Reinventing the Contingency Wheel: Scalable Visual Analytics of Large Categorical Data

Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and M. Eduard Gröller

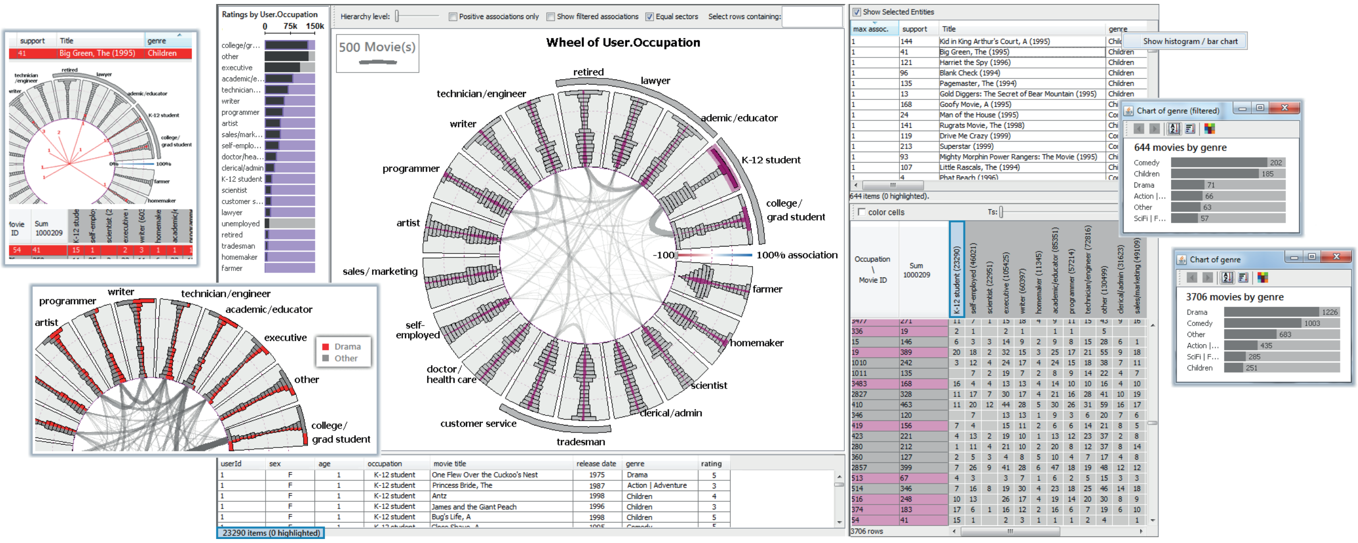


Fig. 1. Contingency Wheel++ uses complementing visual representations and a multi-level overview+detail user interface to enable rich exploratory analysis of large categorical data. The example above shows information about 1 million user ratings on 3706 movies.

Abstract—Contingency tables summarize the relations between categorical variables and arise in both scientific and business domains. Asymmetrically large two-way contingency tables pose a problem for common visualization methods. The Contingency Wheel has been recently proposed as an interactive visual method to explore and analyze such tables. However, the scalability and readability of this method are limited when dealing with large and dense tables. In this paper we present Contingency Wheel++, new visual analytics methods that overcome these major shortcomings: (1) regarding automated methods, a measure of association based on Pearson's residuals alleviates the bias of the raw residuals originally used, (2) regarding visualization methods, a frequency-based abstraction of the visual elements eliminates overlapping and makes analyzing both positive and negative associations possible, and (3) regarding the interactive exploration environment, a multi-level overview+detail interface enables exploring individual data items that are aggregated in the visualization or in the table using coordinated views. We illustrate the applicability of these new methods with a use case and show how they enable discovering and analyzing nontrivial patterns and associations in large categorical data.

Index Terms—Large categorical data, contingency table analysis, information interfaces and representation, visual analytics.

1 Introduction

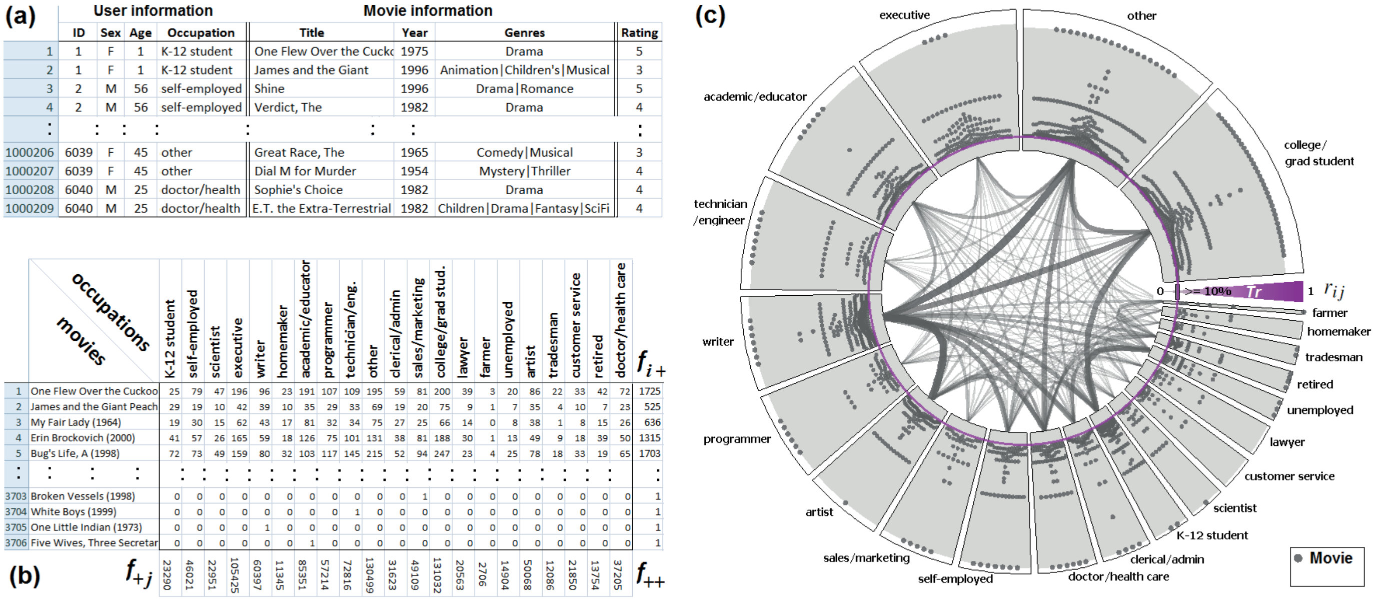
Many problems in scientific domains such as medicine, biology and pharmacology, as well as in business domains such as retail and logistics require analyzing associations between categorical variables. For example, a movie retailer might be interested in associations between movies and users based on sales data with the goal of optimizing marketing strategies. The discrete nature of categorical data and their lack of an inherent similarity measure pose significant challenges to the fields of information visualization [2] and data mining [42]. Contingency tables (also known as crosstabs) are a common way to summarize categorical data as a first step of analysis. A two-way contingency table is an $n \times m$ matrix that records the frequency of observations f_{ij} for each combination of categories of two categorical variables. Many data analysis frameworks such as KNIME [4], WEKA [11] and R [28] offer possibilities to create and analyze contingency tables. One of the best-known statistical tests for the overall association (or independence) between two categorical variables is Pearson's χ^2 test [30]. It assesses the significance of associations between the categories of the two variables. However, it does not provide information about how single pairs of categories are associated.

Several visualization methods were developed to analyze associated categories in contingency tables. As we discuss in Sect. 5, these methods are designed to handle rather small tables having few categories. However, often much larger contingency tables need to be analyzed, which poses a problem to these methods. Figure 2a shows large categorical data from the MovieLens data set [10]. It contains about one million user ratings on movies. For each user, it provides his or her occupation, sex, and age group, and for each movie, its release date and genres. Examples for tables extracted from this data set are:

- A 3706 × 21 table which counts for each movie, how many times it was rated from users of each occupation (figure 2b).
- A 6040 × 17 table which counts for each user, how many times he/she rated movies from each genre (figure 6d).

