RC24168 (W0701-155) January 31, 2007
Computer Science

IBM Research Report

Identifying Bundles of Product Options Using Mutual Information Clustering

Claudia Perlich, Saharon Rosset
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Identifying Bundles of Product Options using Mutual Information Clustering

Claudia Perlich, Saharon Rosset
{perlich,srosset}@us.ibm.com
IBM T.J. Watson Research Center

January 30, 2007

Abstract

Mass-produced goods tend to be highly standardized in order to maximize manufacturing efficiencies. Some high-value goods with limited production quantities remain much less standardized and each sale can be configured to meet the specific requirements of the customer. In this work we suggest a novel methodology to reduce the number of options for complex product configurations by identifying meaningful sets of options that exhibit strong empirical dependencies in previous customer orders. Our approach explores different measures from statistics and information theory to capture the degree of interdependence between the choices for any pair of product components. We use hierarchical clustering to identify meaningful sets of components that can be combined to decrease the number of unique product specifications and increase production standardization. The focus of our analysis is on the influence of different similarity measure - including chi-squared statistics and versions of mutual information - on the ability of the clustering to find meaningful clusters.

1 Introduction and Motivating Example

While bundling of products has received significant attention in the economic literature (e.g., [1, 2, 3]), the bundling of product options is typically limited to considerations of production efficiency and engineering. In order to optimize the tradeoffs between maximizing production efficiencies and making products that meet the individualized requirements of particular customers, manufacturers have developed techniques of combining options into bundles so that batches of similarly customized products may be made together, rather than making each customized product individually. Existing approaches to developing such bundles, however, have been driven by the choices of product designers and have not afforded a systematic way of incorporating customer preference data. Examples of products that offer option bundles can be observed in the car industry. Toyota for instance offers an 'All Weather Guard Package' that includes an Intermittent Rear Window Wiper, Windshield Molding, Heavy-Duty Heater and Rear-Seat Heater Ducts. All of the above components appear related to the requirements of driving under harsh weather. The alternative 'standard package' contains the simpler default components - more suitable for driving in geographies with milder winters.

Even considering all the different options, passenger cars remain a highly standardized product class. In this work we explore the task of finding good sets of components for bundling on the example of truck configurations. Trucks are ordered for a specific use and the customer can specify all major components separately. A truck will only be produced after the customer has made his choices. However, one would suspect a limited set of usage categories and certain recurring patterns in the customer orders. For example, trucks are specialized towards different weight-of-goods categories. Safety requirements for transportation of heavy goods include enhanced chassis support, number of gears, shock absorbers, efficient breaks, and also more powerful engines. Trucks for lighter consumer-goods on the other hand will employ less expensive options for motorization, braking and build. So the objective of our bundling task is to find component combinations that empirically exhibit strong customer-choice dependencies (potentially driven by the objectives of truck usage) and will appeal to future customers of new orders.

An important point here is that we want to bundle components, not choices. That is, we do not want to identify groups of *specific* choices that go together (like, say, Intermittent Rear Window Wiper and Heavy Duty Heater), but groups of *generic* components (like Rear Window Wipers and Heating System), such that their choices tend to be correlated across customer segments. This is important to us because we eventually want to create multiple independent component groups (clusters) such that each customer can select a specific package for each component group. Keeping the component groups fixed guarantees that customers will be able to select from each one independetly of the others. We make this distinction concrete in Section 2, and discuss it further in the following sections.

To address this objective we need to quantify dependencies between components. In order to achieve this goal we will explore potential measures of dependencies between nominal variables in Section 3 and discuss properties of such similarity measures. Given that such measures can capture only pair wise dependencies, we propose in Section 4 the use of hierarchical clustering to find larger sets of components that all exhibit large pair wise dependencies. We will illustrate the issues and results on the example of the truck configuration domain.

