Self-Sufficient Itemsets: An Approach to Screening Potentially Interesting Associations Between Items

GEOFFREY I. WEBB
Monash University, Australia

3

Self-sufficient itemsets are those whose frequency cannot be explained solely by the frequency of either their subsets or of their supersets. We argue that itemsets that are not self-sufficient will often be of little interest to the data analyst, as their frequency should be expected once that of the itemsets on which their frequency depends is known. We present tests for statistically sound discovery of self-sufficient itemsets, and computational techniques that allow those tests to be applied as a post-processing step for any itemset discovery algorithm. We also present a measure for assessing the degree of potential interest in an itemset that complements these statistical measures.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Association discovery, association rules, itemset discovery, itemset screening, statistical evaluation

ACM Reference Format:

Webb, G. I. 2010. Self-sufficient itemsets: An approach to screening potentially interesting associations between items. ACM Trans. Knowl. Discov. Data. 4, 1, Article 3 (January 2010), 20 pages.
DOI = 10.1145/1644873.1644876 http://doi.acm.org/10.1145/1644873.1644876

1. INTRODUCTION

Itemsets are collections of items that co-occur in data. Historically, their primary use in data mining has been as an intermediate step in discovery of association rules [Agrawal et al. 1993]. However, they are often of potential interest in their own right. This is because it can be interesting to discover multiple items that co-occur with unexpected frequency, and the division of those

This research has been supported by the Australian Research Council under grant DP0772238.
Author's address: G. I. Webb, Faculty of Information Technology, Monash University, Clayton, Vic., 3800, Australia; email: webb@infotech.monash.edu.au.

DOI 10.1145/1644873.1644876 http://doi.acm.org/10.1145/1644873.1644876

items into subsets, as required to form a rule, often provides no additional useful information and may result in the discovery of many artifacts (rules) for every correlation of interest. For example, take the classic market-basket example of beer and diapers. If these two products co-occur more frequently than expected, association rule discovery will find two rules $beer \rightarrow diapers$ and $diapers \rightarrow beer$, whereas the essential discovery embodied in these rules might be more succinctly represented by the single itemset {*beer, diapers*}. There is also potential for the representation of correlations as rules to mislead users into believing that the antecedents and consequents have specific causal relationships, even though this has not been established. Further, when there are more than two items, there is potential for users to fail to realize that multiple representations of the one higher-order correlation are related. For example, a user might fail to detect that the two rules $\{flour, eggs, milk\} \rightarrow sugar$ and $\{eggs, milk, sugar\} \rightarrow flour$ both arise from a single multivariate correlation, especially if they are not listed proximately in a system's output.

While there is a large literature about identifying and discovering interesting association rules [Piatetsky-Shapiro 1991; Agrawal et al. 1993; Bayardo and Agrawal 1999; Bayardo et al. 2000; Bastide et al. 2000; Zaki 2000, 2004; Calders and Goethals 2002; Liu et al. 2001; Webb 2007], the literature on identifying and discovering interesting itemsets is relatively sparse.

The current article contributes a set of constraints and statistical tests that can be applied during and after itemset discovery to identify itemsets whose frequency can be explained solely by either higher or lower-order interactions between factors within the data. These techniques have been developed as a result of our extensive experience in practical applications of itemset discovery. We believe that for many applications the itemsets they exclude will be of little interest. The techniques may be applied as a postprocessing step to filter collections of itemsets that have been identified as potentially interesting using application-specific measures.

We also present *itemset leverage*, a measure for degree of potential interest that arises naturally from the tests that we develop.

2. PROBLEM STATEMENT

Given a domain of items I, an itemset S is a set of items $S \subseteq I$. A dataset D is a vector of n records $\langle d_1 \ldots d_n \rangle$. Each record d_i is an itemset. For transactional data, items are atomic terms. For attribute-value data, there exists a set of m attributes $A_1 \ldots A_m$, each attribute A_i has a domain of values $\text{dom}(A_i)$, each item is an attribute-value pair denoted as $A_i = v$, where $v \in \text{dom}(A_i)$, and each record d_i contains at most one item for each attribute. The indices for records are called their *transaction identifiers or TIDs*. The *cover* of an itemset S, $\text{cov}(S)$, is the set of TIDs for the records that contain S,

$$\text{cov}(S) = \{i : 1 \leq i \leq n \wedge S \subseteq d_i\}. \tag{1}$$

Note, we use TIDs here, rather than records, in order to distinguish multiple identical records.

The *support* of an itemset S is the proportion of records in dataset D of which S is a subset:

$$\sup(S) = |\operatorname{cov}(S)|/n. \tag{2}$$

We use $|\cdot|$ to denote the cardinality of a set. We use *support count* to denote the number of records of which S is a subset, $|\operatorname{cov}(S)|$.

Usually we are examining D as a sample drawn from some distribution $\mathcal{D}$ in order to make inferences about $\mathcal{D}$. In particular, we often wish to make inferences about

$$\overset{*}{\sup}(S) = P_{\mathcal{D}}(S), \tag{3}$$

the probability that a random case from $\mathcal{D}$ will contain S.

The itemset discovery task is to find interesting itemsets with reference to D. Which itemsets will be interesting will vary greatly depending upon the specific analytic task. It is not credible that any one criterion will identify exactly the itemsets that will prove interesting for all analytic objectives, but some criteria may prove useful for a wide range of objectives.

There is a subtle but important difference between the objectives of finding interesting itemsets and of finding interesting correlations between variables. The latter has been widely studied in statistics [Tabachnick and Fidell 2001; Agresti 2002]. Itemset discovery seeks interactions between specific attribute-values rather than between variables as a whole. This is the specific focus of many real-world analytic tasks, as the user wishes to know which circumstances lead to either a positive or a negative effect. For example, in the context of screening for adverse drug reactions [DuMouchel 1999], the interest is in knowing which specific specific combinations of drugs and side-effects co-occur with unexpected frequency, rather than simply to identify which variables interact.

In the current research we focus on the task of identifying interesting positive interactions between attribute-values or items. That is, we focus on finding itemsets that co-occur more frequently than expected. Many of the techniques generalize directly to negative interactions, but different search procedures are required in that context.

3. IDENTIFYING UNINTERESTING ITEMSETS

In the absence of prior expectations about the correlations that exist between items, and when positive correlations are sought, an itemset S might be considered potentially interesting if its frequency is greater than would be obtained if the component items were independent of one another,

$$\sup(S) > \prod_{s \in S} \sup(\{s\}). \tag{4}$$

However, this criterion alone is insufficient. Take any itemset S that represents a correlation between its constituent elements, and any further item i that is independent of S. The itemset $S \cup \{i\}$ will pass (4) and hence will be identified as potentially interesting. This is illustrated in Figure 1. Of course, one would not normally make any inferences from so little data. However, for

1: x
2: batteries
3: beer, diapers
4: beer, diapers, batteries

Note: x is used as a placeholder for a record that does not contain any of the itemsets of interest.

