, , 1–30 ()

Beyond Market Baskets: Generalizing Association Rules to Dependence Rules

CRAIG SILVERSTEIN * csilvers@cs.stanford.edu
SERGEY BRIN ** brin@cs.stanford.edu
RAJEEV MOTWANI † motwani@cs.stanford.edu
Department of Computer Science, Stanford University, Stanford, CA 94305

Editor: Usama Fayyad

Abstract. One of the more well-studied problems in data mining is the search for association rules in market basket data. Association rules are intended to identify patterns of the type: "A customer purchasing item A often also purchases item B." Motivated partly by the goal of generalizing beyond market basket data and partly by the goal of ironing out some problems in the definition of association rules, we develop the notion of *dependence rules* that identify statistical dependence in both the presence and absence of items in itemsets. We propose measuring significance of dependence via the chi-squared test for independence from classical statistics. This leads to a measure that is upward-closed in the itemset lattice, enabling us to reduce the mining problem to the search for a border between dependent and independent itemsets in the lattice. We develop pruning strategies based on the closure property and thereby devise an efficient algorithm for discovering dependence rules. We demonstrate our algorithm's effectiveness by testing it on census data, text data (wherein we seek term dependence), and synthetic data.

Keywords: data mining, market basket, association rules, dependence rules, closure properties, text mining

1. Introduction

One particularly well-studied problem in data mining is the search for association rules in market basket data (Agrawal et al., 1993a, Agrawal et al., 1993b, Klemettinen et al., 1994, Mannila et al., 1994, Agrawal and Srikant, 1994, Han and Fu, 1995, Houtsma and Swami, 1995, Park et al., 1995, Srikant and Agrawal, 1995, Savasere et al., 1995, Agrawal et al., 1996, Toivonen, 1996). In this setting, the base information consists of register transactions of retail stores. The goal is to discover buying patterns such as two or more items that are bought together often.[1] The market basket problem has received a great deal of attention in the recent past, partly due to its apparent utility and partly due to the research challenges it presents. The past research has emphasized techniques for improving the performance of algorithms for discovering association rules in large databases of sales information. There has also been some work on extending this paradigm to numeric and geometric data (Fukuda et al., 1996a, Fukuda et al., 1996b).

* Supported by the Department of Defense and an ARCS fellowship, with partial support from NSF Award CCR–9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.
** Supported by an NSF Fellowship.
† Supported by an Alfred P. Sloan Research Fellowship, an IBM Faculty Partnership Award, an ARO MURI Grant DAAH04–96–1–0007, and NSF Young Investigator Award CCR–9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

While Piatetsky-Shapiro and Frawley (Piatetsky and Frawley, 1991) define an *association problem* as the general problem of finding recurring patterns in data, much of the recent work on mining of large-scale databases has concerned the important special case of finding association rules. Association rules, explained below, are primarily intended to identify rules of the type, "A customer purchasing item A is likely to also purchase item B." In general, the development of ideas has been closely linked to the notion of associations expressed via the customer preference example.

Our work is motivated partly by the goal of generalizing beyond market basket data, and partly by the goal of ironing out some problems in the definition of association rules. We develop techniques to mine generalized baskets, which are defined to be a collection of subsets from an item space, such as a corpus of text documents (where the items are words) or census data (where the items are boolean or numeric answers to questions). In this more general setting, association rules are but one of the many types of recurring patterns that could or should be identified by data mining. Consequently, we develop the notion of mining rules that identify dependencies (generalizing associations) taking into consideration both the presence and the *absence* of items as a basis for generating rules.

To give a concrete example, Mosteller and Wallace (Mosteller and Wallace, 1964) studied dependencies in text data to determine the authorship of each essay in the *Federalist Papers*. This collection of essays was written in the late 1700s by John Jay, Alexander Hamilton, and James Madison, but the essays were all signed "Publius." Mosteller and Wallace studied the writing style of each essay to determine authorship. One factor they looked at was word co-occurrence, which is best measured by correlation. In fact, the tests they ran in 1964 are similar to the dependence tests we run on text data in Section 6.

The remainder of this section is organized as follows. In order to place our work in the context of earlier work, in Section 1.1 we review some of the details of the past work on association rules in the market basket application. Then, in Section 1.2, we point out some problems in the current definition of association rules and demonstrate that dependence rules often are better at capturing the patterns that are being sought after in the definition of association rules. Finally, we give an overview of the rest of this paper in Section 1.3.

1.1. Association Rules

We briefly review some of the details of the past work on association rules in market basket data. For this purpose, we define **basket data** in general terms.

Definition 1. Let $I = \{i_1, \ldots, i_k\}$ be a set of k elements, called **items**. Then, **basket data** $B = \{b_1, \ldots, b_n\}$ is any collection of n subsets of I, and each subset $b_i \subseteq I$ is called a **basket** of items.

For example, in the *market basket* application, the set I consists of the items stocked by a retail outlet and each basket is the set of purchases from one register transaction; on the other hand, in the *document basket* application, the set I contains all dictionary words and proper nouns, while each basket is a single document in the corpus (for now we ignore the frequency and ordering of the words in a document).

While it is clear that the simple notion of basket data is powerful and captures a wide variety of settings amenable to data mining, it should be kept in mind that there could be structure in the data (e.g., word ordering within documents) that is lost in this general framework.

An association rule (Agrawal et al., 1993a) in the database B is defined as follows.

Definition 2. We say there is an **association rule** $i_1 \Rightarrow i_2$ if

1. i_1 and i_2 occur together in at least $s\%$ of the n baskets (the **support**);
2. and, of all the baskets containing i_1, at least $c\%$ also contain i_2 (the **confidence**).

This definition extends easily to $I \Rightarrow J$, where I and J are disjoint sets of items instead of single items. Since it is possible to have alternate definitions of association rules, we will henceforth refer to the above definition as the *support-confidence framework* for association rules. It should be noted that the symbol $\Rightarrow$ is a bit misleading since such a rule does not correspond to real implications; clearly, the confidence measure is merely an estimate of the *conditional probability* of i_2 given i_1.

Consider applying the above definition to market basket data from a grocery store. Association rules are then statements of the form: "When people buy tea, they also often buy coffee." In practice, such rules may be used to make a marketing campaign effective or to justify changing product placement in the store. The confidence statistic ensures that "often" is a large enough percentage of the people who buy tea to be potentially interesting. The support statistic, on the other hand, justifies financing the marketing campaign or product placement — these products generate enough sales to be worthy of attention. Support is also used to help ensure statistical significance, because if items are rare, the variance of the confidence statistic may be too large to draw any useful conclusions.

1.2. A Critique of Association Rules

Association rules, and the support-confidence framework used to mine them, are well-suited to the market basket problem. Other basket data problems, while seemingly similar, have requirements that the support-confidence framework does not address. For instance, testing the association BATTERIES $\Rightarrow$ CAT FOOD could not discover a fact such as, "When people buy batteries, they do not usually also buy cat food"; finding such negative implications requires a separate test. While perhaps not as useful to the marketing staff of supermarkets, such implications can be helpful in many other settings. For example, fire code inspectors trying to mine useful fire prevention measures might like to know of any negative dependence between certain types of electrical wiring and the occurrence of fires.

A more serious problem with the support-confidence framework is illustrated in the following example.

Example 1 Suppose we have market basket data from a grocery store, consisting of n baskets. Let us focus on the purchase of tea and coffee. In the following table, x represents the presence of an item, $\overline{x}$ its absence, and the numbers represent percentages of baskets.

	TEA	$\overline{\text{TEA}}$	row-sum
COFFEE	20	70	90
$\overline{\text{COFFEE}}$	5	5	10
col-sum	25	75	100

Let us apply the support-confidence framework to the potential association rule TEA $\Rightarrow$ COFFEE. The support for this rule is 20%, which is quite high. Let $p(x)$ be the probability that the items x appear in a random basket. Consider the customer purchasing a basket chosen uniformly at random from the n baskets. Then, confidence is effectively the conditional probability that the customer buys coffee, given that she buys tea, i.e., $p(\text{TEA} \wedge \text{COFFEE})/p(\text{TEA}) = 20/25 = 0.8$ or 80%, which too is pretty high. At this point, we may conclude that the rule TEA $\Rightarrow$ COFFEE is interesting and useful.

However, consider now the fact that the a priori probability that the customer buys coffee is 90%. In other words, a customer who is known to buy tea is less likely to buy coffee (by 10%) than a customer about whom we have no information. Of course, it may still be interesting to know that such a large number of people who buy tea also buy coffee, but stating that rule by itself is at best incomplete information and at worst misleading. The truth here is that there is a *negative* dependence between buying tea and buying coffee; at least that information should be provided along with the association rule TEA $\Rightarrow$ COFFEE.

One way of measuring the dependence between TEA and COFFEE is to compute

$$p(\text{TEA} \wedge \text{COFFEE})/(p(\text{TEA}) \times p(\text{COFFEE})) = 0.2/(0.25 \times 0.9) = 0.89.$$

The fact that this quantity is less than 1 indicates negative dependence, since the numerator is the actual likelihood of seeing a customer purchase both tea and coffee, and the denominator is what the likelihood would have been in the case where the two purchases completely independent. On the other hand,

$$p(\text{TEA} \wedge \overline{\text{COFFEE}})/(p(\text{TEA}) \times p(\overline{\text{COFFEE}})) = 0.05/(0.25 \times 0.10) = 2.00,$$

indicating a strong positive dependence between the absence of coffee and the presence of tea. (Note that, by contrast, the confidence in the association rule $\overline{\text{COFFEE}} \Rightarrow$ TEA is only $0.05/0.10 = 50\%$, much lower than the confidence in the TEA $\Rightarrow$ COFFEE rule.)

If further analysis found that the dependence between COFFEE and TEA were statistically significant, we could claim the dependence rule, "The purchase of COFFEE and TEA are dependent." Furthermore, we could say, "The major dependence is a positive dependence between the absence of COFFEE and the occurrence of TEA." As a result, the store manager may decide to target non-coffee drinkers in his tea displays. □

If the coffee and tea example seems a bit contrived, consider Russian politicians. Suppose we wish to explore Russian photographs in order to understand power politics in the Kremlin. We can posit that if two people often appear together in photographs, they are allied. Then the items are Russian political figures, and each basket consists of a list of figures in one photograph. It is reasonable to suppose that the Prime Minister appears in 90% of all photographs and the Defense Minister in 25%. These percentages could well break down exactly as in the coffee and tea example, and the same kinds of potentially misleading association rules would result.

In the coffee and tea example, we deduced a dependence, but it is not clear whether our deduction was statistically significant. Testing for significant dependence is a problem that statisticians have been studying for over a century; refer to Lancaster (Lancaster, 1969) for the theory and a history of this problem. A standard test of independence involves the chi-squared statistic, which is both easy to calculate and reliable under a fairly permissive set of assumptions. This test is useful because it not only detects dependence but can distinguish positive dependence (as in the tea and coffee example) and negative dependence (as in the fire code example).