While the work in this paper concentrates on a specific application, we believe that the bundling problem is of general interest and we are not aware of prior work on this formalization as hierarchical clustering under an appropriate similarity measure. Other contributions of this work include the identification of desirable properties in the given context of nominal variables with differing numbers of choices and skewed probability distributions. While there has been substantial work on clustering using chi-square based similarities as well as clustering with mutual information (e.g., [4, 5, 6]), we are not aware of combined methods that incorporate both, the mutual information and the statistical significance as clustering criteria, which we propose below.

2 Notation and Formalization

Formally, a complex product consists of n components $C_1, ..., C_n$. For every component C_j, there is a limited set of k_j possible choices $\{c_{j1}, ..., c_{jk_j}\}$ where the number of choices k_j differs across components. We assume that we have N past observations that indicate for each order the particular choices as a vector $o_1, ..., o_n$ with $o_j \in \{c_{j1}, ..., c_{jk_j}\}$. Note that this setup differs considerably from the typical basket analysis of customer choices that motivated the work on large itemsets and mining of association rules [7]. The notion of components imposes additional constraints:

- all customers have the identical number of n components and
- for each component only one choice is permissible.

While frequent itemsets may be indicators of semantic interdependencies between choices, they do not measure the interdependence of components. Each itemset considers only one particular choice $c_{jg} \in \{c_{j1}, ..., c_{jk_j}\}$ and how often it appears with another choices for another component, but not how much each possible choice $c_{j1}, ..., c_{jk_j}$ for component c_j correlates with the choices for the other component. Another problem with the notion of frequent itemsets is its dependence on the prior probability of a particular choice. In particular, a frequent itemset analysis identifies typically combinations of default values for components with one very common default value and a small set of much less common values. That does not mean that there is any deeper semantic dependency between the components. It is just an artifact of the high skew of the probabilities. While there are measures of the 'unexpectedness' of an itemset, these measures are typically a function of the size of the set, with larger sets exhibiting much more unexpected behavior. We will take a brief look at the results of a frequent itemsets analysis for our application domain in Section 5.1 to highlight the distinction between a component-level analysis and a choice-level analysis offered by frequent itemset mining.

To address our specific bundling objective we need to quantify dependencies between sets of components, not sets of choices. In order to achieve this goal we will explore potential measures of dependencies between nominal variables in the following Section and discuss properties of such similarity measures.

3 Measuring Dependence

The objective in our bundling task is to find sets of components where past customer choices exhibit some form of dependence. So far we have used the term dependence rather loosely in a non-technical sense of some form of a semantic connection. While it is difficult to formalize dependence without a clear prior notion of how things depend on each other, there is a clear statistical notion of the opposite: independence between random variables. We can formalize the observation of a customer choice o_i for a particular component C_i as the outcome of a random experiment over the sample space $\Omega_i = \{c_{i1}, ..., c_{ik}\}$. Formally, two random variables are independent if their joint distribution is equal to the product of their individual distribution functions

$$P(o_h = c_{hp}, o_l = c_{lm}) = P(o_h = c_{hp}) * P(o_l = c_{lm}) \tag{3.1}$$

for all elements of the Cartesian product of the two sample spaces $\Omega_i \times \Omega_j$ (all possible choice pairs for the two components). Independence is defined generally over an arbitrary number of variables and we could attempt to devise a measure of the interdependence within entire sets of components. However, such a strategy will not lead to non-overlapping bundles as desired in our case. In addition, given the somewhat vague business objective, the final choice of bundles is potentially subject to many additional production constraints and considerations. We will therefore restrict our work to pairs of components and employ hierarchical clustering to suggest a hierarchy of potential non-overlapping bundles, from which the domain experts may choose the desired granularity for option bundling.