Fig. 1. Illustration of inadequacy of independence as a sole criterion for identifying uninteresting itemsets.

ease of exposition, in this and the following figure, we assume the few examples presented perfectly reflect the distribution from which they are drawn. In Figure 1, *beer* and *diapers* pass (4), as $\sup(\{beer, diapers\}) = \sup(\{beer\}) = \sup(\{diapers\}) = 0.5 > \sup(\{beer\}) \times \sup(\{diapers\}) = 0.25$. In contrast, $\{batteries\}$ is independent of $\{beer, diapers\}$, as $\sup(\{beer, diapers, batteries\}) = \sup(\{beer, diapers\}) \times \sup(\{batteries\})$. However, $\{beer, diapers, batteries\}$ satisfies (4) and hence is also potentially interesting according to this criterion, as $\sup(\{beer, diapers, batteries\}) = 0.25 > \sup(\{beer\}) \times \sup(\{diapers\}) \times \sup(\{batteries\}) = 0.125$. It seems unacceptable that an itemset should be considered potentially interesting when it contains an item (*batteries*) that is independent of the remaining items.

A stronger condition than (4) is the *productive* criterion, so named because it is the counterpart for itemsets of the productive criterion for rules [Webb 2007]. This specifies that the frequency of the itemset must be greater than that which would be expected under any assumption of independence between any partition of the items into two independent itemsets:

$$\text{productive}(S) = \forall S_1, S_2 : S_1 \subset S \wedge S_2 \subset S \wedge S_1 \cup S_2 = S \wedge S_1 \cap S_2 = \emptyset \implies \sup(S) > \sup(S_1)\sup(S_2). \quad (5)$$

In other words, an itemset is productive if and only if every rule that can be formed from it is productive.

Note that this criterion also covers the case of more than two subsets of the items being independent of one another. Suppose three items a, b, and c are independent of each other. In this case $\sup(\{a, b\}) = \sup(\{a\})\sup(\{b\})$, so $\sup(\{a, b, c\}) = \sup(\{a, b\})\sup(\{c\})$, and hence $\{a, b, c\}$ is not productive.

Continuing our example from Figure 1, the itemset $\{beer, diapers, batteries\}$ does not satisfy (5) (under the binding $S_1 = \{beer, diapers\}$ and $S_2 = \{batteries\}$), and hence can be judged uninteresting. When considering whether S is productive with respect to distribution $\mathcal{D}$ we use

$$\text{productive}^*(S) = \forall S_1, S_2 : S_1 \subset S \wedge S_2 \subset S \wedge S_1 \cup S_2 = S \wedge S_1 \cap S_2 = \emptyset \implies \overset{*}{\sup}(S) > \overset{*}{\sup}(S_1)\overset{*}{\sup}(S_2). \quad (6)$$

However, even this condition fails to handle some important cases. Consider any itemset S and item i, such that $\text{cov}(S) \subset \text{cov}(\{i\})$. For any itemset $R \supseteq S$, $\text{cov}(R) = \text{cov}(R \cup \{i\})$. For example, the coverage of {pregnant, proteinuria} should be identical to the coverage of {female, pregnant, proteinuria}, as the

1: x
2: beer
3: diapers
4: pretzels
5: beer, pretzels
6: beer, diapers
7: diapers, pretzels
8: beer, diapers, pretzels
9: beer, diapers, pretzels
10: beer, diapers, pretzels

Note: x is used as a placeholder for a record that does not contain any of the itemsets of interest.

Fig. 2. Illustration of inadequacy of (5) and (7) as sole criteria for identifying uninteresting itemsets.

item *female* is a generalization of *pregnant* and should apply to all records to which *pregnant* applies. We can thus define *nonredundant* itemsets, a concept directly related to nonredundant rules [Bastide et al. 2000; Zaki 2000].

$$\text{nonredundant}(S) = \forall\, Q \subset R \subset S,\ \text{sup}(Q) > \text{sup}(R). \tag{7}$$

It is credible that, in most applications, redundant itemsets (those that are not nonredundant) will not be of interest, as the addition of the generalization adds no information to the itemset. When considering whether S is nonredundant with respect to distribution $\mathcal{D}$ we use:

$$\text{nonredundant}^*(S) = \forall\, Q \subset R \subset S,\ \overset{*}{\text{sup}}(Q) > \overset{*}{\text{sup}}(R). \tag{8}$$

Note that {female, pregnant} is nonredundant. While {female, pregnant} should have the same support as {pregnant}, (7) requires that there is a proper subset R of S that has identical support to one of its subsets. Thus any superset of {female, pregnant} will be redundant but {female, pregnant} is not. This is appropriate. If it were not known a priori that all pregnant people are female, the fact that they are could be an interesting discovery. However, once this is known, it should be expected that every superset {female, pregnant}$\cup X$ will have the same support as {pregnant} $\cup X$, and hence {female, pregnant}$\cup X$ will not usually be of interest.

The productive and nonredundant criteria both filter more specific itemsets on the basis of their subsets. However, there are further cases where it may be desirable to filter generalizations on the basis of their supersets. To illustrate such a situation, consider the hypothetical data in Figure 2. The following itemsets are all productive and nonredundant, {*beer*, *pretzels*}, {*beer*, *diapers*}, {*diapers*, *pretzels*} and {*beer*, *diapers*, *pretzels*}. Nonetheless, the two-item itemsets are all potentially misleading, as it is only due to the high frequency of the three-item itemset that the frequency of the two-item itemsets is raised. None of the three pairs is correlated except in the context of the remaining item. To identify any one of the two-item itemsets as potentially interesting is potentially misleading, as it fails to specify an important

constraint on the conditions under which the two items are correlated. This situation can be addressed as follows.

