Model-based and Scalable Relationship Discovery in Business Analytics

Jing Shyr [1], Damir Spisic[1], Jane Chu[1], Sier Han[2], Xue Ying Zhang[2]
[1]IBM SPSS Predictive Analytics, 200 W. Madison St. 23th Fl., Chicago, IL 60606, USA
[2] IBM SPSS Predictive Analytics, 2&3 Floor, Building C, Outsourcing Park Phase I, No. 11 Jinye 1st Rd., High Tech Zone, Xi'an, China

Abstract

This paper proposes a model-based and scalable system to produce a series of tabular reports that illustrate the important measure-dimension (or continuous target-categorical predictor) relationships and exhibit strong dimension interactions. The analysis for each report is based on a linear regression model and the interaction detection is an ANOVA test. However, only basic statistics are used to evaluate model accuracy and conduct ANOVA tests without actually fitting models. For datasets with the large number of categorical predictors, it becomes prohibitive to generate and analyze all possible tables. So a structured and scalable search process is applied: all the tables with single dimension are considered first; the tables with two and higher dimensions are considered selectively based on the analysis of the corresponding tables of lower dimension. This ensures that the computational effort needed for generating and analyzing the tables is limited. Furthermore, the top selected tables are further analyzed by detecting any cells, which correspond to the category combinations of dimensions in the table, with high contributions to the significance of the interaction effect.

Key words: ANOVA test, interaction effect, tabular report.

1. Introduction

Data analysts today have to deal with increasingly large volumes of data. Attempting to find insights in large amounts of data (e.g., terabytes, petabytes, etc.), with many possible combinations between fields, is a difficult task. A common business scenario is identifying the relationship and influence of dimensions generated by categorical fields or attributes on a continuous target field or measure. The goal for the data analyst is to determine which of the dimensions are relevant to the measure and among those that are relevant, discerning the magnitude of their impact. Ultimately, the goal is to produce a series of aggregated tabular reports that illustrate measure-dimension relationships. It is from these relationships that analysts derive insights into their businesses. The challenge is trying to navigate through what may possibly be thousands of reports, each representing a possible measure-dimension combination.

This paper proposes a model-based and scalable system (we call it “Relationship Discovery”) to produce a series of tabular reports that illustrate the important measure-dimension (or continuous target-categorical predictor) relationships and exhibit strong dimension interactions. An interaction describes a situation in which the simultaneous influence of two dimensions on the measure is not additive.

The analysis for each aggregate report is based on a linear model including the corresponding target and the categorical predictors determining the table dimensions. And the interaction detection is based on an ANOVA (analysis of variance) test. However, only basic statistics are used to evaluate model accuracy and conduct ANOVA tests without actually fitting models. The detected dimension interactions are ranked according to their strength and reported to the user.

For datasets with the large number of categorical predictors, it becomes prohibitive to generate and analyze all possible aggregate tables, even with a low number of dimensions. For example, data with 100 predictors would generate the total of 166,750 tables with three or fewer dimensions. We apply a structured search where all the tables with single dimension are considered first. The tables with two or three dimensions are considered selectively based on the analysis of the corresponding tables of lower dimension. This ensures that the computational effort needed for generating and analyzing the tables is limited. It is also effective by resulting in the detection of a higher number of relevant tables than by a random search of comparable size.

The top selected tables are further analyzed by detecting any cells with high contribution to the interaction effect. The overall model based summaries as well as the results of the cell-by-cell analyses are made available for output to the user.

The rest of this paper is organized as follows: Section 2 will describe the Relationship Discovery system in details while a few concluding remarks are in Section 3.

2. Relationship Discovery System

Given a data set and a measure of interest, Relationship Discovery system provides a model-based and scalable search for interactions in the multitude of all possible low-dimensional aggregate tabular reports based on the available dimensions. Each dimension is formed by a different categorical attribute with potential to impact the measure. Dimension cells in the reports correspond to the categories of the matching fields. Table cells correspond to the combinations of categories from fields matching different dimensions in the table. In the following sub-sections we first present the overall framework and the functional flow, followed by the detailed description of three layers (data aggregation, search and insight construction) and their interactions.

Figure 1 illustrates the framework and the functional flow chart for the Relationship Discovery system for up to 3 dimensions. Please note that the Relationship Discovery system can be extended to dimensions larger than 3 easily.

