The Share, Coincidence and Dominance Measures for Itemsets: Extended Report

Colin L. Carter, Howard J. Hamilton and Nick Cercone

Technical Report CS-97-01
February, 1997

Regina, Saskatchewan, CANADA
S4S 0A2

ISSN 0828-3494
ISBN 0-7731-0339-2

The Share, Coincidence and Dominance Measures for Itemsets: Extended Report

Colin L. Carter, Howard J. Hamilton and Nick Cercone
Department of Computer Science
University of Regina, Regina, SK, Canada, S4S 0A2
{carter, hamilton, nick}@cs.uregina.ca

Abstract

We introduce the measures *share*, *coincidence* and *dominance* as alternatives to the standard itemset methodology measure of support. We also redefine the confidence measure in this context. An itemset is a group of items bought together in a transaction. The support of an itemset is the ratio of transactions in which an itemset appears to the total number of transactions. The share of an itemset specifies the ratio of the count of items purchased together to the total count of items purchased in all transactions. The coincidence of an itemset is the ratio of the count of items in that itemset to the total of those same items in the database. The dominance of an item in an itemset specifies the extent to which that item dominates the total of all items in the itemset.

Measures based on share have the advantage of reflecting accurately how many units are being moved by a business, a capability that current itemset methodology does not provide. In addition, the share can be extended to give an accurate picture of the financial impact of an itemset on the business bottom line.

1. Introduction

The existence of large amounts of scan code data collected by many businesses represents a potential wealth of information given adequate methods of transforming the data into meaningful information. One class of such data is stored in transaction databases from which all items obtained in a single transaction can be retrieved as a unit. The transactions can then be examined to determine what items customers typically buy together. This in turn gives insight into questions such as how to market these products more effectively, how to group them in store layout or product packages, or which items to offer on sale to boost the sale of other items.

Recent research has focused on determining which groups of items, called *itemsets*, are frequently bought together. From any itemset an *association rule* may be derived which, given the

purchase of a subset of the items in the itemset, predicts the probability of the purchase of the remaining items [1], [2], [4], [6]. Several algorithms have been proposed for finding generalized itemsets from items that are classified by one or more taxonomic hierarchies [3],[8]. The interestingness of the discovered rules and some methods for pruning uninteresting rules have been addressed in [5] and [8].

Data managers are typically interested only in itemsets which are bought in sufficient numbers to form a substantial portion of the business income. This portion is somewhat reflected in the *support* of an itemset which is the ratio of the number of transactions in which the itemset was purchased to the total number of transactions in the database. An association rule of the form $A \rightarrow B$, where A and B are itemsets, is associated with a *confidence* measure which is the ratio of the support of the itemset $A \cup B$ to the support of the itemset A. The confidence quantifies the probability that when A is bought, B will also be bought.

The current definition of the support of an itemset is somewhat limited in its informative feedback. The support of an itemset tells only the number of transactions in which that itemset was purchased. The number of items purchased is unknown and the precise impact of the purchase of that itemset cannot be measured in terms of stock, cost or profit.

In this paper we present several alternative measures that address the need for more informative feedback from an itemset task. Since items are often purchased in multiples, we assume a count is associated with each item in a transaction. The *share* of an itemset is the ratio of the total count of items in the itemset to the total count of all items in the database. The *coincidence* of an itemset is the proportion of the count of the items in the itemset to the total count of the same items in the whole database. Since these measures recognize the possibility of one item being more or less frequent in an itemset than others, the *dominance* of an item in an itemset is a measure that quantifies the total of one item relative to other items in the itemset. Using an example from commercial data, we show that the share measure may give a different view of the relative importance of an itemset than that implied by support.

In Section 2 of this paper, we review the definitions of itemsets, association rules and the measures of support and confidence. In Section 3, we identify some limitations of the support and confidence measures. In Section 4, we present the share, coincidence and dominance measures and modified confidence and frequency measures. In Section 5 we extend the share measure

beyond counts to financial measures and present an example application of these measures to a commercial database. We conclude in Section 6.

2. Itemsets, Support and Confidence

We now summarize itemset methodology formally as follows (adapted from [2]). Let $I = \{I_1, I_2, ..., I_m\}$ be a set of literals, called *items*. Let $T = \{T_1, T_2, \ldots, T_n\}$ be a set (database) of transactions, where each transaction $T_q \in T$, $T_q \subseteq I$, is a set of items purchased together. A set of items $X \subseteq T_q$ is called an *itemset*. A transaction T_q *contains* an itemset X if $X \subseteq T_q$. An *association rule* is an implication of the form $X \rightarrow Y$, where X and Y are itemsets, and $X \cap Y = \varnothing$. Each itemset X is associated with a set of transactions $T_X = \{T_q \in T \mid T_q \supseteq X\}$ which is the set of transactions which contain the itemset X. The support of an itemset X and any association rule derived from X is $sup(X) = |T_X| / |T|$, which is the ratio of the number of transactions which contain X to the number of transactions in the database. The confidence of the association rule $X \rightarrow Y$ derived from itemset $Z = X \cup Y$ is $conf(X \rightarrow Y) = sup(X) / sup(Z)$ and represents the percentage of transactions in T containing X which also contain Y. The user specifies a minimum level of support to separate interesting itemsets from itemsets that do not have enough support to be considered interesting. Itemsets with support above the minimum support are called *frequent itemsets*. The user also specifies a minimum level of confidence which separates interesting association rules from rules that do not have enough predictive strength to be considered interesting.

The determination of itemsets and association rules is usually handled as two consecutive steps. First all frequent itemsets are extracted from the database, and then association rules are generated from these itemsets.

We present a small example database in Table 1 to illustrate the support and confidence measures. There are five transactions in the database, so the support of each itemset is measured relative to 5.

Table 2 shows the supports of the 6 items in the database. The first column lists the item, the second lists the number of transactions in which the item appears, and the third column lists the support of the item. Item A in the first row, for example, appears in 2 transactions in Table 1 (transactions T_1 and T_3). This represents 40% of the 5 total transactions.

Transaction ID	Items
T_1	A, C, D
T_2	B, E
T_3	A, B, C, E
T_4	B, E
T_5	B, D, F

Table 1. A transaction database with 5 transactions

Item	Number of transactions	Support *sup(X)*
A	2	40%
B	4	80%
C	2	40%
D	2	40%
E	3	60%
F	1	20%

Table 2. Single item support

Itemset	Number of Transac-tions	Support *sup(X)*
A, C	2	40%
A, B	1	20%
B, D	1	20%
C, D	1	20%
A, B, C	1	20%
A, B, E	1	20%
A, C, D	1	20%
B, D, F	1	20%
A, B, C, E	1	20%

Table 3. Some itemsets and their supports

Association Rule	Confidence *conf(X→Y)*
A→C	100%
A→B	50%
B→D	25%
A, B→C	100%
A, C→B	50%
B, E→A	33%

Table 4. Some association rules and their confidences

Table 3 shows the support for some itemsets derived from the database. The columns are analogous to those in Table 2. For example, the itemset {A, B} in the second row appears in only one transaction, transaction T_3, which gives it a support of 20%.