We first define the *exclusive domain* of an itemset S as:

$$\text{edom}(S) = \{1 \ldots n\} \setminus \bigcup_{\substack{R \supset S \\ \text{productive}(R) \\ \text{nonredundant}(R)}} \text{cov}(R \setminus S). \tag{9}$$

This is the set of TIDs for records not covered by any of the sets of additional items in any nonredundant productive superset of S. For an itemset S to be *self-sufficient*, it must both be productive and nonredundant, as already defined, as well as being productive with respect to its exclusive domain.

$$\begin{aligned} &\text{ssuf}(S) = \text{productive}(S) \wedge \text{nonredundant}(S) \\ &\quad \wedge \forall S_1, S_2 : S_1 \subset S \wedge S_2 \subset S \wedge S_1 \cup S_2 = S \wedge S_1 \cap S_2 = \emptyset \\ &\quad \Longrightarrow |\text{cov}(S) \cap \text{edom}(S)| > \frac{|\text{cov}(S_1) \cap \text{edom}(S)| \times |\text{cov}(S_2) \cap \text{edom}(S)|}{|\text{edom}(S)|}. \end{aligned} \tag{10}$$

We illustrate this requirement with respect to the itemset {*beer*, *pretzels*} and the data presented in Figure 2. Recall, {*beer*, *pretzels*} is productive because its support is 0.4, which is greater than the product of the supports of its subsets {*beer*} (0.6) and {*pretzels*} (0.6). However, it only attains greater frequency than would be expected if *beer* and *pretzels* were independent of one another due to the frequency with which {*beer, diapers, pretzels*} occurs. This latter itemset is the only productive superset of {*beer, pretzels*}. Hence, the exclusive domain of {*beer*, *pretzels*} $= \{1 \ldots 10\} \setminus \text{cov}(\{beer, diapers, pretzels\} \setminus \{beer, pretzels\}) = \{1 \ldots 10\} \setminus \text{cov}(\{diapers\}) = \{1, 2, 4, 5\}$. We now test whether {*beer*, *pretzels*} is productive relative to this set of TIDs. The cover of {*beer*, *pretzels*} within this set of TIDs is 1. The covers of both {*beer*} and {*pretzels*} within the exclusive domain are 2 and the size of the exclusive domain is 4. So the expected cover of {*beer*, *pretzels*} $= 2 \times 2/4 = 1$. As the actual cover does not exceed the expected, {*beer*, *pretzels*} is not self-sufficient.

When considering whether S is self-sufficient with respect to $\mathcal{D}$ we use:

$$\text{e}^*(x, S) = \neg\exists R \supset S \wedge x \supseteq (R \setminus S) \wedge \text{productive}^*(R) \wedge \text{nonredundant}^*(R) \tag{11}$$

$$\begin{aligned} &\text{ssuf}^*(S) = \text{productive}^*(S) \wedge \text{nonredundant}^*(S) \\ &\quad \wedge \forall S_1, S_2 : S_1 \subset S \wedge S_2 \subset S \wedge S_1 \cup S_2 = S \wedge S_1 \cap S_2 = \emptyset \\ &\quad \Longrightarrow \frac{P_{\mathcal{D}}(x : x \supseteq S \wedge \text{e}^*(x, S))}{P_{\mathcal{D}}(y : \text{e}^*(y, S))} \\ &\qquad > \frac{P_{\mathcal{D}}(x_1 : x_1 \supseteq S_1 \wedge \text{e}^*(x_1, S))}{P_{\mathcal{D}}(y : \text{e}^*(y, S))} \times \frac{P_{\mathcal{D}}(x_2 : x_2 \supseteq S_2 \wedge \text{e}^*(x_2, S))}{P_{\mathcal{D}}(y : \text{e}^*(y, S))}. \end{aligned} \tag{12}$$

Self-sufficient itemsets are similar in spirit to *actionable rules* [Liu et al. 2001]. The latter are association rules that have higher support than can be predicted from either their generalizations or specializations. The current work translates this principal into the itemset context.

4. IDENTIFYING SELF-SUFFICIENT ITEMSETS

We have determined that itemsets that are not self-sufficient are unlikely to be interesting in many contexts. However, this does not necessarily imply that all self-sufficient itemsets will be interesting. Further, the experimental evidence presented in the following suggests that in many applications there will be very large numbers of self-sufficient itemsets. In consequence, it will often be desirable to couple tests for self-sufficiency with methods that highlight application-specific features of an itemset that make it most likely to be interesting. We present here statistical tests for identifying whether an itemset is self-sufficient. These are intended to be applied as a post-processing step to filter itemsets that are identified as potentially interesting on other grounds.

4.1 Statistical Tests for Self-Sufficient Itemsets

False discoveries are a serious potential problem for any pattern discovery system [Webb 2007]. When patterns are inferred from data that are a sample of some distribution, it is almost certain that some patterns S will have support within the sample that is substantially higher than their probability in the distribution, $\sup(S) \gg \sup^*(S)$. If we do not wish to discover large numbers of spurious patterns, it is critical that we employ appropriate statistical tests. Thus, it is necessary to develop suitable statistical tests for (6), (8), and (12).

For (6) and (12) the appropriate form of statistical test is a hypothesis test. These require the formation of a *null-hypothesis*, which is the negation of the hypothesis we wish to accept. The hypothesis test then seeks to reject the null-hypothesis. This is done if it can be established that the probability p that D, or a more extreme set of data than D, would be obtained if the null-hypothesis were true, is less than a significance level α, which is often set to 0.05.

There are three primary approaches to sound statistical testing for pattern discovery. *Randomization tests* [Megiddo and Srikant 1998] repeatedly randomize the data in such a manner as to establish the null-hypothesis for a statistical test. For example, if the null-hypothesis is that no items within the data are correlated, then the columns in the data can be randomized, thereby breaking any correlations between them. Patterns are then discovered from the randomized data, and the distribution is observed of the most extreme value of some measure (such as support) of each set of discoveries. This distribution can be used to determine the statistical significance of an observed value of the measure with respect to the unrandomized data.

Randomization tests do not appear to be suited to the current problem, as there are multiple null hypotheses required and it does not appear possible to instantiate them all through a single randomization of the data. For example, to test whether {*beer*, *diapers*, *pretzels*} is productive it is necessary to test whether *diapers* is independent of {*beer, pretzels*}, *pretzels* is independent of {*beer*, *diapers*} and *beer* independent of {*diapers*, *pretzels*}, in each case keeping any correlation that might exist between the pair but breaking any that might exist with the singleton.

The *within-search testing* approach [Webb 2007] applies a statistical test during the search process, using a Bonferroni adjustment [Shaffer 1995] to the critical-value that is employed. This is achieved by dividing the desired experimentwise critical-value by the size of the search space. This ensures that the risk of any false discovery is no greater than the specified experimentwise critical-value.

The *holdout evaluation* approach [Webb 2007] partitions the available data into *exploratory* and *holdout* sets. Patterns are discovered through analysis of the exploratory data. These are then evaluated via statistical tests with respect to the holdout data. To apply this approach to discover self-sufficient itemsets, we can first find potentially interesting, nonredundant and productive itemsets in a single pass through the exploratory data, and then statistically test their self-sufficiency with respect to the holdout data. We use holdout evaluation rather than within-search testing because previous work [Webb 2007] suggests that it is more powerful, finding greater numbers of significant associations.

