16}$$

is not interesting enough because this association rule exists on its own and cannot be supported by any other association rules. Theoretically, association rules in (16) are meaningfully supported by:

$$breathes(0) \Rightarrow mammal(0). \tag{17}$$

That is, if "breathes using lungs" is a fundamental characteristic of mammals, then logically speaking, an animal that does not breathe using lungs is not a *mammal*; however, the association rule in (17) (with support 20.8 percent) does not occur frequently enough. A detailed study shows that the association rule in (17) is contradicted by the following association rule with a stronger existence:

$$breathes(1) \Rightarrow mammal(0),\quad support = 38.6\%, confidence = 48.8\%. \tag{18}$$

Hence, the association rule in (17) is suppressed from complementing the meaning on a positive association rule in (16). *ChSearch* does not report single positive association rules that cannot be supported by another association rule because this indicates that the association rule,

TABLE 12
Summary of the Statistics on Transaction Records Used

Data-Set	# Items	# Records	Av. Items	Mx. Items
Dn1	489	245,480	9	31
Dn2	985	245,050	10	29
Dn3	1,475	244,980	10	30
Sp1	423	247,490	9	29
Sp2	852	243,840	10	32
Sp3	1,289	245,090	10	32

Av. Items: average number of items per transaction record, Mx. Items: maximum number of items per transaction record.

$breathes(1) \Rightarrow mammal(1)$, is not strong enough to be a pseudoimplication of equivalence.

In addition, based on Table 11, while 41 sets of animals that breath using lungs are associated to *mammal*, this characteristic is also associated to 39 sets of nonmammals (that is, all animals except *fish*). The use of this characteristic (*breathes*(1)) is statistically weak to describe a *mammal*.

The association rule in (18) is also not reported by *ChSearch* because it contradicts the association rule in (16). Logically, if mammals breathe using lungs, then they cannot be described as breathing without lungs. Furthermore, the association rule in (18) is not supported by any other association rule.

6.3 Mining Feasibility without a Minimum Support

Search for coherent rules is more expensive compare to a priori because it needs to consider both the positive and negative association rules. Our proposed algorithm in Section 5.1 can be used to explore large market basket data sets at zero minimum support threshold. This is typically not possible via a priori and meaningless because a large number of redundant association rules will be discovered. Nonetheless, our algorithm discovers logic-based association rules via coherent rules within acceptable time.

We generate a following three dense artificial data sets with increasing complexity using the IBM synthetic data generator [36]. The symbols used in representing a data set are explained below:

- D: number of transactions in 000s,
- T: average items per transaction,
- N: number of unique items,
- L: number of patterns (possible rules) in 000s, and
- I: average length of maximal pattern.

The dense data sets have a low number of patterns (L) available. These dense data sets have an increase number of items: Dn1:D300T8N500L0.5I7, Dn2:D300T8N500L1.0I7, and Dn3:D300T8N500L1.5I7. We also generate sparse data sets with increasing number of items, hence, complexity: Sp1:D300T8N500L5.0I7, Sp2:D300T8N500L10.0I7, and Sp3:D300T8N500L15.0I7.

The data generator based on the above parameters generates six transaction records with the actual unique number of items and records shown in Table 12.

We ran a priori for a given consequence item set on a notebook equipped with Intel Core 2 Duo at 2 GHz and 4 GB of physical memory running Windows Vista Business. Fig. 3 shows that a priori could not report results at preset minimum support thresholds above 0.5 and 0.1 percent on the Dn1 and Sp1 data sets, respectively. A priori exhausted the memory resources and could not report the association rules beyond 14,000 and 51,000 of association rules on respective data sets.