Table 4 shows the confidence measures of several association rules derived from the itemsets in Table 3. The confidence of 100% for the rule A→C means that in every transaction in which A appears, C also appears. The confidence of this rule can be calculated by dividing the number of transactions in which the itemset {A, C} appears, which is 2 (see Table 3), by the number of transactions in which the item A appears, also 2 (Table 2).

3. Limitations of the Support and Confidence Measures

The *support* measure has been used as a foundational concept for determining interesting itemsets. The concept is justified by the intuitive reasonability of considering only itemsets that are purchased in greater than some predetermined percentage of transactions. The support meas-

ure also gives a stable way of comparing itemsets since the support of an itemset is relative to the number of transactions being examined, and this value does not change. However, the notion of support has some limitations based primarily on the assumption that the number of items purchased is either irrelevant or always one.

In a market analysis, it is not only interesting to know the percentage of transactions in which a product is purchased, but it is also important to know how many items in total were purchased. The support measure does not reflect this. For example, some items in a grocery store, such as frozen concentrated juice or carbonated beverages, are typically bought in multiples. Given a minimum support of 4%, one of these items may only have 3% support and therefore be classed as infrequent. In actuality, however, the number of these items purchased might be twice as great as some other item usually purchased singly, but which occurs in 4% of the transactions and is therefore considered interesting.

The support measure also does not allow for accurate financial calculations or comparisons. Masand and Piatetsky-Shapiro [6] note that for target marketing, measures should take into account both the frequency of an item contributing to a predictive rule and the value of the items in the prediction. The support measure allows for neither of these, so measures based on specific numbers of items, such as percentage of gross sales, costs or net profit, cannot be calculated, and business payoff cannot be maximized. For example, an item with 2% support that profits the business 15¢ per purchase is not as interesting as an item with 1.5% support and a profit of 25¢ per purchase, unless the 15¢ profit item is normally purchased in multiples. While the support measure might be used in conjunction with item cost or profit to concentrate on higher valued items, it fails where the number of items plays a significant role in determining the value of the relevant sales.

One might suggest that simply adding some feature to track the number of items in an itemset might be sufficient to alleviate these concerns. However, a new measure of support that simply sums the number of items purchased and divides by the number of transactions will not suffice, because this could lead to supports greater than 100%. For example, in a specialty food store such as a bakery, if 75% of customers bought bread, and two thirds of these (50% of all customers) bought two loaves, then the support of bread under this simplistic extension of the current

definition would be 125% (50% buying 2 loaves and 25% buying one loaf), which lacks intuitive reasonableness.

Any new measure should maintain the properties that the support measure has for items purchased singly, and in addition should increase the information that that can be obtained when items are purchased in multiples. The utility of the increased information should justify the additional computation. The properties that we suggest a new measure should have are:

- The measure should be intuitively reasonable and understandable.
- The measure should be relative to some stable baseline such as the transaction count.
- The measure should take into account the number of items purchased by customers.
- The measure should offer broader capabilities relevant to financial performance than support does.

4. New Measures

4.1 Share

A reasonably intuitive measure which takes into account multiple items is the ratio of the total count of items in an itemset to the total count of all items in the transaction database. We call this measure the *share* of an itemset.

We expand the formalism presented in Section 2. Let $I = \{I_1, I_2, ..., I_m\}$ be a set of m items and $T = \{T_1, T_2, ..., T_n\}$ be a set (database) of n transactions, $T_q \subseteq I$, $T_q \in T$. Let the *transaction count* of item $I_p \in I$ in transaction $T_q \in T$ be $c(I_p,T_q)$, which is the number of item I_p purchased in transaction T_q. Each item I_p has an associated set of transactions $T_{I_p} = \{T_q \in T \mid I_p \in T_q\}$ which are all transactions containing item I_p. Let the *global count* of item I_p in the database T be $C(I_p) = \sum_{T_q \in T_{I_p}} c(I_p, T_q)$, which is the sum of the counts of item I_p in every transaction of the database in which I_p appears. Let the *total item count* of all items in I in the transaction set T be $C = \sum_{p=1}^{m} C(I_p)$ which is the sum of the global counts of the individual items in I, or in other words, the total number of items in the database.

We illustrate these measures using the database shown in Table 5. This is the same database as shown in Table 1 except that each item has a transaction count. For this database, the set of

items I = {A, B, C, D, E, F}, m = 6, and the set of transactions $T = \{T_1, T_2, T_3, T_4, T_5\}$, n = 5. Table 6 shows the transactions counts $c(I_p,T_q)$ for each item. From Table 6 we see that the set of transactions associated with the item A are $T_A = \{T_1, T_3\}$. The global count for item A is:

$$C(A) = \sum_{T_q \in T_A} c(A, T_q) = c(A,T_1) + c(A,T_3) = 1+2 = 3.$$

The remaining global counts for items B to F are listed in the rightmost column of Table 6. The total item count for the database is:

$$C = \sum_{p=1}^{m} C(I_p) = C(A) + C(B) + \ldots + C(F) = 3+4+4+3+5+1 = 20.$$

A *k-itemset* is a set $X = \{x_1, x_2, \ldots, x_k\}$, $X \subseteq I$, $2 \le k \le m$, of k distinct items that are purchased together in one or more transactions in the database. Each itemset X has an associated set of transactions $T_X = \{T_q \in T \mid T_q \supseteq X\}$ which is the set of transactions that contain the itemset X. Let the *local count of an item x_i in the itemset X* be $c(x_i, X) = \sum_{T_q \in T_X} c(x_i, T_q)$, which is the sum of the transaction counts of the item x_i in all transactions which contain the itemset X. An item's local count will always be less than or equal to the global count $C(x_i)$ for the same item since the global count represents the sum of transaction counts of x_i in every transaction in which x_i individually occurs, whether or not the complete itemset occurs in each of these transactions. A single item

Transaction ID	Item	Item count
T_1	A	1
	C	2
	D	2
T_2	B	1
	E	3
T_3	A	2
	B	1
	C	2
	E	1
T_4	B	1
	E	1
T_5	B	1
	D	1
	F	1

Table 5. A transaction database with counts

Item	Transactions counts $c(I_p,T_q)$					Item global counts $c(I_p)$
	T_1	T_2	T_3	T_4	T_5	
A	1	0	2	0	0	3
B	0	1	1	1	1	4
C	2	0	2	0	0	4
D	2	0	0	0	1	3
E	0	3	1	1	0	5
F	0	0	0	0	1	1

Table 6. Transaction counts and global counts for each item

will have a separate local count for each itemset in which the item appears. Thus for some item I_q, its local count $c(I_q,X)$ in the itemset X will be different from the local count $c(I_q,Z)$ of the same item in the itemset Z, $Z \neq X$.

Let the *local count of a k-itemset X* be $c(X) = \sum_{i=1}^{k} c(x_i, X)$, which is the sum of the local counts of each item in the itemset. Let the *global count of the k-itemset X* be $C(X) = \sum_{i=1}^{k} C(x_i)$ which is the sum of the global counts of each item in the itemset.

