Beyond Market Baskets: Generalizing Association Rules to Correlations

Sergey Brin*
Department of Computer Science
Stanford University
Stanford, CA 94305
brin@cs.stanford.edu

Rajeev Motwani†
Department of Computer Science
Stanford University
Stanford, CA 94305
motwani@cs.stanford.edu

Craig Silverstein‡
Department of Computer Science
Stanford University
Stanford, CA 94305
csilvers@cs.stanford.edu

Abstract

One of the most well-studied problems in data mining is mining for association rules in market basket data. Association rules, whose significance is measured via support and confidence, are intended to identify rules of the type, "A customer purchasing item A often also purchases item B." Motivated by the goal of generalizing beyond market baskets and the association rules used with them, we develop the notion of mining rules that identify correlations (generalizing associations), and we consider both the absence and presence of items as a basis for generating rules. We propose measuring significance of associations via the chi-squared test for correlation 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 correlated and uncorrelated itemsets in the lattice. We develop pruning strategies and devise an efficient algorithm for the resulting problem. We demonstrate its effectiveness by testing it on census data and finding term dependence in a corpus of text documents, as well as on synthetic data.

1 Introduction

The term "data mining" has been applied to a broad range of activities that attempt to discover new information from existing information, where usually the original information was gathered for a purpose entirely different from the way it is used for data mining. Typically, the applications involve large-scale information banks such as data warehouses [30] or datacubes [13]. One of the more well-studied problems in data mining is the search for association rules in market basket data [2, 3, 17, 20, 5, 15, 16, 24, 28, 27, 4, 29]. 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 [11, 12].

While Piatetsky-Shapiro and Frawley [26] define an "association problem" as 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, whose significance is measured via support and confidence as explained below, are primarily intended to identify rules of the type, "A customer purchasing item X is likely to also purchase item Y." 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 by the goal of generalizing beyond market baskets and association rules used with them. 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, the type of association rule described above is 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 correlations (generalizing associations) and we also take into consideration the *absence* of items as a basis for generating rules.

We propose measuring significance of rules via the chi-squared test for correlation from classical statistics. This leads to a measure that 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 correlated and uncorrelated itemsets in the lattice. Based on this observation and some pruning strategies we developed, we present efficient algorithms for the resulting problem. We also demonstrate the effectiveness of our algorithms by experiments on census data and finding term dependency in a corpus of text documents.

*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.

‡Supported by the Department of Defense, with partial support from NSF Award CCR–9357849, with matching funds from IBM, Mitsubishi, Schlumberger Foundation, Shell Foundation, and Xerox Corporation.

[1] A classic example is the rule that people who buy diapers in the afternoon are particularly likely to buy beer at the same time [8].

1.1 Association Rules

In order to place our work in the context of earlier work, it will be helpful to first review some of the details of the past work on association rules in the market basket application. For this purpose, we define **basket data** in general terms. Let $I = \{i_1, \ldots, i_k\}$ be a set of k elements, called **items**. Let $B = \{b_1, \ldots, b_n\}$ be a set of n subsets of I. We call each $b_i \subseteq I$ 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 let us 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 [2] is intended to capture a certain type of dependence among items represented in the database B. Specifically, we say that $i_1 \Rightarrow i_2$ 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 are also likely to buy coffee." The confidence statistic ensures that the statement is true often enough to make a marketing campaign effective or to justify changing product placement in the store. 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 significance statistic may be too large to draw any useful conclusions.

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, the support-confidence framework does not support negative implications of the type: "When people buy batteries, they do not usually also buy cat food." 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 correlations between certain types of electrical wiring and the occurrence of fires.

A bigger problem is the support-confidence framework does not work well when correlation is the appropriate measure.

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, rows t and $\overline{t}$ correspond to baskets that do and do not, respectively, contain tea, and similarly columns c and $\overline{c}$ correspond to coffee. The numbers represent percentage of baskets.*

	c	$\overline{c}$	Σ_{row}
t	20	5	25
$\overline{t}$	70	5	75
Σ_{col}	90	10	100

Let us apply the support-confidence framework to the potential association rule $t \Rightarrow c$. The support for this rule is 20%, which is fairly high. The confidence is defined to be the conditional probability that a customer buys coffee, given that she buys tea, i.e., $P[t \wedge c]/P[t] = 20/25 = 0.8$ or 80%, which too is pretty high. At this point, we may conclude that the rule $t \Rightarrow c$ is a valid rule.