To test (6) with respect to itemset S we need to assess:

$$\overset{*}{\text{sup}}(S) > \overset{*}{\text{sup}}(S_1) \times \overset{*}{\text{sup}}(S_2) \tag{13}$$

for each S_1 and S_2, such that $S_1 \subset S \wedge S_2 \subset S \wedge S_1 \cup S_2 = S \wedge S_1 \cap S_2 = \emptyset$.

The Fisher exact test [Agresti 1992] can be used to evaluate (13). For each S, S_1, and S_2, calculate the probability of observing the observed number or more occurrences of S given the number of observed occurrences of S_1 and S_2 if $P(S) = P(S_1) \times P(S_2)$.

To calculate this probability, let $a = |\operatorname{cov}(S)|$, $b = |\operatorname{cov}(S_1) \setminus \operatorname{cov}(S)|$, $c = |\operatorname{cov}(S_2) \setminus \operatorname{cov}(S)|$ and $d = |\{1 \ldots n\} \setminus (\operatorname{cov}(S_1) \cup \operatorname{cov}(S_2))|$.

$$p = \sum_{i=0}^{min(b,c)} \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!(a+i)!(b-i)!(c-i)!(d+i)!}. \tag{14}$$

Here, $n!$ denotes the factorial of n. See Agresti [1992] for more details.

As this is an exact test, the p-value is reliable even with infrequent data. (12) is assessed using the same test, except that $a = |\operatorname{cov}(S) \cap \operatorname{edom}(S)|$, $b = |(\operatorname{cov}(S_1) \cap \operatorname{edom}(S)) \setminus \operatorname{cov}(S)|$, $c = |(\operatorname{cov}(S_2) \cap \operatorname{edom}(S)) \setminus \operatorname{cov}(S)|$ and $d = |\operatorname{edom}(S) \setminus (\operatorname{cov}(S_1) \cup \operatorname{cov}(S_2))|$.

We calculate (14) with respect to every division of S into subsets S_1 and S_2, and take the highest of the resulting p-values. We must then perform a statistical adjustment for the number of itemsets tested against the holdout data. We use the Holm [1979] procedure, which is more powerful than the better known Bonferroni adjustment [Shaffer 1995], while still guaranteeing that the experimentwise risk of any type-1 error is no more than critical-value α. We take the highest p-value associated with each of the x itemsets and order them from the lowest, p_1, to the highest, p_x. We then set $\kappa = \max(p_i : \forall 1 \leq j \leq i, p_j \leq \alpha / (x - j + 1))$, that is the highest p_i such that all p_j up to and including p_i pass the test $p_j \leq \alpha / (x - j + 1)$. If no p_i satisfies this test then $\kappa = 0.0$.

Testing for nonredundancy (8) is more problematic. Traditional hypothesis test statistics cannot address situations where the null-hypothesis is a test for

equality. On the face of it, the negation of (8) is

$$\exists\, Q \subset R \subset S,\ \overset{*}{\sup}(Q) \leq \overset{*}{\sup}(R), \tag{15}$$

which does not involve an equality test. However, the support of a subset cannot be less than the support of its superset. Hence, (15) implies:

$$\exists\, Q \subset R \subset S,\ \overset{*}{\sup}(Q) = \overset{*}{\sup}(R), \tag{16}$$

which requires a test for equality. For any $Q \subset R$, Q must cover all cases covered by R. Hence, the issue is whether Q covers any cases not covered by R. For any $Q \subset R \subset S$, a single example of a case covered by Q but not R, establishes that $P(\sup^*(Q) = \sup^*(R)) = 0.0$, allowing us to reject the null-hypothesis with respect to that Q and R. On the other hand, if there is any $Q \subset R \subset S$, $\sup(Q) = \sup(R)$, we cannot assess the probability of D given (15) and hence cannot reject the null hypothesis. In consequence, we accept as nonredundant (with respect to $\mathcal{D}$) all and only itemsets that are nonredundant (satisfy 7) with respect to D, and do not apply a statistical test.

4.2 Computational Considerations

While the Fisher exact test has a reputation for high computational overheads, it is actually polynomial with respect to the size of the data and the computation is minor compared to many other facets of the pattern discovery process.

However, the number of S_1, S_2 pairs that must be examined for any S when assessing each of (6) and (12) is $O(2^{|S|})$. This may prove prohibitive if very long itemsets need testing. Furthermore, the test for self-sufficiency requires identification of every productive superset of each productive itemset S. This requires $O(q^2)$ comparisons, where q is the number of productive itemsets. In consequence, the process may not scale well with respect to the number or size of the itemsets to be post-processed. We explore the practical implications for the scalability of the approach in Section 7.3.

5. QUANTIFYING THE DEGREE OF POTENTIAL INTEREST

The proposed method for testing whether an itemset is productive suggests a strategy for assessing the degree to which an itemset is productive. The test assesses whether an itemset S has higher support than would be expected under any assumption of independence between subsets of S. This suggests that a direct manner to quantify degree of productivity is to evaluate how much greater is the support of S than would be expected under any such assumption.

$$\mathrm{lev}(S) = \sup(S) - \operatorname*{argmax}_{\substack{U \subset S,\, V \subset S \\ U \cap V = \emptyset \\ U \cup V = S}} (\sup(U) \times \sup(V)). \tag{17}$$

We call this measure *itemset leverage*, as it equals the minimum leverage [Piatetsky-Shapiro 1991; Webb and Zhang 2005] of any rule that can be formed using all items in S.

Itemset leverage has some similarity to DuMouchel and Pregibon's [2001] *excess*-2 measure. Excess-2 seeks to measure the difference in frequency between that observed for the itemset and that accountable solely in terms of

second-order (two-item) interactions within the itemset. Itemset leverage seeks to measure the difference in frequency between that observed for the itemset and that accountable for by any lower-level interactions within the data.

6. RELATED APPROACHES

The standard approach to itemset discovery is to find all frequent itemsets [Agrawal et al. 1993]. These are itemsets that occur frequently in the data, or to express this another way, have high support. However, frequent itemsets will often be of little interest, as very frequent items should be expected to co-occur frequently, and hence being frequent is insufficient evidence for assuming that an itemset occurs more frequently than should be expected [Webb 2007]. This section surveys a number of existing approaches to identifying potentially interesting itemsets.

6.1 Closed Itemsets and the Condensed Frequent Pattern Base

One manner in which itemsets can be filtered is by deleting those that are not *closed* [Zaki and Hsiao 2002; Bastide et al. 2000]. An itemset S is closed if and only if

$$\neg\exists R \supset S, \sup(R) = \sup(S). \tag{18}$$

Deleting itemsets that are not closed allows a reduced set of itemsets to be found from which all frequent itemsets and their precise statistics can be inferred. These techniques have an important role to play in efficient approaches to itemset discovery.