Table 7 lists several 2-itemsets from the example database, the transactions which contain each itemset, the local counts of each item and the local count of each itemset. If the itemset X = {A,C} (row 1 in the table), then X is associated with the set of transactions $T_X = \{T_1, T_3\}$ (see Table 5). The local count of item A in the itemset X is $c(A,X) = \sum_{T_q \in T_X} c(A, T_q) = c(A,T_1) + c(A,T_3) =$ 1+2 = 3 as shown in column 3. The local count of the item C in the itemset X is $c(C,X) = \sum_{T_q \in T_X} c(C, T_q) = c(C,T_1) + c(C,T_3) = 2+2 = 4$ (column 4). The local count of the itemset X is therefore $c(X) = \sum_{i=1}^{2} c(x_i, X) = 3+4 = 7$ as shown in column 5 of Table 7.

The support of an itemset X, $sup(X)$, can now be restated as $sup(X) = |T_X|/m$, which is the ratio of the number of transactions in which the itemset X appears to the number of transactions in the database ($m = |T|$).

Definition 1. The share of an item x_i in the itemset X: $share(x_i, X) = c(x_i, X)/C$

The share of the item x_i in the itemset X is the ratio of the local count of the item x_i in X to the total item count. Table 8 lists the same 2-itemsets that are in Table 7, repeats the local counts of

Itemset	Containing transactions	First item's local count	Second item's local count	Itemset local count
X	T_X	$c(x_1,X)$	$c(x_2,X)$	$c(X)$
A, C	**$\{T_1, T_3\}$**	**3**	**4**	**7**
A, B	**$\{T_3\}$**	2	1	3
B, D	**$\{T_4\}$**	1	1	2
C, D	**$\{T_1\}$**	2	2	4

Table 7. Some 2-itemsets and their local counts

each item and itemset, and shows the shares of each item in the itemset. For the sample database, the total item count in the database is $C = 20$. The item *A* in the itemset {A, C} (first row of Table 8) has a local count of 3 (see column 2) and therefore a share of 3/20 = 15% (column 5).

Definition 2. The share of an itemset *X*: $share(X) = c(X)/C$

The share of the itemset *X* is the ratio of the local count of the itemset *X* to the total item count, or in other words, the ratio of the count of items in the itemset to the total count of items in the database. (This could be calculated as a marginal frequency.) In Table 8, the itemset {A, C} has a local count of 7 (column 4) and a share of 7/20 = 35% (column 7). Since a single item can be considered as a trivial itemset of size 1, the share of an individual item I_p in the itemset $\{I_p\}$ is $share(\{I_p\}) = c(\{I_p\})/C = C(I_p)/C$. That is, the share of a single item in the database is the ratio of the global count of the item to the total number of items in the database. For simplicity, we denote this as $share(I_p)$. The shares of the individual items in the sample database are shown in Table 10.

Definition 3. The average share of the k-itemset *X*: $share_{avg}(X) = share(X)/k$

The average item share in the k-itemset *X* is the local share of the itemset divided by the number of items in the itemset. The 2-itemset {A, C} has a share of 35% (column 7 of Table 8) and an average share of 35% / 2 = 17.5% (column 8). Since the share of an itemset may increase with the addition of new items to the itemset, the average item share is useful for comparing itemsets of different sizes.

An association rule is an implication of the form $X \rightarrow Y$ derived from an itemset *Z*, where *X* and *Y* are itemsets, $X \cup Y = Z$, and $X \cap Y = \varnothing$. The support of the association rule $X \rightarrow Y$ in the transaction set *T* is $sup(Z)$, the support of the itemset *Z*. The share of the association rule $X \rightarrow Y$

Itemset X	First item's local count $c(x_1,X)$	Second item's local count $c(x_2,X)$	Itemset local count $c(X)$	First item's share $share(x_1,X)$	Second item's share $share(x_2X)$	Itemset share $share(X)$	Average share $share_{avg}(X)$
A, C	**3**	**4**	**7**	**15%**	**20%**	**35%**	**17.5%**
A, B	2	1	3	10%	5%	15%	7.5%
B, D	1	1	2	5%	5%	10%	5%
C, D	2	2	4	10%	10%	20%	10%

Table 8. Some 2-itemsets, their counts and shares

Item I_p	Number of transactions $\|T_{Ip}\|$	Support $\|T_{Ip}\| / 5$	Global count $C(I_p)$	Share $share(I_p) = C(I_p) / 20$
A	2	40%	3	15%
B	4	80%	4	20%
C	2	40%	4	20%
D	2	40%	3	15%
E	3	60%	5	25%
F	1	20%	1	5%

Table 10. Counts, supports and shares for single items

in the transaction set *T* is *share(Z)*, the share of the itemset *Z.*

In Table 10 we compare the support and shares of the items in the sample database. Column 2 lists the number of transactions in which each item appears, and Column 3 lists the support derived by dividing the value from Column 2 by the total number of transactions (5). The number of transactions and support values are the same as those in Table 2. Column 4 lists the global count of each item, and the share of the item in Column 5 is calculated by dividing the global count of the item by the total item count $C = 20$. There is no local count for single items since local counts refer only to counts within an itemset, and itemsets are of size 2 and larger. From Table 10, item B has the greatest support of 80% and E has 60% support. However, from the share column, item E in fact represents 25% of the total items in the database, while B represents only 20% of the items. If items B and E were equally profitable, then item B would be erroneously identified by support as the more significant item.

Itemset	Number of transactions	Support	Itemset local count	Share	Average Share
X	$\|T_X\|$	$supp(X) = \|T_X\| / 5$	$c(X)$	$share(X) = c(X) / 20$	$share_{avg}(X) = share(X) / k$
A, C	2	40%	7	35%	17.5%
A, B	1	20%	3	15%	7.5%
B, D	1	20%	2	10%	5.0%
C, D	1	20%	4	20%	10.0%
A, B, C	1	20%	5	25%	8.3%
A, B, E	1	20%	4	20%	6.7%
A, C, D	1	20%	5	25%	8.3%
B, D, F	1	20%	3	15%	5.0%
A, B, C, E	1	20%	6	30%	7.5%

Table 9. Support and shares for various itemsets in the database

Table 9 shows the support and share values for various itemsets in the database. Not all possible itemsets are listed. The *Number of transactions* and *Support* columns are the same as in Table 10. All itemsets but one have a support of 20%, but the shares of the itemsets range from 10% to 30%. Decision makers would not know which of the itemsets with equal support should be focused on, but a decision based on share would be much clearer.

The average item share for each itemset is given in the rightmost column of Table 9. This measure is useful for determining the relative interest of itemsets of different sizes. For example, the itemsets {A, B} and {B, D, F} both have a share of 15%. Which is more significant? The itemset {A, B} has an average item share of 7.5%, and the items {B, D, F} have an average item share of 5%. In general, then, the items A and B in {A, B} represent a larger proportion of the total items in the database than the items B, D and F in {B, D, F}. The balance of these two factors will need to be weighed by each individual data manager to decide which itemset is more significant.