We can now measure dependence in terms of the degree of violation of the equality 3.1 over all pairs of choices c_{hp}, c_{lm}) for a pair of components (C_h, C_l). This requires initially the estimation of the distribution for all possible components and their choices $P(o_l = c_{lm})$ and choice pairs $p(c_{hp}, c_{lm})$. We will simplify the notation and use $p(c_{hp})$ to denote $P(o_h = c_{hp})$ and $p(c_{hp}, c_{lm})$ for $P(o_j = c_{hp}, o_l = c_{lm})$ respectively. Note that for the posed business problem, we do not have a clear evaluation metric for the quality of bundling. Otherwise we could hope to derive (either implicit or explicitly) an appropriate similarity measure leading to optimal bundling performance. Our results will depend very much on the particular choice of similarity. We will therefore discuss in more detail some desirable and useful properties and frame existing measures with respect to these properties. While there are many possible choices of a similarity measure $D([0,1]^s, [0,1]^s) \to \mathbb{R}$ (where $s = k_h * k_l$ is the number of choice pairs), reasonable candidates can be constructed from an 'atomic' measure of similarity $D_0([0,1],[0,1]) \to \mathbb{R}$ of the elements (c_{hp}, c_{lm}) of the Cartesian product over the sample spaces and aggregates $A(\mathbb{R}^s) \to \mathbb{R}$ over all the atomic similarities.

In order to be suitable for our bundling task, we would like the similarity to exhibit three other desirable properties:

- It has to be **symmetric** with $D(C_h, C_l) = D(C_l, C_h)$, since there is no special order on the components;
- It should to be **comparable** across component pairs. In particular, it should be rather insensitive to the specific size of the Cartesian product of the sample spaces;
- It should be **robust** towards estimation errors of the distribution. Given a limited sample of prior customer orders and a large sample space for some components with many possible values, the estimation quality of the probabilities will be limited. This problem is particularly dominant for rare choices.

The issue of assessing dependence has been considered in different fields including the analysis of contingency tables in statistics and information theoretical work on the information content of signals.

3.1 Chi-Square Based Similarity

3.1 Chi-Square Based Similarity Measures of association have a long history in the context of the analysis of contingency tables. For an extensive overview consider [8]. However, the majority of association measures is not suitable for our task for due to a lack of symmetry, and focus on the conditional mode of the distribution while ignoring less common choices. One standard approach to evaluate the significance of statistical dependencies of two nominal random variables (C_h and C_l) is based on a Chi-square test.

(3.2)
$$\chi^2(C_h, C_l) = N \sum_{i=1}^{k_h} \sum_{j=1}^{k_l} \frac{(p(c_{hp}, c_{lm}) - p(c_{hp})p(c_{lm}))^2}{p(c_{hp})p(c_{lm})}$$

Note that this formulation uses an 'atomic' Euclidean similarity and a sum as aggregation function where the denominator reflects the expected probability of observing a pair under the null-hypothesis of independence. Let us make a few observations that contradict two of our desirable properties for the bundling task - comparability and robustness:

- The measure from Equation 3.2 follows (under certain assumptions) approximately a Chi-square distribution with $d = (k_h - 1)(k_l - 1)$. This means that its expected value is a function of the sizes of the sample spaces and renders a comparison across component pairs impossible.
- The Chi-square statistic is known to be sensitive to small number of expected observations in the denominator. The Fisher exact test is correcting for this problem but is only applicable for 2x2 tables.

One can consider several solutions for both issues. To address the dependence on the degrees of freedom, we can either convert the statistic into the corresponding p-value or correct it based on the Normal approximation. The p-value is derived from the cumulative distribution with the appropriate degrees of freedom and reflects the probability of such a Chi-square occurring by chance. However, as we will see in the experiments, this correction eliminates most of the information. Given the comparably large size of our dataset, almost all observed values are significant with high probability and most of the p-values are indistinguishable from 0. The second correction takes advantage of the fact that a Chi-square with large number of degrees of freedom $d = (k_h - 1)(k_l - 1)$ is approximately normally distributed with a mean equal to d and a variance equal to $2*d$. We can therefore use the following correction:

(3.3)
$$N_{\chi^2}(C_h, C_l) = \frac{\chi^2(C_h, C_l) - d}{\sqrt{2d}}$$

To address the issue of small expectations, we combine multiple rare component choices into a new value 'other'. Note that a replacement with 'other' can artificially create dependencies and should be taken with a grain of salt: the fact that for two components some cases both have the value 'other' is likely to indicate that the customers are picky and always want something special, not that this choice of one component affects the other.