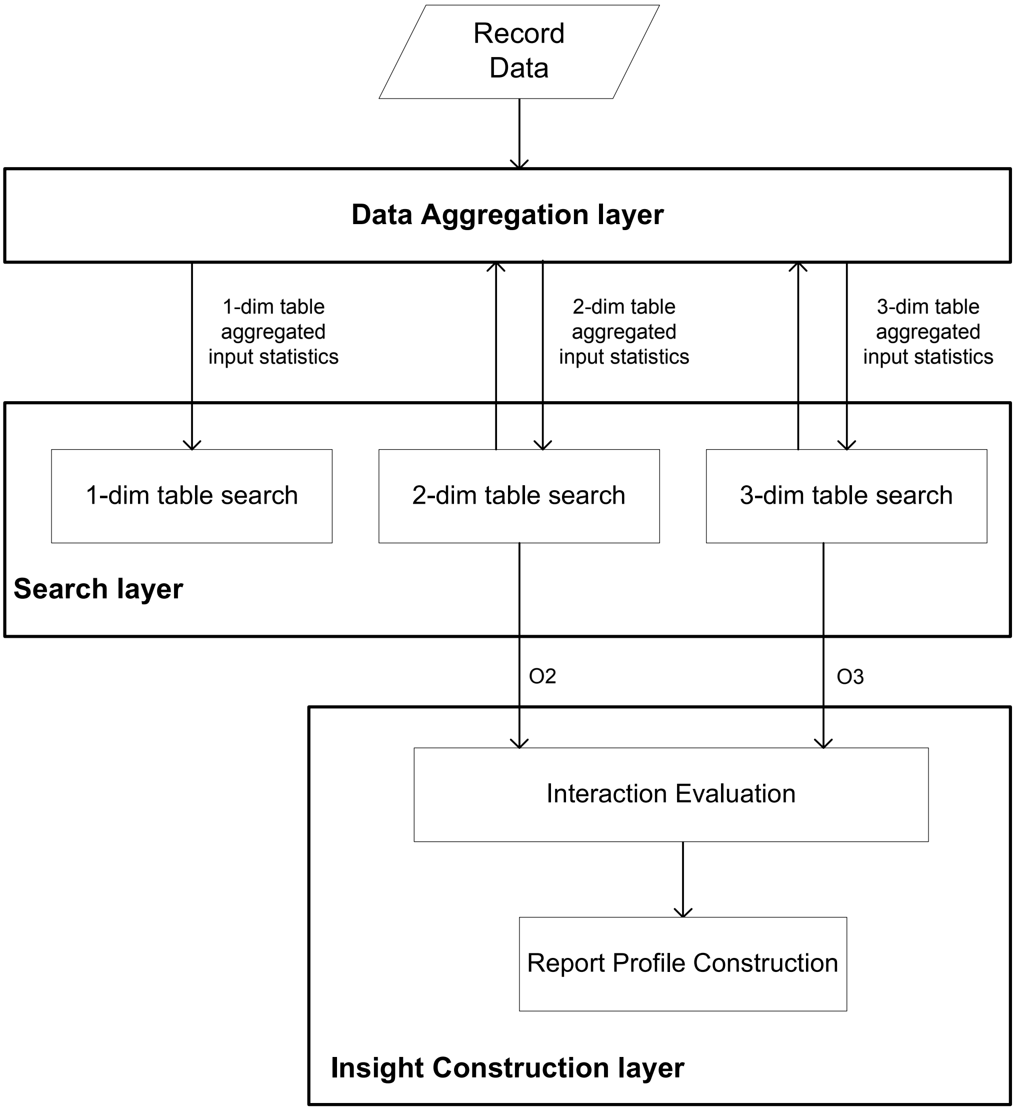
Record Data
Data Aggregation layer
1-dim table aggregated input statistics
2-dim table aggregated input statistics
3-dim table aggregated input statistics
1-dim table search
2-dim table search
3-dim table search
Search layer
O2
O3
Interaction Evaluation
Report Profile Construction
Insight Construction layer

Fig 1. The flow chart of the Interaction Discovery system

Record Data contains the measure of interest and a potentially large number of dimensions and would be the input for Data Aggregation layer.

Data Aggregation layer processes records from the Record Data unit and generates aggregated input statistics for the specified combinations of dimensions (e.g., for 1-dimenational tables, …, for 3-dimenional tables). These statistics can be generated in a single processing of the Record Data for the specified dimensional tables.
The number of all possible tables can overwhelm machine resources when a large number of dimensions is presented. The Search layer's task is to generate consecutive lists of limited number of tables for which the aggregated input statistics are generated by the Data Aggregation layer. There are three units in the Search layer. The goal of 1-dim table search unit is to find the most promising 1-dimensional tables for extension. Extension is described as processes of augmenting tables with an additional dimension. The goal of 2-dim table search unit is to find the most interesting 2-dimensional tables for

output and extension. The goal of 3-dim table search unit is to find the most interesting 3-dimensional tables for output. The details of each unit in the Search layer are described in Section 2.1.

After receiving all the available 2-dim and 3-dim tables from the Search layer, Interaction Evaluation unit in the Insight Construction layer will conduct ANOVA-based interaction effect tests, compute interaction effect size values, sort the tables with the significant interaction effect by the interaction effect size and export the top selected tables. Then Report Profile Construction unit in the Insight Construction layer generates summaried statistics, and interpretation with insights for selected 2 and 3 dimensional tables. The insights include detection of any table cells with high contribution to the interaction effect. Sections 2.2 and 2.3 describe Interaction Evaluation unit and Report Profile Construction unit, respectively.

2.1. Search Layer

The Search layer consists of the 1, 2, and 3-dim table search units. Each unit performs a search over 1, 2 and 3-dimensional tables generated by different categorical fields respectively. The inputs to all units are the aggregated input statistics from Data Aggregation layer. The statistics for each table are listed in Table 1 (assuming that there are *K* table cells in a table).

