Algorithm description:

An algorithm is significant at level i+1 if it has support at i and is not correlated at i but passes the chi-squared test at i+1.

The computation of the chi-squared heuristic seems to take time 2^i

This can be fixed as follows:

$$\sum_{r \varepsilon R} \frac{(O(r) - E[r])^2}{E[r]} = \sum_r \frac{O(r)}{E[r]}(O(r) - E[r]) + \sum_r E[r]$$

The last term equals n and the first term is 0 if O(r) is 0. Thus we can compute the chi-squared value based only on the non-zero terms (and there can only be n of them).

Algorithm chi-squared support

input: a chi-squared significance level a, 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, count O(i)

2. initialize CAND <- 0, SIG <- 0, NOTSIG <- 0

3. For each pair i_a, i_b such that $O(i_a) > s$ and $O(i_b) > s$, add $\{i_a, i_b\}$ to CAND.

4. NOTSIG <- 0

5. If CAND is empty, return SIG, end

6. for each itemset in CAND, construct the contingency table, if less than p percent of the cells have count s, then go to 8

7. if the chi-squared value exceeds a threshold, then add itemset to SIG, else add it to NOTSIG.

8. continue to next itemset in CAND. If there are no more items in CAND, set CAND to be the set of all S such that S-1 itemsets are not significant. go to step 4.

Algorithms for correlation rules:

Finding correlations is identical to finding a border in the itemset lattice.

Borders can be very large, and therefore, we need to identify interesting itemsets in the border.

Use support to develop a level-wise algorithm that identifies interesting borders at i+1 itemsets from i itemsets.

The notion of support must be changes since we also take into account negation.

We redefine support of items S as s at p% level if atleast p% of the cells in the contingency table for S have the value s.

Other pruning techniques are anti-support.

Prune based on chi-squared values (prune off high values under the assumption that they are probably obvious).

Measures of interest:

The statistical definition of dependence between two sets A and B is

P(A and B) / (P(A) * P(B))

The dependence of the cell is defined to be its interest.

In contingency table notation,

I(r) = O(r) / E(r).

The farther this number is from 1, the higher the dependence, and interest.

Example:

The interest values for the previous example are:

	i2	i2'
i7	1.07	0.44
i7'	0.89	1.91

This indicates the interesting cells!

Example:

i2: never served in military
i7: no more than 40 years old.

	i2	i2'	
i7	17918	911	18829
i7'	9111	2430	11541
	27029	3341	30370

The chi-squared value is 2006.34

(independence hypothesis is discarded).

The biggest value comes from i2'i7'

(i.e. people serve in the military and people over 40 are strongly correlated.)

For comparison, if we run the support-confidence framework on this, we get

i2' => i7' i2 => i7
i7' => i2 i7 => i2
(at confidence of 0.5 and support of 1%) These rules make no sense.

Example:

	i8	i8'	SUM(row)
i9	1	2	3
i9'	4	2	6
SUM	5	4	9

E(i9) = O(i9) = 3,
E(i8) = O(i8) = 5

The Chi-squared value is

K(i8,i9) + K(i8,i9') + K(i8',i9) + K(i8'i9')

can be computed to be 0.9

Since 0.9 < 3.84 at the 95% confidence interval, we do not reject the independence hypothesis.

Chi-squared test

$E(i_j) = O(i_j)$ for a single item.

$E(i_j') = n - O(i_j)$

$E(r) = n * (E(r_1) / n) * (E(r_2) / n) * ... * (E(r_k)/n))$

Chi^2 = SUM_{r in R} (O(r) - E(r))^2 / E(r)

i.e. it is the normalized deviation from expectation.

We can restrict the range of r if we need to compute the value for a subset.

- calculate the value of the chi-squared statistic
- if all variables are independent, they would all be 0.
- if it is higher than a cutoff, we reject the independence assumption.

It can be shown that chi-squared test is upward closed!

The chi-squared test

Let R = {i1,i1'} X {i2,i2'} x ... x {ik,ik'}

and r = r1...rk in R.

R is the set of all basket values.

r is a single basket value.

Each value of r denotes a cell (in a k-dimensional space)

Let O(r) denote the number of baskets falling in cell r.

To test if a given cell is dependent, we must determine if the actual count in r differs sufficiently from the expectation.

In the chi-squared test, expectation is calculated under the assumption of independence.

The border of correlation:

Example (implication does not have a border):

d	c	c'	sum	d'	c	c'	sum
t	8	1	9	t	10	2	12
t'	40	2	42	t'	35	2	37
sum	48	3	51	sum	45	4	49

P(c AND d) = 48, P(c) = 93; therefore c => d has a confidence of 0.52

on the other hand P(c AND d AND t) = 8 and P(t AND c) = 18 therefore

c,t => d has a confidence of 0.44.

For a confidence cutoff of 0.5,

c => d has the confidence, but c,t => d does not.

Border properties are extremely useful for developing fast algorithms.

The closure property:

Upward closure is a constructive procedure.

We start with the assumption that no (i+1) itemset is correlated.
Finding an i itemset that is correlated, we can determine all i+1 itemsets that are correlated.

Unlike in the former where we generate false positives, here we generate
false negatives (ignored i+1 itemsets)

Because of this, upward closure is most useful when we are looking for an
unwanted property.

Then we are looking to prune some (i+1) itemset; and all that happens if
we miss some correlated itemsets is that our pruning is less efficient.

Thus we look for minimal correlated itemsets.

Correlation Rules:

Let P(A) be the probability of event A and P'(A) = 1 - P(A)
(the probability that A does not occur)

P(AB) = probability that both A and B occur
P(A'B) = probability that A does not occur but B occurs.

Events A and B are independent if P(AB) = P(A)*P(B)

If A and B are not independent, then they are said to be dependent.

A and B are said to be correlated if any of P(AB), P(A'B), P(AB') or
P(A'B') are dependent.

If we have n trials, we denote the number of occurances as O(A)

P(A) = O(A) / n

In the context of itemsets, {ia1, ia2, ..., iam} is a correlation rule if the occurances are correlated.

An important property of correlations is that it is upward closed.

If a set of items S is correlated, so is every superset of S.
(i.e. by adding items to a set, you cannot get rid of the correlation!)