- *Bilal Alsallakh, Wolfgang Aigner, Silvia Miksch, and Eduard Gröller are with Vienna University of Technology, E-mail: bilal@cvast.tuwien.ac.at, {aigner, miksch}@ifs.tuwien.ac.at, and groeller@cg.tuwien.ac.at*

*Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*



(a)

	User information				Movie information			
	ID	Sex	Age	Occupation	Title	Year	Genres	Rating
1	1	F	1	K-12 student	One Flew Over the Cuckc	1975	Drama	5
2	1	F	1	K-12 student	James and the Giant	1996	Animation|Children's|Musical	3
3	2	M	56	self-employed	Shine	1996	Drama|Romance	5
4	2	M	56	self-employed	Verdict, The	1982	Drama	4
:	:	:	:	:	:	:	:	:
1000206	6039	F	45	other	Great Race, The	1965	Comedy|Musical	3
1000207	6039	F	45	other	Dial M for Murder	1954	Mystery|Thriller	4
1000208	6040	M	25	doctor/health	Sophie's Choice	1982	Drama	4
1000209	6040	M	25	doctor/health	E.T. the Extra-Terrestrial	1982	Children|Drama|Fantasy|SciFi	4

(b)

	occupations / movies	K-12 student	self-employed	scientist	executive	writer	homemaker	academic/educator	programmer	technician/eng.	other	clerical/admin	sales/marketing	college/grad stud.	lawyer	farmer	unemployed	artist	tradesman	customer service	retired	doctor/health care	f_{i+}
1	One Flew Over the Cuckoo	25	79	47	196	96	23	191	107	109	195	59	81	200	39	3	20	86	22	33	42	72	1725
2	James and the Giant Peach	29	19	10	42	39	10	35	29	33	69	19	20	75	9	1	7	35	4	10	7	23	525
3	My Fair Lady (1964)	19	30	15	62	43	17	81	32	34	75	27	25	66	14	0	8	38	1	8	15	26	636
4	Erin Brockovich (2000)	41	57	26	165	59	18	126	75	101	131	38	81	188	30	1	13	49	9	18	39	50	1315
5	Bug's Life, A (1998)	72	73	49	159	80	32	103	117	145	215	52	94	247	23	4	25	78	18	33	19	65	1703
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:		:
3703	Broken Vessels (1998)	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
3704	White Boys (1999)	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
3705	One Little Indian (1973)	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3706	Five Wives, Three Secretar	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	f_{+j}	23290	46021	22951	105425	60397	11345	85351	57214	72816	130499	31623	49109	131032	20563	2706	14904	50068	12086	21850	13754	37205	f_{++}

Fig. 2. (a) Categorical variables of the MovieLens data set [10] showing about one million user ratings on 3706 movies, (b) the contingency table of the variables “movie title” and “user occupation”, (c) the Contingency Wheel of the table in (b): Sectors represent occupations and dots represent movies positively associated with them. Thicker arcs show which occupations share more movies highly associated with both of them.

The Contingency Wheel [1] has been introduced as an interactive visual method for exploring positive associations in asymmetrically large tables. The column categories are visualized as sectors of a ring chart and the table cells are depicted as dots in these sectors (figure 2c). The dot for cell (i, j) is placed in sector i at a radial distance from the ring's inner circle proportional to the strength of association r_{ij} between row i and column j. A layout algorithm calculates the angular positions of the dots in each sector to reduce occlusion. It copes with a large number of rows by visualizing only the cells that represent significant associations r_{ij}, determined by adjustable thresholds. An arc is drawn between two sectors if one or more rows have dots in both sectors. This arc is thicker if more such rows exist and if their dots represent higher associations with both sectors. User interaction enables analyzing different types of associations in large tables.

Scalability is one of the major challenges visual analytics aims to address [39]. The wheel metaphor explained above has several shortcomings which degrade its readability and scalability, especially with large and dense tables (Sect. 2). In this paper we propose Contingency Wheel++: new visual analytics methods that tackle the issues of the original wheel. Our methods (described in Sect. 3) address its computational component, visual representation and interactive interface, and intertwine these three components to enable scalable analysis of categorical data. The new methods encompass:

- *Automated methods*: a new association measure results in a better distribution of the dots to sectors of different sizes. This is important when analyzing large tables that often exhibit high skewness in the distribution of their frequencies.
- *Visualization methods*: a frequency-based abstraction of the dots eliminates overlapping which allows showing all the cells, instead of just small subsets thereof. This enables analyzing and querying both positive and negative associations.
- *Interactive exploration environment*: an overview+detail interface allows exploring individual items aggregated in the visualization or in the table, and analyzing their attributes.

In Sect. 4 we present a use case to illustrate how our new methods can be used to explore the MovieLens data set. We show how nontrivial patterns and associations in the data can be discovered. In Sect. 5 we compare our approach with other methods for visualizing categorical data and elaborate on its scalability.

2 Limitations of the Contingency Wheel

Based on a pilot evaluation study of the Contingency Wheel [23], we identified several issues that limit its readability and scalability. In particular, we focus on issues related to the conceptual design of the wheel and its interpretability rather than usability issues:

Data mapping: The Contingency Wheel visualizes association values r_{ij} that represent deviations from expected values (Sect. 3.1) rather than absolute frequencies f_{ij}. Many users did not have sufficient background on statistical association measures to interpret that correctly.

Visual mapping: Users agreed that the visualization provides a quick overview of the distribution of dots within sectors as compared with a tabular view. However, they found it difficult to accurately interpret the meaning of these dots at the beginning. They expected absolute frequencies f_{ij} rather than association values. It was confusing that the dot size and its radial position convey the same information. The angular position of the dots was even more confusing since it bears no meaning. It was also confusing that dots in different sectors can represent the same entities. Though arcs are intended to clarify this fact, users realized it only after selecting a dot (which also highlights all dots in other sectors that represent the same row).

Interaction: Dots closer to the center were often too small and overlapping, which made them difficult to identify. The same issue applies to arcs between small sectors. Also, filtering the dots by moving a slider became clear only after the users understood the data representation. Some users forgot that parts of the dots were filtered out and drew wrong conclusions about the data.

Most of the above-mentioned readability issues are related to dots. Dots as representations of individual table cells suffer inherently from limited scalability: Only a few hundred dots can be shown at once without overlapping. The Contingency Wheel reduces the large number of dots by filtering out cells (i, j) with $r_{ij} \leq T_r$ (where T_r is the association threshold) and by filtering out entire rows with $f_{i+} < T_s$ (where T_s is the support threshold) [1]. However, filtering limits the ability to gain insights into the whole dataset and it does not work well for dense tables with large f_{ij} values.

Contingency Wheel++ [1] improves both on the readability and on the scalability issues mentioned above by employing visual analytics methods as presented in the next section.

[1] A prototype implementation of Contingency Wheel++ is available at http://www.cvast.tuwien.ac.at/wheel