Table 1: Aggregated input statistics

Measure Y Table Dimension(s)	• The number of records corresponding to each table cell ($N_k, k=1,\dots,K.$) • The mean value of Y for the records corresponding to each table cell ($\bar{y}_k, k=1,\dots,K.$) • The sum of squares of Y for the records corresponding to each table cell ($C_k, k=1,\dots,K.$)

The search and sorting strategy employed in the 1, 2, and 3-dim table search units rely largely on the linear regression model based search statistics *Goodness of Fit*. The search statistics are computed based on the aggregated input statistics and described in Table 2.

Table 2: Model based search statistics

Goodness of Fit	$R^2 = 1 - SSE/SST.$
Summary statistics	• SST, the total sum of the squares in a linear model and it is computed as $SST = \sum_{k=1}^{K} C_k - \left(\sum_{k=1}^{K} N_k \bar{y}_k\right)^2 \Big/ \sum_{k=1}^{K} N_k.$ • SSE, the residual sum squares in a linear model and it is computed as $SSE = \sum_{k=1}^{K} C_k - \sum_{k=1}^{K} N_k \bar{y}_k^2.$

Other criterion, such as adjusted R^2, AICC (Akaike information criterion), etc., can be used. The detailed search and sorting strategy for 1, 2 and 3-dim table search units is illustrated in Figures 2, 3 and 4, respectively, with some descriptions provided after Figures.

2.1.1. 1-dimensional table search unit

Figure 2 illustrates the framework and the functional flow chart for the 1-dimensional table search unit.

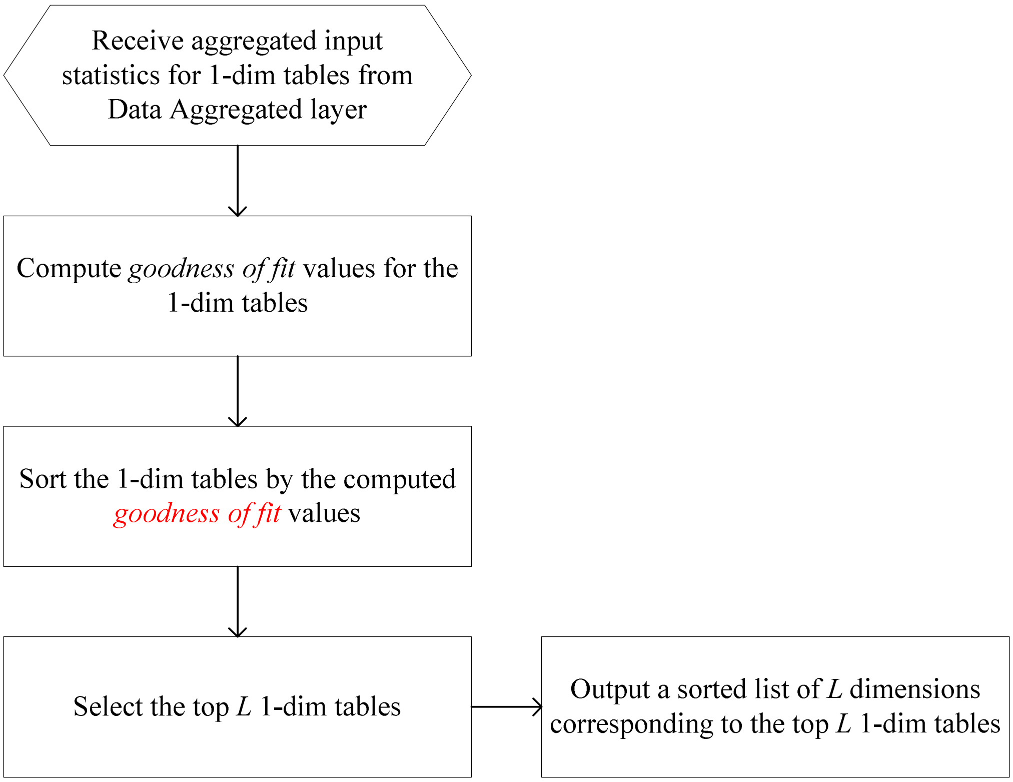
Receive aggregated input statistics for 1-dim tables from Data Aggregated layer
Compute *goodness of fit* values for the 1-dim tables
Sort the 1-dim tables by the computed *goodness of fit* values
Select the top *L* 1-dim tables
Output a sorted list of *L* dimensions corresponding to the top *L* 1-dim tables

Fig 2. The flow chart of the 1-dimensional table search unit

In Figure 2, the inputs to the 1-dim table search unit are aggregated input statistics listed in the Table 1 for each dimension. First, *Goodness of Fit* of the linear model is computed for each 1-dim table. These tables are then sorted by the *Fit* values. A sorted list of dimensions corresponding to the top *L* 1-dim tables is sent to the 2-dim table search unit. The number *L* is chosen so that the number of considered tables remains limited to conserve time and memory but it should be as large as possible for accuracy purposes.

2.1.2. 2-dimensional table search unit

Figure 3 illustrates the framework and the functional flow chart for the 2-dim table search unit.