However, consider now the fact that the a priori probability that a 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 *correlation between buying tea and buying coffee; at least that information should be provided along with the association rule $t \Rightarrow c$. One way of measuring correlation is to compute $P[t \wedge c]/(P[t] \times P[c]) = 0.2/(0.25 \times 0.9) = 0.89$. The fact that this quantity is significantly less than 1 indicates negative correlation, 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 when the two purchases are completely independent.*

In the coffee and tea example, we calculated a correlation value but could not tell whether it was statistically significant. Testing for significant correlation is a problem statisticians have been studying for over a century; refer to Lancaster [18] for the theory and a history of this problem. The preferred test for correlation 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 captures correlation (as in the tea and coffee example) but can also detect negative implication (as in the fire code example).

In the rest of this paper, we develop our notion of correlation rules based on the chi-squared statistic by describing the theoretical underpinnings, efficient algorithms, implementations, and experiments with real basket data. In Section 2 we discuss the generalization of association rules to correlations. We also study some properties of correlation rules, particularly closure properties with respect to the itemset lattice. Consequently, we reduce the data mining task as the problem of computing a border (consisting of the minimally correlated itemsets) in the lattice. In Section 3 we study the chi-squared test for the correlation rules and provide some illustrative examples. We point out that the chi-squared test needs to be augmented with a measure of interest and provide a plausible candidate. We also contrast our approach with the support-confidence framework for association rules and discuss the the limitations of our framework. In Section 4 we describe *level-wise* algorithms for performing the task of identifying correlation rules. A pruning strategy is developed to increase the efficiency of the algorithm. We present the results from our experiments with two extremely different real world basket data sets in Section 5. Finally, in Section 6 we make concluding remarks. Appendix A gives some of the theoretical basis for the chi-squared test in statistics.

2 Correlation Rules

Let $p(A)$ be the probability that event A occurs and $p(\overline{A}) = 1 - p(A)$ the probability that event A does not occur. Likewise, $p(AB)$ is the probability that both event A and event B occur together, while $p(\overline{A}B)$ is the probability that B occurs but A does not. The events A and B are said to be *independent* if $p(AB) = p(A)p(B)$. Similarly, if $p(ABC) = p(A)p(B)p(C)$, then A, B, and C are *3-way independent.* If a set of events is not independent, it is *dependent.* If any of AB, $\overline{A}B$, $A\overline{B}$, $\overline{AB}$ are dependent, then A and B are said to be *correlated.* Likewise, if any of the eight combinations of A, B, C, and their complements are dependent, then A, B, and C are correlated.

If we have a series of n trials, we denote the number of times event A occurs as $O_n(A)$, or just $O(A)$ when n is understood. We can estimate $p(A)$ by $O_n(A)/n$. We can also estimate whether $p(AB) \neq p(A)p(B)$; our confidence in this estimate depends on n and, to a lesser extent, the observed counts.

To put this in the context of mining association rules, let I be a set of items, and B be a set of subsets of I. We say $\{i_{a_1}, \ldots, i_{a_m}\}$ is a *correlation rule* if the occurrences of the items $i_{a_1}, \ldots, i_{a_m}$ are correlated.

One important property of correlation is that it is upward closed: If a set of items S is correlated, so is every superset of S. Intuitively, it is clear that adding items to a correlated set cannot magically "cancel out" the correlation. This is easy to show formally by contradiction. Suppose A and B are correlated but A, B, and C are not. Then $p(AB) = p(ABC) + p(AB\overline{C}) = p(A)p(B)p(C) + p(A)p(B)p(\overline{C}) = p(A)p(B)$, where the middle equality follows because A, B, and C are independent. We can derive similar formulas for $p(A\overline{B})$, $p(\overline{A}B)$, and $p(\overline{AB})$. Together these imply A and B are not correlated, which is a contradiction. Of course, if $n < \infty$ we can never be *certain* S is actually correlated. However, the closure property also holds at any significance level α, in that if S is correlated with significance level α, any superset of S is also correlated with significance level α.[2] (See Appendix A for a proof.)