A further extension to this approach is the condensed frequent pattern base [Pei et al. 2002; Xin et al. 2005; Cheng et al. 2006]. This is a subset of the frequent itemsets from which all frequent itemsets can be inferred and from which every frequent itemset's support can be inferred to within a user-specified precision.

While these approaches can greatly reduce the number of itemsets that are provided to the user, they do not directly address the issue of whether the discovered patterns are more frequent than expected, and hence go only part way toward identifying the interesting itemsets.

6.2 Nonderivable Itemsets

Non-derivable itemsets [Calders and Goethals 2002] provide an alternative approach. An itemset is *derivable* if its support can be inferred from the supports of its subsets using the inclusion-exclusion principle. This is done as follows. Let:

$$\sigma_S(R) = \sum_{R \subseteq R' \subset S} (-1)^{|S \setminus R'|+1} \sup(R') \tag{19}$$

$$l = \max\{\sigma_S(R) \mid R \subset S, \text{even}(|R|)\} \tag{20}$$

$$u = \min\{\sigma_S(R) \mid R \subset S, \text{odd}(|R|)\}, \tag{21}$$

```
{} [8124,1.00]
{52} [3516,0.43]
{29} [2160,0.27]
{101} [1632,0.20]
{29, 101} [1584,0.19]
{52, 101} [1344,0.17]
{29, 52} [1296,0.16]
{29, 52, 101} [1296,0.16]
```

Fig. 3. An example of a derivable itemset {29, 52, 101} that may be interesting.

$\sup(S) = l = u$ if and only if $l = u$. If $\sup(S) = l = u$ (and hence $\sup(S) = l = u$), then S is derivable.

The support of all itemsets can be inferred from the support of the nonderivable itemsets, and so the set of nonderivable itemsets and their supports contains all the support information contained by the set of all itemsets and their supports. Nonderivable itemsets tend to be fewer in number than open itemsets [Calders and Goethals 2002]. Nonderivable itemset techniques have an important role to play in techniques for efficient itemset discovery. If the analytic objective is to find a compact representation of the support of all itemsets then nonderivable itemsets provide an effective and elegant solution.

It is also, at least on the face of it, credible that they address to some degree the issue of whether itemsets are more frequent than expected, as if an itemset is derivable then its support can be derived from the support of its subsets. However, it does not follow from the existence of an inference mechanism that can determine the support that its existence will necessarily be apparent to a human observer. Further, it is credible that in some circumstances it is the more specific itemset, the superset, that will most succinctly capture the interesting interdependencies in the data, rather than the more general ones, even if the supports of the more general ones can be used to derive the support of the more specific.

These points are illustrated by the itemsets in Figure 3. These itemsets are derived from the mushroom.dat dataset donated by Roberto Bayardo to the FIMI repository.[1] The itemset {29, 52, 101} is derivable because its support can be inferred from the support of its subsets, which are listed in Figure 3 (support count and support listed within [-] brackets). sup({29, 52}) sets an upper bound on sup({29, 52, 101}). A lower bound is established by sup({29, 101}) + sup({52, 101}) – sup({101}). As the lower and upper bounds both equal 0.16, sup({29, 52, 101}) must be 0.16. However, it is far from clear that the fact that {29, 52, 101} is derivable determines that it is likely to be uninteresting. One view of why it is derivable is due to all transactions containing {29, 52} also containing 101. This fact may be interesting, and is perhaps captured most succinctly by the two itemsets {29, 52} [1296, 0.16] and {29, 52, 101} [1296, 0.16]. Thus, a derivable itemset might be part of a succinct summary of the interesting interactions between items.

On the other hand, some nonderivable itemsets might not be very interesting. For example, if two items 0 and 1 both have support 0.5 and the support of

[1]http://fimi.cs.helsinki.fi/data/

itemset {0, 1} is 0.25, then the latter is nonderivable, but is unlikely to be interesting, as its support indicates that the two items are independent of one another. Nonderivable itemsets provide a useful approach for reducing the set of itemsets to a subset that is richer in itemsets that are likely to be of interest, but it does not provide a solution that will always identify all and only, the interesting itemsets for any given application.

6.3 Itemsets Associated with Numeric Values

There are a number of techniques that have been developed for contexts where records have associated numeric values, such as the profit derived from a transaction [Chan et al. 2003; Yao and Hamilton 2006; Aumann and Lindell 1999; Webb 2001; Zhang et al. 2004]. These techniques find itemsets that optimize a variety of statistics with respect to the associated values of the records that the itemset covers. The current work differs from these approaches in that it addresses a situation where such associated values are not present and utilizes only the frequency of itemsets to assess their potential interest.

6.4 Interesting Itemset Discovery

One approach that does directly address the problem of identifying itemsets that occur more often than expected is *interesting itemset discovery* [Jaroszewicz and Simovici 2004; Cooley et al. 1999]. This approach uses a user-specified Bayesian network to determine the expected frequency of itemsets and then identifies itemsets that differ significantly from their expected frequency. This is a very powerful approach. The current work seeks to build upon the philosophy that underlies this approach, presenting techniques that allow discovery of interesting itemsets in contexts where a Bayesian network representing known interactions between variables is not available.

A Bayesian network specifies interdependencies between items. Itemsets whose frequency is consistent with known interdependencies are unlikely to be interesting. However, when there is little or no apriori knowledge of interdependencies, an alternative approach has to be pursued. The approach we pursue is to treat as likely to be uninteresting, any itemset whose frequency cannot be accounted for by the interdependencies inherent in other identified itemsets.

6.5 Closed Interesting Itemsets

One approach to this end calculates interestingness measures developed for rule discovery [Malik and Kender 2006]. These are applied during the itemset search process, treating the current k-itemset as the antecedent and the candidate addition to that itemset as the consequent. This approach may result in different itemsets being identified as interesting, depending on the path traversed through the search space.

The current work is complementary to interestingness measures such as these. Such interestingness measures can be used to rank itemsets on inherent properties that are likely to make them more or less interesting for some specific application. Our statistical tests can then be used to screen these to remove

itemsets that are likely to not be as interesting as the inherent properties suggest because of their relationship to either more specific, or more general, variants.

We also present the itemset leverage measure for assessing the level of potential interest of an itemset, a measure that is independent of the search order by which itemsets are explored.