3.2 Mutual Information [9] measures the information about one component that is shared by another. If the components are independent, then one contains no information about the other and vice versa, so their mutual information is zero. Formally, the mutual information MI of two random variables for components C_h and C_l is defined as:

(3.4)
$$MI(C_h, C_l) = \sum_{i=1}^{k_h} \sum_{j=1}^{k_l} p(c_{hp}, c_{lm}) \log \frac{p(c_{hp}, c_{lm})}{p(c_{hp}) * p(c_{lm})}$$

In the case of mutual information the aggregation function is again a weighted sum and the 'atomic' similarity is the log of the ratio of the expected and observed probability. While this measure both symmetric and robust to small expectations due to the log transformation, it is not comparable across pairs of components. If the sample space of the two variables is identical, the maximum mutual information under complete dependence is equal to the entropy. Entropy however is a function of the sample size. In particular, a tight upper bound on the mutual information [10] is given by

$$MI(C_h, C_l) \leq \frac{H(C_h) + H(C_l)}{2} \tag{3.5}$$

where $H(C_h)$ is the entropy [11] of component C_h defined as

$$P(C_h) = \sum_{i=1}^{k_h} p(c_{hi}) \log(\frac{1}{p(c_{hi})}) \tag{3.6}$$

We therefore define a normalized mutual information as suggested by [10] as

$$NMI(C_h, C_l) = \frac{2MI(C_h, C_l)}{H(C_h) + H(C_l)}. \tag{3.7}$$

3.3 Combining Mutual Information and Significance While both measures work on the same underlying information, the objective for which they were developed is very different. The goal of the Chi-square measure is to assess significance relative to the null-hypothesis of independence. This means in particular, that it matters how many observations are provided. The power of a test is a function of the number of observations and as the sample becomes very large, almost every small deviation becomes significant. We can see the relevance of the sample size N in Equation 3.2. Information theory ([11, 9]) on the other hand takes a different perspective. Mutual information is completely independent of the sample size N and in does not assess whether the observed amount of information could have been observed by random chance. So mutual information is a closer measure of the quantity we are interested in, the degree of dependence, but does not take randomness into account and whether the observed quantities are significant. We therefore propose a similarity measure that incorporates both statistical considerations of significance and the amount of information:

$$SIM(C_h, C_l) = NMI(C_h, C_l) * (1 - p(C_h, C_l)) \tag{3.8}$$

where $p(C_h, C_l)$ is the p-value of the appropriate Chi-square and can be calculated from the cumulative density function for the Chi-square distribution with $(k_h - 1)(k_l - 1)$ degrees of freedom as $p(C_h, C_l) = 1 - cdf(\chi^2(C_h, C_l), (k_h - 1)(k_l - 1))$. This similarity measure weights the observed amount of shared information by the probability of it not being random. The multiplicative weighting can motivated by a somewhat simplistic expected value calculation of two possible states. In one state with a probability equal to the p-value the true relationship is actually random (mutual information equal to 0) and in the other with probability 1 minus the p-value the dependence it is not random and the estimate of the mutual information is assumed to be correct.

4 Hierarchical Clustering

Clustering and cluster analysis (e.g., [12, 13]) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective groups. Rather than finding a fixed number of clusters in the data, agglomerative hierarchical clustering as proposed by Johnson [12] proceeds iteratively by combining existing clusters and may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. Examples of such dendrograms are given Figure 1,2,3. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm. The similarities between the nodes reflect the relative similarities of the clusters. Given a set of n items to be clustered, and an $n * n$ similarity matrix, the basic process of hierarchical clustering [12] is this:

1. Start by assigning each of the n component to

its own cluster. Let the similarities between the clusters be the same as the similarities between the items they contain.