2.1 The Closure Property

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, than all its subsets also have support. Researchers have taken advantage of this closure property in devising algorithms. Level-wise algorithms [2] find all items with a given property among itemsets of size i (i-itemsets), and use this knowledge to explore itemsets of size $i+1$ (($i+1$)-itemsets). Another class of algorithms, random walk algorithms [14], generated 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. (It turns out that the random walk algorithm has a natural implementation in terms of a datacube [13]; a connection we intend to explore in a later paper.) Both level-wise and random walk algorithms use knowledge of a set and its closure properties to make inferences about its supersets.

Downward closure is a pruning property. To use it, we start out with the assumption that all $(i+1)$-itemsets are supported (to use a concrete example of a downward closed property). 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. We start with the assumption that no $(i+1)$-itemset is, say, correlated. Looking at an i-itemset, we can say that if it is correlated, all its supersets are also correlated. This gives us a list of correlated $(i+1)$-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 correlated $(i+1)$-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 correlated itemsets is that our pruning is less effective. It is for this reason we concentrate on *minimal* correlated itemsets, that is, itemsets that are correlated though no subset of them is correlated. Then, finding correlation is really a pruning step: We prune all the parents of a correlated i-itemset because they are not minimal.

2.2 The Border of Correlation

An advantage of upward closure is that it means the itemsets of interest form a **border**. That is, we can list a set 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 correlation data as the algorithm proceeds. This time- and space-saving shortcut does not work for confidence, which is not upward closed. If we combine correlation 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*

[2] A significance level of α means that, under the null hypothesis (in this case, that S is not correlated), $\chi^2_S < \chi^2_\alpha$ with probability α.

baskets including doughnuts, while the second is for baskets lacking doughnuts.

d	c	$\overline{c}$	Σrow
t	8	1	9
$\overline{t}$	40	2	42
Σ_{col}	48	3	51

$\overline{d}$	c	$\overline{c}$	Σrow
t	10	2	12
$\overline{t}$	35	2	37
Σ_{col}	45	4	49

Observe that $P[c \wedge d] = 48$, $P[c] = 93$, *so the rule* $c \Rightarrow d$ *has confidence* 0.52. *On the other hand,* $P[t \wedge c \wedge d] = 8$, $P[t \wedge c] = 18$, *so the rule* $c, t \Rightarrow d$ *has confidence* 0.44. *For a reasonable confidence cutoff of* 0.50, $c \Rightarrow d$ *has confidence but its superset* $c, t \Rightarrow d$ *does not.*

The border property is incredibly useful. 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. It can then do a local analysis of the border near the crossing.

While upward closure seems superior to downward closure because of the border property, in reality it is not necessary to choose between them. We discuss efficient ways of combining the two closure properties in Section 4.

3 The Chi-squared Test for Independence

Let R be $\{i_1, \overline{i_1}\} \times \cdots \times \{i_k, \overline{i_k}\}$ and $r = r_1 \ldots r_k \in R$. Here R is the set of all possible basket values, and r is a single basket value. Each value of r denotes a *cell* — this 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, expectation is calculated under the assumption of independence. Thus, $E[i_j] = O_n(i_j)$ for a single item, $E[\overline{i_j}] = n - O_n(i_j)$, and $E[r] = n \times E[r_1]/n \times \cdots \times E[r_k]/n$. 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 correlated, 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. If all the variables were really independent, the chi-squared value would be 0 (allowing for fluctuations if $n < \infty$). If it is higher than a cutoff value (3.84 at the 95% significance level) we reject the independence assumption. Note that the cutoff value for any given significance level can be obtained from widely available tables for the chi-squared distribution.

In Theorem 1 (Appendix A), we prove that the chi-squared test at a given significance level is upward closed.

Example 3 *Consider the census data presented in Table 1. The contingency table for* i_8 *and* i_9 *would be as follows:*

	i_8	$\overline{i_8}$	Σrow
i_9	1	2	3
$\overline{i_9}$	4	2	6
Σ_{col}	5	4	9

Now $E[i_9] = O(i_9) = 3$, *while* $E[i_8] = O(i_8) = 5$; *note that* $E[i_9]$ *is the sum of row 1, while* $E[i_8]$ *is the sum of column 1. The chi-squared value is*

$$\frac{(1 - 3 \times 5/9)^2}{3 \times 5/9} + \frac{(2 - 3 \times (9-5)/9)^2}{3 \times (9-5)/9}$$
$$+ \frac{(4 - (9-3) \times 5/9)^2}{(9-3) \times 5/9} + \frac{(2 - (9-3) \times (9-5)/9)^2}{(9-3) \times (9-5)/9}$$
$$= 0.267 + 0.333 + 0.133 + 0.167 = 0.900$$