4.2 Coincidence

Definition 4. The coincidence of an itemset *X*: $coinc(X) = c(X)/C(X)$

The *coincidence* of an itemset *X* is the ratio of the local count of the itemset to the global count of the itemset. Intuitively, a coincidence of 60% for an itemset states that 60% of all the items in that itemset are bought together. Table 11 shows some 2-itemsets from the sample database, their local and global counts, and their coincidences. The local counts are from Table 7 and the global counts from Table 6. The 100% coincidence for the itemset {A, C} indicates that all occurrences of A were bought in transactions that also had C and vice versa. The next most coincidental pair is {C, D} which has a coincidence of 4/7 or 57%, which states that out of all the

Itemset	Local counts			Global counts			Coincidence
	1st item	2nd item	Itemset	1st item	2nd item	Itemset	*coinc(X)*=
X	$c(x_1,X)$	$c(x_2,X)$	$c(X)$	$C(x_1)$	$C(x_2)$	$C(X)$	$c(X) / C(X)$
A, C	3	4	7	3	4	7	100%
A, B	2	1	3	3	4	7	43%
B, D	1	1	2	4	3	7	29%
C, D	2	2	4	4	3	7	57%

Table 11. Some 2-itemsets and their coincidences

Itemset *X*	Local count *c(X)*	Global count *C(X)*	Coincidence *c(X) / C(X)*
A, B, C	5	11	45%
A, B, E	4	12	33%
A, C, D	5	10	50%
B, D, F	3	8	38%
A, B, C, E	6	16	38%

Table 12. Coincidence measures for itemsets in Table 3.

items C and D sold, 57% of these were sold in conjunction with each other. Table 12 shows the local and global counts of several larger itemsets and their coincidences.

While the share measures an itemset's importance relative to all items sold, the coincidence measures the importance of an itemset in terms of all items of that itemset sold. Two itemsets may have the same share, but a higher coincidence in one of them indicates a stronger relationship between its items. For example, the itemsets {A, C, D} and {A, B, C} (Table 12) both have a share of 25%, but the coincidence of {A, C, D} is 50% and of {A, B, C} is 45%. Therefore, relatively more of the first set appear together than the second set, making the first set a more likely target for marketing. On the other hand, itemset {A, C} has a share of 35% (see Table 9) and a coincidence of 100%. If market researcher were looking to boost the coincidence of products sold together, there is no sense in promoting the pair {A, C} since all of them are already sold together.

4.3 Dominance

In an itemset, the item local counts may not contribute equally to the itemset count. For example, for the itemset {B, D} in Table 11, the local counts for B and D are the both 1. However, the local counts of A and C in {A, C} are different, being 3 and 4 respectively. We measure the proportion of the item's local count to the itemset count with a *dominance* measure.

Several variations of the dominance measure are possible. The dominance could represent the simple percentage of the local count of the item relative to the local count of the itemset. For example, for the itemset {A, B} in Table 11, the item A with a local count of 2 represents 66% of the itemset's local count of 3. However, this measure would always have to be compared to the size of the itemset to be fully appreciated.

Alternatively, we can normalize the dominance of an item by multiplying the simple percentage of the local count of an item by the number of items in the itemset.

Definition 5. The locally weighted dominance of an item x_i in the k-itemset *X:*

$$ldom(x_i, X) = c(x_i, X)/c(X) * k$$

The locally weighted dominance of item x_i in the k-itemset *X* normalizes the item proportions according to the assumption of uniform distribution of items in the itemset. In a k-itemset *X* with local count *c(X)*, therefore, items with a count of *c(X)/k* would have a locally weighted dominance of 1.0 regardless of the value of *k*. Items dominating the count of an itemset of size *k* would have a locally weighted dominance approaching *k*. For example, the items in the itemset {B, D, F} each have local counts of 1, and therefore each item has a locally weighted dominance of 1.0. However, in the itemset {A, C}, which has an itemset count of 7, the item A has a local count of 3 and C has a local count of 4 (Table 11). The locally weighted dominance of A would be 3/7 * 2 = 0.86, and C would be 4/7 * 2 = 1.14. Items with a locally weighted dominance less than 1.0 have a lower than average count, and those with greater than 1.0 have a greater than average count.

Since we have the global count of each item in the database, however, we might assume that the ratios between items in itemsets might normally be proportional to the ratio of same items in the database. We can therefore normalize the dominance of an item in accordance with its ratio to the set of items globally.

Definition 6. The globally weighted dominance of an item x_i in the k-itemset *X:*

$$gdom(x_i,X) = \frac{c(x_i, X)/c(X)}{C(x_i)/C(X)}$$

Items that match the global ratio precisely would have a globally weighted dominance of 1.0. This is the case with the itemset {A, C} in Table 11 since the local counts of A and C (3 and 4) are the same as the global counts of the items. However, for the itemset {A, B}, items A and B have local counts of 2 and 1 respectively and global counts of 3 and 4. The globally weighted dominance of the item A in the itemset {A, B} is (2/3) / (3/7) = 1.55. The globally weighted dominance of B is (1/3) / (4/7) = 0.58. Intuitively, in this itemset, one would expected more B's than A's since there are more B's than A's globally. However, there are more A's. The item A, therefore, receives a globally weighted dominance greater than 1.0 and B receives less than 1.0.

The higher the global weighted dominance of a given item, the more unusual its count is relative to global proportions. Itemset tasks can therefore more easily detect and quantify deviations from usual distributions. These deviations are generally considered as more interesting than patterns conforming to the norm. This capability cannot be provided using support as a measure.

4.4 Share-based Confidence

In standard itemset methodology, the confidence of an association rule $X \rightarrow Y$ is the strength of prediction the antecedent X has in reference to the consequent Y. The higher the confidence, the more likely the items in the consequent Y will be purchased when the items in the antecedent X are also purchased. This same intuition can be carried into itemsets where counts of the items are available.

Definition 7. The share-based confidence of the association rule $X \rightarrow Y$ derived from an itemset $Z = \{z_1, z_2, \ldots, z_k\} = X \cup Y$: $conf_{share}(X \rightarrow Y) = \sum_{z_i \in X} c(z_i, Z) \Big/ c(X)$,

The share-based confidence of the association rule $X \rightarrow Y$ is the ratio of the sum of the local counts of the items comprising X in Z to the local count of the itemset X. Note that the local counts $c(z_i,Z)$ of the items common to both X and Z are counts of the items z_i in the context of the itemset Z, while the local count $c(X)$ of the itemset X is the sum of the counts of the same items in the context of the itemset X.

For example, from Table 9, consider the itemsets X = {A, C} with local item counts of 3 and 4 respectively and a local itemset count of 7, and Z = {A, C, D} with local item counts 1, 2 and 1 respectively. The sum of the local counts of items A and C in Z is 3, and the local count of the itemset X is 7. The share-based confidence of the association rule {A, C} $\rightarrow$ {D} is therefore 3/7 = 43%. In contrast, the support-based confidence is 50% because {A, C, D} has a support of 1, and {A, C} has a support of 2, which gives a confidence of 1/2.