2. Find the closest pair of clusters and merge them into a single cluster, so that now you have one cluster less.

3. Compute similarities between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size n.

Aside from the similarity measure, the criterion to define 'closest' in step 2 is one of the major components of the clustering algorithm and can affect the results significantly. Different criteria include:

- **Minimum**: Similarity between clusters is the smallest similarity from any member of one cluster to any member of the other cluster.

- **Average**: Similarity is the average over the similarities from any member of one cluster to any member of the other cluster. Alternatively, one can consider the median, which is more robust to similarity outliers.

- **Maximum**: Similarity between clusters is the largest similarity from any member of one cluster to any member of the other cluster.

5 Dataset and Empirical Results

Our experiments are based on legacy transaction records of a truck manufacture. We selected (based on the recommendation of the manufacturer) a small subset of 31 important components for the illustration of this work and included in the analysis a total of 3500 recent orders. An overview of the components is provided in Table 1. The table also provides some information about the statistical properties including the size of the sample space for each component (Size) and the empirical prevalence of the mode for each component (Mode).

5.1 Analysis of Most Frequent Itemsets As we discussed in Section 2, frequent itemsets do not directly address our problem of component bundling, because they only identify bundles of choices (component values). However, we wanted to investigate whether the most frequent itemsets could give us information about component bundles, by identifying dominant option bundles, that can be a basis for component bundling. We used APRIORI to identify frequent itemsets on the order database. Each record is a truck specification consisting of 31 component choices. Let us consider bundles that combine between 3 components. APRIORI finds 4 itemsets that appear in more than 95% of truck orders covering 3 components as shown in Table 2.

Code	Component	Size	Mode
C01	Model	8	0.80
C06	Exhaust Package	15	0.64
C08	Brake Package	4	0.92
C03	Dead Axle Package	11	0.98
C10	Engine	54	0.14
C12	Retarder Driveline	12	0.65
C18	Clutch	717	0.51
C21	LH Fuel Tank	15	0.45
C22	RH Fuel Tank	17	0.45
C26	Radiator	8	0.40
C29	BatteryBox	7	0.98
C34	Transmission	82	0.11
C36	PTO Engine Front	4	0.96
C32	PTO Transmission	22	0.80
C40	Axle Front	21	0.20
C44	Brake Front	9	0.48
C42	Axle Rear Drive	57	0.21
C41	Axle Ration	49	0.15
C43	Brake Rear	13	0.12
C54	Wheelbase	164	0.19
C56	Frame Rail	10	0.27
C55	Frame Overhang	115	0.23
C57	Fifthwheel	33	0.91
C62	Suspension Front	12	0.40
C63	Suspension Rear	73	0.11
C68	SleeperCab	2	0.99
C82	Cab Size	7	0.74
C84	Business Segment	30	0.33
C85	Vehicle Service	14	0.67
C92	Trailer Type	12	0.86
C93	Body Type	30	0.44

Table 1: Component Codes and Definitions for the Example Domain. The size column represents the number of possible choices for the component (size of the sample space) and the last column presents the probability of the most common choice (Mode) as an indicator of the skew in the probabilities.

In particuar, these results would suggest to combine any subset of three from the four components: C29, C68, C03 and C36. The only thing that these components have in common, is a very dominant Mode option (see last column of Table 1 that is prevalent in more than 98% of all orders. However, there is no "meaningful" relationship between the Battery Box, the Sleeper Cab, and the Dead Axle Package. Furthermore, if you consider the actual choices that appear in the frequent

3 Componet-Choice Set	**Coverage**
C03-998 C68-998 C36-998	(98.4%)
C29-017 C68-998 C36-998	(98.3%)
C29-017 C03-998 C36-998	(97.1%)
C29-017 C03-998 C68-998	(97.1%)

Table 2: Most frequent itemsets with 3 components from the APRIORI algorithm.

itemsets, they are mostly default values of the form 9**, typically indicating at times "None".

These results reflect the previously suggested shortcomings of the frequent itemset approach in this particular context: the results are heavily influenced by the distribution of the choices within components; they ignore the frequencies of alternative choices for the same set of components, and finally do not identify interesting dependences between components.