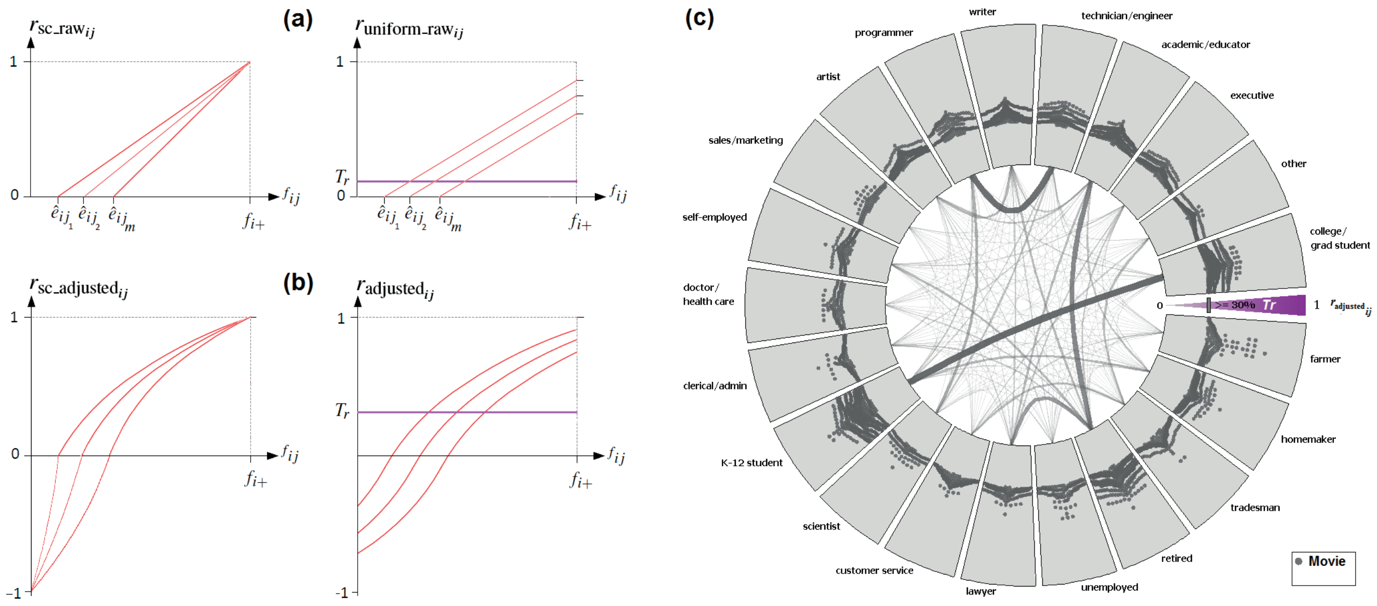


Fig. 3. (a) raw residuals and (b) adjusted residuals plotted as a function of f_{ij} for different values of f_{+j} with both nonuniform- (left) and uniform scaling (right), (c) the same data plotted in figure 2c using uniformly-scaled adjusted residuals instead of raw residuals (with $T_r = 30\%$ and $T_s = 1$).

3 Contingency Wheel++

In the following, f_{ij} denotes the frequency in cell (i,j), $f_{i+} = \sum_{j=1}^{m} f_{ij}$ and $f_{+j} = \sum_{i=1}^{n} f_{ij}$ are the marginal row- and column frequencies, and f_{++} is the sum of all table frequencies (figure 2b). We first address the data mapping employed by Contingency Wheel++ (Sect. 3.1). Then we propose a frequency-based visual representation which abstracts the dots (Sect. 3.2). In Sect. 3.3 we show how an interactive visual interface integrates additional table views to bridge the gap between the data representation and the visual representations and to support a flexible visual exploration process.

3.1 Mapping Frequencies to Associations

The main goal of Contingency Wheel++ is to reveal how the row categories of a contingency table are associated with its column categories. For this purpose, it uses a statistical measure r_{ij} that computes the association between row i and column j based on f_{ij} and takes value in the range $[-1,1]$. This measure is usually based on statistical residuals between the actual frequency f_{ij} and expected frequencies $\hat{e}_{ij}$. The frequency in cell (i,j) predicted under the null hypothesis H_0, i.e., assuming no association, is [37]:

$$\hat{e}_{ij} = \frac{f_{i+} \cdot f_{+j}}{f_{++}} \tag{1}$$

If $f_{ij} = \hat{e}_{ij}$ holds for cell (i,j), its share f_{ij}/f_{i+} of the marginal row frequency is equal to the column's share f_{+j}/f_{++} of all table frequencies. This means that row i is neither positively nor negatively associated with column j, and corresponds to a zero association value $r_{ij} = 0$. Cells with $f_{ij} > \hat{e}_{ij}$ indicate a positive association between row i and column j. Statistical residuals r_{ij} can be used to quantify this association. They can be designed to incorporate a priori information about the data and their distribution. In the following we describe the originally-used residuals and our improvements on them.

3.1.1 Raw residuals

The association measure used originally by the Contingency Wheel is based on raw residuals $(f_{ij} - \hat{e}_{ij})$ [1]. To generate association values $r_{ij} \leq 1$, the raw residual for cell (i,j) is divided by the maximum value it can take $(f_{i+} - \hat{e}_{ij})$:

$$r_{\text{sc\_raw}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{f_{i+} - \hat{e}_{ij}} \tag{2}$$

This measure maps frequencies linearly to association values (figure 3a-left). The maximum association $r_{ij} = 1$ is reached when all cells of row i have zero frequencies except for cell (i,j). For such a row, only one dot is created on the outer boundary of sector j. A cell with $r_{ij} = 0$ creates a dot on the inner boundary of sector j (assuming no thresholds). Cells with negative associations are ignored. The above-mentioned normalization is not uniform with respect to the columns: For row i, different scaling factors are used in different columns, because the expected frequency $\hat{e}_{ij}$ is larger for columns with larger f_{+j}. This makes better use of the sector area for revealing the distribution of dots along the radial dimension. Also, rows i that are fully associated with column j ($f_{ij} = f_{i+}$) can be easily found as dots at the outer boundary. However, the different scaling factors result in a bias especially when f_{+j} varies largely between sectors. This impacts the comparison of associations between different sectors and reduces the expressivity of the arcs. A uniform scaling factor for all columns can be used instead:

$$r_{\text{uniform\_raw}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{f_{i+}} \tag{3}$$

Figure 3a-right shows how this scaling maps frequencies to associations. For cells with $f_{ij} = f_{i+}$, Eq. 3 evaluates to $1 - f_{+j}/f_{++}$ which is independent of i. Such cells are hence mapped to the same radial distance within a sector (figure 2c). The sectors are scaled by their marginal frequencies. Sectors with larger f_{+j} values attract more dots than sectors with smaller f_{+j} values, due to an inherent statistical bias that raw residuals suffer from (even with uniform scaling).

3.1.2 Adjusted residuals

Standardized Pearson residuals [37] avoid the bias of raw residuals by adjusting the variance of the r_{ij} values to $N(0,1)$:

$$r_{\text{pearson}_{ij}} = \frac{f_{ij} - \hat{e}_{ij}}{\sqrt{\hat{e}_{ij} \cdot (1 - f_{i+}/f_{++}) \cdot (1 - f_{+j}/f_{++})}} \tag{4}$$

We use a logarithmic scale for the visual mapping of these residuals to better reveal their distribution along the radial dimension (where cte is a constant computed from the table to ensure $-1 \leq r_{ij} \leq 1$):

$$r_{\text{adjusted}_{ij}} = \frac{\operatorname{sgn}(r_{\text{pearson}_{ij}})}{cte} \cdot \ln\left(1 + \left| r_{\text{pearson}_{ij}} \right|\right) \tag{5}$$

Figure 3b-right, shows how this measure maps frequencies to associations. Figure 3c shows the same data as in figure 2c using $r_{ij} = r_{\text{adjusted}_{ij}}$ with $T_r = 30\%$ and with equal sectors. The dots are distributed more uniformly among the sectors. This results in arcs that suggest other similarities between occupations. The logarithmic scale amplifies smaller raw residuals, giving them more visual prominence. This potentially generates more dots, and hence a higher value for T_r is needed to filter out insignificant associations. Cells with $f_{ij} = f_{i+}$ are mapped to different radial distances in sector j, depending on f_{i+}. This makes the arcs more robust to changes in T_s since rows with smaller f_{i+} values contribute less to the arcs. On the other hand, these cells are somewhat difficult to locate. The following nonuniform scaling stretches r_{ij} to the range $[-1,1]$:

$$r_{\text{sc\_adjusted}_{ij}} = \frac{r_{\text{adjusted}_{ij}}}{\max\left(s_{ij} \cdot r_{\text{adjusted}_{ij}}|_{f_{ij}=f_{i+}}, s_{ij} \cdot r_{\text{adjusted}_{ij}}|_{f_{ij}=0}\right)} \qquad (6)$$

where $s_{ij} = \operatorname{sgn}(r_{\text{adjusted}_{ij}})$ and $r_{\text{adjusted}_{ij}}|_{f_{ij}=x}$ is the value $r_{\text{adjusted}_{ij}}$ would take if $f_{ij} = x$. Figure 3b-left depicts how this scaling maps frequencies. As can be seen, rows with $f_{ij} = f_{i+}$ are always mapped to the largest radial distance. Also, if the visualization can include negative associations, rows with $f_{ij} = 0$ are always mapped to the lowest radial distance. Nonuniform scaling, however, re-introduces a small bias in the associations, toward columns with larger f_{+j}.

3.2 Visualizing the Contingency Table

The visualization aims to reveal how the row categories of a contingency table are associated with its column categories, based on the association measure used. Our new visual representation makes use of the advantages of uniformly adjusted residuals (Sect. 3.1.2). It provides a clearer and more intuitive visualization of the table, as compared to the original wheel design [1]. Moreover, depending on the user's choice, it enables showing all associations or positive associations only as we describe in the following subsections.