1.3. Overview of Paper

In the following sections, we show how dependence can be used as a basis for mining general basket data. We use the chi-squared measure in place of the support-confidence framework to generate what we call *dependence rules* which overcome the problems with association rules discussed above. Furthermore, we demonstrate how dependence rules can be computed efficiently.

We begin in Section 2 with some preliminary definitions and notation. Our definition of dependence rules is presented in Section 3. We show that the set of dependence rules is upward-closed in the lattice of subsets of the item space, enabling us to reduce the mining problem to the search for a border between dependent and independent itemsets in the lattice. In Section 4 we propose that the chi-squared test from classical statistics be used to measure the significance of dependence rules. We also define a measure of interest for a dependence rule. Our framework is contrasted with the support-confidence framework for association rules and we argue that there are several advantages to using our framework. We also comment on some limitations of our approach.

Based on the upward-closure property of dependence rules, and some pruning strategies we develop, in Section 5 we present efficient algorithms for the discovery of dependence rules. In Section 6 we demonstrate the effectiveness of our algorithms by experiments on census data and finding term dependency in a corpus of text documents. Finally, in Section 7 we make concluding remarks. Appendix A gives some of the theoretical basis for the chi-squared test in statistics.

2. Preliminaries

In Section 1.1 we defined basket data in terms of a collection of baskets, where each basket was a set of items. It will be convenient to also have an alternate view of basket data in terms of the boolean indicator variables for items, as follows.

Definition 3. Let $I_1, \ldots, I_k$ be a set of k boolean variables called **attributes**. Then, a set of baskets $B = \{b_1, \ldots, b_n\}$ is a collection of n k-tuples from $\{\text{TRUE}, \text{FALSE}\}^k$ which represent a collection of value assignments to the k attributes.

Assigning a true value (TRUE) to an attribute variable I_j represents the presence of item i_j. A k-tuple from $\{\text{TRUE}, \text{FALSE}\}^k$ denotes the set of items present in a basket in the obvious way.

We adopt the following notational convention with regard to an attribute variable A representing some item a. The event a denotes $A = \text{TRUE}$, or, equivalently, the presence of the corresponding item a in a basket. The complementary event $\overline{a}$ denotes $A = \text{FALSE}$, or the absence of the item a from a basket. There is an overloading of notation in that lower-case letters are used to represent both items and the event that the item is present in a basket, but the meaning will always be clear from the context; on the other hand, upper-case letters will always represent variables corresponding to items. Finally, we will use x and y to refer to events that could be either positive or negative, unlike a and b which refer to purely positive events.

Definition 4. We define $p(a) = P[A = \text{TRUE}]$ to be the probability that item a appears in a random basket. Likewise, $p(\overline{a}) = P[A = \text{FALSE}] = 1 - p(a)$. Joint probability is defined in a similar way. For instance, $p(a\overline{b}) = P[A = \text{TRUE}, B = \text{FALSE}]$ is the probability that item a is present while item b is absent.

Note that the probability space underlying these definitions is some hypothetical space from which the baskets are assumed to have been drawn and whose structure is desired to be captured via either association rules or dependence rules.

We define independence and dependence of events and variables.

Definition 5.

1. Two events x and y are **independent** if $P[x \wedge y] = P[x]P[y]$.

2. Two variables A and B are **independent** if $P[A = v_a \wedge B = v_b] = P[A = v_a]P[B = v_b]$ for all possible values $v_a, v_b \in \{\text{TRUE}, \text{FALSE}\}$.

3. Events, or variables, that are not independent are **dependent**.

The definition of independence extends in the obvious way to the independence of three or more events or variables. Note that the test for k-way independence of variables involves 2^k combinations of event independence. Finally, observe that the independence of two variables A and B implies the independence of the events a and b (as well as $\overline{a}$ and b, etc.), but the converse is not true in general.

We now briefly discuss how we estimate event probabilities for the market basket problem. If we have n baskets, let $O_n(a)$ be the number of baskets that include item a. Likewise, $O_n(\overline{a})$ is the number of baskets not including a. We estimate $p(a)$ by $O_n(a)/n$. This is the maximum likelihood estimate of $p(a)$, which is a standard estimate used in the statistics and data mining community. An alternate approach, based on Bayesian techniques, would be to start with a prior value of $p(a)$ and modify it based on $O_n(a)$ and n.

3. Dependence Rules

The definition of dependence is all that is needed to define dependence rules.

Definition 6. Let I be a set of attribute variables. We say that the set I is a **dependence rule** if I is dependent.

This definition is simple but powerful. Part of its power comes from the fact that I also includes $2^{|I|}$ possible patterns of event dependence.

In Section 4 we talk about techniques both for determining if I is dependent, and for measuring the power of its dependence. We also discuss how to measure the relative power of the various patterns of event dependence aggregated into a dependence rule. But first, in the rest of this section, we identify some crucial properties of dependence that hold regardless of how dependence is determined or measured.

3.1. The Closure Property

An important feature of our definition is that the property of being dependent is upward-closed in the lattice of itemsets, where the notion of upward-closure is defined as follows.

Definition 7. Consider the lattice $\mathcal{L}$ of all possible itemsets from the universe of items I. A property $\mathcal{P}$ is said to be **upward-closed** with respect to the lattice $\mathcal{L}$ if for every set with property $\mathcal{P}$, all its supersets also have property $\mathcal{P}$. Similarly, property $\mathcal{P}$ is said to be **downward-closed** if for every set with property $\mathcal{P}$, all its subsets also have property $\mathcal{P}$.

We now prove that our notion of dependence rules forms an upward-closed property.

THEOREM 1 *If a set of variables I is dependent, so is every superset of I.*

Proof: If a set of variables is dependent, then some set of events associated with the variables must also be dependent. Suppose, without loss of generality, that the variables are A and B and that the events a and b are dependent. Assume, then, that some superset ABC is independent. Then all events associated with ABC must be independent. In particular, we must have $p(abc) = p(a)p(b)p(c)$ and $p(ab\overline{c}) = p(a)p(b)p(\overline{c})$. Then $p(ab) = p(abc) + p(ab\overline{c}) = p(a)p(b)p(c) + p(a)p(b)p(\overline{c}) = p(a)p(b)$. Thus, a and b are independent after all, contrary to our hypothesis, so ABC must actually be dependent. ■

In Section 4 we will propose using the χ^2 test for independence to identify dependence rules. We show in Appendix A that the χ^2 statistic is also upward-closed. That is, if a set I of items is deemed dependent at significance level α, all supersets I are also dependent at significance level α. The proof is considerably helped by the fact that, if all variables are boolean, the degrees of freedom for the chi-squared test is 1 regardless of the number of variables.

Definition 8. If an itemset I is dependent but no subset of I is dependent, we say I is **minimally dependent**.

To understand the significance of closure, let us examine how mining for association rules is implemented. Using the support-confidence test, the problem is usually divided into two parts: First finding supported itemsets, and then discovering rules in those itemsets that have large confidence. Almost all research has focused on the first of these tasks. One reason is that finding support is usually the more expensive step, but another reason is that rule discovery does not lend itself as well to clever algorithms. This is because confidence possesses no closure property. Support, on the other hand, is downward-closed: If a set of items has support, then all its subsets also have support.

Researchers have taken advantage of the downward-closure of support in devising efficient algorithms for association rules. Level-wise algorithms (Agrawal et al., 1993a) operate level-by-level, bottom-up, in the itemset lattice. At the i^{th} level of the lattice, all itemsets are of size i and are called *i-itemsets*. The level-wise algorithms start with all i-itemsets satisfying a given property, and use this knowledge to explore $(i + 1)$-itemsets. Another class of algorithms, random walk algorithms (Gunopulos et al., 1997), generate a series of random walks, each of which explores the local structure of the border. A random walk is a walk up the itemset lattice. It starts with the empty itemset and adds items one at a time to form a larger itemset. It is also possible to walk down the itemset lattice by deleting items from an initial, full itemset. Both level-wise and random walk algorithms use the closure property to make inferences about the supersets of an itemset.

It is clear that the upward- and downward-closure are two faces of the same coin. In particular, if a property $\mathcal{P}$ is upward-closed, then not having the property is downward-closed. Thus, an upward-closed property could be turned into a downward-closed property by "turning the lattice upside-down." However, if there are two or more conditions that itemsets need to satisfy, some upward-closed and others downward-closed, then it might be necessary to simultaneously deal with both forms of closure. In this case, turning the lattice upside-down does not really change anything. We briefly discuss the slightly different ways in which we can exploit the two kinds of closure properties to speed up an algorithm.

Downward-closure is a *pruning* property. That is, it is capable of identifying objects that *cannot* have a property of interest. To use it, we start out with all $(i + 1)$-itemsets as candidates for being, say, supported. As we examine i-itemsets, we cross out some $(i+1)$-itemsets that we know cannot have support. We are, in effect, using the contrapositive of the support definition, saying, "If any subset of an $(i + 1)$-itemset does not have support, then neither can the $(i + 1)$-itemset." After crossing out some items, we go through the remaining list, checking each $(i + 1)$-itemset to make sure it actually does have the needed support.

Upward-closure, on the other hand, is *constructive*, in that it identifies objects that *must* have a property of interest. For instance, we may start with the belief that no $(i+1)$-itemset is, say, dependent. Looking at an i-itemset, we can say that if it is dependent, all its supersets are also dependent. This gives us a list of dependent $(i + 1)$-itemsets. Unlike in the pruning case, where we generate false positives ($(i + 1)$-itemsets that do not really have support), here we generate false negatives (ignored dependent $(i + 1)$-itemsets). Because of this, upward-closure is most useful if the property we are looking for is an *unwanted* one. Then, we are finding $(i+1)$-itemsets to prune, and all that happens if we miss some dependent itemsets is that our pruning is less effective. It is for this reason we concentrate on minimal dependent itemsets, that is, itemsets that are dependent though no subset of them is. The minimality property allows us to prune all the parents of a dependent i-itemset, since clearly no superset of a dependent set can be minimally dependent.

3.2. The Border of Dependence

An advantage of an upward-closed property is that closure means the itemsets of interest form a **border** in the itemset lattice. That is, we can list a collection of itemsets such that every itemset above (and including) the set in the item lattice possesses the property, while every itemset below it does not.

Because of closure, the border encodes all the useful information about the interesting itemsets. Therefore, we can take advantage of the border property to prune based on dependence data as the algorithm proceeds. This time- and space-saving shortcut does not work for confidence, which is not upward-closed. If we combine dependence with support, we can prune using both tests simultaneously. In support-confidence, on the other hand, confidence testing has to be a post-processing step.

To show that confidence does not form a border, we present an example where an itemset has sufficient confidence while a superset of it does not.