Consider also the itemset Z = {A, B, C} with item local counts of 2, 1 and 2. Again the itemset {A, C} has a local count of 7, while the local counts of items A and C in the itemset Z have a sum of 4. The association rule {A, C} $\rightarrow$ {B} has a share-based confidence of 4/7 = 57% and a

support-based confidence of 50%. The share-based confidence values can therefore be more or less than those based on support.

One might argue that the confidence measure does not seem to rely on the local count of the consequent *Y*, and therefore the measure is not valid. However, the presence of the consequent *Y* in the itemset $Z = X \cup Y$ limits the values of the local counts of the items in the antecedent *X* in comparison to the unrestricted itemset *X*. The result of the confidence measure, therefore, is the ratio of those items appearing with *Z* to the same items appearing with or without *Z*. This is the same as confidence based on support except that the counts of items are taken into account.

The advantage of a count based confidence is increased precision. Confidence values based on support for the preceding two examples were both 50% and allowed no differentiation between the two. Confidence values based on share were 43% and 57%, which would allow a data manager to more precisely rank the value of the two association rules.

4.5 Frequent Itemsets

In standard itemset methodology, an itemset is only considered interesting if it is a *frequent* itemset, that is, if it falls above a user specified minimum support value. An association rule is considered interesting if its confidence falls above a second threshold, the minimum confidence. When using the share as the base measure of an itemset, we continue to use the minimum confidence for association rules. However, an itemset's frequency must be redefined in terms of a minimum share threshold.

The minimum support specifies the minimum percentage of transactions in which an itemset must be contained to be considered frequent and therefore interesting [1]. As the itemset size grows, the support of the itemset never increases and usually decreases. All subsets of a frequent itemset are also guaranteed to be frequent [1],[2]. In fact, a (k+1)-itemset is only generated as a potentially frequent itemset if all of its k-itemset subsets were frequent in the previous round of processing [2].

When using the share measure, however, as the size of an itemset increases, the share of the itemset may also grow. If minimum share is measured against the share of the itemset, then it would be possible to have an itemset above the minimum share whose component itemsets may not fall above the minimum share. For example, given a minimum itemset share of 25%, the itemset {A, B, C, E} in Table 9 would be considered frequent since its share is 30%. One com-

ponent itemset, {A, B, E}, however, has a share of only 20%. With this definition of minimum share, {A, B, E} would be eliminated from the frequent itemsets after the third round of processing, and {A, B, C, E} would not be generated as a candidate for the fourth round. Defining minimum support as referring to the total itemset share, therefore, would necessitate using infrequent itemsets to generate candidate itemsets for the next round of processing. This would lead to an undesirable decrease in the performance of itemset algorithms.

Instead, we define the minimum share to refer to each item in an itemset rather than the itemset as a whole. We consider an itemset X to be *frequent* if $\forall x \in X$, $share(x) \geq minshare$, a user defined minimum share value. Intuitively, we are only interested in items that are individually bought frequently and also bought frequently with other frequent items. Using this definition of frequency, the individual shares of the items will follow a similar pattern to the support measure, never increasing and generally decreasing with increased itemset size. Only those itemsets, therefore, that include relatively significant items will be considered interesting. For example, given a minimum share of 10% (a count of 2) in our example database, only itemsets {A, C} and {C, D} of the 2-itemsets listed in Table 9 would be frequent since their component items each have a share of 10% or more.

5. Extensions of Share Based Measures

The primary distinction of share based measures is that these measures take into account the numbers of items bought as opposed to simply the number of transactions in which the items appear. This may result in certain items that are typically bought in multiples being increased in importance relative to other items that have higher support but a lower share. A related capability of share based measures is that they can be extended to include other data, particularly financial figures, to give a more informative feedback about the relative importance of various items. For example, assume a transaction table has the following simplified schema:

transactions(transaction_id, item, item_count, item_profit)

The transaction_id, item and item_count fields are all as previously described. The item_profit represents the net profit generated by each item sold. An itemset task which takes into account the amount of money generated will rate more profitable items higher than less profitable ones and give data managers a more realistic view of what products are driving the business.

To test this intuition, we ran an itemset task on commercial data from one of our corporate sponsors. We implemented the Apriori algorithm [1],[2] with both support and two share based measures. The share measures reflected both the number of items sold and the gross income generated from these. Net profit figures would have been more useful but were not available in our data set. The database included approximately 3.3 million records representing close to half a million customer accounts and 2200 items. Each record represented a piece of equipment rented or service subscribed to by a customer, the number of items and the cost of each item. Since financial information was relevant to this task, only those items with an associated cost were included in the data retrieval, eliminating about half of the input records. A minimum support and share of 0.25% was used to determine frequent items and itemsets. An itemset was considered as frequent if it was above the minimum support, if the share based on count of each component item was above 0.25% of the total number of items in the retrieval, or if each item's financial value was above 0.25% of the total income represented in the database.

We present three figures (Figure 1 - Figure 3) which rank the top 20 individual items in terms of support, share of count and share of income. The rear series of each table represents the sup-