5.2 Similarity Measures Following the discussion in Section 3 we now shift our analysis to the estimation of dependences between components using the 7 different similarity measure at our disposal:

N_{χ^2}:	Chi-sqare corrected for degrees of freedom by Normal approximation as defined in Equation 3.3
$N_{\chi_r^2}$:	Chi-sqare without rare options (occurrence below 20) corrected for degrees of freedom by Normal approximation
$p(\chi^2)$:	p-values of Chi-square
$p(\chi_r^2)$:	p-values of Chi-square without rare options
MI:	Mutual information as defined in Equation 3.4
NMI:	Normalized mutual information as defined in Equation 3.7
SIM:	Combined mutual information and p-value as defined in Equation 3.8

Table 3 shows the correlation (which implicitly assumes a linear relationship) between the measures.

	N_{χ^2}	$N_{\chi_r^2}$	$p(\chi^2)$	$p(\chi_r^2)$	MI	NMI	SIM
N_{χ^2}	**1.00**	**0.80**	0.20	0.16	0.41	0.63	0.63
$N_{\chi_r^2}$	**0.80**	**1.00**	0.18	0.16	0.59	**0.84**	**0.84**
$p(\chi^2)$	0.20	0.18	**1.00**	0.58	0.18	0.22	0.24
$p(\chi_r^2)$	0.16	0.16	0.58	**1.00**	0.16	0.20	0.20
MI	0.41	0.59	0.18	0.16	**1.00**	**0.88**	**0.88**
NMI	0.63	**0.84**	0.22	0.20	**0.88**	**1.00**	**0.99**
SIM	0.63	**0.84**	0.24	0.20	**0.88**	**0.99**	**1.00**

Table 3: Correlation of the different similarity measures.

We can clearly identify three groups: measures based on mutual information (MI, NMI and SIM), measure based on the p-values and the two Chi-square measures. The fact that the p-values are only very vaguely correlated with the Chi-square measures is due to the inherent non-linearity of the cumulative density function. Replacing rare values has a moderate effect both in the case of p-values and the Chi-square measures. The normalization of the mutual information clearly has an effect, much more so than the weighting by the p-value. The only exception to the nice separation of the measures into 3 groups is the high correlation between the Chi-square adjusted for rare values and the two normalized mutual information measures of 0.84.

As pointed out earlier, the measures using a p-value only reflect whether the observed degree of dependence could be random. We have a fairly large dataset and both measures assign a value of 0 to 93% of all pair wise distances. This renders it unusable as a similarity measure for the clustering objective. The only pairs that show values above 0 involve typically components with a very high probability for the mode (e.g., components C03, C68, and C36).

5.3 Clustering Results We used Pajek [14] to perform the hierarchical clustering using average criterion and the visualization of the corresponding dendrograms. Given our earlier analysis of the similarity measures we consider for clustering only SIM, $N_{\chi_r^2}$, and N_{χ^2}.

Figures 1,2,3 show the dendrograms for the three similarity measures. The relative lengths of the horizontal lines in the dendograms reflect the decrease of similarity. The further left two branches join, the more similar the two clusters. We can use each of the dendograms to identify component clusters. We (somewhat arbitrarily) decide on a vertical cutoff in the dendrogram and consider all clusters left of it. Table 4 identifies clusters in the dendograms that can be suggested to a domain expert as potential component bundles. **Bold** indicates sets that are common across all 3 measures, *italic* indicates sets that occur in at least 2 dendograms. The clusters are listed in the order they appear top to bottom in the respective figures. While the order of the components top-to-bottom varies across the dendograms (it is determined by the clustering algorithm), we find fairly consistent results across all three measures in terms of potential bundles. Note that only one cluster under the Chi-square measures (C29,C68) consists of components that were identified in the frequent itemset analysis. Let's consider the four bold clusters that are common across all measures in more detail. In particular, we will select the four most common option combinations for each of the proposed bundle and see what percentage of the orders they cover.