3.2.1 Visualizing columns

Like in the original wheel metaphor, columns are drawn as sectors of a ring chart. The main difference is that they are drawn with equal size. This has several advantages: First, this is in accordance with the fact that adjusted residuals result in a more uniform distribution of the cells to the sectors. Second, by using a frequency-based representation (Sect. 3.2.2), the distribution of the associations can be compared between different sectors. Third, the arcs become evenly distributed in the central area, unlike the arcs in figure 2c which overlap more near small sectors. Finally, column categories are treated equally from a visual point of view, in the same way as the dimensions of a star plot [12]. This simplifies the visualization and eliminates confusion about the meaning of different sector sizes. The information of different column marginal frequencies f_{+j} is conveyed in a linked bar chart (Sect. 3.3). Incorporating it in the wheel representation would not contribute to the goal of Contingency Wheel++, i.e., to explore associations.

3.2.2 Visualizing row-column associations

The radial dimension of the ring chart linearly encodes the association values r_{ij} computed by one of the association measures. The outer boundary corresponds to $r_{ij} = 1$. The inner boundary corresponds to $r_{ij} = -1$ if showing all associations, and to $r_{ij} = 0$ if showing positive associations only. Instead of the dot representation originally used, we suggest a frequency-based representation to visualize the row-column associations. A histogram H_j is created in each sector j to show the distribution of the associations r_{ij} along the radial dimension. An adjustable number b of equal bins is used for all histograms, initially determined by Scott's normal reference rule [34]. Each bin k in sector j aggregates the rows i having associations in the interval $I_k = [l_k, l_{k+1}[$. The interval boundaries l_k are equally spaced between $[-1,1]$:

$$l_k = \frac{2(k-1)-b}{b} \qquad (7)$$

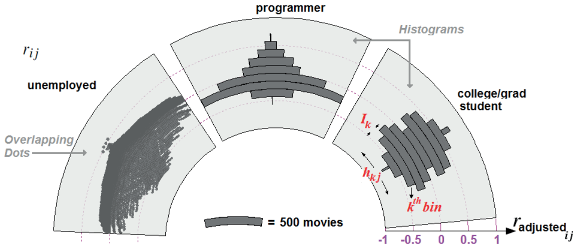


Fig. 4. Dot vs. histogram representation of row-column associations. The dimensions of a histogram bin are annotated (Eq. 8).

A closed interval $I_b = [l_b, 1]$ is used for the last bin to account for $r_{ij} = 1$. Hence, the number of items h_{kj} in the k^{th} bin of sector j is:

$$h_{kj} = \left|\left\{1 \le i \le n : f_{i+} \ge T_s \wedge r_{ij} \in I_k\right\}\right| \qquad (8)$$

Each bin k of histogram H_j is visualized as a track in sector j. This track occupies the radial position which corresponds to its interval I_k. The length of this track is proportional to h_{kj}. A uniform or sector-specific scaling factor ensures that all tracks fit in their sectors. Tracks are centered in their sectors, following the Gestalt principle of symmetry [41]. This avoids artificial asymmetry along the angular dimension in the sectors and makes it easier to compare their histograms. Figure 4 shows both dot and histogram representations for some sectors of figure 3c. The histograms show how 3706 movies are associated with 2 occupations. Both positive and negative associations are included.

Rows whose associations with sector j lie in a specific interval can be inspected individually along with the attributes of their entities, as explained in Sect. 3.3. The distribution of a numerical or categorical attribute of these entities can be shown by coloring the histograms instead of coloring individual dots. This provides a clearer understanding of the attribute distribution at different radial distances. Figure 5a shows the release-date distribution of movies positively associated with specific occupation categories. Figure 5b shows the genres of the movies. Movies highly associated with the "Retired" category tend to be old. The opposite holds for the "K-12 student" category which also tends to be highly associated with "Children" movies. Movies highly associated with "Technician/Engineer" are more likely to have "Sci-Fi / Fantasy" genres. These tendencies seem stronger as compared to the distribution of both attributes among all movies (figure 5c).

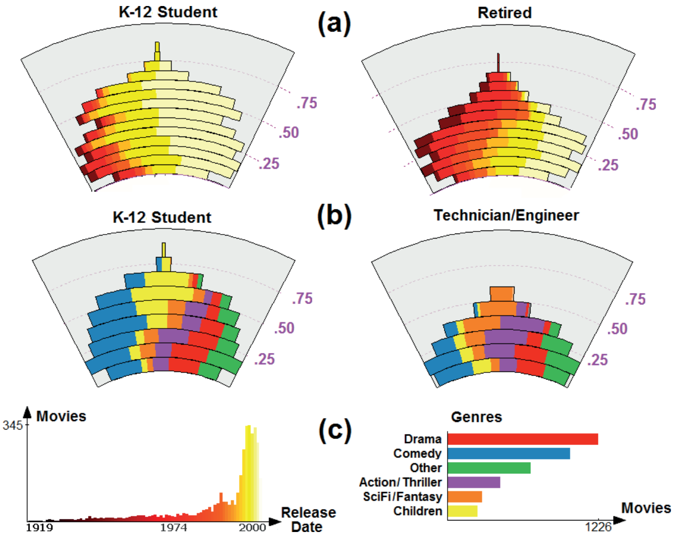
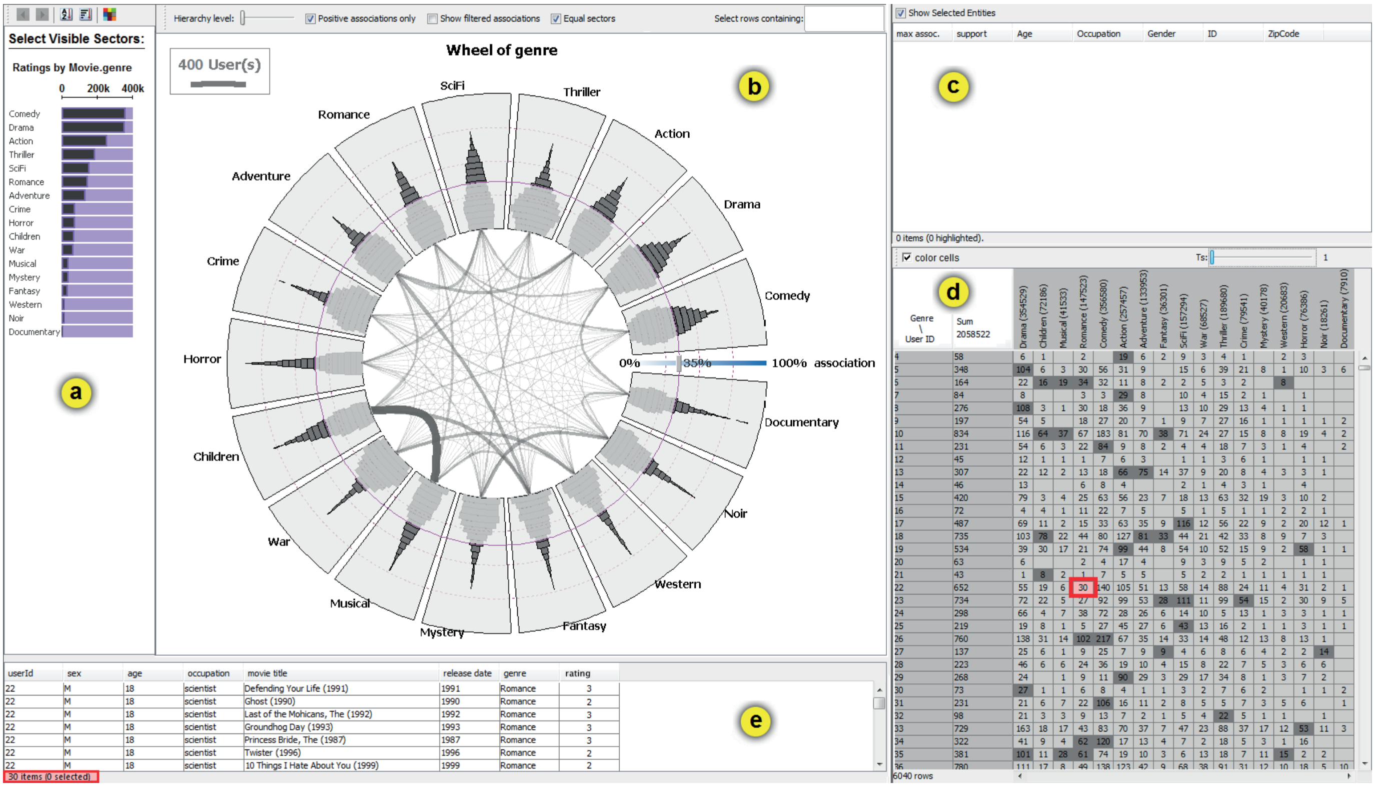


Fig. 5. Distributions of (a) a numerical attribute (release date) or, (b) a categorical attribute (genre) of the movies in the histograms. (c) The global distributions of release date and genre among all movies.