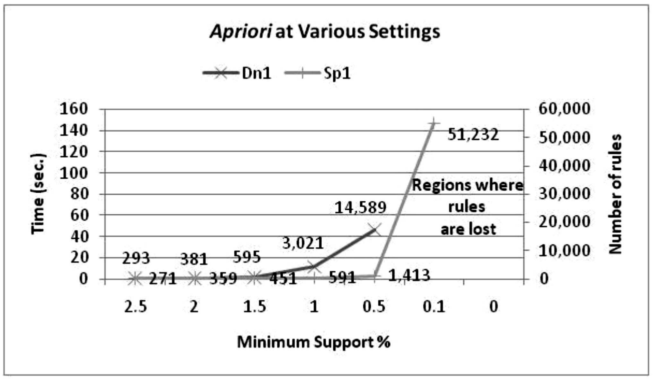


Fig 3. Search time and the number of rules on decreasing minimum supports on dense and sparse data sets.

The experiment using *ChSearch* shows that search for coherent rules is feasible for the generated data sets (Fig. 4). The graph suggests that the search time also increases linearly with the number of items.

7 Conclusion

We used mapping to logical equivalences according to propositional logic to discover all interesting association rules without loss. These association rules include item sets that are frequently and infrequently observed in a set of transaction records. In addition to a complete set of rules being considered, these association rules can also be reasoned as logical implications because they inherit propositional logic properties. Having considered infrequent items, as well as being implicational, these newly discovered association rules are distinguished from typical association rules. These new association rules reduce the risks associated with using an incomplete set of association rules for decision making, as following:

- Our new set of association rules avoids reporting that item A is associated with item B if there is a

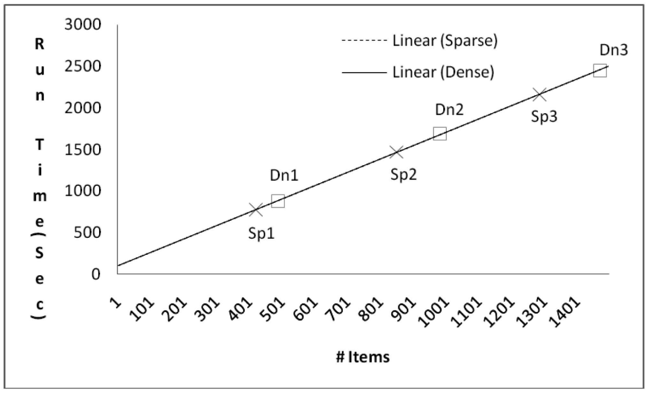


Fig 4. Search time on an increase complexity on dense and sparse data sets.

stronger association between item A and the *absence* of item B. Using prior association rules that do not consider this situation could lead a user to erroneous conclusions about the relationships among items in a data set. Again, identifying the strongest rule among the same items will promote information correctness and appropriate decision making.

- The risks associated with incomplete rules are reduced fundamentally because our association rules are created without the user having to identify a minimum support threshold. Among the large number of association rules, only those that can be mapped to logical equivalences according to propositional logic are considered interesting and reported.

In this paper, we introduced our contributions in novel frameworks: a generalized framework to discover association rules that have the properties of propositional logic, and a specific framework (*Coherent Rules Mining Framework*) with a basic algorithm to generate coherent rules from a given data set. The discovery of coherent rules is important because through coherent rules, a complete set of interesting association rules that are also implicational according to propositional logic can be discovered. The search for coherent rules does not require a user to preset a minimum support threshold. In contrast, an association rule is typically not implicational according to propositional logic, and the many approaches used to output association rules have lost rules involving infrequent item sets. A *coherent rules mining framework* can thus be appreciated for its ability to discover rules that are both implicational and complete according to propositional logic from a given data set.

ACKNOWLEDGMENTS

The authors thank the anonymous referees for their comments. A.T.H. Sim acknowledges Monash University and Universiti Teknologi Malaysia for scholarships.

REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *SIGMOD Record,* vol. 22, pp. 207-216, 1993.

[2] C. Longbing, "Introduction to Domain Driven Data Mining," *Data Mining for Business Applications,* L. Cao, P.S. Yu, C. Zhang, and H. Zhang, eds., pp. 3-10, Springer, 2008.