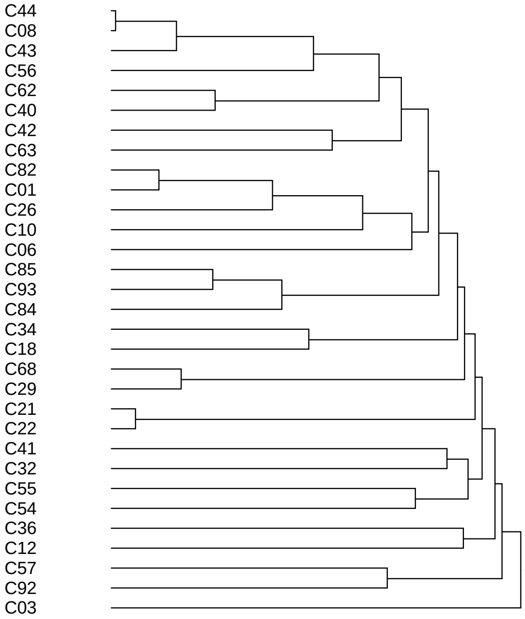

Figure 1: Dendrograms for hierarchical clustering using the N_{χ^2} measure.

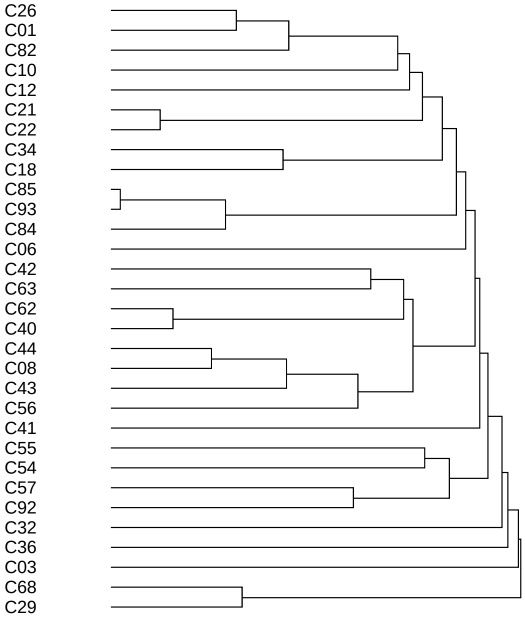

Figure 2: Dendrograms for hierarchical clustering using the $N_{\chi_r^2}$ measure.

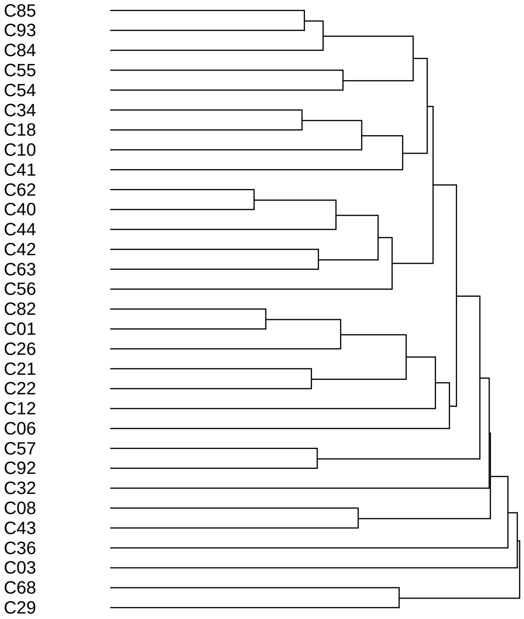

Figure 3: Dendrograms for hierarchical clustering using the SIM measure.

(C21,C22) determines the left and right fuel tank. Given the typically symmetrical form of trucks it seems very reasonable that there is a strong dependence between the two. The coverage of the four most common option combinations for this bundle is 60%.

(C42,C63) relates to the rear specification of the truck - the axle rear drive and the rear suspension with a coverage of 22% for the four most common option combinations. While the coverage seems low, note that the expected coverage of the most common option pairs under the independence assumption is only 2% since the most common choice for C63 occurs in only 11% of orders and the mode of C42 appears in 20%. In the light of this, 22% is still impressive.