Example 2 Below we summarize some possible market basket data for coffee, tea, and doughnuts. The first table is for baskets including doughnuts, while the second is for baskets lacking doughnuts.

DOUGHNUTS	TEA	$\overline{\text{TEA}}$	row-sum
COFFEE	8	40	48
$\overline{\text{COFFEE}}$	1	2	3
col-sum	9	42	51

$\overline{\text{DOUGHNUTS}}$	TEA	$\overline{\text{TEA}}$	row-sum
COFFEE	10	35	45
$\overline{\text{COFFEE}}$	2	2	4
col-sum	12	37	49

Observe that $p(\text{COFFEE} \wedge \text{DOUGHNUTS}) = 0.48$, $p(\text{COFFEE}) = 0.93$, so the rule COFFEE $\Rightarrow$ DOUGHNUTS has confidence 0.52. On the other hand, $p(\text{TEA} \wedge \text{COFFEE} \wedge \text{DOUGHNUTS}) = 0.08, p(\text{TEA} \wedge \text{COFFEE}) = 0.18$, so the rule COFFEE, TEA $\Rightarrow$ DOUGHNUTS has confidence 0.44. For a reasonable confidence cutoff of 0.50, COFFEE $\Rightarrow$ DOUGHNUTS has confidence but its superset COFFEE, TEA $\Rightarrow$ DOUGHNUTS does not. □

The border property makes practical a wide range of association rule algorithms. Level-wise algorithms can stop early if the border is low (as is often the case in practice). Random walk algorithms hold promise, since a given walk can stop as soon as it crosses the border. The algorithm can then do a local analysis of the border near the crossing.

4. The Chi-squared Test for Independence

Let $R = \{i_1, \overline{i_1}\} \times \cdots \times \{i_k, \overline{i_k}\}$ be the Cartesian product of the event sets corresponding to the presence or absence of items in a basket. An element $r = r_1 \ldots r_k \in R$ is a single basket value, or an instantiation of all the variables. Each value of r denotes a *cell* — this

Table 1. I for a collection of census data. This set of items was formed by arbitrarily collapsing a number of census questions into binary form.

item	var name	a signifies...	$\overline{a}$ signifies...
i_0	SOLO-DRIVER	drives alone	does not drive, carpools
i_1	FEW-KIDS	male or less than 3 children	3 or more children
i_2	NOT-VETERAN	never served in the military	military veteran
i_3	ENGLISH	native speaker of English	not a native speaker
i_4	NOT-CITIZEN	not a U.S. citizen	U.S. citizen
i_5	BORN-US	born in the U.S.	born abroad
i_6	MARRIED	married	single, divorced, widowed
i_7	UNDER-40	no more than 40 years old	more than 40 years old
i_8	MALE	male	female
i_9	HOUSEHOLDER	householder	dependent, boarder, renter, etc.

terminology comes from viewing R as a k-dimensional table, called a **contingency table**. Let $O(r)$ denote the number of baskets falling into cell r. To test whether a given cell is dependent, we must determine if the actual count in cell r differs sufficiently from the expectation.

In the chi-squared test, the expected count of an event is calculated under the assumption of independence. For a single event, we use the maximum likelihood estimators $E(i_j) = O_n(i_j)$ and $E(\overline{i_j}) = n - O_n(i_j)$. For sets of events, we use the independence assumption to calculate $E(r) = n \times E(r_1)/n \times \cdots \times E(r_k)/n$. Then the chi-squared statistic is defined as follows:

$$\chi^2 = \sum_{r \in R} \frac{(O(r) - E(r))^2}{E(r)}.$$

In short, this is a normalized deviation from expectation. Refer to Appendix A for a discussion of the theoretical underpinnings of the chi-squared statistic which leads to the above formula.

The chi-squared statistic as defined will specify whether all k items are k-way independent. In order to determine whether some subset of items are dependent, for instance i_1, i_2, and i_7, we merely restrict the range of r to $\{i_1, \overline{i_1}\} \times \{i_2, \overline{i_2}\} \times \{i_7, \overline{i_7}\}$.

No matter how r is restricted, the chi-squared test works as follows: Calculate the value of the chi-squared statistic. Corresponding to this value and a degrees of freedom count (always 1, for boolean variables) is a p value.[2] This value, between 0 and 1, indicates the probability of witnessing the observed counts were the variables really independent. If this value is low (say, less than 0.05), we reject the hypothesis that the variables are independent. We say a set of items is dependent at **significance level** α if the p value of the set is at most $1 - \alpha$.

To put χ^2 values in perspective, for a p value of 0.05 with one degree of freedom, the χ^2 cutoff value is 3.84. Thus, any set of items with a χ^2 value of 3.84 or more is significant at the $1 - 0.05 = 95\%$ confidence level.

Table 2. B for a collection of census data. There are actually 30370 baskets, but we show only the first 9 entries here. Person 1, for instance, either does not drive or carpools, is male or has less than 3 children, is not a veteran, speaks English natively, and so on. Person 5 fits the same set of attributes, so $O(i_1, i_2, i_3, \overline{i_4}, i_5, \overline{i_6}, i_7, \overline{i_8}, i_9) = 2$.

basket	items	basket	items	basket	items
1	$i_1\ i_2\ i_3\ i_5\ i_7\ i_9$	4	$i_1\ i_2\ i_3\ i_5\ i_7\ i_8$	7	$i_1\ i_2\ i_3\ i_5\ i_7\ i_8$
2	$i_1\ i_2\ i_3\ i_7$	5	$i_1\ i_2\ i_3\ i_5\ i_7\ i_9$	8	$i_1\ i_2\ i_3\ i_5\ i_7\ i_8$
3	$i_1\ i_2\ i_3\ i_5\ i_7\ i_8\ i_9$	6	$i_1\ i_2\ i_3\ i_5\ i_7$	9	$i_1\ i_3\ i_5\ i_7\ i_8$

Example 3 Consider the census data introduced in Table 1. For this example we restrict our attention to the nine baskets in Table 2. The contingency table for MALE and HOUSEHOLDER is as follows:

	MALE	$\overline{\text{MALE}}$	row-sum
HOUSEHOLDER	1	2	3
$\overline{\text{HOUSEHOLDER}}$	4	2	6
col-sum	5	4	9

Now $E(\text{HOUSEHOLDER}) = O(\text{HOUSEHOLDER}) = 3$, while $E(\text{MALE}) = O(\text{MALE}) = 5$; note that $E(\text{HOUSEHOLDER})$ is the sum of row 1, while $E(\text{MALE})$ is the sum of column 1. The chi-squared value is

$$\frac{(1-3\times 5/9)^2}{3\times 5/9} + \frac{(2-3\times 4/9)^2}{3\times 4/9} + \frac{(4-6\times 5/9)^2}{6\times 5/9} + \frac{(2-6\times 4/9)^2}{6\times 4/9}$$

$$= 0.267 + 0.333 + 0.133 + 0.167 = 0.900$$

Since 0.900 is less than 3.84, we do not reject the independence hypothesis at the 95% confidence level. □

The next example, also based on census data detailed in Section 6, helps to indicate how dependence rules may be more useful than association rules in certain settings.

Example 4 Consider the census data presented in Table 1. We focus on testing the relationship between military service and age.[3] Using the full census data, with $n = 30370$, we obtain the following contingency table:

	UNDER-40	$\overline{\text{UNDER-40}}$	row-sum
NOT-VETERAN	17918	9111	27029
$\overline{\text{NOT-VETERAN}}$	911	2430	3341
col-sum	18829	11541	30370

We can use row and column sums to obtain expected values, and we get a chi-squared value of 2006.34, which is significant at the 95% significance level. Furthermore, the largest contribution to the χ^2 value comes from the bottom-right cell, indicating that the dominant dependence is being a veteran and being over 40. This matches our intuition.

For comparison, let us try the support-confidence framework on this data, with support at 1% (i.e., count 304) and confidence at 50%. All possible rules pass the support test, but only half pass the confidence test. These are

- $\overline{\text{NOT-VETERAN}} \Rightarrow \overline{\text{UNDER-40}}$,
- $\text{NOT-VETERAN} \Rightarrow \text{UNDER-40}$,
- $\overline{\text{UNDER-40}} \Rightarrow \text{NOT-VETERAN}$, and
- $\text{UNDER-40} \Rightarrow \text{NOT-VETERAN}$.

These statements correspond to the following claims: "Many people who have served in the military are over 40," "Many people who have never served in the military are 40 or younger," "Many people over 40 have never served in the military," and "Many people 40 or younger have never served in the military." Taken together, these statements do not carry much useful information. A traditional way to rank the statements is to favor the one with highest support. In this example, such a ranking leaves the first statement — the one which the chi-squared test identified as dominant — in last place. □

The following theorem shows that the chi-squared statistic is closed and can therefore be used for pruning and for locating the border. It is proved in Appendix A.

THEOREM 2 *In the binomial case, the chi-squared statistic is upward-closed.*

4.1. Measures of Interest

In the last example, as indeed in the first example on coffee and tea, we wanted to find the dependence of a given cell, in order to give a more precise characterization of the dependence.

Definition 9. We define the **interest** of two events x and y to be

$$I(xy) = \frac{p(xy)}{p(x)p(y)},$$

with the obvious extension to more than two events.

By considering k events, each associated with one of the k items, we obtain the interest of a single cell of a k-dimensional contingency table. We denote the interest of a cell r by $I(r)$. Note that dependence rules refer to variables, and therefore an entire contingency table, while interest applies to events and therefore a single cell of the contingency table.

In contingency table notation, $I(r) = O(r)/E(r)$ since $p(a)p(b) = E(ab)/n$ and $p(ab) = O(ab)/n$. We can show that the cell with the interest value farthest from 1 is, in some sense, the most dependent of any cell in the contingency table.

LEMMA 1 *For a given contingency table, let r be the cell with interest value $I(r)$ maximizing $|I(r) - 1|$. This cell contributes most to the χ^2 value of the contingency table.*

Proof: By definition, the deviation of the interest from 1 is $|O(r)/E(r) - 1|$. The cell that maximizes this quantity also maximizes $|O(r) - E(r)|/E(r)$, and thus maximizes $(O(r) - E(r))^2/E(r)$. This is exactly the contribution of cell r to χ^2. ■

Interest values above 1 indicate positive dependence, while those below 1 indicate negative dependence. While the absolute number is meaningless, most comparative measures are not. For instance, if the second-highest interest value is close to the first, then the corresponding cell has almost as much dependence, though it is dangerous to try to quantify the difference. Comparing interest values from one contingency table to interest values from another is meaningless.

Example 5 Consider the census data from Example 4. The corresponding interest values are

	UNDER-40	$\overline{\text{UNDER-40}}$
NOT-VETERAN	1.07	0.89
$\overline{\text{NOT-VETERAN}}$	0.44	1.91

The bottom-right cell has the most extreme interest, agreeing with the conclusion from Example 4 based on contribution to χ^2. The other cell values are meaningful as well; for instance, there is a large negative dependence (0.44) between being 40 or younger and being a veteran.