Since 0.900 *is less than* 3.84, *we do not reject the independence assumption at the* 95% *confidence interval.*

The next example, also based on census data detailed in Section 5, helps to indicate how correlation 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] *This corresponds to items* i_2 *and* i_7. *Using the full census data, with* $n = 30370$, *we obtain the following contingency table:*

	i_2	$\overline{i_2}$	Σrow
i_7	17918	911	18829
$\overline{i_7}$	9111	2430	11541
Σ_{col}	27029	3341	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* $\overline{i_2} \Rightarrow \overline{i_7}$, $i_2 \Rightarrow i_7$, $\overline{i_7} \Rightarrow i_2$, *and* $i_7 \Rightarrow i_2$. *This allows for 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 statment — the one which the chi-squared test identified as dominant — in last place.*

[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 indicate actual use.

item	attribute	possible non-attribute values
i_0	drives alone	does not drive, carpools
i_1	male or less than 3 children	3 or more children
i_2	never served in the military	veteran
i_3	native speaker of English	not a native speaker
i_4	not a U.S. citizen	U.S. citizen
i_5	born in the U.S.	born abroad
i_6	married	single, divorced, widowed
i_7	no more than 40 years old	more than 40 years old
i_8	male	female
i_9	householder	dependent, boarder, renter

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

Table 1: I and B for a collection of census data. We formed I by arbitrarily collapsing a number of census questions into binary form. B actually has size 30370, 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 $O(i_1, i_2, i_3, \overline{i_4}, i_5, \overline{i_6}, i_7, \overline{i_8}, i_9) = 2$.

3.1 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 determine the cause of correlation. The statistical definition of dependence of two sets A and B is

$$\frac{P[A \wedge B]}{P[A]P[B]},$$

with the obvious extension to more than two sets. The further the value is from 1, the more the dependence. Note that dependence applies to a single cell of a contingency table, while correlation applies to the entire table.

In the context of contingency tables, we define the dependence of a cell r to be its **interest**, denoted $I(r)$. In contingency table notation, $I(r) = O(r)/E[r]$ (since $P[A]P[B] = E[AB]/n$ and $P[A \wedge B] = O(AB)/n$). The farther $I(r)$ is from 1, the higher the dependence of items in cell r. In fact, the r with the most extreme interest value is the one that contributes most to the χ^2 value. By construction, this cell maximizes $\left|\frac{O(r)}{E[r]} - 1\right| = \left|\frac{O(r)-E(r)}{E[r]}\right|$, and thus maximizes $\frac{(O(r)-E[r])^2}{E[r]}$. 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.

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

	i_2	$\overline{i_2}$
i_7	1.07	0.44
$\overline{i_7}$	0.89	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.*

3.2 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 involving correlation. 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.

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 correlation, 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.

3.3 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, statistics texts (such as Moore [22]) recommend 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 3 dimensions will have 343 million cells, and, as the sum of the expected cell values is only $n = 1$ million, not all cells can have expected value greater than 1.

The solution to this problem 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. In the meantime, we merely ignore cells with small expected value. We justify this with a support argument: If a set of items would have been correlated because of the contribution of the very small cell, then the correlation would involve very rare events. In many applications, an event that has expectation less than 1 can be ignored as uninteresting. See Section 4 for a discussion of combining χ^2 with support.

4 Algorithms for Correlation Rules

As we have mentioned, finding correlation 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 set of 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 mine for 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: A set of items S has 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 support downward closed. Note that values in the contingency table are observed values, not expected values.

One weakness of this support definition is that, unless p is larger than 50%, all items have 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. $p > 0.25$ 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 algorithms may be used, besides support. 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 correlations 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 [14], might be appropriate for this kind of pruning.

Combining the chi-squared correlation rule with pruning via support, we obtain the algorithm in Figure 1. We say that an itemset is *significant* if it is supported and minimally correlated. The key observation is that an itemset at level $i+1$ can be significant only if all its subsets at level i have support and none of its subsets at level i are correlated. Thus, for level $i+1$, all we need is a list of the supported but uncorrelated itemsets from level i. This list is held in NOTSIG. The list SIG, which holds the supported and correlated 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 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 support. If so, we add it to either SIG or NOTSIG for level $i+1$, depending on its χ^2 value.