(C82,C01,C26) determine the model, the cab size and the radiator with a coverage of 82%.

(C85,C84,C93) are 2 broad vehicle (business segment and vehicle service) categories and the body type. The coverage of the four most frequent combinations is 42%.

Recall that we selected a small subset of 31 components for this analysis that are important and do not include any small parts such as wiper blades, for which

Measure	Componet Set
$N_{\chi_r^2}$	*C44,C18,C43,C56*
	C62,C40
	C42,C63
	C82,C01,C26
	C85,C93,C84
	C68,C29
	C21,C22
N_{χ^2}	**C82,C01,C26**
	C21,C22
	342,180
	C85,C93,C84
	C42,C63
	C62,C40
	C44,C18,C43,C56
	C57,C92
	C68,C29
SIM	**C85,C93,C84**
	552,545
	342,180,101
	C62,C40,C44
	C42,C63
	C82,C01,C26
	C21,C22
	C57,C92
	C18,C43

Table 4: Component clusters in the dendograms that can be suggested to a domain expert as potential component bundles. **Bold** indicates sets that are common across all 3 measures, *italic* indicates sets that occur in at least 2 dendograms.

we would expect to see stronger dependencies. While the final decision about the appropriateness of clusters has to be done by a domain expert, we feel confident that our methodology can provide a good set of candidate component bundles for closer examination.

6 Discussion and Conclusion

We presented an analytical approach that can guide the design of appropriate bundles of components for complex products such as trucks. While the task is very relevant in practice, there is no clear measure of performance and the validity of the results can only be assessed based on domain specific information or by a domain expert. We suggest a novel approach that compines hierarchical clustering and a similarity measure based on mutual information, adjusting for the number of options and combining it with statistical significance. While the adjustment for the number of options shows a strong effect on the similarity measure, incorporating the p-values has a minor effect if the dataset is large. In this case most p-values are close to zero. Using our similarity measure to assess dependencies between customer choices we can identify meaningful candidate sets of components. We are not aware of studies that investigate issues of similarity scaling and distribution in the context of different clustering approaches and hope to address this topic in future work.

References

[1] Eppen, G.D., Hanson, W.A.: Bundling-new products, new markets, low risk. Sloan Management Review **32**(4) (1991) 7–14

[2] Adams, W.J., Yellen, J.L.: Commodity bundling and the burden of monomoly. Quaterly Journal of Economics **90** (1976) 475–498

[3] Bakos, Y., Brynjolfsson, E.: Bundling information goods: Pricing, profits and efficiency. Management Science **45**(12) (1999)

[4] Strehl, A., Ghosh, J., Mooney, R.: Impact of similarity measures on web-page clustering. In: Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 July 2000, Austin, Texas, USA, AAAI (2000) 58–64

[5] Kraskov, A., Stogbauer, H., Andrzejak, R.G., Grassberger, P.: Hierarchical clustering based on mutual information. Europhysics Letters **70**(2) (2005) 278–284

[6] Slonim, N., Atwal, G.S., Tkacik, G., Bialek, W.: Information based clustering: Supplementary material. Proceedings of the National Academie of Sciences **102**(51) (2005) 18297–18302

[7] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: International Conference of Very Large Data Bases (VLDB). (1994) 487–499

[8] Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. Springer-Verlag,New York (1979)

[9] Kullback, S.: Information Theory and Statistics. Dover, New York (1968)

[10] Strehl, A.: Relationship-based clustering and cluster ensembles for high-dimensional data mining, phd thesis, the university of texas at austin (2002)

[11] Shannon, C.: A mathematical theory of communication. The Bell system technical journal **27** (1948) 379–423

[12] Johnson, S.C.: Hierarchical clustering schemes. Psychometrika **2** (1967) 241–254

[13] Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York (1990)

[14] Batagelj, V., Mrvar, A.: Pajek - program for large network analysis. Connections **21**(2) (1998) 47–57