Looking back at the raw cell counts in Example 4, we see that the cells with high interest have low counts. Nevertheless, since the chi-squared value for this example is well above the 95% significance threshold, we have confidence that these interest values are statistically significant. □

4.2. *Comparison of Interest and Correlation*

While interest is simple to calculate and interpret, and is closely tied to the chi-squared test and contingency tables, it is not the normal statistic used to measure the power of dependence. Instead, the **correlation coefficient** is normally used. The correlation coefficient of a set of items is defined to be the covariance of the items, normalized by dividing with the product of the standard deviations of the items. This value is always between -1 and 1. Because of the normalization by the standard deviations, it is possible to meaningfully compare the correlation coefficients of different itemsets. Such comparisons using interest measures, as we have already noted, are meaningless.

The correlation coefficient, however, is not really appropriate for dependence rules. One major problem is that covariance is calculated as an aggregate over the range of values of the random variables. In fact, the value of the correlation coefficient is a weighted sum of the dependence between the events associated with itemset. Therefore, a positive correlation coefficient near 1 indicates that either a and b are highly dependent, or $\overline{a}$ and $\overline{b}$

are highly dependent, or both, but does not provide any more detailed understanding of the dependence.

Another serious problem is that since the correlation coefficient conflates the dependence judgments of many events, it is possible for the coefficient to be 0 even when variables are dependent (though such a "false zero" is not possible when all variables are boolean). In general, the correlation coefficient is useful for identifying *linear functional dependence* between random variables, but is poor at capturing other kinds of dependencies or handling the case of categorical variables.

Despite these problems with the correlation coefficient, it does have the advantage that it allows us to compare disparate itemsets. Thus, the correlation coefficient could be used in some cases to infer that the dependence rule $\{\text{COFFEE}, \text{TEA}\}$ has higher dependence than the rule $\{\text{DOUGHNUTS}, \text{TEA}\}$, although interest would be needed to identify the events that contribute the most to each rule.

4.3. Contrast with Support-Confidence Framework

Example 4 demonstrated how the chi-squared test could be more useful than support-confidence for a wide range of problems. We list some of the advantages of the χ^2-interest framework over the support-confidence framework.

1. The use of the chi-squared significance test is more solidly grounded in statistical theory. In particular, there is no need to choose ad-hoc values of support and confidence. While the significance level is an arbitrary value, it is not ad-hoc in that its value can be chosen in a meaningful way, with results that can be predicted and interpreted by statistical theory.

2. The chi-squared statistic simultaneously and uniformly takes into account all possible combinations of the presence and absence of the various attributes being examined as a group.

3. The interest measure is preferable as it directly captures dependence, as opposed to confidence which considers directional implication (and treats the absence and presence of attributes non-uniformly).

4. The experimental data suggests that using chi-squared tests combined with interest yields results that are more in accordance with our a priori knowledge of the structure in the data being analyzed.

4.4. Limitations of the Chi-squared Test

The chi-squared statistic is easy to calculate, which in the world of statistics is a sure tip-off that it is an approximation. In this case, the chi-squared test rests on the normal approximation to the binomial distribution (more precisely, to the hypergeometric distribution). This approximation breaks down when the expected values are small. As a rule of thumb, Moore (Moore, 1986) recommends the use of chi-squared test only if

- all cells in the contingency table have expected value greater than 1;

- and, at least 80% of the cells in the contingency table have expected value greater than 5.

For association rules, these conditions will frequently be broken. For a typical application, $|I|$ may be 700 while $n = 1000000$. Even a contingency table with as few as 20 of the 700 possible dimensions will have over a million cells, and, as the sum of the expected cell values is only 1 million, not all cells can have expected value greater than 1.

One solution to this problem is to only consider k-itemsets where $k \ll \log_2 n$. In most cases this is probably sufficient: it is not clear that a dependence involving dozens of items can be easily interpreted, even if it can be constructed. An alternate solution is to use an exact calculation for the probability, rather than the χ^2 approximation. The establishment of such a formula is still, unfortunately, a research problem in the statistics community, and more accurate approximations are prohibitively expensive.

Even in low dimensions, many contingency tables may have some cells with small counts. For these cells, small inaccuracies in the expected count will greatly affect the χ^2 value. For this reason, for cells with expectation less than 1 we reassign $E(r) = O(r)$. This is the most conservative course of action possible in this case, and it helps ensure that we will not make a judgment of dependence because of the contribution of a cell with very low support. See Section 5 for further discussion of combining χ^2 with support.

Finally, it is tempting to use the value of the χ^2 statistic to indicate the degree of dependence. This is dangerous, because when the independence hypothesis is false, the calculated χ^2 value tends to infinity as the sample size increases. While comparing χ^2 values within the same data set may be meaningful, comparing values of different data sets will almost certainly not be.

5. Pruning-based Algorithms for Dependence Rules

As we have mentioned, finding dependence rules is equivalent to finding a border in the itemset lattice. How big can this border be? In the worst case, when the border is in the middle of the lattice, it is exponential in the number of items. Even in the best case the border is at least quadratic. If there are 1000 items, which is not unreasonable, finding the entire border can be prohibitively expensive. Thus, it is necessary to provide some pruning function that allows us to ignore "uninteresting" itemsets in the border. This pruning function cannot merely be a post-processing step, since this does not improve the running time. Instead, it must prune parts of the lattice as the algorithm proceeds.

Consider the level-wise algorithms, which first determine the significant (and interesting) nodes among the itemsets of size 2, and then considers the itemsets of size 3, and so on. Then for the pruning criterion to be effective, it must be closed, so we can determine potentially interesting nodes at the next level based on nodes at the current level. An obvious pruning function fitting this criterion is support.

We need a different definition of support, however, than the one used in the support-confidence framework, because unlike in the support-confidence framework we also seek negative dependence. In other words, the support-confidence framework only looks at the top-left cell in the chi-squared contingency table. We extend this definition of support as follows:

Definition 10. A set of items S has **contingency table support (CT-support)** s at the $p\%$ level if at least $p\%$ of the cells in the contingency table for S have value s.

By requiring that p be a percent, rather than an absolute number, we make our definition of CT-support downward-closed.

THEOREM 3 *The property of having CT-support s at the $p\%$ level is a downward-closed property.*

Proof: Suppose a set of items S has CT-support s at the $p\%$ level. Consider, without loss of generality, a subset $T = S \setminus \{i\}$. Then each cell of T has a value equal to the sum of two cells in S. In particular, cell I of T is equal to them sum of cells $I \cup \{i\}$ and $I \cup \{\bar{i}\}$ in S. Since s is an absolute number, if either of the two cells in S has support s, so will the cell in T. In terms of counting supported cells, the worst case is if both $I \cup \{i\}$ and $I \cup \{\bar{i}\}$ are supported. In this case, there are two supported cells in S corresponding to a single supported cell in T, causing the number of supported cells in T to be half that of S. But the number of total cells in T is half that of S, so the percent of supported cells cannot decrease. ■

Note that values in the contingency table are observed values, not expected values.

One weakness of this CT-support definition is that, unless p is larger than 50%, all items have CT-support at level 1. Thus, pruning at level 1 is never productive, and a quadratic algorithm looms. If p is larger than 25%, though, we can do special pruning at level 1. Observe that $p > 0.25$ means that at least two cells in the contingency table will need support s. If neither item i_1 or i_2 occurs as often as s, this amount of support is impossible: only $\overline{i_1 i_2}$ could possibly have the necessary count. If there are many rare items — a similar argument holds if there are many very common items — this pruning is quite effective.

Other pruning strategies may be used, besides support-based pruning. One possibility is anti-support, where only rarely occurring combinations of items are interesting. This may be appropriate in the fire code example mentioned in Section 1, for instance, since fires — and the conditions leading up to them — are rare. Anti-support cannot be used with the chi-squared test at this time, however, since the chi-squared statistic is not accurate for very rare events. Another possible pruning method is to prune itemsets with very *high* χ^2 values, under the theory that these dependencies are probably so obvious as to be uninteresting. Since this property is not downward-closed, it would not be effective at pruning in a level-wise algorithm. A random walk algorithm, for instance (Gunopulos et al., 1997), might be appropriate for this kind of pruning.

5.1. The Algorithm

Combining the chi-squared dependence rule with pruning via CT-support, we obtain the algorithm in Figure 1.

Definition 11. We say that an itemset is **significant** if it is CT-supported and minimally dependent.

Algorithm Dependence Rules
Input: A chi-squared significance level α, support s, support fraction $p > 0.25$. Basket data B.
Output: A collection of minimal dependent itemsets, from B.

1. **For** each item $i \in I$, **do** count $O(i)$. We can use these values to calculate any necessary expected value, as explained in Section 4.

2. Initialize CAND $\leftarrow \emptyset$, SIG $\leftarrow \emptyset$, NOTSIG $\leftarrow \emptyset$.

3. *For* each pair of items $i_a, i_b \in I$ such that $O(i_a) > s$ and $O(i_b) > s$, **do** add $\{i_a, i_b\}$ to CAND.

4. NOTSIG $\leftarrow \emptyset$.

5. **If** CAND is empty, **then return** SIG and terminate.

6. **For** each itemset in CAND, **do** construct the contingency table for the itemset. **If** less than p percent of the cells have count s, **then goto** Step 8.

7. **If** the χ^2 value for the contingency table is at least χ^2_α, **then** add the itemset to SIG, **else** add the itemset to NOTSIG.

8. **Continue** with the next itemset in CAND. **If** there are no more itemsets in CAND, **then** set CAND to be the collection of all sets S such that every subset of size $|S| - 1$ of S is in NOTSIG. **Goto** Step 4.

Figure 1. The algorithm for determining significant (i.e., dependent and CT-supported) itemsets. It hinges on the fact that significant itemsets at level $i + 1$ are supersets of CT-supported but independent sets at level i. Step 8 can be implemented efficiently using hashing.

The key observation is stated in the following theorem; the proof follows from the preceding discussion.

THEOREM 4 *An itemset at level* $i + 1$ *can be significant only if all its subsets at level* i *have CT-support and none of its subsets at level* i *are dependent.*

Proof: If some subset of I fails to have CT-support, then I also must fail to have CT-support, since CT-support is downward closed. If some subset of I is dependent, then I cannot be minimally dependent by definition. ■

Thus, for level $i + 1$, all we need is a list of the CT-supported but independent itemsets from level i. This list is held in NOTSIG. The list SIG, which holds the CT-supported and dependent itemsets, is the output set of interest.