6.6 Log-Linear Modeling

Log-linear modeling is a modern statistical approach to identifying interdependencies in multivariate categorical data [Agresti 2002]. Hierarchical log-linear modeling finds the highest-order multivariate interactions that cannot be accounted for solely in terms of lower-order interdependencies. This is often exactly what is sought by a data analyst interested in positive interactions between items. However, log-linear analysis has a number of limitations that render its application infeasible in many data mining tasks. Most critically, no combination of attribute-values should have an expected frequency below 1.0, and so it is not applicable to sparse data. In general, it requires at least five records for each possible combination of attribute values. For example, for market basket data for 1000 products (quite a modest number in practice) $5 \times 2^{1000} \approx 10^{301}$ transactions would be required.

DuMouchel and Pregibon [2001] present an approach to using log-linear modeling that finds higher-order interactions that cannot be accounted for in terms of second-order interactions. Thus, a three-itemset {*beer*, *diapers*, *pretzels*} will only be considered if the three items occur more frequently than can be accounted for by the frequencies with of each of the pairs {*beer*, *diapers*}, {*beer*, *pretzels*} and {*diapers*, *pretzels*}.

A limitation of this approach is that if a three-itemset is found, as only two-item interactions are considered when evaluating a higher-order itemset, many supersets of the three-itemset are also likely to be found. This is because those supersets will also occur more frequently than can be accounted for by the two-item interactions even though this may be due only to the three-item interaction.

Wu et al. [2003] seek to overcome this limitation by applying full hierarchical log-linear analysis. Because it is not feasible to apply log-linear analysis to large numbers of variables, they first partition the items into *components*. First all pairs of items are tested for interdependence. A dependence network is then formed in which the items are nodes and there is an edge between each pair of itemsets that are interdependent. A component is a maximal set of items such that there is a path traced through the dependence network from every item in the component to every other item in the component. Log-linear analysis is then applied to each component in an attempt to keep the number of factors considered in each log-linear analysis tractable.

Our assessment suggests that this approach may have limitations on many real-world applications. Specifically, our experiments in Section 7.5 show that many real-world applications lead to large components that will not be susceptible to log-linear analysis.

Table I. Datasets

Dataset	Records	Items	Description
BMS-WebView-1	59,602	497	E-commerce clickstream data
Covtype	581,012	125	Geographic forest vegetation data
IPUMS LA 99	88,443	1,874	Census data
KDDCup98	52,256	19,662	Mailing list profitability data
Letter Recognition	20,000	74	Image recognition data
Mush	8,124	127	Biological data
Retail	88,162	16,470	Retail market-basket data
Shuttle	58,000	34	Space shuttle mission data
Splice Junction	3,177	243	Gene sequence data
TICDATA 2000	5,822	689	Insurance policy holder data

7. EXPERIMENTS

It is valuable to assess the impact of applying the proposed filters on discovered patterns. This serves two purposes. On the one hand it can reveal how greatly the filtering can reduce the number of patterns that the analyst must consider. On the other hand, it can indicate how many potentially misleading patterns might be discovered if the proposed statistical filtering is not employed. We also seek to assess the computational feasibility of the process.

The approach we have developed is a postprocessing technique that can be applied to any itemset discovery algorithm or preference function. We apply it here in the context of the dominant approach to itemset discovery: frequent itemset discovery.

7.1 Datasets

We employed the same ten datasets used in previous related research [Webb 2007]. These include eight of the largest attribute-value datasets from the UCI machine learning [Newman et al. 2006] and KDD [Hettich and Bay 2006] repositories together with the BMS-WebView-1 [Zheng et al. 2001] and Retail [Brijs et al. 1999] datasets. These datasets are described in Table I.

Each dataset was randomly divided into exploratory and holdout sets as close as possible to equal size. Numeric attributes were discretized into 3 bins, each containing as close as possible to one third of the exploratory data.

7.2 Assessment of Filters

All experiments were performed using the Magnum Opus association discovery system [Webb 2009] on an AMD64 939.3000 (1.8 GHz) Linux system. Magnum Opus uses the OPUS search algorithm [Webb 1995] to find top-k (also known as k-optimal) itemsets.

Magnum Opus was first applied to each exploratory set to discover the minimum values for minimum-support, that produced no more than 10,000 itemsets. These values, expressed as counts, are shown in the the column headed “MinS” in Table II.

Magnum Opus was then used to find all patterns with each specified minimum support from the exploratory data (“Total” in Table II). The column headed “0&1” shows the number of these itemsets that contained 0 or 1 items only. This

Table II. Productive, Self-Sufficient and Nonderivable Itemsets for Ten Large Datasets

Dataset	MinS	Total	0&1	NR	Prod	PNR	SS	Clsd	ND
BMS-WebView-1	27	9343	350	8981	8955	8594	4599	8555	8925
Covtype	288052	9993	22	9993	22	22	22	9993	191
IPUMS LA 99	26880	10000	33	2202	532	291	133	1278	556
KDDCup98	26880	9908	20	1052	56	47	35	384	267
Letter Recognition	571	9956	49	9796	2540	2427	934	9795	9769
Mush	914	9844	40	1344	2866	530	316	894	713
Retail	38	9650	2528	9628	3436	3434	3203	9591	7130
Shuttle	673	9984	30	5781	3092	1566	609	5085	5225
Splice Junction	112	9899	244	9403	729	557	417	9373	9394
TICDATA 2000	2861	9946	25	2065	55	53	49	196	233

is of interest, as such itemsets are both nonredundant and productive by definition, but may not be very interesting, as many users will have ready access to information about the frequency of individual items.

The meanings of the remaining columns are:

—*NR*. the number of nonredundant itemsets;

—*Prod*. the number of itemsets that are productive when tested against the hold-out data;

—*PNR*. the number of nonredundant itemsets that are productive when tested against the holdout data;

—*SS*. the number of nonredundant itemsets that are self-sufficient when tested against the holdout data;

—*Clsd*. the number of itemsets that are closed;

—*ND*. the number of itemsets that are nonderivable, determined by running the NDI software Goethals [2007] over the exploratory data.

The proportion of itemsets that were nonredundant varied from all for the Covtype dataset to under 5% for Mush at the lower minimum-support level.

For all but one dataset (BMS-WebView 1), less than half of the discovered patterns were productive. For half the datasets, less than 10% of discovered patterns were productive. For 30% of datasets, less than 1% of discovered patterns were productive. Note that for the Covtype dataset, the 22 productive patterns were the emptyset and 21 single-item itemsets. As already stated, such itemsets must be productive by definition, but will often be of little interest to the user. For every dataset other than Mush, the proportion of itemsets that were productive decreased as minimum-support decreased, suggesting that higher support itemsets are more likely to be productive.

The proportion of productive itemsets that were self-sufficient also varied greatly. For the Mush data, approximately 11% of productive itemsets were self-sufficient, while for Covtype all were. However, all the productive Covtype itemsets were found to be self-sufficient, as none of their supersets were productive, and hence there were no reference supersets against which they might be found to not be self-sufficient.