[3] G.I. Webb and S. Zhang, "k-Optimal Rule Discovery," *Data Mining and Knowledge Discovery,* vol. 10, no. 1, pp. 39-79, 2005.

[4] E. Babbie, F. Halley, and J. Zaino, *Adventures in Social Research, Data Analysis Using SPSS 11.0/11.5 for Windows.* Pine Forge Press, 2003.

[5] C. Frankfort-Nachmias and A. Leon-Guerrero, *Social Statistics for a Diverse Society.* Pine Forge Press, 2006.

[6] J.F. Healey, E.R. Babbie, and J. Boli, *Exploring Social Issues: Using SPSS for Windows 95, Versions 7.5, 8.0, or Higher.* Pine Forge Press, 1999.

[7] B. Liu, W. Hsu, and Y. Ma, "Mining Association Rules with Multiple Minimum Supports," *Proc. ACM SIGKDD,* pp. 337-341, 1999.

[8] Y.-H. Hu, "An Efficient Algorithm for Discovering and Maintenance of Frequent Patterns with Multiple Minimum Supports," master's thesis, Dept. of Information Management, Nat'l Central Univ., 2003.

[9] Y.S. Koh, N. Rountree, and R.A. O'Keefe, "Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse," *Int'l J. Data Warehousing and Mining,* vol. 2, pp. 38-54, 2006.

[10] H. Mannila, "Database Methods for Data Mining," *Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (Tutorial),* 1998.

[11] Y.-H. Hu and Y.-L. Chen, "Mining Association Rules with Multiple Minimum Supports: A New Mining Algorithm and a Support Tuning Mechanism," *Decision Support Systems,* vol. 42, pp. 1-24, 2006.

[12] W.-Y. Lin, M.-C. Tseng, and J.-H. Su, "A Confidence-Lift Support Specification for Interesting Associations Mining," *Proc. Sixth Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining (PAKDD),* pp. 148-158, 2002.

[13] H. Yun, D. Ha, B. Hwang, and K.H. Ryu, "Mining Association Rules on Significant Rare Data Using Relative Support," *J. Systems Software,* vol. 67, pp. 181-191, 2003.

[14] S. Brin, R. Motwani, and C. Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations," *Proc. 1997 ACM SIGMOD,* pp. 265-276, 1997.

[15] G.I. Webb, "Association Rules," *The Handbook of Data Mining,* pp. 26-39. Mahwah, 2003.

[16] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, 2006.

[17] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," *SIGMOD Record,* vol. 26, pp. 255-264, 1997.

[18] T. Scheffer, "Finding Association Rules That Trade Support Optimally against Confidence," *Intelligent Data Analysis,* vol. 9, pp. 381-395, 2005.

[19] J. Li and Y. Zhang, "Direct Interesting Rule Generation," *Proc. Third IEEE Int'l Conf. Data Mining,* pp. 155-162, 2003.

[20] X. Wu, C. Zhang, and S. Zhang, "Efficient Mining of Both Positive and Negative Association Rules," *ACM Trans. Information Systems,* vol. 22, pp. 381-405, 2004.

[21] C.C. Aggarwal and P.S. Yu, "A New Framework for Itemset Generation," *Proc. 17th ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems (PODS '98),* pp. 18-24, 1998.

[22] J. Blanchard, F. Guillet, H. Briand, and R. Gras, "Assessing Rule Interestingness with a Probabilistic Measure of Deviation from Equilibrium," *Proc. 11th Int'l Symp. Applied Stochastic Models and Data Analysis (ASMDA '05),* pp. 191-200, 2005.

[23] M.-L. Antonie and O.R. Zaïane, "Mining Positive and Negative Association Rules: An Approach for Confined Rules," *Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD '04),* pp. 27-38, 2004.

[24] G. Piatetsky-Shapiro, "Discovery, Analysis, and Presentation of Strong Rules," *Knowledge Discovery in Databases,* pp. 229-248, AAAI/MIT Press, 1991.

[25] A.M. Liebetrau, *Measures of Association.* Sage Publications, 1983.