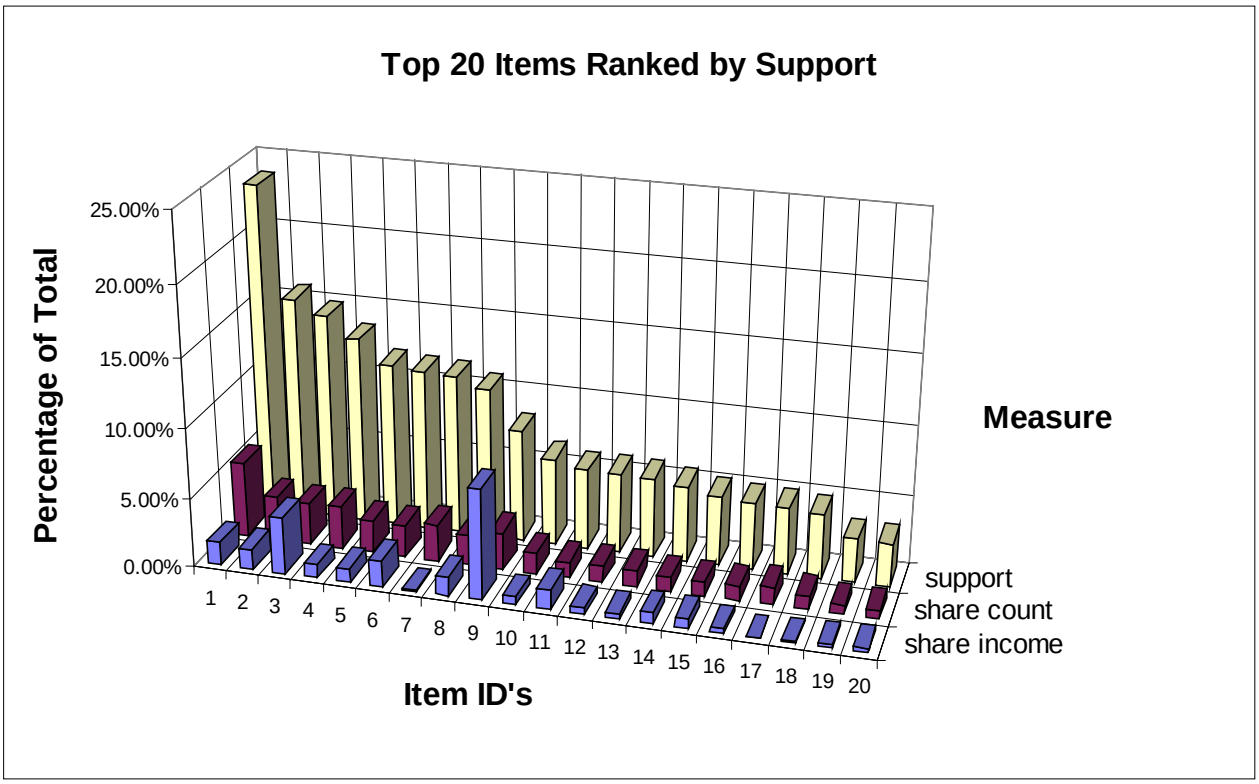

Figure 1. Top 20 items ranked by support

port of each item. The middle series represents the share of each item based on counts. The front series represents the share of each item based on income. To maintain confidentiality of the data, we ranked the items by support from greatest to least and replaced each item ID by an integer representing its standing in this ranking. That is, the item with highest support was assigned the new ID of 1, the second highest 2, and so on. There were 110 frequent items detected. We eliminated one of these items since it was so frequent as to be purchased by almost all customers and therefore not interesting from a marketing perspective.

Figure 1 shows the top 20 items as ranked by support. The height of each column represents the percentage of support or share of each item. Item ID's are listed along the front of the graph.

Our first observation is that the support of an item over-represents the actual frequency of that item in the database, both in terms of the number of items and the profitability of each item. The support of the most frequent item is almost 25% even though this item represents only about 5% of the number of items and 2% of the value of all items. The trend of the top 20 items ranked by support seems at first to parallel the trend of the items measured by share of count. The shares of income of these 20 items, however, shows some substantial variations from both support and

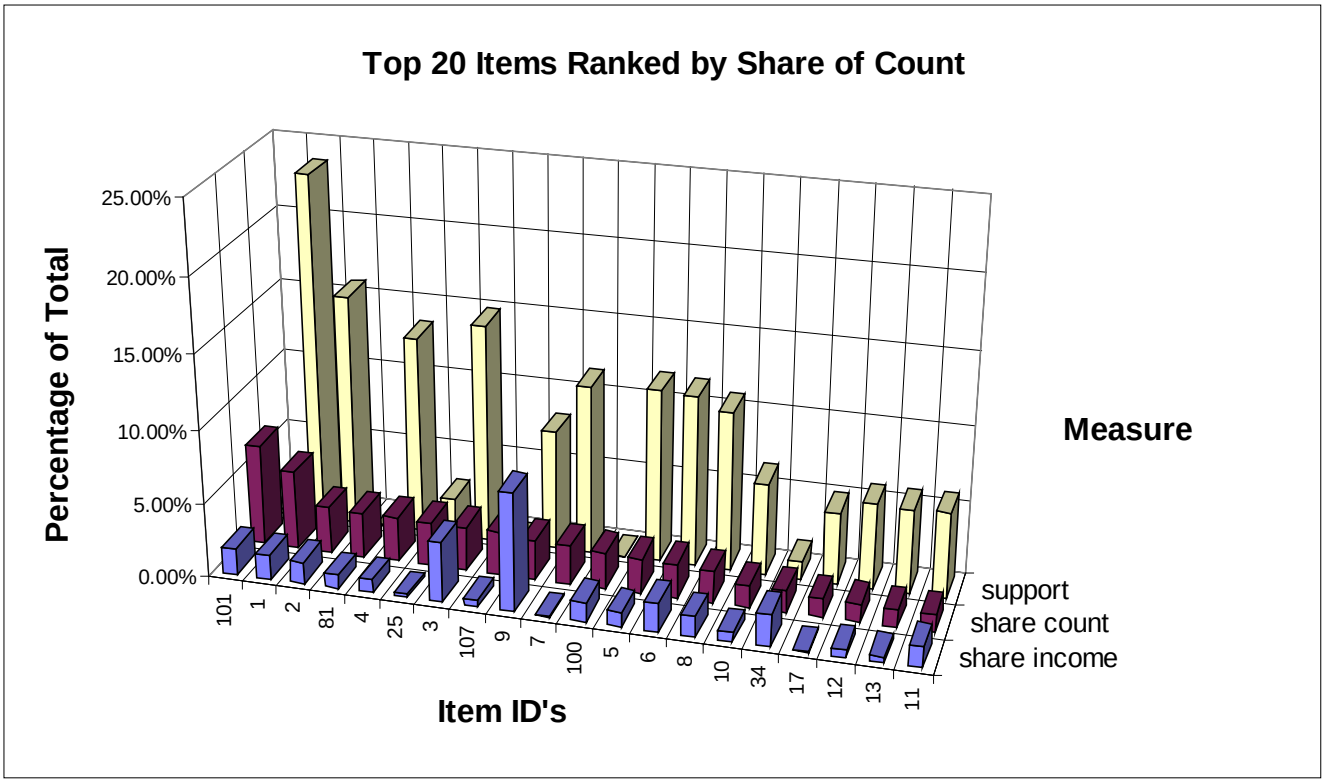

Figure 2. Top 20 items ranked by share of count

share of count.

Figure 2 shows the top 20 items as ranked by share of count. The item ID's remain the same as assigned by support to allow for the difference in ranking to be seen. 14 of the items that were ranked highest by support, those with ID's below 20, are also ranked in the top 20 by share of count. 6 items, however, (101, 81, 25, 107, 100, 34) are seen to be much more significant when ranked by the number of items actually sold (share of count) as opposed to the number of transactions in which they appeared (support). The items 100, 101 and 107 are especially noteworthy in that there were only 109 frequent items ranked. The support measure considered these three as close to the least significant of the whole set, but the share measure ranked them in the top 20.

Figure 3 shows the top 20 items as ranked by share of income. Again, the original ID's assigned by support ranking are retained. Now only 9 items that were ranked in the top 20 by support are placed in the top 20 when ranked by financial impact of the items. Also, 9 items which were ranked by support in the lower half of the most frequent 109 items (from 55-109) are shown to be in the top 20 by income generated. 12 items occur in both the top 20 as ranked by count and income (1-6, 8, 9, 34, 81, 100, 101), but this means that 8 items that were ranked as most impor-