The most expensive part of the algorithm is Step 8. We propose an implementation based on perfect hash tables (see [10, 7] for a description of the perfect hash function we used). 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, $O(n|\text{CAND}|) \in O(nk^i)$.

Algorithm χ^2-support
Input: A chi-squared significance level α, support s, support fraction $p > 0.25$. Basket data B.
Output: A set of minimal correlated 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 3.
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 set 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., correlated and supported) itemsets. It hinges on the fact that significant itemsets at level $i+1$ are supersets of supported but uncorrelated sets at level i. Step 8 can be implemented efficiently using hashing.

Overall, the running time for level i is $O(n \cdot |\text{CAND}| \cdot \min\{n, 2^i\} + i \cdot |\text{NOTSIG}|^2)$.

It is instructive to compare the algorithm in Figure 1 to the hash-based algorithm of Park, Chen, and Yu [24] 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 [24] 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. Our algorithm fails if we allow collisions, since we need hash table lookup; it is an open problem to modify our algorithm for very large databases.

5 Experimental Results

There is a wide range of problems for which correlation is the measure of interest and correlation 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 5.1), text data from newsgroups (Section 5.2), and synthetic data (Section 5.3). The first two are useful for illustrating the conceptual aspect of the correlation tests, and the last shows the effect of our pruning strategies on the performance of the algorithm.

Census data, such as that in Table 1, readily lends itself to correlation calculations. Since the chi-squared test extends easily to non-binary data, we can analyze correlations 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 correlations (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 correlations may also be useful, such as the discovery that a document consisting of recipes contains the word "fatty" less often than would be expected.

5.1 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 and are reproduced below for convenience. We show results for both the χ^2-interest test (Table 2) and the support-confidence test (Table 3).

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 2. Since so many pairs are correlated, we are struck by $\{i_1, i_4\}$ and $\{i_1, i_5\}$, which are not. We are even more surprised when we see that i_1 concerns number of children and i_4 and i_5 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 $\{i_4, i_8\}$ and $\{i_5, i_8\}$ to explore this. These are both significant, and the interest figures show there is indeed a dependency between

[4] A danger is that as the number of cells increases, problems with accuracy of the χ^2 statistic increase as well.

a b	χ^2	$I(ab)$	$I(\overline{a}b)$	$I(a\overline{b})$	$I(\overline{a}\overline{b})$
i_0 i_1	**37.15**	1.025	0.995	**0.773**	1.050
i_0 i_2	**244.47**	0.934	1.015	**1.554**	0.879
i_0 i_3	0.94	1.004	0.999	0.966	1.007
i_0 i_4	**4.57**	**0.901**	1.022	1.007	0.998
i_0 i_5	0.05	0.999	1.000	1.008	0.998
i_0 i_6	**737.18**	**1.574**	0.874	0.807	1.042
i_0 i_7	**153.11**	0.880	1.026	**1.192**	0.958
i_0 i_8	**138.13**	**1.155**	0.966	0.866	1.029
i_0 i_9	**746.28**	**1.404**	0.912	0.722	1.061
i_1 i_2	**296.55**	0.989	1.104	1.094	**0.135**
i_1 i_3	**24.00**	0.997	1.030	1.026	**0.759**
i_1 i_4	1.60	1.009	0.917	0.999	1.006
i_1 i_5	1.70	0.999	1.008	1.007	0.933
i_1 i_6	**352.31**	0.939	**1.562**	1.021	0.811
i_1 i_7	**2010.07**	1.067	0.385	0.892	**1.988**
i_1 i_8	**2855.73**	1.109	**0.000**	0.906	1.863
i_1 i_9	**229.07**	0.965	**1.317**	1.024	0.782
i_2 i_3	**82.02**	0.994	1.053	1.051	**0.576**
i_2 i_4	**190.71**	1.103	**0.140**	0.993	1.061
i_2 i_5	**176.05**	0.991	1.075	1.077	**0.355**
i_2 i_6	**993.31**	0.892	**1.901**	1.036	0.697
i_2 i_7	**2006.34**	1.070	0.414	0.887	**1.942**
i_2 i_8	**3099.38**	0.881	**1.994**	1.103	0.142
i_2 i_9	**819.90**	0.931	**1.573**	1.047	0.606
i_3 i_4	**9130.58**	0.271	**6.823**	1.052	0.588
i_3 i_5	**11119.28**	1.073	0.417	0.372	**6.016**
i_3 i_6	**110.31**	0.963	**1.294**	1.012	0.901
i_3 i_7	**62.22**	0.987	1.101	1.020	**0.838**
i_3 i_8	**21.41**	0.990	**1.081**	1.009	0.930
i_3 i_9	0.10	1.001	0.994	0.999	1.004
i_4 i_5	**18504.81**	0.000	1.071	**9.602**	0.391
i_4 i_6	**189.66**	**1.512**	0.964	0.828	1.012
i_4 i_7	**76.04**	1.148	0.989	**0.762**	1.017
i_4 i_8	**14.48**	**1.088**	0.994	0.924	1.005
i_4 i_9	3.27	0.953	1.003	1.032	0.998
i_5 i_6	**312.15**	0.940	**1.512**	1.020	0.827
i_5 i_7	**10.62**	0.995	1.043	1.008	**0.930**
i_5 i_8	**12.95**	0.992	**1.065**	1.007	0.944
i_5 i_9	2.50	0.996	1.032	1.003	0.978
i_6 i_7	**2913.05**	0.579	1.142	**1.677**	0.772
i_6 i_8	**66.49**	**1.087**	0.971	0.925	1.025
i_6 i_9	**186.28**	**1.163**	0.945	0.888	1.038
i_7 i_8	**98.63**	1.048	**0.922**	0.958	1.067
i_7 i_9	**4285.29**	0.643	**1.574**	1.246	0.605
i_8 i_9	**12.40**	**1.026**	0.977	0.982	1.016