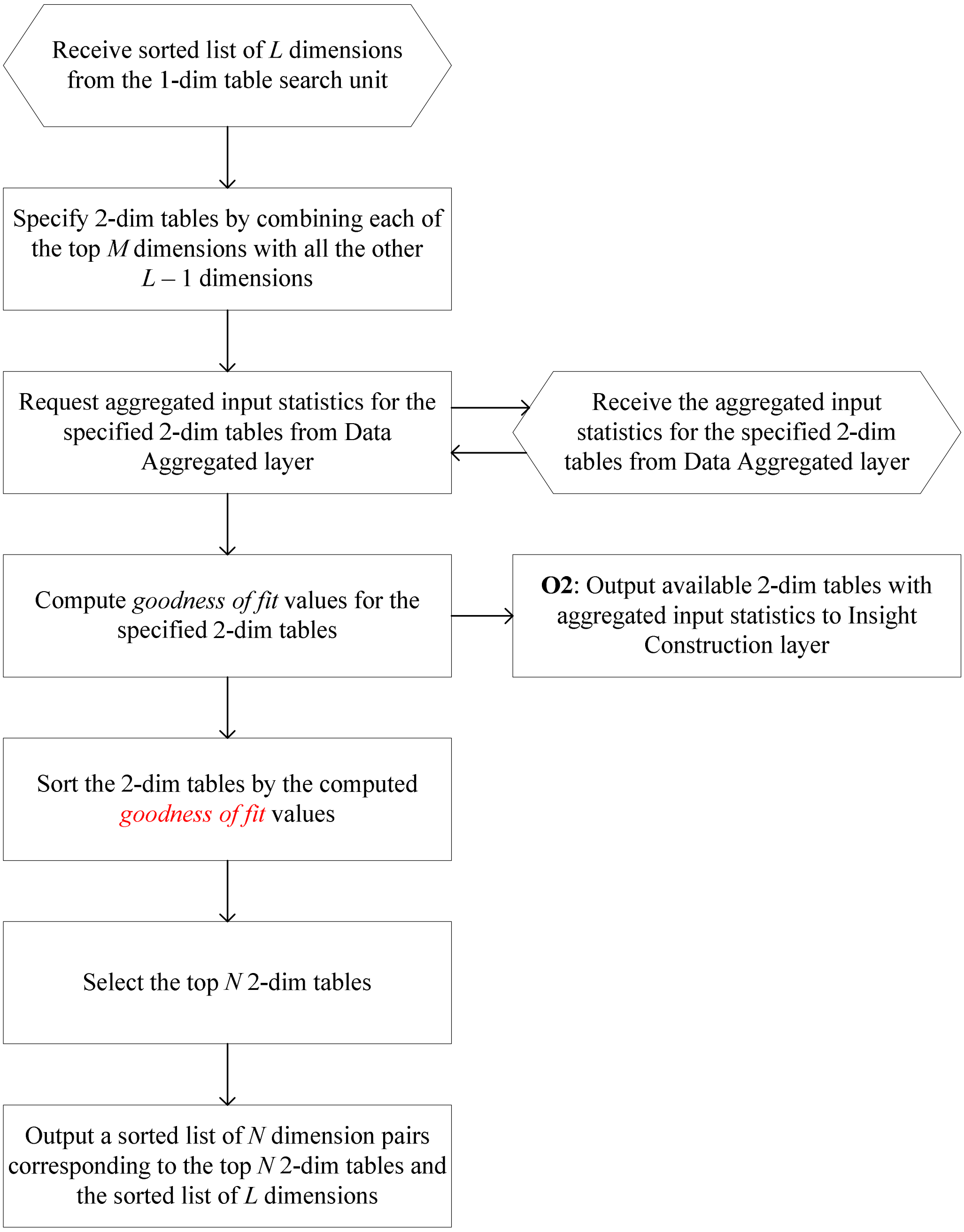
Receive sorted list of L dimensions from the 1-dim table search unit
Specify 2-dim tables by combining each of the top M dimensions with all the other $L - 1$ dimensions
Request aggregated input statistics for the specified 2-dim tables from Data Aggregated layer
Receive the aggregated input statistics for the specified 2-dim tables from Data Aggregated layer
Compute *goodness of fit* values for the specified 2-dim tables
O2: Output available 2-dim tables with aggregated input statistics to Insight Construction layer
Sort the 2-dim tables by the computed *goodness of fit* values
Select the top N 2-dim tables
Output a sorted list of N dimension pairs corresponding to the top N 2-dim tables and the sorted list of L dimensions

Fig 3. The flow chart of the 2-dimensional table search unit

In Figure 3, the 2-dim table search unit receives a sorted list of L dimensions from the 1-dim table search unit. Tables are specified by combining each of the top M dimensions with all the other $L - 1$ dimensions. The number M is normally smaller than L and it is chosen so that the total number of specified tables is limited to conserve computational resources. The 2-dim table search unit then request aggregated input statistics for the specified 2-dim tables from the Data Aggregation layer. The aggregated statistics needed are listed in Table 1. After receiving the aggregated statistics, *Goodness of Fit* values for all the specified 2-dim tables are computed according to the method in Table 2. The 2-

dim tables with aggregated input statistics are passed to the Insight Construction layer as illustrated by **O2**. Within the 2-dim table search unit, all available 2-dim tables are then sorted by *Fit*. The sorted list of dimension pairs corresponding to the top *N* 2-dim tables and the sorted list of *L* dimensions from the 1-dim table search unit are passed to the 3-dim search unit. The number *N* is chosen so that the number of considered 3-dim tables remains limited.

2.1.3. 3-dimensional table search unit

Figure 4 illustrates the framework and the functional flow chart for the 3-dim table search unit.