In many cases the number of itemsets that were both nonredundant and productive was only very slightly smaller than the number that were productive, but for Mush less than 20% of the productive itemsets were also nonredundant.

Table III. Compute Times in CPU Seconds

Dataset	No holdout	Productive	Self-sufficient
BMS-WebView-1	4.9	3.9	10.9
Covtype	28.6	28.6	28.9
IPUMS LA 99	45.5	44.9	44.8
KDDCup98	47.4	47.7	48.6
Letter Recognition	1.3	1.3	1.6
Mush	3.4	3.4	3.6
Retail	14.0	14.0	14.4
Shuttle	6.5	6.6	7.0
Splice Junction	0.3	0.3	0.3
TICDATA 2000	0.8	0.9	0.8

Relative to the full set of itemsets, in all cases fewer than 50% were self-sufficient. For all datasets there were substantially fewer self-sufficient than open or nonderivable itemsets. These results suggest that the proposed techniques are an effective way to reduce the number of itemsets that the analyst must consider.

7.3 Assessment of Computational Overheads

The running times of Magnum Opus without holdout evaluation and with each type of holdout evaluation are presented in Table III. Note that such results are subject to substantial variance and should be treated as indicative only. As can be seen, in only one case is there a substantial computational overhead for holdout evaluation, on the BMS-Webview-1 data, for which it increases compute time by 6 CPU seconds. This is probably a result of the large number of non-redundant and productive rules for this dataset, as all such rules need to be compared to all others to determine whether one is a superset of the other, as part of the test for self-sufficiency.

To assess the scalability of the approach, we took the data set with the worst computational profile, BMS-Webview-1, and ran it at successive minimum support levels of 27, 22, 20, 19, and 18, to produce collections of varying numbers of itemsets for postprocessing. Figure 4 plots the number of itemsets processed (in thousands) against compute time (in minutes). From this and Table III it seems apparent that the computational overheads of the approach become an issue when the number of itemsets to be processed is in the hundreds of thousands, with the processing time for 214,638 itemsets approaching 2 hours, albeit on a slow processor.

7.4 Assessment of the Itemset Leverage Measure

The earlier set of experiments were repeated, but this time the itemsets were selected that maximized itemset leverage rather than support. For ease of comparison, the same number of itemsets were sought for each dataset as were discovered for that dataset in the previous experiment. These itemsets were then assessed for productivity and self-sufficiency relative to the holdout data. The results are presented in Table IV. As the emptyset and single-item itemsets all have itemset leverage of 0.0, none of the results contained any of these.

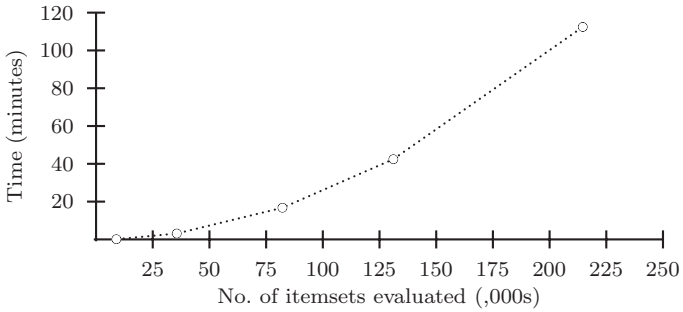

Fig. 4. Plot of evaluation time against number of itemsets evaluated.

Table IV. Nonredundant and Self-Sufficient Itemsets when Maximizing Itemset Leverage

Dataset	Total	Productive & non-redundant	Self-sufficient
BMS-WebView-1	9343	9234	3816
covtype	9993	9993	918
ipums.la.99	10000	10000	106
kddcup98	9908	9908	3860
letter-recognition	9956	9720	1137
mush	9844	9844	237
retail	9650	3248	2666
shuttle	9984	5163	319
splice	9899	1203	212
ticdata2000	9946	9629	1840

For all datasets other than retail, more high-leverage itemsets are productive and nonredundant than is the case for high-support itemsets. Retail has more productive and nonredundant high-support itemsets because of the large number of high-support single-item itemsets, all of which are productive and nonredundant according to (6) and (8).

For several datasets, searching for high-support itemsets results in more self-sufficient itemsets, despite there often being fewer productive and nonredundant itemsets, for example, BMS-WebView-1, retail, and splice. On the other hand, there are also several datasets for which searching for high-leverage itemsets results in more self-sufficient itemsets, for example, covtype, kddcup98, and ticdata2000. We hypothesize that this is due to search-by-support finding smaller itemsets. As a result, for each itemset there will be fewer productive and nonredundant supersets found against which an itemset might be deemed not self-sufficient. Whether there is a qualitative difference between the self-sufficient itemsets found when searching for high leverage or high support itemsets requires domain-specific judgement on an application by application basis, and hence falls outside the scope of this article, although the quantitative evidence that more productive and nonredundant itemsets are found when searching by itemset leverage might suggest that these itemsets are likely to be of greater value.

Table V. Size of Largest Component

Dataset	Size
BMS-WebView-1	497
covtype	125
ipums.la.99	1,584
kddcup98	19,586
letter-recognition	74
mush	117
retail	16,377
shuttle	34
splice	243
ticdata2000	576

7.5 Assessment of Feasibility of Log-Linear Analysis

To assess the feasibility of Wu et al.'s [2003] approach to discovering itemsets through log-linear analysis, described in Section 6.6, we identify the largest component for each of our datasets. There are shown in Table V. Of all the datasets, the smallest of these largest components is 34. As log-linear analysis requires an average frequency of at least 5 per cell and there are 2^{34} cells, at least 10^{10} examples would be required for log-linear analysis in this case. Such data quantities are often not available, restricting the applicability of this approach.

8. CONCLUSIONS

Itemsets are self-sufficient if their frequency is greater than can be accounted for by either the frequency of their subsets or of their supersets alone. We argue that itemsets that are not self-sufficient will often be of little interest to the data analyst, as their frequency should be expected once that of the itemsets on which their frequency depends is known. We have presented statistical tests for statistically sound discovery of self-sufficient itemsets, and computational techniques for application of those tests as a postprocessing step that may be applied with any itemset discovery algorithm. We also present a new measure, itemset leverage, for quantifying the productivity of an itemset.

Experiments demonstrate that these statistical techniques are computationally efficient for collections of itemsets numbering in the thousands, but that the computational burden increases considerably when hundreds of thousands of itemsets are to be processed.