[26] Encyclopedia Britannica, "Analytic Proposition," http://www.britannica.com/eb/article-9007348, 2007.

[27] N. Rescher, *Conditionals.* MIT Press, 2007.

[28] Open University (Mathematics Foundation Course Team), *Logic II: Proof,* vol. 17. The Open Univ. Press, 1971.

[29] W.V.O. Quine, *Mathematical Logic,* second ed. Harper & Row Publishers, 1951.

[30] Encyclopedia Britannica, "Condition," http://www.britannica.com/eb/article-9025123, 2007.

[31] R. Forsyth, "Zoo Data Set," Orange, AI Lab, http://magix.fri.uni-lj.si/orange/doc/datasets/zoo.htm, 1990.

[32] C. Borgelt, "A Priori—Association Rule Induction/Frequent Item Set Mining," http://www.borgelt.net/apriori.html, 2008.

[33] S.-J. Yen and Y.-S. Lee, "Mining Interesting Association Rules: A Data Mining Language," *Advances in Knowledge Discovery and Data Mining,* pp. 172-176, Springer, 2002.

[34] A. Das, D.K. Bhattacharyya, and J.K. Kalita, "Horizontal versus Vertical Partitioning in Association Rule Mining: A Comparison," *Proc. Sixth Int'l Conf. Computational Intelligence and Natural Computation (CINC),* pp. 1617-1620, 2003.

[35] S. Chiu, W.-k. Liao, and A. Choudhary, "Design and Evaluation of Distributed Smart Disk Architecture for IO-Intensive Workloads," *Proc. Int'l Conf. Computational Science (ICCS '03),* pp. 230-241, 2003.

[36] IBM, "Quest Synthetic Data Generation Code," http://www.almaden.ibm.com/cs/projects/iis/hdb/Projects/data_mining/datasets/syndata.html, 2010.

Alex Tze Hiang Sim received the BSc (Hons.) and Master of Information Technology degrees from the Universiti of Teknologi Malaysia (UTM), and the Master of Business Systems degree and the PhD degree in computer science from Monash University. He worked for six years in the field related to costing and information technology support after his first graduation. He is currently with the Universiti of Teknologi Malaysia. His current interest is in data mining techniques and its applications.

Maria Indrawan received the doctorate degree from Monash University in 1998. She has been working with the Faculty of Information Technology, Monash University, Melbourne, Australia, since 1998. She has been involved in program committees and the chair of several International conferences and workshops in the area of Internet technology, multimedia retrieval, grid computing, and provenance in e-science.

Samar Zutshi received the BSc degree from Madurai Kamaraj University, in 1999, and the MIT and PhD degrees from Monash University, in 2002 and 2008, respectively. He is a lecturer in business, management and technology at Swinburne University of Technology, Australia. His current research interests include content-based multimedia retrieval and personalized learning-based pedagogy. He is a member of the ACM and the IEEE.

Bala Srinivasan received the Gold Medal in the Bachelor of Engineering (Honors) degree in electronics and communication engineering from Guindy Engineering College, University of Madras, India, and the master's and PhD degrees in computer science from the Indian Institute of Technology, Kanpur, India. He is the senior most professor in the Faculty of Information Technology, Monash University, Melbourne, Australia, where he joined in 1984 as a lecturer and progressively promoted to the chair in Information Technology in 1993. Until now, he has supervised to completion either as a principal or joint supervisor of 41 research theses of which 27 of them are PhDs. Two of those supervised PhDs got University awards for the best thesis in the faculty. His success in research supervision was recognized by the Monash University by awarding him the Vice-Chancellors Medal for Excellence in Research Training in 2002. He is a founding chairman of the Australiasian database conference which is now being held annually. He has authored and/or jointly edited six technical books and authored and/or coauthored around 50 journal articles and nearly 200 conference and book chapters. His areas of research interests are varied and wide and include Multimedia Retrieval Systems, Distributed databases, Mobile Computing, and Data Mining.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.