Genre \ User ID	Sum 2058522	Drama (354529)	Children (72186)	Musical (41533)	Romance (147523)	Comedy (356580)	Action (257457)	Adventure (133953)	Fantasy (36301)	SciFi (157294)	War (68527)	Thriller (189680)	Crime (79541)	Mystery (40178)	Western (20683)	Horror (76386)	Noir (18261)	Documentary (7910)
4	58	6	1		2		19	6	2	9	3	4	1		2	3		
5	348	104	6	3	30	56	31	9		15	6	39	21	8	1	10	3	6
6	164	22	16	19	34	32	11	8	2	2	5	3	2		8			
7	84	8			3	3	29	8		10	4	15	2	1		1		
8	276	108	3	1	30	18	36	9		13	10	29	13	4	1	1		
9	197	54	5		18	27	20	7	1	9	7	27	16	1	1	1	1	2
10	834	116	64	37	67	183	81	70	38	71	24	27	15	8	8	19	4	2
11	231	54	6	3	22	84	9	8	2	4	4	18	7	3	1	4		2
12	45	12	1	1	1	7	6	3		1	1	3	6	1		1	1	
13	307	22	12	2	13	18	66	75	14	37	9	20	8	4	3	3	1	
14	46	13			6	8	4			2	1	4	3	1		4		
15	420	79	3	4	25	63	56	23	7	18	13	63	32	19	3	10	2	
16	72	4	4	1	11	22	7	5		5	1	5	1	1	2	2	1	
17	487	69	11	2	15	33	63	35	9	116	12	56	22	9	2	20	12	1
18	735	103	78	22	44	80	127	81	33	44	21	42	33	8	9	7	3	
19	534	39	30	17	21	74	99	44	8	54	10	52	15	9	2	58	1	1
20	63	6			2	4	17	4		9	3	9	5	2		1	1	
21	43	1	8	2	1	7	5	5		5	2	2	1	1	1	1	1	
22	652	55	19	6	30	140	105	51	13	58	14	88	24	11	4	31	2	1
23	734	72	22	5	27	92	99	53	28	111	11	99	54	15	2	30	9	5
24	298	66	4	7	38	72	28	26	6	14	10	5	13	1	3	3	1	1
25	219	19	8	1	5	27	45	27	6	43	13	16	2	1	1	3	1	1
26	760	138	31	14	102	217	67	35	14	33	14	48	12	13	8	13	1	
27	137	25	6	1	9	25	7	9	9	4	6	8	6	4	2	2	14	
28	223	46	6	6	24	36	19	10	4	15	8	22	7	5	3	6	6	
29	268	24		1	9	11	90	29	3	29	17	34	8	1	3	7	2	
30	73	27	1	1	6	8	4	1	1	3	2	7	6	2		1	1	2
31	231	21	6	7	22	106	16	11	2	8	5	5	7	3	5	6		1
32	98	21	3	3	9	13	7	2	1	5	4	22	5	1	1		1	
33	729	163	18	17	43	83	70	37	7	47	23	88	37	17	12	53	11	3
34	322	41	9	4	62	120	17	13	4	7	2	18	5	3	1	16		
35	381	101	11	28	61	74	19	10	3	6	13	18	7	11	15	2	2	
36	780	111	17	8	49	138	123	42	9	68	38	91	31	12	10	18	5	10

userId	sex	age	occupation	movie title	release date	genre	rating
22	M	18	scientist	Defending Your Life (1991)	1991	Romance	3
22	M	18	scientist	Ghost (1990)	1990	Romance	2
22	M	18	scientist	Last of the Mohicans, The (1992)	1992	Romance	3
22	M	18	scientist	Groundhog Day (1993)	1993	Romance	3
22	M	18	scientist	Princess Bride, The (1987)	1987	Romance	3
22	M	18	scientist	Twister (1996)	1996	Romance	2
22	M	18	scientist	10 Things I Hate About You (1999)	1999	Romance	2

Fig. 6. Five levels of abstraction to explore the user-genre table and the underlying information: (a) a bar chart of the column categories (genres), (b) the wheel view showing sectors for the items selected in (a), (c) detail view for items selected in (b) (currently empty), (d) the contingency table with cells in active parts in (a) colored in dark gray, (e) the categorical data summarized in the cell highlighted in red in (d).

The frequency-based representation has several advantages over the dot representation: First, the angular dimension now has a clear meaning (frequency of associations at different radial distances in the sectors). Second, the artifacts and overlaps caused by showing separate dots are eliminated. Third, histograms are familiar visualizations that are easy to interpret. They better emphasize that the visualization is showing a distribution of the row associations in each sector, and not individual entities. This avoids the confusion due to multiple dots representing the same row. Finally, the redundancy of double-coding the association using both dot size and dot location is also eliminated.

Bended histograms embedded in a ring chart suffer from visual illusions in perceiving different arc lengths at different radial distances. This effect can be accounted for computationally and is minimized when arcs are short that are perceptually flattened [31] (figure 6).

3.2.3 Visualizing column similarities

We compute similarities between the columns of the contingency table based on their row associations. A similarity value $rc_{j_1 j_2}$ is computed for every pair of columns (j_1, j_2), to assess how similar the two distributions r_{ij_1} and r_{ij_2} are. Only active rows in both sectors are included in the computation. Active rows in sector j have sufficient support f_{i+} and associations r_{ij} higher than T_r, and are defined as follows:

$$A_j = \left\{1 \leq i \leq n : r_{ij} \geq T_r \wedge f_{i+} \geq T_s\right\} \tag{9}$$

Active rows in each sector are depicted in dark gray in the respective histogram (figure 6b). The column similarities are computed as follows:

$$rc_{j_1 j_2} = \frac{1}{|A_{j_1}| + |A_{j_2}|} \cdot \sum_{i \in A_{j_1} \cap A_{j_2}} r_{ij_1} \cdot r_{ij_2} \tag{10}$$

Between each pair of sectors (j_1, j_2), an arc is drawn whose thickness and opacity are determined by $rc_{j_1 j_2}$. A thick arc means that the active rows in both sectors tend to have similar associations with the two columns j_1 and j_2. Changing the T_r value results in smaller or larger active parts, and hence influences the thicknesses. By checking the arcs with different T_r values, the user can examine in which ranges and to which extent the column similarities hold.

Arcs showing column similarities based on row associations is a unique feature of the Contingency Wheel and one of the main reasons of adopting a circular layout for the visualization. This layout provides a compact representation to show and compare column similarities. Furthermore, arcs are useful in creating a user-controlled hierarchical grouping of the column categories based on their similarities: A right-click on an arc merges the two affected sectors into one sector. The resulting wheel is built from the contingency table that results by merging the corresponding columns into one column, by summing up the frequencies cell-by-cell. The new sector is inserted at its appropriate position according to the sector ordering scheme currently in use (alphabetical, by size, or user-defined sector ordering). Successively merging pairs of sectors connected by thick arcs enables abstracting the visualization by reducing the number of visual items. Moreover, it enables analyzing similarities between groups of similar columns and not only between pairs of columns, as the use case shows (Sect. 4).