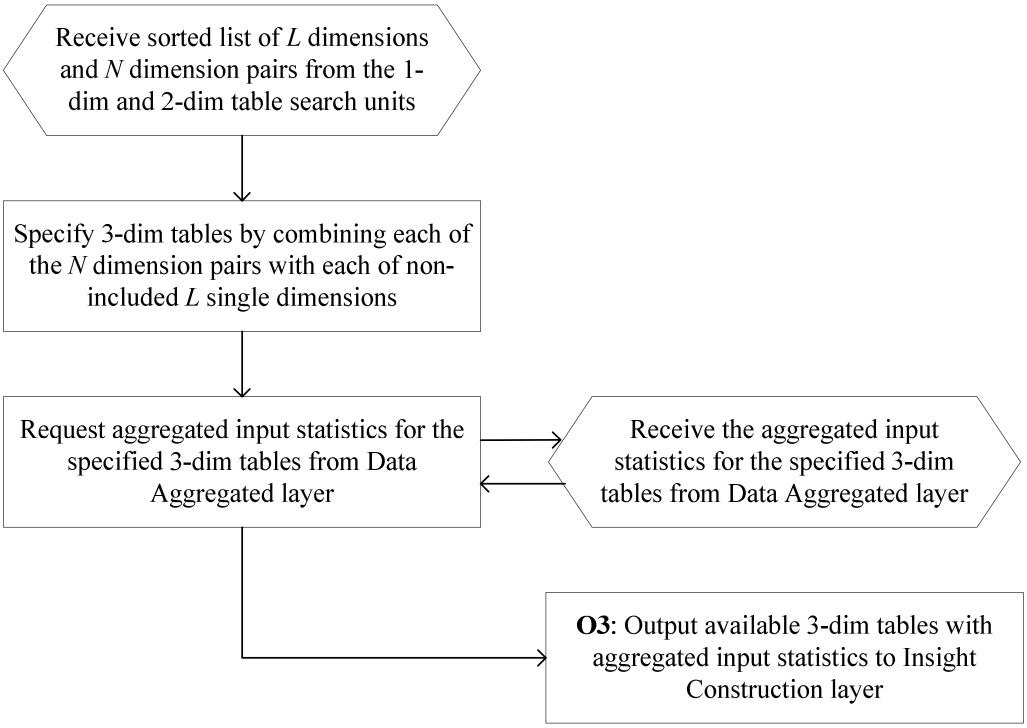
Receive sorted list of *L* dimensions and *N* dimension pairs from the 1-dim and 2-dim table search units
Specify 3-dim tables by combining each of the *N* dimension pairs with each of non-included *L* single dimensions
Request aggregated input statistics for the specified 3-dim tables from Data Aggregated layer
Receive the aggregated input statistics for the specified 3-dim tables from Data Aggregated layer
O3: Output available 3-dim tables with aggregated input statistics to Insight Construction layer

Fig 4. The flow chart of the 3-dimensional table search unit

In Figure 4, the 3-dim table search unit receives the sorted list of *L* dimensions from the 1-dim search unit and the sorted list of *N* dimension pairs from the 2-dim table search unit. The 3-dim tables are generated by combining each of the *N* dimension pairs with each of non-included *L* single dimensions. The non-included dimensions are ones that are not already in the 2-dimensional tables. The 3-dim table search unit then sends request to the Data Aggregation layer for aggregated input statistics for these 3-dim tables. The aggregated input statistics required are listed in the Table 1. The 3-dim tables with aggregated input statistics are passed to the Insight Construction layer, as illustrated by O3.

If K-dimensional tables are requested, where K > 3, then the 3-dim table search unit will also find some promising 3-dimensional tables for extension. The extension part for the 3 or higher dimensional tables is similar to that in the 2-dim table search unit.

2.2. Insight Construction Layer

The Insight Construction layer consists of the Interaction Evaluation unit and the Report Profile Construction unit. Section 2.2.1 describes the flow chart and the interaction indices used in the Interaction Evaluation unit. The model based summaries and the results of cell-by-cell analyses generated in the Report Profile Construction unit are given in Section 2.2.2.

2.2.1. Interaction Evaluation unit

Figure 5 illustrates the framework and the functional flow chart for the interaction evaluation unit within the Insight Construction layer.