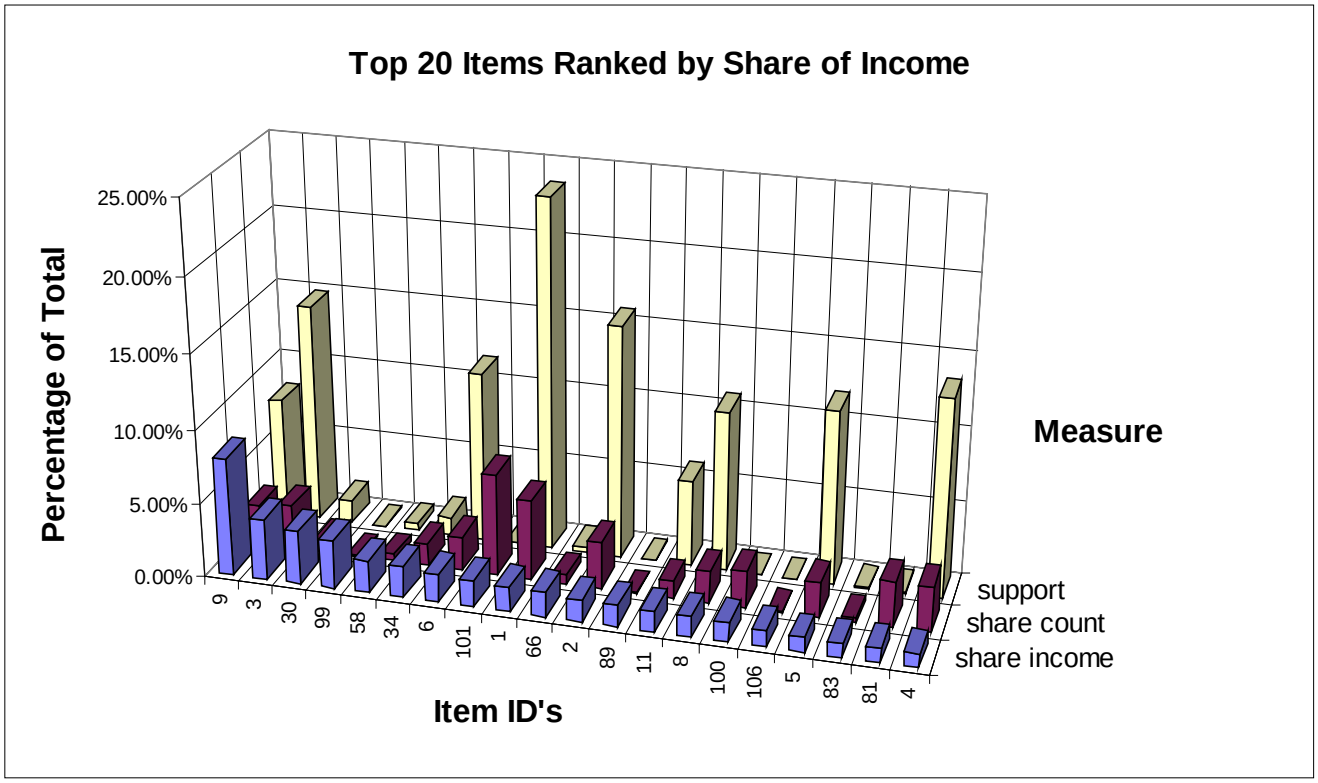

Figure 3. Top 20 items ranked by share of income

tant by financial contribution were not ranked in the top 20 when ranked by a simple share of count alone. The financial contribution of these items, however, is ultimately what makes them of greatest interest to those making decisions that affect the financial performance of the business.

The patterns seen in itemsets are similar to individual items. Table 13 shows the rankings of three sets of 2-itemsets (itemsets with 2 members) ranked by support, share of count and share of income. Each set of three columns represents 20 itemsets ranked by some criterion. We assigned each of 351 frequent 2-itemsets three different ordinals representing its ranking by support, share of count and share of income respectively. We then sorted the itemsets by support and extracted the rankings of the top 20 of these. This is shown in the first three columns of the table. For example, the 2-itemset ranked as most frequent by support was ranked 4th by share of count and also 1st by share of income (see the first row in the top 20 ranked by support). Thus, while this itemset did not represent the most frequent itemset sold in terms of raw count, it was both purchased by the most number of customers and represented the most gross income of all 2-itemsets. On the other hand, the itemset ranked 10th by support was ranked 49th by share of count and 109th

Top 20 ranked by support			Top 20 ranked by share of count			Top 20 ranked by share of income		
support ranking	count ranking	income ranking	support ranking	count ranking	income ranking	support ranking	count ranking	income ranking
1	**4**	**1**	**306**	**1**	**18**	1	4	1
2	13	3	341	2	38	293	8	2
3	17	9	324	3	27	2	13	3
4	19	12	1	4	1	305	45	4
5	20	5	294	5	23	5	20	5
6	22	11	316	6	32	288	121	6
7	27	28	291	7	24	75	80	7
8	35	33	293	8	2	287	206	8
9	47	59	307	9	29	3	17	9
10	**41**	**109**	301	10	31	**336**	**350**	**10**
11	49	41	339	11	22	6	22	11
12	51	37	328	12	69	4	19	12
13	54	42	2	13	3	318	37	13
14	55	58	343	14	77	304	348	14
15	62	97	349	15	76	300	248	15
16	61	90	340	16	67	314	108	16
17	67	101	3	17	9	337	31	17
18	68	132	323	18	44	306	1	18
19	69	98	4	19	12	138	57	19
20	65	154	5	20	5	54	60	20

Table 13. 2-Itemsets ranked by support, share of count and share of amount

by share of income.

The second and third sets of three columns follow a similar pattern, except the second set shows the top 20 itemsets ranked by share of count, and the third set shows the top 20 itemsets ranked by share of income. Notice that the item ranked as most frequent by share of count was ranked only as 306th by support. This is evidently an item that is typically bought in multiples, making it a more frequently bought item than would be evidenced by the simple number of transactions alone. In fact, 15 of the 20 top items ranked by share of count all ranked below the 291st item as ranked by support. From the third set of three columns, the item that was ranked 10th by share of income was ranked 336th by support and 350th by share of count. This is a costly item that although it is not bought as frequently as some other items, makes the company a relatively large amount of money in comparison to some other more frequently purchased items.

Table 14 shows the top 20 2-itemsets as ranked by share of income and the coincidence measures associated with these itemsets. For example, the most frequent itemset generated about 6.5%

Itemset Ranking	Share of Income	Coincidence
1	6.58%	30.93%
2	5.09%	99.95%
3	4.84%	87.35%
4	4.50%	98.01%
5	4.47%	20.78%
6	4.09%	75.49%
7	4.05%	35.05%
8	4.04%	95.34%
9	3.97%	19.34%
10	3.94%	89.75%
11	3.88%	18.52%
12	3.81%	18.52%
13	3.78%	89.55%
14	3.75%	98.36%
15	3.57%	31.81%
16	3.57%	95.30%
17	3.34%	88.25%
18	3.03%	97.34%
19	2.95%	29.26%
20	2.87%	33.45%