3.2.4 Visual aids

We provide several visual aids to facilitate understanding. Three association levels evenly spaced between the inner and the outer sector boundaries are shown to allow an easier interpretation of the radial distances. An additional circle in pink shows the current value of the association threshold T_r, which can be adjusted using the slider embedded in the ring chart. Inactive parts of the histograms (Eq. 9) are visually de-emphasized. A color gradient is shown in the background of the T_r slider to reflect the association levels. It uses either a diverging or a sequential color scale [13], depending on whether negative associations are included or not. Arcs outside the ring chart indicate sector groups (figure 1). Finally, a legend shows the scale used in the histograms by depicting an arc of average length.

3.3 Interactive Exploration Environment

The original Contingency Wheel may result in a cluttered visualization especially for large data because it creates dots for individual row entities. These dots need to be selected individually to obtain details about the corresponding entities [1]. To improve on these shortcomings, our new methods follow Shneiderman's visual information-seeking mantra [36]: The visualization first shows an overview of the data using histograms. The user can filter the data interactively and select entities she is interested in exploring. Then, details about these entities can be obtained in linked views. Contingency Wheel++ offers an overview visualization of an asymmetrically-sized contingency table. Likewise, the contingency table offers a summarization of a larger data set by cross-tabulating two categorical dimensions. We designed the user interface to enable exploring the data at these multiple levels of abstraction as explained in the following.

3.3.1 Multiple Views

Whenever we explain Contingency Wheel++ to new users, our first step is to show the underlying contingency table. This allows explaining the basic concepts like row- and column marginal frequencies, actual- and expected frequencies (Eq. 1), and row-column associations (Eq. 2-5). We are thus showing both the wheel visualization and the underlying table side-by-side in one interface. This combination bridges the gap between the visual representations and the data representation (i.e., association values) computed by the automatic methods. The main user interface (UI) of our prototype is divided into five coordinated views:

A *bar chart* shows the column categories and their marginal frequencies f_{+j} (figure 6a). Columns selected in this view define the sectors of the wheel view. The user can thus focus on selected columns. Also, if the number of columns exceeds the limits for the wheel, smaller subsets of columns can still be visualized.

The *wheel view* is the central part of the interface (figure 6b). It provides an overview of the data and existing associations within. Several interactions are possible to find interesting patterns in the data and select specific row entities for further analysis. The association threshold T_r can be adjusted interactively via the slider embedded in the ring chart. Also, this view enables setting several parameters by means of its toolbar and context menu.

A *list view* shows details about the row entities selected in the wheel view (figure 6c and figure 7d). Beside the attributes of these entities (available from the data set), their marginal row frequencies f_i+ and associations r_{ij} with a specific column j are listed. The entities can be sorted according to one of the columns, and histograms or bar charts can be created for a specific column in the list.

A *tabular view* shows the contingency table and the marginal frequencies (figure 6d). By hovering the mouse pointer over a cell (i, j), a tooltip shows the expected frequency $\hat{e}_{ij}$ and the association value r_{ij} according to the measure used. If cell coloring is enabled via a checkbox, the cell is shown in dark gray if it corresponds to an active part in the visualization (i.e., $i \in A_j$). Also, the support threshold T_s can be adjusted via a slider to filter out entire rows i with $f_{i+} < T_s$.

A *second list view* shows details about selected items from the tabular view (figure 6e). By double-clicking on a cell, a row, or a column in the tabular view, cross-tabulated data items are shown in this list view along with their attributes. The items can be sorted and the distributions of the values in a specific column can be explored using a histogram or a bar chart.

These views make it easier to explain to new users how the data are visualized in Contingency Wheel++. Even more importantly, they constitute a multi-level overview+detail exploration interface. This allows experienced users to perform elaborate analysis workflows by having quick access to all information available in the data. Hence, associations can be detected and investigated further in relation to other attributes. The incorporation of analytical methods in the visual interface enables a visual analytics process following Keim's mantra [20]: Analyze first – show the important – zoom, filter and analyze further – details on demand. After computing the row-column associations (Sect. 3.1) and the columns similarities (Sect. 3.2.3), the visualization shows the important results, i.e., strong associations or high similarities. Using different interactions, the user can change the thresholds T_r and T_s, merge columns, or set a different association measure. This causes the analytical methods to recompute the associations and similarities which are then visualized interactively. Details on selected items in the wheel or in the tabular view can be obtained on demand.

3.3.2 Linking and Brushing

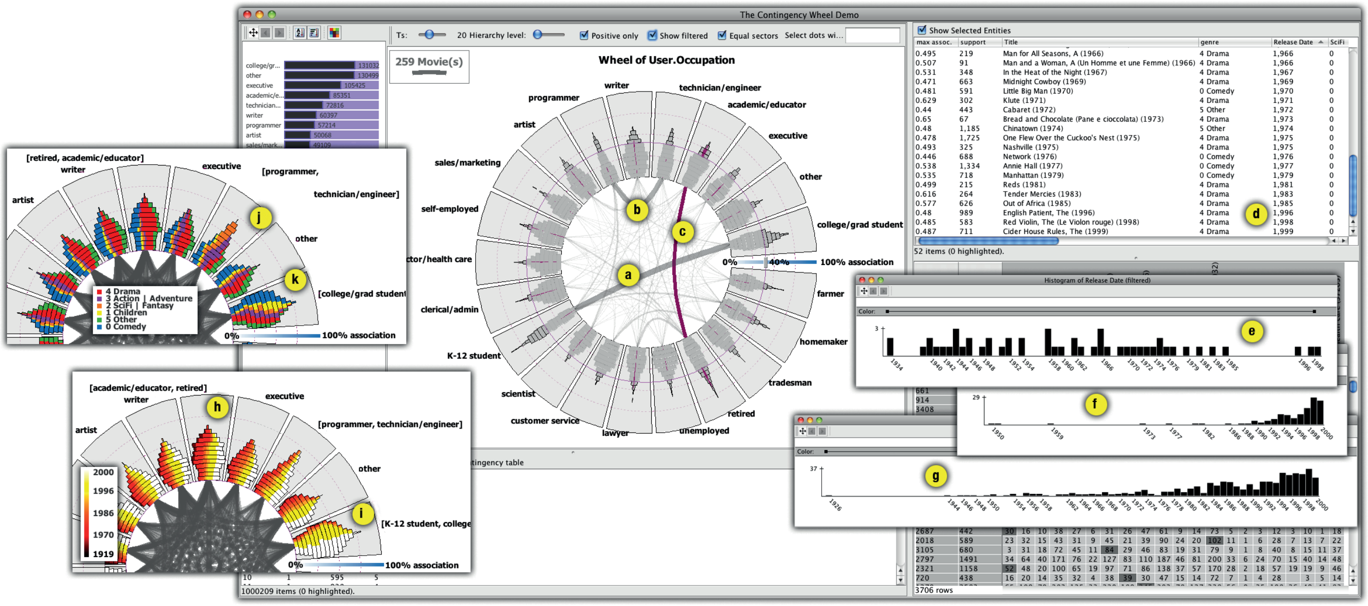
Contingency Wheel++ offers multiple ways to brush the visualized row categories. One way is by clicking on a bar in the histograms, which selects the rows it aggregates (Sect. 3.2.2). Another way is using the sector marquee tool to define a radial interval I in a sector j using the mouse (figure 1). This selects the rows i with $r_{ij} \in I$. Clicking on sector j selects the rows A_j that are currently active (Sect. 3.2.3). Also, clicking on an arc selects rows active in both sectors it connects (the items that define this arc). Rows i with $r_{ij} \leq T_r$ for all columns j can be selected using a menu command. When T_r is positive but small, this command selects rows that do not exhibit a high association with any column. Finally, rows can be selected using an external query, like the instant search box at the top of the view. This box selects row categories containing a specific text.

The top-right list view (figure 7d) shows the selected rows defined either by filtering, brushing, selection, or the search box query. When an item in this list is clicked, the tabular view scrolls to and highlights the corresponding row i which shows the frequencies f_{ij} (figure 10e). Also, a star graph [12] of the associations r_{ij} can be shown in the wheel view, labeled with these frequencies. Selected row categories are highlighted in the histograms of all sectors. The original histograms become desaturated and new sub-histograms are drawn centered on top of them showing the selected portion using color (figure 7c). Likewise, the original arcs are desaturated and the parts corresponding to the selected items are highlighted. Three modes are offered for performing brushing operations in the wheel, depending on keyboard modifiers:

- Set union: the new selection is added to an existing selection.
- Set intersection: the new selection is intersected with the existing selection. This enables creating nested queries on the data. For example, in the wheel showing the movies-occupation table, the user can select movies highly associated with the categories "programmer" and "scientist" but negatively associated with the category "executive". This is done by drawing ranges at the corresponding radial distances in each sector while the CTRL key is pressed. TimeSearcher uses a similar brushing technique for time-series data by means of timebox widgets [16].
- If no modifier is defined or if brushing is performed using an external query, the active selection is replaced by the new one.