item	attribute values	non-attribute values
i_0	drives alone	does not drive, carpools
i_1	male or less than 3 children	3 or more children
i_2	never served in the military	veteran
i_3	native speaker of English	not a native speaker
i_4	not a U.S. citizen	U.S. citizen
i_5	born in the U.S.	born abroad
i_6	married	single, divorced, widowed
i_7	no more than 40 years old	more than 40 years old
i_8	male	female
i_9	homeowner	dependent, boarder, renter

Table 2: We consider all possible pairs of census items for the χ^2-interest test. Bold χ^2 values are significant and the significance level is 95%. Bold interest values are the most extreme values; there is no bold interest value if χ^2 is not significant.

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 non-correlation we observed. A further jarring note for our explanation is the pair $\{i_1, i_3\}$. This pair includes native language, another marker of immigration. But $\{i_1, i_3\}$ *is* significant, which would lead us to believe immigration is correlated with family size. Furthermore, i_3 is just as dependent on i_8 (sex) as the other two markers of immigration. Perhaps, then, our assumption that i_3, i_4, and i_5 are good markers of immigration is flawed. Table 2 gives us much to mull on.

We invite the reader to attempt a similar analysis with the support-confidence data in Table 3. 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 i_0 and i_9 are correlated, and the 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 correlations, 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 i_6 and i_8, we have both the rules, "If you are married you are likely to be male" and "If you are male you are likely not to be married." These two statements are not inconsistent, but they are confusing; among other things, they seem to imply not very many people are married. What is more worrisome, every pair of items has the maximum four supported rules. Someone mining this data using support-confidence would conclude that all item pairs have all sorts of valid 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 i_1 and i_4, turn out not to be correlated.

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.

5.2 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 4. This left us with 416 distinct words.

One would expect words to be highly correlated, and

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**	$0.08p$	**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 3: We consider all possible pairs of census items for the support-confidence test. Bold values are significant. Support values are given in percents and the support cutoff is 1%. The confidence cutoff is 0.5. Note that confidence values are not bold unless the corresponding support value is significant.

indeed this turned out to be the case. Of the $\binom{416}{2} = 86320$ word pairings, there were 8329 correlated pairs, i.e., 10% of all word pairs are correlated. More than 10% of all triples of words are correlated. 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 correlated itemsets is presented in Table 4. We show not only the correlated words but the major dependence in the data. We see some obvious correlations: "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 correlated itemsets, so no subset of a triple is itself correlated. Thus "Burundi," "commission," and "plan" are 3-way correlated, though "commission" and "plan," say, are not. Since the major dependence has "commision" 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 correlated, 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 correlated 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 correlates.

5.3 Synthetic Data

The real data presented above suffers from a Goldilocks problem. The census data is too small, and its border too low, to study the effectiveness of the pruning techniques. 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 just right for exploring the effectiveness of our algorithm, we turn to synthetic data from IBM's Quest group [1].