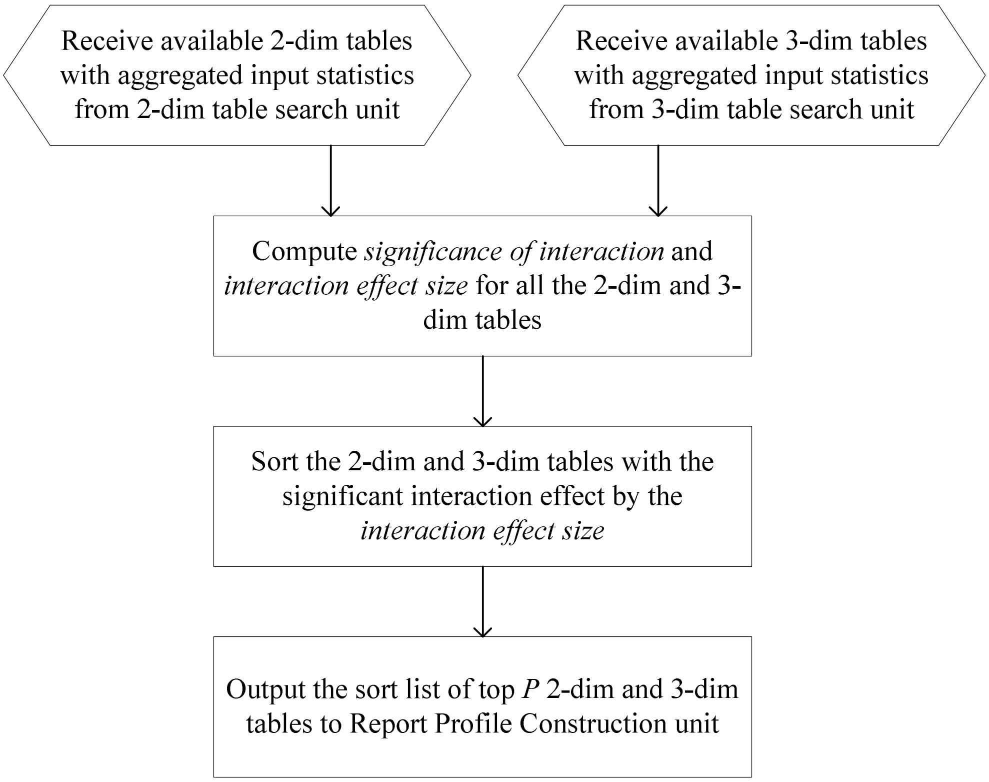
Receive available 2-dim tables with aggregated input statistics from 2-dim table search unit
Receive available 3-dim tables with aggregated input statistics from 3-dim table search unit
Compute *significance of interaction* and *interaction effect size* for all the 2-dim and 3-dim tables
Sort the 2-dim and 3-dim tables with the significant interaction effect by the *interaction effect size*
Output the sort list of top *P* 2-dim and 3-dim tables to Report Profile Construction unit

Fig 5. The flow chart of the Inaction Evaluation unit

In Figure 5, the inputs to the Interaction Evaluation unit are all the available tables with aggregated input statistics from the 2-dim table search unit and all the available tables with aggregated input statistics from the 3-dim table search unit. Upon reception of all inputs, *Significance of interaction* and *Interaction effect size* for all 2-dim and 3-dim tables are computed. Details are described in the Table 3. Tables with the significant interaction effect are then sorted by *Interaction effect size.* Top *P* tables are sent to the Report Profile Construction unit.

Interaction indices employed in the Interaction Evaluation unit are *Significance of interaction*, and *Interaction effect size* which are based on ANOVA tests. They are computed and applied to the 2 and 3-dimensional tables considered in the Insight Construction layer. Table 3 specifies the interaction indices in terms of various summaries derived from the aggregated input statistics. See Appendix to illustrate the computation of F statistic for a 2-dimensional table.

Table 3: Model based interaction indices

Index	Statistics
Significance of interaction	The p-value = $1-prob\left(F_{df_{interaction},df_e} \le F\right)$, where $F_{df_{interaction},df_e}$ is an F distribution with $df_{interaction}$ and df_e as numerator and denominator degrees of freedom, respectively; and $F=\dfrac{SS_{interaction}/df_{interaction}}{SSE/df_e}$, • $SS_{interaction}$ is the sum of squares for the interaction effect. See the iterative process below for details. • $df_{interaction}$ is degrees of freedom corresponding to $SS_{interaction}$. Typically, it is the product of the number of cells reduced by one in each dimension. • df_e is degrees of freedom corresponding to SSE. It equals the total number of records minus the number of cells in the table.
Interaction effect size	Eta squared = $SS_{interaction}/SST$.

To compute $SS_{interaction}$ based on aggregated input statistics we will follow the method used in Sarawagi et al. (1998) or Chen (1999). It is an iterative process which is described as follows:

(1) Compute the initial $SS_{interaction}$ from the means of Y and the number of records of all table cells.
(2) Compute marginal mean of Y for each cell of a single dimension by averaging Y over all table cells containing the same category in the given dimension.
(3) Update the cell means of Y in the whole table by subtracting the corresponding marginal mean from each cell.
(4) Repeat the steps (2) and (3) above for each dimension in the table.
(5) For tables with 3 dimensions, repeat the steps (2) and (3) for each pair of dimensions representing a marginal sub-table.
(6) Compute the current $SS_{interaction}$ from the updated cell means of Y, by multiplying the squared mean for each cell with the cell number of records to obtain SS_{cell} and summing over all cells in the table.
(7) Repeat steps (2) to (6) until the difference of $SS_{interaction}$ in two successive iterations is smaller than a preset tolerance value, and output the approximated $SS_{interaction}$ of the final iteration.

See Appendix to illustrate the iterative process for a 2-dimensional table.

2.2.2. Report Profile Construction unit

The top selected tables are further analyzed to determine the level of cell contribution to the interaction effect in Report Profile Construction unit. We call such analysis "influential cells detection".