The experiments also demonstrate that, in addition to focusing attention on itemsets that are most likely to be interesting, they can also substantially reduce the numbers of itemsets that an analyst need consider. It is important to note, however, that unlike closed and nonderivable itemset approaches, the set of self-sufficient itemsets will not provide sufficient information to derive the support of all frequent itemsets. They are intended to screen out classes of itemsets that are unlikely to be of interest to the user, rather than to provide a succinct summary of the support for all frequent itemsets.

We hope that the development of these techniques will assist in the application of itemset discovery in the place of rule discovery in contexts where

the division of items into an antecedent and consequent is not relevant to the discovery task.

REFERENCES

AGRAWAL, R., IMIELINSKI, T., AND SWAMI, A. 1993. Mining associations between sets of items in massive databases. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*. 207–216.

AGRESTI, A. 1992. A survey of exact inference for contingency tables. *Statis. Sci. 7*, 1, 131–153.

AGRESTI, A. 2002. *Categorical Data Analysis*. Wiley-Interscience, New York.

AUMANN, Y. AND LINDELL, Y. 1999. A statistical theory for quantitative association rules. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*. 261–270.

BASTIDE, Y., PASQUIER, N., TAOUIL, R., STUMME, G., AND LAKHAL, L. 2000. Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the 1st International Conference on Computational Logic (CL'00)*. Springer-Verlag, Berlin, 972–986.

BAYARDO, JR., R. J. AND AGRAWAL, R. 1999. Mining the most interesting rules. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'99)*. 145–154.

BAYARDO, JR., R. J., AGRAWAL, R., AND GUNOPULOS, D. 2000. Constraint-based rule mining in large, dense databases. *Data Min. Knowl. Disc. 4*, 2/3, 217–240.

BRIJS, T., SWINNEN, G., VANHOOF, K., AND WETS, G. 1999. Using association rules for product assortment decisions: A case study. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM Press, New York, NY, 254–260.

CALDERS, T. AND GOETHALS, B. 2002. Mining all non-derivable frequent itemsets. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02)*. Springer, Berlin, 74–85.

CHAN, R., YANG, Q., AND SHEN, Y. D. 2003. Mining high utility itemsets. In *Proceedings of the 3rd IEEE International Conference on Data Mining*. 19–26.

CHENG, J., KE, Y., AND NG, W. 2006. δ-tolerance closed frequent itemsets. In *Proceedings of the 6th International Conference on Data Mining*. 139–148.

COOLEY, R., TAN, P.-N., AND SRIVASTAVA, J. 1999. Discovery of interesting usage patterns from Web data. In *Proceedings of the International WEBKDD'99 Workshop*. Springer, Berlin, 163–182.

DUMOUCHEL, W. 1999. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *Americ. Statis. 53*, 3, 177–190.

DUMOUCHEL, W. AND PREGIBON, D. 2001. Empirical Bayes screening for multi-item associations. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. ACM Press, New York, NY, 76–76.

GOETHALS, B. 2007. NDI. Software. http://www.adrem.ua.ac.be/ goethals/software/.

HETTICH, S. AND BAY, S. D. 2006. The UCI KDD archive. Department of Information and Computer Science. University of California, Irvine, CA. http://kdd.ics.uci.edu.

HOLM, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian J. Statis. 6*, 65–70.

JAROSZEWICZ, S. AND SIMOVICI, D. A. 2004. Interestingness of frequent itemsets using Bayesian networks as background knowledge. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*. R. Kohavi, J. Gehrke, and J. Ghosh, Eds. ACM Press, New York, NY, 178–186.

LIU, B., HSU, W., AND MA, Y. 2001. Identifying non-actionable association rules. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. 329–334.

MALIK, H. H. AND KENDER, J. R. 2006. High quality, efficient hierarchical document clustering using closed interesting itemsets. In *Proceedings of the 6th IEEE International Conference on Data Mining*. IEEE, 991–996.

MEGIDDO, N. AND SRIKANT, R. 1998. Discovering predictive association rules. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*. AAAI Press, Menlo Park, US, 27–78.

NEWMAN, D. J., HETTICH, S., BLAKE, C., AND MERZ, C. J. 2006. UCI repository of machine learning databases. [Machine-readable data repository]. Department of Information and Computer Science, University of California, Irvine, CA.

PEI, J., DONG, G., ZOU, W., AND HAN, J. 2002. On computing condensed frequent pattern bases. In *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM'02)*. 378–385.

PIATETSKY-SHAPIRO, G. 1991. Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and J. Frawley, Eds. AAAI/MIT Press, Menlo Park, CA., 229–248.

SHAFFER, J. P. 1995. Multiple hypothesis testing. *Annual Rev. Psych. 46*, 561–584.

TABACHNICK, B. G. AND FIDELL, L. S. 2001. *Using Multivariate Statistics*. Allyn and Bacon, Boston, MA.

WEBB, G. I. 1995. OPUS: An efficient admissible algorithm for unordered search. *J. Arti. Intell. Res. 3*, 431–465.

WEBB, G. I. 2001. Discovering associations with numeric variables. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*. The ACM, New York, NY, 383–388.

WEBB, G. I. 2007. Discovering significant patterns. *Mach. Learn. 68*, 1, 1–33.

WEBB, G. I. 2009. Magnum Opus Version 4.3. Software, G. I. Webb & Associates, Melbourne, Aust.

WEBB, G. I. AND ZHANG, S. 2005. K-optimal rule discovery. *Data Mining Knowl. Discov. 10*, 1, 39–79.

WU, X., BARBARÁ, D., AND YE, Y. 2003. Screening and interpreting multi-item associations based on log-linear modeling. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 276–285.

XIN, D., HAN, J., YAN, X., AND CHENG, H. 2005. Mining compressed frequent-pattern sets. In *Proceedings of the International Conference on Very Large Databases (VLDB'05)*. 709–720.

YAO, H. AND HAMILTON, H. J. 2006. Mining itemset utilities from transaction databases. *Data Knowl. Engin. 59*, 603–626.

ZAKI, M. J. 2000. Generating non-redundant association rules. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00)*. ACM, New York, NY, 34–43.

ZAKI, M. J. 2004. Mining non-redundant association rules. *Data Mining Knowl. Discov. 9*, 3, 223–248.

ZAKI, M. J. AND HSIAO, C. J. 2002. CHARM: An efficient algorithm for closed itemset mining. In *Proceedings of the 2nd SIAM International Conference on Data Mining*. 457–473.

ZHANG, H., PADMANABHAN, B., AND TUZHILIN, A. 2004. On the discovery of significant statistical quantitative rules. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD'04)*. ACM Press, New York, NY, 374–383.

ZHENG, Z., KOHAVI, R., AND MASON, L. 2001. Real world performance of association rule algorithms. In *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining (KDD'01)*. ACM, New York, NY, 401–406.

Received March 2008; revised March 2009; accepted May 2009