The final list is CAND, which builds candidate itemsets for level $i + 1$ from the NOTSIG list at level i. Let S be a set of size $i + 1$ for which every subset of size i is in NOTSIG. Then S is not ruled out by either CT-support pruning or significance pruning and is added to CAND. Once CAND has been constructed, we are done processing itemsets at level i. To start level $i+1$, we examine each set $S \in$ CAND to see if it actually does have the necessary CT-support. If so, we add it to either SIG or NOTSIG for level $i+1$, depending on its χ^2 value.

5.2. Implementation Details and Analysis

We now specify some of the key implementation details of our algorithm and obtain bounds on its running time.

The most expensive part of the algorithm is Step 8. We propose an implementation based on perfect hash tables (Fredman et al., 1984, Dietzfelbinger et al., 1988). In these hash tables, there are no collisions, and insertion, deletion, and lookup all take constant time. The space used is linear in the size of the data. Both NOTSIG and CAND are stored in hash tables. Elements of SIG can be stored in an array, or output as they are discovered and not stored at all.

To construct candidates for CAND using hash tables, we consider each pair of elements in NOTSIG. Suppose A and B are itemsets in NOTSIG. If $|A \cup B| = i+1$, $A \cup B$ might belong in CAND. To test this, we consider all $i-1$ remaining subsets of $A \cup B$ which have size i. We can test each one for inclusion in NOTSIG in constant time. If all subsets are in NOTSIG, we add $A \cup B$ to CAND, otherwise we ignore it. The total time for this operation is $O(|\text{NOTSIG}|^2 i)$.

Calculation of χ^2, at first blush, seems to take time $O(2^i)$ at level i, since we need to consider every cell in the contingency table. We can reduce the time to $O(\min\{n, 2^i\})$ by storing the contingency table sparsely, that is, by not storing cells where the observed count is 0. The problem is that cells with count 0 still contribute to the χ^2 value. Thus we massage the χ^2 formula as follows:

$$\sum_{r \in R} \frac{(O(r) - E(r))^2}{E(r)} = \sum_r \frac{O(r)}{E(r)}(O(r) - 2E(r)) + \sum_r E(r).$$

Now $\sum_r E(r) = n$, and $\frac{O(r)}{E(r)}(O(r) - 2E(r))$ is 0 if $O(r)$ is 0. We can calculate χ^2 values based only on occupied cells, and there can be at most n of these.

One expensive operation remains. To construct the contingency table for a given itemset, we must make a pass over the entire database. In the worst case, this requires k^i passes at level i. An alternative is to make one pass over the database at each level, constructing all the necessary contingency tables at once. We need one contingency table for each element of CAND. This requires $O(k^i)$ space in the worst case, though pruning will reduce the space requirements significantly. At level 2, which usually requires the most space in practice, the space requirement of $O(k^2)$ is probably not onerous, especially since storing an entire 2-dimensional contingency table requires only 4 words. The time required at level i is, in both cases, $O(n|\text{CAND}|) \in O(nk^i)$.

The preceding discussion yields the following theorem.

THEOREM 5 *The running time of Algorithm Dependence Rules for level i is*

$$O(n|\text{CAND}| \min\{n, 2^i\} + i|\text{NOTSIG}|^2).$$

It is instructive to compare the algorithm in Figure 1 to the hash-based algorithm of Park, Chen, and Yu (Park et al., 1995) for the support-confidence framework. Their algorithm also uses hashing to construct a candidate set CAND, which they then iterate over to verify the results. One difference is that verification is easier in their case, since they only need to test support. We also need to test chi-squared values, a more expensive operation that makes careful construction of CAND more important. Another difference is we use perfect hashing while Park, Chen, and Yu (Park et al., 1995) allow collisions. While collisions reduce the effectiveness of pruning, they do not affect the final result. The advantage of allowing collisions is that the hash table may be smaller. Hashing with collisions is necessary when the database is much larger than main memory. While we can afford collisions when constructing the candidate set — with the result of less accurate pruning — we need perfect hashing for NOTSIG. NOTSIG grows with the dimensionality and with the number of items. It is an open problem to modify our algorithm for databases with many items.

6. Experimental Results

There is a wide range of problems for which dependence rules are appropriate. In this section, we describe the results of the experiments we performed with three different kinds of data: boolean/numeric census data (Section 6.1), text data from newsgroups (Section 6.2), and synthetic data (Section 6.3). While the first two are useful for illustrating the conceptual aspect of the dependence rules, the last shows the effect of our pruning strategies on the performance of the algorithm.

Census data, such as that in Tables 1 and 2, readily lends itself to dependence analysis. Since the chi-squared test extends easily to non-binary data, we can analyze dependencies between multiple-choice answers such as those found in census forms.[4] Even when collapsing the census results to binary data, as we have chosen to do, we can find useful dependencies (see Example 4).

Another important application is the analysis of text data. In this case, each basket is a document, and each item is a word that occurs in some document. If the documents are newspaper articles, for instance, mining may turn up two company names that occur together more often than would be expected. We could then examine these two companies and see if they are likely to merge or reach an operating agreement. Negative dependencies may also be useful, such as the discovery that a document consisting of recipes contains the word FATTY less often than would be expected.

6.1. Census Data

The first data set we tested was the census data set, with $n = 30370$ baskets and $k = 10$ binary items. The items are as in Table 1. We show results for both the χ^2-interest test (Table 3) and the support-confidence test (Table 4). For the χ^2-interest test, we report p values as well as χ^2 values. The p value associated with a χ^2 value is the probability that independent variables would produce data yielding the χ^2 value seen (or a larger one). This means no χ^2 value is significant at a level above $1 - p$. Thus, χ^2 scores that are significant at the 95% significance level are exactly those with p value below 0.05.

a	b	χ^2	p value	$I(ab)$	$I(\overline{a}b)$	$I(a\overline{b})$	$I(\overline{a}\overline{b})$
SOLO-DRIVER	FEW-KIDS	**37.15**	**0.0000**	1.025	0.995	**0.773**	1.050
SOLO-DRIVER	NOT-VETERAN	**244.47**	**0.0000**	0.934	1.015	**1.554**	0.879
SOLO-DRIVER	ENGLISH	0.94	0.3323	1.004	0.999	0.966	1.007
SOLO-DRIVER	NOT-CITIZEN	**4.57**	**0.0325**	**0.901**	1.022	1.007	0.998
SOLO-DRIVER	BORN-US	0.05	0.8231	0.999	1.000	1.008	0.998
SOLO-DRIVER	MARRIED	**737.18**	**0.0000**	**1.574**	0.874	0.807	1.042
SOLO-DRIVER	UNDER-40	**153.11**	**0.0000**	0.880	1.026	**1.192**	0.958
SOLO-DRIVER	MALE	**138.13**	**0.0000**	**1.155**	0.966	0.866	1.029
SOLO-DRIVER	HOUSEHOLDER	**746.28**	**0.0000**	**1.404**	0.912	0.722	1.061
FEW-KIDS	NOT-VETERAN	**296.55**	**0.0000**	0.989	1.104	1.094	**0.135**
FEW-KIDS	ENGLISH	**24.00**	**0.0000**	0.997	1.030	1.026	**0.759**
FEW-KIDS	NOT-CITIZEN	1.60	0.2059	1.009	0.917	0.999	1.006
FEW-KIDS	BORN-US	1.70	0.1923	0.999	1.008	1.007	0.933
FEW-KIDS	MARRIED	**352.31**	**0.0000**	0.939	**1.562**	1.021	0.811
FEW-KIDS	UNDER-40	**2010.07**	**0.0000**	1.067	0.385	0.892	**1.988**
FEW-KIDS	MALE	**2855.73**	**0.0000**	1.109	**0.000**	0.906	1.863
FEW-KIDS	HOUSEHOLDER	**229.07**	**0.0000**	0.965	**1.317**	1.024	0.782
NOT-VETERAN	ENGLISH	**82.02**	**0.0000**	0.994	1.053	1.051	**0.576**
NOT-VETERAN	NOT-CITIZEN	**190.71**	**0.0000**	1.103	**0.140**	0.993	1.061
NOT-VETERAN	BORN-US	**176.05**	**0.0000**	0.991	1.075	1.077	**0.355**
NOT-VETERAN	MARRIED	**993.31**	**0.0000**	0.892	**1.901**	1.036	0.697
NOT-VETERAN	UNDER-40	**2006.34**	**0.0000**	1.070	0.414	0.887	**1.942**
NOT-VETERAN	MALE	**3099.38**	**0.0000**	0.881	**1.994**	1.103	0.142
NOT-VETERAN	HOUSEHOLDER	**819.90**	**0.0000**	0.931	**1.573**	1.047	0.606
ENGLISH	NOT-CITIZEN	**9130.58**	**0.0000**	0.271	**6.823**	1.052	0.588
ENGLISH	BORN-US	**11119.28**	**0.0000**	1.073	0.417	0.372	**6.016**
ENGLISH	MARRIED	**110.31**	**0.0000**	0.963	**1.294**	1.012	0.901
ENGLISH	UNDER-40	**62.22**	**0.0000**	0.987	1.101	1.020	**0.838**
ENGLISH	MALE	**21.41**	**0.0000**	0.990	**1.081**	1.009	0.930
ENGLISH	HOUSEHOLDER	0.10	0.7518	1.001	0.994	0.999	1.004
NOT-CITIZEN	BORN-US	**18504.81**	**0.0000**	0.000	1.071	**9.602**	0.391
NOT-CITIZEN	MARRIED	**189.66**	**0.0000**	**1.512**	0.964	0.828	1.012
NOT-CITIZEN	UNDER-40	**76.04**	**0.0000**	1.148	0.989	**0.762**	1.017
NOT-CITIZEN	MALE	**14.48**	**0.0001**	**1.088**	0.994	0.924	1.005
NOT-CITIZEN	HOUSEHOLDER	3.27	0.0706	0.953	1.003	1.032	0.998
BORN-US	MARRIED	**312.15**	**0.0000**	0.940	**1.512**	1.020	0.827
BORN-US	UNDER-40	**10.62**	**0.0011**	0.995	1.043	1.008	**0.930**
BORN-US	MALE	**12.95**	**0.0003**	0.992	**1.065**	1.007	0.944
BORN-US	HOUSEHOLDER	2.50	0.1138	0.996	1.032	1.003	0.978
MARRIED	UNDER-40	**2913.05**	**0.0000**	0.579	1.142	**1.677**	0.772
MARRIED	MALE	**66.49**	**0.0000**	**1.087**	0.971	0.925	1.025
MARRIED	HOUSEHOLDER	**186.28**	**0.0000**	**1.163**	0.945	0.888	1.038
UNDER-40	MALE	**98.63**	**0.0000**	1.048	**0.922**	0.958	1.067
UNDER-40	HOUSEHOLDER	**4285.29**	**0.0000**	0.643	**1.574**	1.246	0.605
MALE	HOUSEHOLDER	**12.40**	**0.0004**	**1.026**	0.977	0.982	1.016