Table 14. Top 20 2-itemsets ranked by share of amount, showing coincidence

Itemset ID based on Income Rank	Item 1 Income	Item 2 Income	Share of Income	Item 1 locally weighted dominance	Item 2 locally weighted dominance	Item 1 globally weighted dominance	Item 2 globally weighted dominance
1	$1,097,719	$256,343	6.58%	1.62	0.37	0.87	2.42
2	$675,254	$372,353	5.09%	1.28	0.71	0.99	1.00
3	$250,432	$746,943	4.84%	0.50	1.49	0.95	1.01
4	$667,447	$259,202	4.50%	1.44	0.55	1.00	0.98
5	$597,209	$323,156	4.47%	1.29	0.70	0.71	3.94
6	**$675,811**	**$167,108**	**4.09%**	**1.60**	**0.39**	**1.32**	**0.50**
7	$337,775	$496,104	4.05%	0.81	1.18	0.58	1.90
8	$157,120	$675,811	4.04%	0.37	1.62	0.83	1.04
9	$668,445	$149,147	3.97%	1.63	0.36	0.85	3.99
10	$584,922	$227,240	3.94%	1.44	0.55	0.96	1.10
11	**$568,486**	**$230,070**	**3.88%**	**1.42**	**0.57**	**0.76**	**4.49**
12	$619,434	$165,268	3.81%	1.57	0.42	0.82	4.43
13	$623,954	$155,453	3.78%	1.60	0.39	1.03	0.89
14	$105,840	$667,447	3.75%	0.27	1.72	0.97	1.00
15	$100,038	$635,664	3.57%	0.27	1.72	0.19	2.95
16	**$644,586**	**$90,664**	**3.57%**	**1.75**	**0.24**	**1.00**	**0.99**
17	$584,922	$103,889	3.34%	1.69	0.30	0.98	1.12
18	$257,672	$367,230	3.03%	0.82	1.17	0.98	1.01
19	$336,872	$271,158	2.95%	1.10	0.89	0.70	2.10
20	$548,127	$43,056	2.87%	1.85	0.14	1.00	0.99

Table 15. Top 20 2-itemsets ranked by share of income and dominance measures

of the gross income of the company, and of all instances of the two items sold by the company, about 30% of these were sold together. The two items that comprise this itemset, therefore, may have room for increased coincidence through a marketing campaign. However, the second most frequent itemset had a coincidence of almost 100% of those items sold. It would be very unprofitable to promote the sale of these two items in a product package to boost their coincidence since they are already almost always sold together anyway. Other similar patterns are observable between other pairs of itemsets further down the table, for example, itemsets 4 and 5 which have similar shares but coincidences of 98% and 20% respectively. Without the coincidence measure, data managers would not have enough information to make an informed choice about this issue. The standard itemset measure of support is not able to provide information of this nature.

Table 15 again shows the top 20 2-itemsets as ranked by share of income and breaks down the income according to the contribution of each item. Columns 2 and 3 show the income generated by each item. The share of total income is shown in column 4. The locally weighted dominance for each item is shown in columns 5 and 6. The globally weighted dominance for each item is shown in the rightmost 2 columns. All dominance figures are with respect to share of income.

In general, we might argue that we do not want one item of an itemset to dominate another item to a great degree. For example, in itemset 16, the first item contributes close to 90% of the total income of the itemset. This is reflected in a relatively high locally weighted dominance of 1.75 for the first item and a low locally weighted dominance of 0.24 for the second. However, the globally weighted dominance of both items are 1.0 and 0.99 respectively, showing that the proportions in the itemset are almost identical to the proportions of the items in the whole database. Similar patterns can be seen in itemsets 14 and 20. The weighting in the itemsets, therefore, is somewhat predictable from global proportions of the component itemsets, and therefore, these itemsets may not be particularly interesting.

Itemset 6 illustrates a similar pattern to itemset 16 in the locally weighted dominance, with the first item contributing about 80% of the total income. The locally weighted dominance of the first item is therefore 1.60 compared to 0.39 for the second item. However, the globally weighted dominance of the first item is 1.32, indicating that its proportion to the second item in the itemset is greater than its proportion to the same item in the database. We conclude, therefore, that more of the first item are found in conjunction with the second item than might have been otherwise expected, possibly indicating that a purchase of item 2 leads to a higher probability of the purchase of item 1. We should therefore look at the association rule {Item 2} $\rightarrow$ {Item 1}.

Itemset 11 illustrates a third possibility. Item 1 dominates item 2 with a ratio of about 2:1. The globally weighted dominance of item 2, however, is almost 4.5, indicating that many more of item 2 than would be expected according to global patterns were actually purchased with item 1. This indicates a particularly strong correlation of item 2 with item 1 and therefore the association rule {Item 1} $\rightarrow$ {Item 2} is potentially interesting in a marketing campaign. Similar patterns can be seen in itemsets 5, 9 and 12.

With only locally weighted comparisons on which to make a decision, we might discard any of these itemsets because of the perceived imbalance, but the globally weighted comparison al-

lows us to analyze the patterns more thoroughly and understand our data more. None of these capabilities, however, are available using only the support measure which does not take into account the number or profitability of items sold. We therefore concur with Masand and Piatetsky-Shapiro in [6] that measures maximizing business payoff be taken into account in knowledge discovery tasks.

6. Conclusion

We have presented a new set of *share-based* measures to augment the standard measure of support. The support of an item or itemset is proportional to the number of transactions in which an item occurs. The share of an item or itemset takes into account the number of items purchased relative to the total number of items in the database. We also defined the related share measures of coincidence and dominance. The coincidence of an itemset is the ratio of those items that are purchased together to the total number of the same items purchased in the database. The coincidence measure is useful to determine which of several itemsets with approximately the same share are more interesting. Itemsets with high coincidence may not be interesting since most of them are bought together anyway and a marketing campaign would therefore not boost sales substantially. Itemsets with a large share and moderate coincidence are good candidates for marketing in that there is room for coincidence to grow. The dominance of an item in an itemset is a measure of how its count dominates the count of the itemset as a whole. We may not be interested in itemsets which are overly dominated by some item. The locally weighted dominance gives an indication of when this is the case. The globally weighted dominance, however, shows if the proportions in the itemset are relatively similar to proportions of the same items in the database. An itemset whose proportions deviate substantially from database patterns may be of interest.

Share based measures are both intuitively reasonable and understandable. They are based on the stable, unchanging baseline of the total number of items sold in the given set of transactions. Since they into account the number of items purchased by customers, they allow the capability of relating the number of items to the financial impact of the sales. This financial extension is ultimately the most relevant measure when making marketing decisions.

References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proceedings of the 20th VLDB Conference,* Santiago, Chile, 1994, 487-499.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A.I. Verkamo, "Fast Discovery of Association Rules," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data Mining,* AAAI Press / MIT Press, Menlo Park, CA, 1996, 307-328.

[3] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in *Proceedings of the 21st VLDB Conference,* Zurich, Switzerland, 1995.

[4] M. Houtsma and A. Swami, "Set-Oriented Mining for Association Rules in Relational Databases," in *Proceedings of IEEE International Conference on Data Engineering,* March, 1995, 25-33.

[5] 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 Third International Conference on Information and Knowledge Management,* Gaithersburg, Maryland, Nov., 1994.

[6] B. Masand and G. Piatetsky-Shapiro, "A Comparison of Approaches for Maximizing Business Payoff of Prediction Models," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* Portland, Oregon, August, 1996, 195-201.

[7] J. S. Park, M. S. Chen and P. S. Yu, "An Effective Hash Based Algorithm for Mining Association Rules," in *SIGMOD Record,* **24:2,** 1995, 175-186.

[8] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," in *Proceedings of the 21st VLDB Conference,* Zurich, Switzerland, 1995.