4 Use Case

To demonstrate the applicability of our approach, we present a use case along the fictitious character Jane, who is an analyst at a large movie rental service. The use case is based on 10 analysis sessions conducted over the course of a week and added up to a total of 8 hours time. For the analysis, the MovieLens data set [10] has been used as introduced in Sect. 1. Jane's goal is to get insights into the massive amount of data they have collected about their customers who rented, watched, and rated movies using their service. Based on the insights gained from this analysis she plans to make decisions and take actions related to the ongoing marketing strategies and recommender algorithms they have in place. Jane uses mainly two different tables for her analysis: first, *occupations* (movies × user occupations) and second, *genres* (users × movie genres). By exploring the interactive wheels and the associated views and diagrams, Jane gains a number of insights, some of which were expected but also some surprising ones. Before Jane starts the analysis, she asks herself about the semantics of the data – what do associations between the entities user and movie actually mean? A user and a movie are associated if a user has entered a rating for a movie in the system which in turn implies that he or she has watched the movie. However, an association does not express how much they liked a movie.



☑ Show Selected Entities

max assoc.	support	Title	genre	Release Date	SciFi
0.495	219	Man for All Seasons, A (1966)	4 Drama	1,966	0
0.507	91	Man and a Woman, A (Un Homme et une Femme) (1966)	4 Drama	1,966	0
0.531	348	In the Heat of the Night (1967)	4 Drama	1,967	0
0.471	663	Midnight Cowboy (1969)	4 Drama	1,969	0
0.481	591	Little Big Man (1970)	0 Comedy	1,970	0
0.629	302	Klute (1971)	4 Drama	1,971	0
0.44	443	Cabaret (1972)	5 Other	1,972	0
0.65	67	Bread and Chocolate (Pane e cioccolata) (1973)	4 Drama	1,973	0
0.48	1,185	Chinatown (1974)	5 Other	1,974	0
0.478	1,725	One Flew Over the Cuckoo's Nest (1975)	4 Drama	1,975	0
0.493	325	Nashville (1975)	4 Drama	1,975	0
0.446	688	Network (1976)	0 Comedy	1,976	0
0.538	1,334	Annie Hall (1977)	0 Comedy	1,977	0
0.535	718	Manhattan (1979)	0 Comedy	1,979	0
0.499	215	Reds (1981)	4 Drama	1,981	0
0.616	264	Tender Mercies (1983)	4 Drama	1,983	0
0.577	626	Out of Africa (1985)	4 Drama	1,985	0
0.48	989	English Patient, The (1996)	4 Drama	1,996	0
0.485	583	Red Violin, The (Le Violon rouge) (1998)	4 Drama	1,998	0
0.487	711	Cider House Rules, The (1999)	4 Drama	1,999	0

52 items (0 highlighted).

Fig. 7. Visual exploration of movies associated to user occupations: (a-c) major overlaps between user groups, (d) details of selected items in (c), (e-g) histograms of movie release date for different subsets of movies, (h, i) wheel view colored by movie release dates to reveal its relation with different user groups, (j, k) wheel view colored by movie genre to reveal dominant genres in the movie preferences of different user groups.

4.1 Categories & Characteristics

As a first step, Jane aims for getting an overview of the categories to get a feeling for the data and overall (dis)similarities, to explore characteristics of single categories as well as to possibly simplify the data by merging categories.

User occupations Jane starts her analysis by creating a wheel based on a contingency table that displays the different occupations (i.e., jobs) of the users as sectors and the related movies as histogram bars within these sectors (figure 7). After opening the initial wheel, she adjusts the association slider to >40% in order to focus on higher associations and similarities between sectors. By studying the thickness of the connections inside the wheel she notices that there is a lot of overlap (same movies rated) between “K-12 student” and “college/grad student ” (figure 7a) as well as between “programmer” and “technician/engineer” (figure 7b) which seems plausible to her. However, a more surprising aspect is that there is also a higher degree of overlap between “academic/educator” and “retired” (figure 7c). To investigate this connection in more detail, Jane selects the arc (figure 7c) and takes a look at the selected entities in the list view in the upper right of the UI (figure 7d). She sorts the movies by release date by clicking on the respective table header and discovers that they seem to be mostly older movies. Only three out of 52 movies are from the 90’s, the rest are older movies. Based on the mentioned similarities, she merges “K-12 student” with “college/grad student”, “programmer” with “technician/engineer”, and “academic/educator” with “retired” by right-clicking on the corresponding arcs in order to simplify the further analysis.

Then, Jane continues her exploration of release dates of movies highly associated to the group [academic/educator, retired]. For this, she brings up a histogram using the context menu of the release-date column-header (figure 7e). This provides details about the distribution of movies over time. For comparison, she also brings up release-date histograms for [college/grad student, K-12 student] (figure 7f), as well as for all movies that were rated (figure 7g). This confirms that the distribution concerning [academic/educator, retired] is quite different. Moreover, Jane finds out that there seems to be a peak of watched and rated movies in the mid 80s followed by a valley at the beginning of the 90s and another peak at the end of the 90s. To get a further overview of release dates of categories, she colors the wheel by release date which clearly shows that [academic/educator, retired] more often watch older movies than others (larger portion of dark parts than average, figure 7h). [K-12 student, college/grad student] watch more often newer movies than others (larger portion of bright parts than average, figure 7i). Jane concludes her exploration by coloring the occupation wheel by genre. This reveals that [programmer, technician/engineer] contain a much larger portion of highly associated “SciFi/Fantasy” movies (orange, figure 7j) and [college/grad student, K-12 student] have a much larger portion of highly associated “Children” and “Comedy” movies (blue and yellow, figure 7k) than on average.

Movie genres Jane switches to the genre wheel and finds high overlaps of “Musical” and “Children”, “Action” and “Adventure”, and “War” and “Western” which seem to be reasonable to her. More surprisingly, she finds no particularly high overlap between [War, Western] and “Crime” which she would have suspected. Besides, she observes that “Horror” seems inversely related to many other genres. Based on her observations she merges genres into [Children, Musical, and Fantasy], [Noir, Mystery, Thriller], [Adventure, Action, SciFi], and [War, Western] (figure 8a-b). Jane colors the wheel by age using a diverging color scheme (figure 8a). Looking at this, she finds it surprising that the age distribution of users watching [Musical, Children, Fantasy] is not very different from others. Overall, age-group distributions seem to be quite similar in all genres.

Further, Jane wants to inspect possible gender differences and turns on coloring by gender in the genre wheel (figure 8b). As a general observation, she recognizes that there are a lot more men than women rating movies. As anticipated, Jane finds the genre “Romance” as an exception where the most highly associated users are female. Surprisingly “Horror” does not show less women than other genres such as “Documentary”, [Adventure, Action, SciFi], [Noir, Mystery, Thriller], or “Crime”. After that, Jane takes a closer look on different genres using histograms of gender, age, and occupations (figure 8c). For “Children” movies she notices that there is an almost equal distribution between male and female viewers, and that most viewers are in the age group of 18–24 years. Particularly, the last fact is somewhat surprising to Jane, since she thought that the majority of users watching “Children” films would be younger. Having a look at “War” movies she spots that there are by far more men present which are often executives and in the age group of 35–44 years. “Western” movies show a quite similar picture, except that even older age groups watch and rate these movies.