Table 3. The χ^2 test on census data. Bold χ^2 values are significant at the 95% significance level. ($1-p$ indicates the maximum level for which this χ^2 value is significant.) Bold interest values are the most extreme.

a b	$s_{a\cup b}$	$s_{\overline{a}\cup b}$	$s_{a\cup\overline{b}}$	$s_{\overline{a}\cup\overline{b}}$	$a\Rightarrow b$	$\overline{a}\Rightarrow b$	$a\Rightarrow\overline{b}$	$\overline{a}\Rightarrow\overline{b}$	$b\Rightarrow a$	$b\Rightarrow\overline{a}$	$\overline{b}\Rightarrow a$	$\overline{b}\Rightarrow\overline{a}$
i_0 i_1	**16.6**	**73.6**	**1.4**	**8.5**	**0.92**	**0.90**	0.08	0.10	0.18	**0.82**	0.14	**0.86**
i_0 i_2	**15.0**	**74.3**	**3.0**	**7.7**	**0.83**	**0.91**	0.17	0.09	0.17	**0.83**	0.28	**0.72**
i_0 i_3	**16.0**	**72.9**	**1.9**	**9.2**	**0.89**	**0.89**	0.11	0.11	0.18	**0.82**	0.17	**0.83**
i_0 i_4	**1.1**	**5.5**	**16.9**	**76.5**	0.06	0.07	**0.94**	**0.93**	0.16	**0.84**	0.18	**0.82**
i_0 i_5	**16.1**	**73.5**	**1.9**	**8.5**	**0.90**	**0.90**	0.10	0.10	0.18	**0.82**	0.18	**0.82**
i_0 i_6	**7.1**	**18.1**	**10.8**	**64.0**	0.40	0.22	**0.60**	**0.78**	0.28	**0.72**	0.14	**0.86**
i_0 i_7	**9.7**	**51.9**	**8.2**	**30.2**	**0.54**	**0.63**	0.46	0.37	0.16	**0.84**	0.21	**0.79**
i_0 i_8	**9.6**	**36.7**	**8.3**	**45.3**	**0.54**	0.45	0.46	**0.55**	0.21	**0.79**	0.16	**0.84**
i_0 i_9	**10.3**	**30.5**	**7.7**	**51.6**	**0.57**	0.37	0.43	**0.63**	0.25	**0.75**	0.13	**0.87**
i_1 i_2	**79.6**	**9.7**	**10.6**	0.1	**0.88**	**0.99**	0.12	0.01	**0.89**	0.11	**0.99**	0.01
i_1 i_3	**79.9**	**9.0**	**10.3**	0.8	**0.89**	**0.92**	0.11	0.08	**0.90**	0.10	**0.93**	0.07
i_1 i_4	**6.0**	0.6	**84.2**	**9.2**	0.07	0.06	**0.93**	**0.94**	**0.91**	0.09	**0.90**	0.10
i_1 i_5	**80.7**	**8.9**	**9.5**	1.0	**0.90**	**0.90**	0.10	0.10	**0.90**	0.10	**0.91**	0.09
i_1 i_6	**21.3**	**3.9**	**68.9**	**6.0**	0.24	0.39	**0.76**	**0.61**	**0.85**	0.15	**0.92**	0.08
i_1 i_7	**59.3**	**2.3**	**30.9**	**7.5**	**0.66**	0.24	0.34	**0.76**	**0.96**	0.04	**0.80**	0.20
i_1 i_8	**46.3**	0.0	**43.8**	**9.8**	**0.51**	0.00	0.49	**1.00**	**1.00**	0.00	**0.82**	0.18
i_1 i_9	**35.5**	**5.3**	**54.7**	**4.6**	0.39	**0.54**	**0.61**	0.46	**0.87**	0.13	**0.92**	0.08
i_2 i_3	**78.9**	**10.0**	**10.4**	0.7	**0.88**	**0.94**	0.12	0.06	**0.89**	0.11	**0.94**	0.06
i_2 i_4	**6.5**	0.1	**82.8**	**10.6**	0.07	0.01	**0.93**	**0.99**	**0.99**	0.01	**0.89**	0.11
i_2 i_5	**79.3**	**10.3**	**10.0**	0.4	**0.89**	**0.96**	0.11	0.04	**0.89**	0.11	**0.96**	0.04
i_2 i_6	**20.1**	**5.1**	**69.2**	**5.6**	0.22	0.48	**0.78**	**0.52**	**0.80**	0.20	**0.93**	0.07
i_2 i_7	**58.9**	**2.7**	**30.4**	**8.0**	**0.66**	0.26	0.34	**0.74**	**0.96**	0.04	**0.79**	0.21
i_2 i_8	**36.5**	**9.9**	**52.9**	0.8	0.41	**0.92**	**0.59**	0.08	**0.79**	0.21	**0.98**	0.02
i_2 i_9	**33.9**	**6.9**	**55.4**	**3.8**	0.38	**0.64**	**0.62**	0.36	**0.83**	0.17	**0.94**	0.06
i_3 i_4	**1.6**	**5.0**	**87.3**	**6.1**	0.02	0.45	**0.98**	**0.55**	0.24	**0.76**	**0.93**	0.07
i_3 i_5	**85.4**	**4.2**	**3.4**	**7.0**	**0.96**	0.37	0.04	**0.63**	**0.95**	0.05	0.33	**0.67**
i_3 i_6	**21.6**	**3.6**	**67.3**	**7.5**	0.24	0.33	**0.76**	**0.67**	**0.86**	0.14	**0.90**	0.10
i_3 i_7	**54.1**	**7.6**	**34.8**	**3.6**	**0.61**	**0.68**	0.39	0.32	**0.88**	0.12	**0.91**	0.09
i_3 i_8	**40.8**	**5.6**	**48.1**	**5.6**	0.46	**0.50**	**0.54**	0.50	**0.88**	0.12	**0.90**	0.10
i_3 i_9	**36.2**	**4.5**	**52.6**	**6.6**	0.41	0.40	**0.59**	**0.60**	**0.89**	0.11	**0.89**	0.11
i_4 i_5	0.0	**89.6**	**6.6**	**3.8**	0.00	**0.96**	**1.00**	0.04	0.00	**1.00**	**0.64**	0.36
i_4 i_6	**2.5**	**22.7**	**4.1**	**70.7**	0.38	0.24	**0.62**	**0.76**	0.10	**0.90**	0.05	**0.95**
i_4 i_7	**4.7**	**57.0**	**1.9**	**36.4**	**0.71**	**0.61**	0.29	0.39	0.08	**0.92**	0.05	**0.95**
i_4 i_8	**3.3**	**43.0**	**3.3**	**50.4**	**0.50**	0.46	0.50	**0.54**	0.07	**0.93**	0.06	**0.94**
i_4 i_9	**2.6**	**38.2**	**4.0**	**55.2**	0.39	0.41	**0.61**	**0.59**	0.06	**0.94**	0.07	**0.93**
i_5 i_6	**21.2**	**4.0**	**68.4**	**6.4**	0.24	0.38	**0.76**	**0.62**	**0.84**	0.16	**0.91**	0.09
i_5 i_7	**54.9**	**6.7**	**34.6**	**3.7**	**0.61**	**0.64**	0.39	0.36	**0.89**	0.11	**0.90**	0.10
i_5 i_8	**41.2**	**5.1**	**48.4**	**5.3**	0.46	0.49	**0.54**	**0.51**	**0.89**	0.11	**0.90**	0.10
i_5 i_9	**36.4**	**4.4**	**53.2**	**6.0**	0.41	0.42	**0.59**	**0.58**	**0.89**	0.11	**0.90**	0.10
i_6 i_7	**9.0**	**52.7**	**16.2**	**22.2**	0.36	**0.70**	**0.64**	0.30	0.15	**0.85**	0.42	**0.58**
i_6 i_8	**12.7**	**33.6**	**12.5**	**41.2**	**0.50**	0.45	0.50	**0.55**	0.27	**0.73**	0.23	**0.77**
i_6 i_9	**11.9**	**28.8**	**13.3**	**46.0**	0.47	0.39	**0.53**	**0.61**	0.29	**0.71**	0.22	**0.78**
i_7 i_8	**29.9**	**16.4**	**31.7**	**22.0**	0.49	0.43	**0.51**	**0.57**	**0.65**	0.35	**0.59**	0.41
i_7 i_9	**16.1**	**24.6**	**45.5**	**13.8**	0.26	**0.64**	**0.74**	0.36	0.40	**0.60**	**0.77**	0.23
i_8 i_9	**19.4**	**21.4**	**27.0**	**32.3**	0.42	0.40	**0.58**	**0.60**	0.48	**0.52**	0.45	**0.55**

Table 4. Support/Confidence applied to census data. Bold values in the first block correspond to support (at the 1% cutoff); bold values in the second block correspond to confidence (at the 0.5 cutoff) as well.

To generate the χ^2 values for this data, we ran the algorithm in Figure 1 on a 90 MHz. Pentium running Linux 1.2.13. The machine has 32 Meg. of main memory. The program was written in C and compiled using `gcc` with the `-O6` compilation option. The entire database fit into main memory. The program took 3.6 seconds of CPU time to complete.

Let us illustrate how data mining could be performed on the results in Table 3. With so many pairs dependent, we are immediately struck by {FEW-KIDS, NOT-CITIZEN} and {FEW-KIDS, BORN-US}, which are not. We are even more surprised when we see that FEW-KIDS concerns number of children and NOT-CITIZEN and BORN-US concern markers for immigrants. This is surprising because conventional wisdom has it that immigrants are much more likely to have large families than native-born Americans. Perhaps, we conjecture, we are led astray by the category definition, since males are lumped together with women having few children. Perhaps it is not that immigrants have few children, but rather that they are preponderantly male. We look at the data for {NOT-CITIZEN, MALE} and {BORN-US, MALE} to explore this. These are both significant, and the interest figures show there is indeed a dependency between being male and being born abroad or not being a U.S. citizen. The interest values are fairly close to 1, though, indicating the bias is not strong. It does not seem strong enough to account for the independence we observed. A further jarring note for our explanation is the pair {FEW-KIDS, ENGLISH}. This pair includes native language, another marker of immigration. But {FEW-KIDS, ENGLISH} *is* significant, which would lead us to believe immigration is dependent on family size. Furthermore, ENGLISH is just as dependent on MALE as the other two markers of immigration. Perhaps, then, our assumption that ENGLISH, NOT-CITIZEN, and BORN-US are good markers of immigration is flawed. Table 3 gives us much to mull on.

We invite the reader to attempt a similar analysis with the support-confidence data in Table 4. For a special challenge, ignore the last seven columns, which are not typically mined in support-confidence applications. We find that it is much harder to draw interesting conclusions about census data from the support-confidence results.