We generated market basket data with 99997 baskets and 870 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 5.

Note that unlike with the text data, the number of correlations at level 3 is much smaller than the number of correlations at level 2. Though we do not show the numbers, it is again the case that the 3-way correlations have much lower χ^2 values than the average 2-way correlation, with no 3-way correlation 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.

6 Conclusions

We have introduced a generalization of association rules, called correlation 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. Correlation 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 correlation 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 correlated 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 correlation 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 area of research is in pruning criteria besides support. 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 correlated. It might be fruitful to explore the behavior of data sets where the border is exponential in the number of items.

Acknowledgements

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.

References

[1] R. Agrawal, A. Arning, T. Bollinger, M. Mehta, J. Shafer, and R. Srikant. The Quest Data Mining System. In *Proceed-*

correlated words	χ^2	major dependence includes	major dependence omits
area province	24.269	area province	
area secretary war	6.959	area war	secretary
area secretary they	7.127	area they	secretary
country men work	4.047	country men work	
deputy director	9.927	deputy director	
members minority	4.230	members minority	
authorities black official	4.366	authorities official	black
burundi commission plan	5.452	commission plan	burundi
african men nelson	5.935	african men nelson	
liberia west	48.939	liberia west	
mandela nelson	91.000	mandela nelson	
government is number	9.999	is number	government

Table 4: Some word correlations in the clari.world.africa news articles. Sometimes the correlations are suggestive, but not always; the last itemset above is one of the many confusing itemsets.

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

Table 5: 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 |SIG| is one measure of the effectiveness of correlation pruning considered by itself.

ings of the Second International Conference on Knowledge Discovery in Databases and Data, August 1996.

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

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

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo. Fast Discovery of Association Rules. In Fayyad et al [9], pages 307–328, 1996.

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, September 1994.

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

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

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

[9] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthrusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, CA, 1996.

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

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

[12] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama. Mining optimized association rules for numeric data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 13-24, 1996.

[13] J. Gray, A. Bosworth, A. Layman, and H. Pirahesh. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Microsoft Technical Report MSR–TR–95–22, 1995.

[14] D. Gunopulos, H. Mannila, and S. Saluja. Discovering all most specific sentences by randomized algorithms. In *Proceedings of the 6th International Conference on Database Theory*, to appear, January 1997.

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

[16] M. Houtsma and A. Swami. Set-oriented mining of association rules. In *Proceedings of the International Conference on Data Engineering*, pages 25–34, 1995.

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

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

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

[20] H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Efficient algorithms for discovering association rules. In *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, pages 144–155, July 1994.

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

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

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

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

[25] K. Pearson. 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, 1900.

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

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

[28] R. Srikant and R. Agrawal. Mining generalized association rules. In *Proceedings of the 21st International Conference on Very Large Data Bases*, pages 407–419, September 1995.

[29] H. Toivonen. Sampling large databases for finding association rules. In *Proceedings of the 22nd International Conference on Very Large Data Bases*, pages 134–145, September 1996.

[30] J. Widom. Research problems in data warehousing. In *Proceedings of the 4th Conference on Information and Knowledge Management*, November 1995.

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 [18] 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 [21] and Laplace [19] 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

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

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 [25] 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 $CT(\mathbf{r})$ in the table is a count of the number of trials, out of N independent trials, where the event $(X^1 = \mathbf{r}_1, \ldots, X^k = \mathbf{r}_k)$ occurs. We define the χ^2 value as

$$\chi^2 = \sum_{\mathbf{r}\in\{0,1\}^k} \frac{(CT(\mathbf{r}) - E[CT(\mathbf{r})])^2}{E[CT(\mathbf{r})]}$$

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 show that in the binomial case, the chi-squared statistic is upward closed.

Theorem 1 *In the binomial case, the chi-squared statistic is upward closed.*

Proof: The key observation in proving this is that no matter what k is, 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 item 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 events A, B, and C. The χ^2-statistic for the events A and B is defined as follows:

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

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, the expression for S_{AB} changes as follows:

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

Therefore, in $S_{ABC} - S_{AB}$, we have the terms $\frac{(x-X)^2}{x} + \frac{(y-Y)^2}{y} - \frac{(E-O)^2}{E}$ which, after some manipulation, simplifies to $\frac{(xY-yX)^2}{xy(x+y)}$, which is always positive. This implies S_{ABC} is always greater than S_{AB}.

$\square$