The detection is based on a hypothesis test for each cell in the table: H_0: the k^{th} cell has no contribution to significance of the interaction effect; vs. H_A: the k^{th} cell has some high contribution to significance of the interaction effect. The test statistic is

$$W_k = \frac{SS_{cell,k}^{(t)}}{s^2},$$

where $SS_{cell,k}^{(t)}$ is the part from $SS_{interaction}^{(t)}$ corresponding to the k^{th} cell and $SS_{interaction}^{(t)}$ is the exported value from the above procedure; and s^2 is the estimated variance as

$$s^2 = \frac{SS_{interaction}^{(t)}}{K},$$

where K is the number of table cells. The test statistic has an asymptotic chi-squared distribution with 1 degree of freedom, i.e., $W_k \sim \chi_1^2$. See Appendix to illustrate the explanation of test statistic having a chi-squared distribution with 1 degree of freedom for a 2-dimensional table.

Then the p-value can be computed as

$$p = 1 - prob\left(\chi_1^2 \le W_k\right).$$

The k^{th} cell is influential if the p-value is smaller than a given threshold which is illustrated along with other model based summaries as the report profile template in Table 4.

Table 4: Report profile template

Aspect	Statistics	Interpretation and insights
Interaction effect	*Significance*	Interaction effect significance: $p <$ c$_{\text{eff-sig}}$: interaction is significant $p \geq$ c$_{\text{eff-sig}}$: interaction is not significant.
	Effect size	Interaction effect size: *Effect* < c$_{\text{eff-1}}$: strength is weak c$_{\text{eff-1}}$ ≤ *Effect* < c$_{\text{eff-2}}$: strength is moderate c$_{\text{eff-2}}$ ≤ *Effect* : strength is strong.
Table cell	*Cell influence*	Cell influence: $p \leq$ c$_{\text{cell-sig}}$: cell is influential $p >$ c$_{\text{cell-sig}}$: cell is not influential.

3. Conclusion

The proposed Relationship Discovery system has some unique features.

(1) It provides a structured and scalable search among any number of predetermined dimensions for interactions in the multitude of all possible low-dimensional aggregate tabular reports given a data set and a measure of interest.
(2) It provides model-based and efficient discovery of the strongest interaction effects in large data sets with a large number of dimensions by generating statistical models (ANOVA) for analyzing aggregate tables with two or more dimensions and the measure summaries; applying goodness-of-fit to select the best candidate tables and generate tables with additional dimensions; and computing the significance of interaction and interaction effect size values among the table dimensions with respect to the measure.
(3) It not only covers dimension reduction, but also detects interaction effects and influential table cells based on only basic statistics.

Appendix

We will use a 2-dimensional table as an example to illustrate (1) the computation of the interaction effect ANOVA test statistic, (2) the iterative process of computing $SS_{interaction}$, and (3) the explanation of test statistic having a chi-squared distribution with 1 degree of freedom in influential cells detection.

For 2-dimensional tables, suppose the first dimension $\boldsymbol{X}_1$ has *R* categories and the second dimension $\boldsymbol{X}_2$ has *S* categories. The full model would have two main effects, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and an interaction effect, $\boldsymbol{X}_1 \times \boldsymbol{X}_2$, while the reduced model has only $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The interaction effect ANOVA test has the null hypothesis: $H_0 : \boldsymbol{\beta}_3 = \mathbf{0}$, where $\boldsymbol{\beta}_3$ is the parameter vector corresponding to the interaction $\boldsymbol{X}_1 \times \boldsymbol{X}_2$ effect and the test statistic is

$$F = \frac{SS_{interaction} \big/ df_{interaction}}{SSE / df_e},$$

where $SS_{interaction} = SS_{e,reduced} - SS_{e,\,full}$, and $SS_{e,reduced}$ and $SS_{e,\,full}$ are residual sum of squares for the reduced and full models, respectively. We can fit two models to compute these two residual sums of squares values, but it would not be feasible or efficient when there are a large number of dimensions. That's why we propose to use the aggregated input statistics to compute $SS_{interaction}$ in an iterative process (see below) for each 2-dim table. SSE is residual sum of squares for the full model, so $SSE = SS_{e,\,full} = \sum_{i=1}^{R}\sum_{j=1}^{S} C_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{S} N_{ij}\bar{y}_{ij}^2$, where N_{ij}, $\bar{y}_{ij} = \frac{1}{N_{ij}}\sum_{\ell=1}^{N_{ij}} y_{ij,\ell}$ and $C_{ij} = \sum_{\ell=1}^{N_{ij}} \left(y_{ij,\ell}\right)^2$ are the number of records, the mean of *Y* and the sum of squares of *Y* in each table cell (*i*, *j*), $i = 1,\ldots,R$ and $j = 1,\ldots,S$, respectively. Two degrees of freedom values are $df_e = N - RS$ and $df_{interaction} = \left(R-1\right)\left(S-1\right)$, where *N* is number of total records, $N = \sum_{i=1}^{R}\sum_{j=1}^{S} N_{ij}$. Please note that both degrees of freedom are under the assumption that none of table cells is empty, otherwise they would be subtracted by the number of empty table cells.

The iterative process of computing $SS_{interaction}$ is as follows:

(1) Input values for T (maximum number of iterations) and ε (tolerance level of stopping criterion).
Compute the initial $SS_{interaction}$ as $SS_{interaction}^{(0)} = \sum_{i=1}^{R}\sum_{j=1}^{S} N_{ij}\left(\overline{y}_{ij}^{(0)}\right)^2$, where $\overline{y}_{ij}^{(0)}$ is the mean of Y corresponding to the table cell (i, j).
Set the iteration number $t = 1$.