Another interesting result is that SOLO-DRIVER and MARRIED are dependent, and the strongest dependence is between being married and driving alone. Does this imply that non-married people tend to carpool more often than married folk? Or is the data skewed because children cannot drive and also tend not to be married? Because we have collapsed the answers "does not drive" and "carpools," we cannot answer this question. A non-collapsed chi-squared table, with more than two rows and columns, could find finer-grained dependency. Support-confidence cannot easily handle multiple item values.

The magnitude of the χ^2 value can also lead to fruitful mining. The highest χ^2 values are for obvious dependencies, such as being born in the United States and being a U.S. citizen. These values often have interest levels of 0, indicating an impossible event (for instance, having given birth to more than 3 children and being male).

Results from support-confidence framework tend to be harder to understand. Considering BORN-US and MARRIED, we have both the rules, "If you are born in the U.S. you are likely to be married," and "If you are not married you are likely to be born in the U.S." These two statements are not inconsistent, but they are confusing. What is more worrisome, every pair of items has the maximum four supported rules. A good number would continue to support three or four rules even if the confidence level were raised to 75%. Someone mining this data using support-confidence would conclude that all item pairs have all sorts of interesting associations, when a look at the χ^2 values shows that some associations cannot be statistically justified. Furthermore, some of the pairs with the largest support and confidence values, such as FEW-KIDS and NOT-CITIZEN, turn out not to be dependent.

Note that, for this data set, no rule ever has adequate confidence but lacks support. This is not surprising since we examine only itemsets at level 2, where support is plentiful.

6.2. *Text Data*

We analyzed 91 news articles from the clari.world.africa news hierarchy, gathered on 13 September 1996. We chose only articles with at least 200 words (not counting headers), to filter out posts that were probably not news articles. A word was defined to be any consecutive sequence of alphabetic characters; thus "s" as a possessive suffix would be its own word while numbers would be ignored. To keep the experiment a reasonable size, we pruned all words occurring in less than 10% of the documents; this is a more severe type of pruning than the special level 1 pruning discussed in Section 5. This left us with 416 distinct words.

One would expect words to be highly dependent, and indeed this turned out to be the case. Of the $\binom{416}{2} = 86320$ word pairings, there were 8329 dependent pairs, i.e., 10% of all word pairs are dependent. More than 10% of all triples of words are dependent. Because of the huge amount of data generated, thorough analysis of the results is very difficult. We provide some anecdotal analysis, however, to give a taste of the effectiveness of the chi-squared test on text data.

A list of 12 dependent itemsets is presented in Table 5. We show not only the dependent words but the major dependence in the data. We see some obvious dependencies: AREA appears often with PROVINCE, which is not surprising since the two terms are clearly related. The largest single χ^2 value relates NELSON to MANDELA, again hardly surprising.

While some pairs of words have large χ^2 values, no triple has a χ^2 value larger than 10. Remember that we report minimal dependent itemsets, so no subset of a triple is itself dependent. Thus BURUNDI, COMMISSION, and PLAN are 3-way dependent, though COMMISSION and PLAN, say, are not. Since the major dependence has COMMISSION and PLAN but lacks BURUNDI, we might suspect that there are fewer commission making plans in Burundi than other African nations. Likewise, AFRICAN, MEN, and NELSON, are dependent, though AFRICAN and MEN alone are not, leading us to posit that articles including Nelson Mandela might disproportionately refer to African men. Another major dependence has OFFICIAL and AUTHORITIES occurring without the word BLACK. Could that be because race is not mentioned when discussing authority figures, or perhaps because non-black authority figures are given more prominence?

We include the threesome GOVERNMENT, IS, and NUMBER because it has the highest χ^2 value of any triple of words. Like many of the dependent triples, of which there are well over a million, this itemset is hard to interpret. Part of the difficulty is due to the word IS, which does not yield as much context as nouns and active verbs. In practice, it may make sense to restrict the analysis to nouns and active verbs to prune away such meaningless dependencies.

It is important to note that, with so many dependencies identified, some are bound to be incorrect. At a 95% significance level, we would expect 5% of all itemsets identified as

Table 5. Some word dependencies in the clari.world.africa news articles. Sometimes the dependencies are suggestive, but not always; the last itemset is one of the many confusing itemsets.

dependent words	χ^2	p value	major dependence includes	major dependence omits
area province	24.269	0.0000	area province	
area secretary war	6.959	0.0083	area war	secretary
area secretary they	7.127	0.0076	area they	secretary
country men work	4.047	0.0442	country men work	
deputy director	9.927	0.0016	deputy director	
members minority	4.230	0.0397	members minority	
authorities black official	4.366	0.0367	authorities official	black
burundi commission plan	5.452	0.0195	commission plan	burundi
african men nelson	5.935	0.0148	african men nelson	
liberia west	48.939	0.0000	liberia west	
mandela nelson	91.000	0.0000	mandela nelson	
government is number	9.999	0.0016	is number	government

dependent to be actually independent. The typical way to handle this problem is to raise the significance level, based on the number of itemsets we examine, so the expected number of misidentifications is low. When considering hundreds of thousands of itemsets, as in the text example, this approach is not feasible.

6.3. Synthetic Data

The census data is too small, and its border too low, to study the effectiveness of the pruning techniques. On the other hand, the text data is too big: we were forced to prune words with low support even before starting our mining algorithm. To get data that is the appropriate size for exploring the effectiveness of our algorithm, we turn to synthetic data from IBM's Quest group (Agrawal et al., 1996).

We generated market basket data with 99997 baskets and 871 items. We set the average basket size to be 20, and the average size of large itemsets to be 4. To generate the χ^2 values for this data, we ran the algorithm in Figure 1 on a Pentium Pro with a 166 MHz. processor running Linux 1.3.68. The machine has 64 Meg. of memory and the entire database fit into main memory. The program took 2349 seconds of CPU time to complete.

To analyze the effectiveness of the pruning, we look at several factors. One is the number of itemsets that exist at each level, i.e., the number of itemsets we would have to examine without pruning. The next is the size of CAND; this is the number of itemsets we actually examine. Each itemset in CAND is either added to SIG, added to NOTSIG, or discarded. The smaller the number of items discarded, the more effective our pruning techniques. We summarize these figures for the Quest data in Table 6.

Note that unlike with the text data, the number of dependencies at level 3 is much smaller than the number of dependencies at level 2. Though we do not show the numbers, it is again the case that the 3-way dependencies have much lower χ^2 values than the average 2-way dependence, with no 3-way dependence having $\chi^2 > 8.7$. In this case, both support and significance provide pruning, though the effect of support seems to be much more pronounced.

level	$\lvert$itemsets$\rvert$	$\lvert$CAND$\rvert$	$\lvert$CAND discards$\rvert$	$\lvert$SIG$\rvert$	$\lvert$NOTSIG$\rvert$
2	378015	8019	323	4114	3582
3	109372340	782	647	17	118
4	23706454695	0	0	0	0

Table 6. The effectiveness of pruning on reducing the number of itemsets examined. Two measures of pruning quality are the size of CAND and the number of CAND discards. The lower these two quantities are, the better. Note that itemsets in SIG would not be pruned by a support-confidence test, so $|\text{SIG}|$ is one measure of the effectiveness of dependence pruning considered by itself.

7. Conclusions and Further Research

We have introduced a generalization of association rules, called dependence rules, that are particularly useful in applications going beyond the standard market basket setting. In addition, these rules have some advantages over the use of standard association rules. Dependence rules seem useful for analyzing a wide range of data, and tests using the chi-squared statistic are both effective and efficient for mining.

Our work raises many important issues for further research. First, there is the question of identifying other measures and rule types that capture patterns in data not already captured by association rules and dependence rules. For example, in the case of documents, it would be useful to formulate rules that capture the spatial locality of words by paying attention to item ordering within the basket. In addition, it would be interesting to explore the class of measures and rules that lead to upward-closure or downward-closure in the itemset lattice, since closure appears to be a desirable property both from the conceptual and the efficiency points of view. We have also suggested another algorithmic idea, random walks on the lattice, for dependence rules that may apply in other settings. It is easy to verify that a random walk algorithm has a natural implementation in terms of a datacube of the count values for contingency tables, and we hope to explore this connection in a later paper.

With regard to the chi-squared test itself, a significant problem is the increasing inaccuracy of the chi-squared test as the number of cells increase. An efficient, exact test for dependence would solve this problem, though other computational solutions may be possible. In lieu of a solution, more research is needed into the effect of ignoring cells with low expectation. Though ignoring such cells can skew results arbitrarily on artificially constructed data sets, it is not clear what the impact is in practice.

Another major source of error, already mentioned in Section 6.2, is involved in the significance level cutoff. As the cutoff changes, so does the total number of false positives (independent itemsets with χ^2 value above the cutoff) and false negatives (dependent itemsets with χ^2 value below the cutoff). Further research is necessary to determine how the optimal cutoff value varies from application to application.

Another area of research is in non-support-based pruning criteria. If these criteria are not downward-closed, a non-level-wise algorithm will probably be necessary to keep the computation efficient. For example, it would be interesting to experiment with the random walk algorithm.

All of the data we have presented have small borders because most small itemsets are dependent. It might be fruitful to explore the behavior of data sets where the border is exponential in the number of items.

Finally, as we mentioned in Section 5.2, our algorithm requires that all the non-significant itemsets at a level be stored, and therefore it is not scalable to databases with many items. This problem becomes particularly acute when the border is high, for instance when 10-itemsets both exist and are useful for the experimenter to discover. Besides being important from a practical point of view, further research in this area may also yield more insight into properties of the border and of closure. Such a result would be useful in its own right.

Acknowledgments

We are grateful to Jeff Ullman for many valuable discussions. We would also like to thank members of the Stanford Data Mining group, particularly Shalom Tsur, for helpful discussions. We give our sincere thanks to Arnie Rosenthal and Cliff Bingham, who provided helpful and detailed comments, and to the editor and anonymous reviewers, who did the same.

Appendix A

The Theory of Chi-squared Distributions

Intuitively, the chi-squared statistic attempts to measure the degree of independence between different attributes by comparing their observed patterns of occurrence with the expected pattern of occurrence under the assumption of complete independence and a normal distribution on the occurrence of each attribute. Note that the normal distribution assumption is justified for a large value of m, as a reasonable distribution will approach normality asymptotically.

We briefly review the theoretical justification for employing the chi-squared statistic in this setting. This is classical work in statistics that goes back at least to the last century. Refer to the book by Lancaster (Lancaster, 1969) for the history and theory of the chi-squared test for independence.

Let X be a Bernoulli random variable that denotes the number of successes in N independent trials where the probability of success in any given trial is p. The expected number of successes is Np and the variance is $Np(1-p)$. The classical work of de Moivre (de Moivre, 1733) and Laplace (de Laplace, 1878) has established that the random variable $\chi = \frac{X-Np}{\sqrt{Np(1-p)}}$ follows the standard normal distribution. The square of this random variable χ is given by

$$\begin{aligned}
\chi^2 &= \frac{(X-Np)^2}{Np(1-p)} \\
&= \frac{(X-Np)^2}{Np} + \frac{((N-X)-N(1-p))^2}{N(1-p)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(X_1 - Np)^2}{Np} + \frac{(X_0 - N(1-p))^2}{N(1-p)} \\
&= \frac{(X_1 - E(X_1))^2}{E(X_1)} + \frac{(X_0 - E(X_0))^2}{E(X_0)},
\end{aligned}
$$

where X_1 denotes the number of successes and X_0 denote the number of failures in the N trials. Note that, by definition, the χ^2 random variable is asymptotically distributed as the square of a standard normal variable.

Pearson (Pearson, 1900) extended the definition to the multinomial case, where X can take on any value in a set U. The modified formula is

$$
\chi^2 = \sum_{r \in U} \frac{(X_r - E(X_r))^2}{E(X_r)}
$$

and yields a χ^2 distribution with $|U|-1$ degrees of freedom (we lose one degree of freedom due to the constraint $\sum_{r \in U} X_r = N$).

We can further generalize the χ^2 variable to the case of multiple random variables. We consider the binomial case, though the multinomial case extends in the expected way. Let $X^1, \ldots, X^k$ denote *k independent*, binomially distributed random variables. We can define a *contingency table* or *count table* CT that is a k-dimensional array indexed by $\{0,1\}^k$. Each index refers to a unique *cell* of the contingency table. The cell $CT(r_1, \ldots, r_k)$ in the table is a count of the number of trials, out of N independent trials, where the event $(X^1 = r_1, \ldots, X^k = r_k)$ occurs. We define the χ^2 value as

$$
\chi^2 = \sum_{r_1 \in \{0,1\}, \ldots, r_k \in \{0,1\}} \frac{(CT(r_1, \ldots, r_k) - E(CT(r_1, \ldots, r_k)))^2}{E(CT(r_1, \ldots, r_k))}
$$

This has 1 degree of freedom — we have two values in each row of the contingency table and one constraint in that the row sum is fixed. In the general multinomial case, if X^i can have u_i different values, there are $(u_1 - 1)(u_2 - 1) \cdots (u_k - 1)$ degrees of freedom.

We now prove the theorem stated in Section 4.

THEOREM 2 *In the binomial case, the chi-squared statistic is upward-closed.*

Proof: The key observation in proving this is that regarldess of the dimensionality, the chi-squared statistic has only one degree of freedom. Thus, to show upward-closure it is sufficient to show that if a set of items has χ^2 value S, then any superset of the itemset has χ^2 value at least S. We show this for itemsets of size 2, though the proof easily generalizes to higher dimensions.

Consider variables A, B, and C. The χ^2-statistic for the variables A and B is defined as follows:

$$
\begin{aligned}
S_{AB} = \; & \frac{(E(AB) - O(AB))^2}{E(AB)} + \frac{(E(A\overline{B}) - O(A\overline{B}))^2}{E(A\overline{B})} + \\
& \frac{(E(\overline{A}B) - O(\overline{A}B))^2}{E(\overline{A}B)} + \frac{(E(\overline{AB}) - O(\overline{AB}))^2}{E(\overline{AB})}
\end{aligned}
$$

Now, let E denote the value $E(AB)$ and O the value $O(AB)$. Define $x = E(ABC)$ and $y = E(AB\overline{C})$. Likewise, define $X = O(ABC)$ and $Y = O(AB\overline{C})$. Note that $E = x + y$ and $O = X + Y$. Then, in the χ^2-statistic S_{ABC} for the triple A, B, and C, we replace the term

$$\frac{(E-O)^2}{E}$$

in S_{AB} by the terms

$$\frac{(x-X)^2}{x} + \frac{(y-Y)^2}{y}.$$

Therefore, in $S_{ABC} - S_{AB}$, we have the terms

$$\begin{aligned}
&\frac{(x-X)^2}{x} + \frac{(y-Y)^2}{y} - \frac{(E-O)^2}{E} \\
&\quad = \frac{y(x+y)(x-X)^2 + x(x+y)(y-Y)^2 - xy[(x+y)-(X+Y)]^2}{xy(x+y)} \\
&\quad = \frac{xy(x-X)^2 + xy(y-Y)^2 + y^2(x-X)^2 + x^2(y-Y)^2}{xy(x+y)} - \\
&\qquad \frac{xy[(x-X)^2 + (y-Y)^2 + 2(x-X)(y-Y)]}{xy(x+y)} \\
&\quad = \frac{y^2(x-X)^2 + x^2(y-Y)^2 - 2xy(x-X)(y-Y)}{xy(x+y)} \\
&\quad = \frac{[y(x-X) - x(y-Y)]^2}{xy(x+y)} \\
&\quad = \frac{(xY - yX)^2}{xy(x+y)}
\end{aligned}$$

This term is never negative, implying that $S_{ABC} \geq S_{AB}$ always. ■

Appendix B

The data sets used in this paper are accessible via the following URL:

`http://www.research.microsoft.com/datamine`

Notes

1. A classic, albeit apocryphal, example is the rule that people who buy diapers in the afternoon are particularly likely to buy beer at the same time (Ewald, 1994).
2. The *p* value can be easily calculated via formulas, or obtained from widely available tables for the chi-squared distribution.
3. In reality we would mine this data rather than query for it. We present the material in this way in order to compare two testing techniques, not to illustrate actual use.
4. A danger is that as the number of cells increases, problems with accuracy of the χ^2 statistic increase as well.

References

Agrawal, R., Arning, A., Bollinger, T., Mehta, M., Shafer, J., and Srikant, R. 1996. The Quest Data Mining System. Proceedings of the Second International Conference on Knowledge Discovery in Databases and Data.

Agrawal, R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on the Management of Data, pp. 207–216.

Agrawal, R., Imielinski, T., and Swami, A. 1993. Database mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5:914–925.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., and Verkamo, A.I. 1996. Fast Discovery of Association Rules. Fayyad et al. (Fayyad et al., 1996), pp. 307–328.

Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499.

Agresti, A. 1992. A survey of exact inference for contingency tables. Statistical Science, 7:131-177.

Dietzfelbinger, M., Karlin, A., Mehlhorn, K., Meyer, F. auf der Heide, Rohnert, H., and Tarjan, R. 1988. Dynamic perfect hashing: Upper and lower bounds. Proceedings of the 18th IEEE Symposium on Foundations of Computer Science, pp. 524–531.

Ewald, R. 1994. Keynote address. The 3rd International Conference on Information and Knowledge Management.

Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P., and Uthrusamy, R. 1996. Advances in Knowledge Discovery and Data Mining, AAAI Press.

Fredman, M., Komlós, J., and Szemerédi, E. 1984. Storing a sparse table with $O(1)$ worst case access time. Journal of the ACM, 31(3):538–544.

Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T. 1996. Mining Optimized Association Rules for Numeric Attributes. Proceedings of the Fifteenth ACM Symposium on Principles of Database Systems.

Fukuda, T., Morimoto, Y., Morishita, S., and Tokuyama, T. 1996. Mining optimized association rules for numeric data. Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 13–24.

Gunopulos, D., Mannila, H., and Saluja, S. 1997. Discovering all most specific sentences by randomized algorithms. Proceedings of the 6th International Conference on Database Theory.

Han, J. and Fu, Y. 1995. Discovery of multiple-level association rules from large databases. Proceedings of the 21st International Conference on Very Large Data Bases, pp. 420–431.

Houtsma, M. and Swami, A. 1995. Set-oriented mining of association rules. Proceedings of the International Conference on Data Engineering, pp. 25–34.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I. 1994. Finding interesting rules from large sets of discovered association rules. Proceedings of the 3rd International Conference on Information and Knowledge Management, pp. 401–407

Lancaster, H.O. 1969. The Chi-squared Distribution, John Wiley & Sons, New York.

de Laplace, P.S. 1878. Oeuvres complétes de Laplace publiées sous les auspices de l'Académie des Sciences par M.M. les secrétaires perpétuels, Gauthier-Villar, Paris, 1878/1912.

Mannila, H., Toivonen, H., and Inkeri Verkamo, A. 1994. Efficient algorithms for discovering association rules. Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, pp. 144–155.

de Moivre, A. 1733. Approximatio ad summam terminorum binomii $(a+b)^n$ in seriem expansi, Supplement to Miscellanea Analytica, London.

Moore, D.S. 1986. Tests of chi-squared type. Goodness-of-Fit Techniques, R.B. D'Agostino and M.A. Stephens (Eds.), Marcel Dekker, New York, pp. 63–95.

Mosteller, F. and Wallace, D. 1964. Inference and Disputed Authorship: The Federalists, Addison-Wesley.

Park, J.S., Chen, M.S., and Yu, P.S. 1995. An effective hash based algorithm for mining association rules. Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 175–186.

Pearson, K. 1900. On a criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag., 5:157–175.

Piatetsky, G. and Frawley, W. 1991. Knowledge Discovery in Databases, AAAI/MIT Press.

Savasere, A., Omiecinski, E., and Navathe, S. 1995. An efficient algorithm for mining association rules in large databases. Proceedings of the International Conference on Very Large Data Bases, pp. 432–444

Srikant, R. and Agrawal, R. 1995. Mining generalized association rules. Proceedings of the 21st International Conference on Very Large Data Bases, pp. 407–419.

Toivonen, H. 1996. Sampling large databases for finding association rules. Proceedings of the 22nd International Conference on Very Large Data Bases, pp. 134–145.

Craig Silverstein obtained an A.B. degree in Computer Science from Harvard University and is currently a Ph.D. candidate in Computer Science at Stanford University. He is a recipient of a National Defense Science and Engineering Graduate fellowship and an Achievement Awards for College Scientists fellowhip. His research interests include information retrieval on natural language queries and discovering causality via data mining.

Sergey Brin received his B.S. degree in Mathematics and Computer Science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University, where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include data mining of large hypertext collections and time series.

Rajeev Motwani received a B.Tech. degree in Computer Science from the Indian Institute of Technology (Kanpur) in 1983. In 1988 he obtained a Ph.D. in Computer Science from the University of California at Berkeley. Since 1988 he has been at the Computer Science Department of Stanford University, where he now serves as an Associate Professor. He is a recipient of an Arthur P. Sloan Research Fellowship and the NSF National Young Investigator Award from the National Science Foundation. In 1993 he received the Bergmann Memorial Award from the US-Israel Binational Science Foundation, and in 1994 he was awarded an IBM Faculty Development Award. He is a Fellow of the Institute of Combinatorics. Dr. Motwani is a co-author of the book Randomized Algorithms published by Cambridge University Press in 1995. He has authored scholarly and research articles on a variety of areas related to theoretical computer science: combinatorial optimization and scheduling theory; design and analysis of algorithms including approximation algorithms, online algorithms and randomized algorithms; complexity theory; computational geometry; compilers; databases; and robotics.