(2) Compute marginal means of Y with respect to $\boldsymbol{X}_1$ (row marginal means) as $\overline{y}_{i\bullet}^{(t-1)} = \sum_{j=1}^{S} N_{ij}\overline{y}_{ij}^{(t-1)} \Big/ \sum_{j=1}^{S} N_{ij}$, $i = 1,\ldots,R.$

(3) Update the cell means of Y as $\overline{y}_{ij}^{*} = \overline{y}_{ij}^{(t-1)} - \overline{y}_{i\bullet}^{(t-1)}$, $i = 1,\ldots,R.$

(4) Compute marginal means of Y with respect to $\boldsymbol{X}_2$ (column marginal means) as $\overline{y}_{\bullet j}^{(t-1)} = \sum_{i=1}^{R} N_{ij}\overline{y}_{ij}^{*} \Big/ \sum_{i=1}^{R} N_{ij}$, $j = 1,\ldots,S$; and update the cell means of Y as $\overline{y}_{ij}^{(t)} = \overline{y}_{ij}^{*} - \overline{y}_{\bullet j}^{(t-1)}$, $j = 1,\ldots,S.$

(5) Compute the current $SS_{interaction}$ as $SS_{interaction}^{(t)} = \sum_{i=1}^{R}\sum_{j=1}^{S} SS_{cell,ij}^{(t)} = \sum_{i=1}^{R}\sum_{j=1}^{S} N_{ij}\left(\overline{y}_{ij}^{(t)}\right)^2.$

(6) If $\left|SS_{interaction}^{(t)} - SS_{interaction}^{(t-1)}\right| < \varepsilon$ or $t \geq T$, then stop and output $SS_{interaction}^{(t)}$. Otherwise, set $t = t + 1$ and go back to step (2).

The expected mean for each table cell in the reduced model can be written as $\hat{\alpha}_i + \hat{\beta}_j$, $i = 1,\ldots,R$ and $j = 1,\ldots,S$, where $\hat{\alpha}_i$ corresponds to $\boldsymbol{X}_1$ and $\hat{\beta}_j$ corresponds to $\boldsymbol{X}_2$. In fact, $\hat{\alpha}_i$ and $\hat{\beta}_j$ would be the sums of row marginal means and column marginal means over all the iterations from the above iterative process and they are equivalent to those computed by the traditional least square method. On the other hand, the expected mean for each table cell in the full model is the observed mean of Y, y_{ij}, $i = 1,\ldots,R$ and $j = 1,\ldots,S.$ Based on them, we give an explanation why the test statistic in the influential cells detection process has a chi-squared distribution with 1 degree of freedom.

The hypothesis test (H_0: the k^{th} cell has no contribution to significance of interaction effect; vs. H_A: the k^{th} cell has some high contribution to significance of interaction effect) can be translated to a more formal form:

$$H_0 : E\left(\overline{y}_{ij}\right) = \alpha_i + \beta_j,\ i = 1,\ldots,R \text{ and } j = 1,\ldots,S, \text{ vs.}$$
$$H_A : \text{ at least one cell such that } E\left(\overline{y}_{ij}\right) \neq \alpha_i + \beta_j.$$

For a linear model, each Y value is assumed to have a normal distribution with a constant variance σ^2. Further, if H_0 holds, then

$$\overline{y}_{ij} \sim N\left(\alpha_i + \beta_j, \frac{\sigma^2}{N_{ij}}\right) \Rightarrow \frac{N_{ij}\left(\overline{y}_{ij} - \left(\alpha_i + \beta_j\right)\right)^2}{\sigma^2} \sim \chi_1^2.$$

In general, σ^2, α_i and β_j are not known, we replace them with their estimated values, denoted as s^2, $\hat{\alpha}_i$ and $\hat{\beta}_j$, respectively. Then we notice from the above iterative process of computing $SS_{interaction}$ that

$$N_{ij}\left(\overline{y}_{ij}-\left(\hat{\alpha}_i+\hat{\beta}_j\right)\right)^2=N_{ij}\left(\overline{y}_{ij}^{(t)}\right)^2=SS_{cell,ij}^{(t)} \text{ and}$$

$$s^2=\frac{1}{R\times S}\sum_{i=1}^{R}\sum_{j=1}^{S}N_{ij}\left(\overline{y}_{ij}-\left(\hat{\alpha}_i+\hat{\beta}_j\right)\right)^2=\frac{1}{R\times S}SS_{interaction}^{(t)}.$$

Then the test statistic would be

$$W_{ij}=\frac{SS_{cell,ij}^{(t)}}{s^2},$$

and its asymptotic distribution is χ_1^2.

References

Chen, Q. (1999), *Mining Exceptions and Quantitative Association Rules in OLAP Data CUBE*. Unpublished master thesis, the School of Computing Science, Simon Fraser University, Canada.

Sarawagi, S., Agrawal, R. and Megiddo, N. (1998), *Discovery-driven Exploration of OLAP Data Cubes*. Research Report, Almaden Research Center, IBM Research Division.