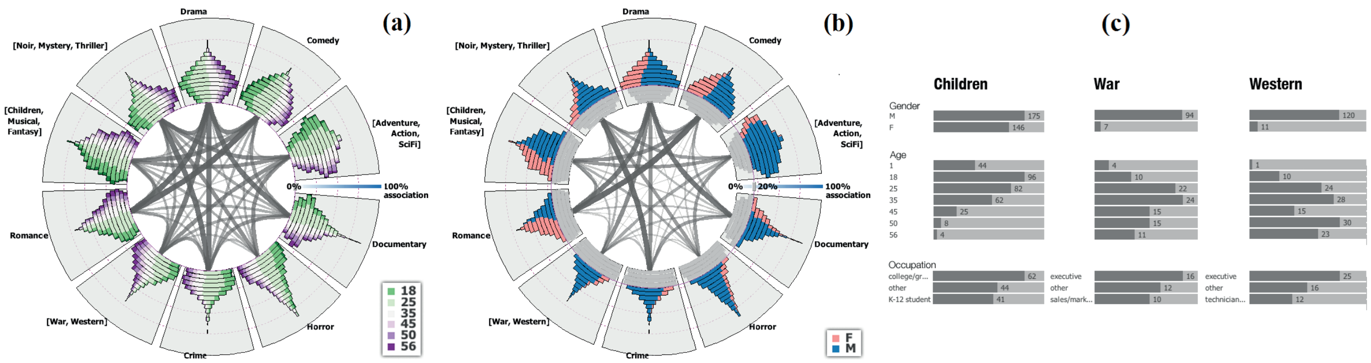


Fig. 8. Associations between users and movie genres (a) colored by age, (b) colored by gender. (c) Details about selected genres.

4.2 Single Movies

After her top-down exploration of occupations and genre categories, Jane has a couple of movies in mind she wants to inspect further for potentially interesting findings in a bottom-up manner.

Mainstream erotic films She takes a look at the two movies Basic Instinct (1992) and Nine 1/2 Weeks (1986) and compares the star plots in the wheel views. Interestingly, both movies are highly associated to "technician/engineer" but negatively associated to "programmer" (figure 9a-d) which are otherwise quite similar as she had found out earlier.

The Godfather trilogy Next, Jane remembers that The Godfather movies (1972, 1974, 1990) were quite big hits at her movie rental service in the last years. She uses the search box (figure 10a) to find them. The movies matching the query are shown in the detail list in the upper right of the UI. She selects the first movie of the trilogy in the list (figure 10b) which brings up a star plot in the center of the wheel view showing individual associations for the different occupations. She can see that the movie has the highest associations with the occupations "executive" and "lawyer" (figure 10c,d). When she selects the second movie, the picture is quite similar, however, the third movie is somewhat different. Jane sees that it is highly associated to "executive" again but that it is negatively associated to "lawyer" and highly associated to "sales/marketing". Further, Jane would like to inspect how the three movies were rated among executives. For this, she double clicks on the "executive" column in the table (figure 10e) which brings up the ratings in the list view on the lower left of the UI (figure 10f). Via a context menu, she displays the rating histograms (figure 10g-i) and spots that they are quite positive and similar for the first two but much lower for the third movie.

4.3 Hypotheses and Specific Questions

During the visual exploration, Jane generated some hypotheses and specific questions she is trying to answer subsequently.

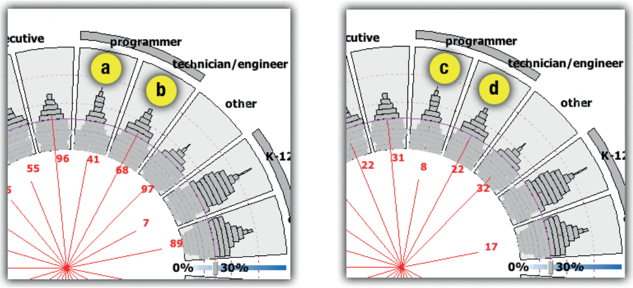


Fig. 9. Associations of mainstream erotic films to "programmer" (a, c) and "technician/engineer" (b, d) – left: Basic Instinct (1992), right: Nine 1/2 Weeks (1986).

Association vs. rating behavior One hypothesis Jane had in mind is to check whether it is true that very high associations of movies correspond to more positive ratings. Using the occupation wheel, she is probing rating histograms of highly associated movies (>75%) with "college/grad student", such as Transformers (1986, good ratings) and Teenage Mutant Turtles II (1991, bad ratings). As a result, she finds evidence that her hypothesis does not hold.

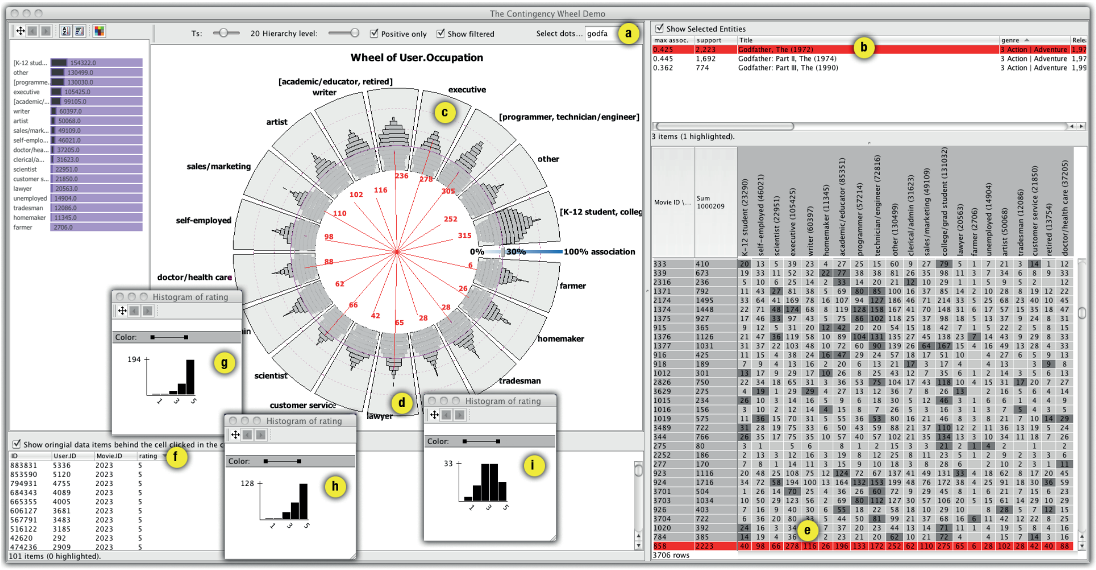
Rating Another question Jane wants to inspect is whether there are differences in the general rating behavior of different user occupations, i.e., are particular groups more or less critical than others in general? By using the occupation wheel and comparing the grading histograms of selected sectors, Jane observes that the rating behavior is strikingly similar among groups. She can only spot subtle differences such as that unemployed persons tend to give lower ratings whereas retired persons do not tend to give many low ratings.

4.4 Decisions and Actions Planned

Visually exploring the vast data collection of her movie rental service helped Jane to better understand her customers and unearth commonalities as well as differences between groups of users and movies. Based on the gained insights, decisions are taken and actions are planned that are intended to make her business more successful: The merging of some user categories and movie genres can simplify the internal recommender engine. New SciFi & Fantasy releases will be presented particularly to programmers and technicians/engineers. As there were more movies watched and rated from the mid 1980s, there will be a campaign highlighting some of these. Suggestions concerning Children movies will no longer be focused on younger customers but concentrate on the age group of 18–24 years. War and Western movies will be recommended more intensively to male executives older than 35 years. Finally, Horror movies will be suggested to users who already watched those without restricting suggestions to men.

4.5 Improvements of Contingency Wheel++

Jane benefited from the improvements in the new design and gained insights that would not have been possible using the original Contingency Wheel. Due to the fact that dots have been replaced by histograms, she was able to represent all movies without filtering steps which would have been necessary to avoid overlaps. The distribution of an attribute of the movies (e.g., release date) can now be inspected using colored histograms which allowed for complex insights involving multiple data attributes. Because of the multi-level overview+detail exploration environment, Jane had easy on-demand access to all available data, such as movie details, contingency table, and raw data. This allowed for drilling down to clarify and check findings from an aggregate level. Further, she was able to create bar charts and histograms of selected elements from different attributes, such as movie release dates or ratings, and compare the results with the global distributions. On top of that, the ability to merge sectors enabled the detection of patterns between groups of sectors that could not be detected when looking at single columns.



[K-12 stud...	154322.0
other	130499.0
[programme...	130030.0
executive	105425.0
[academic/...	99105.0
writer	60397.0
artist	50068.0
sales/mark...	49109.0
self-emplo...	46021.0
doctor/hea...	37205.0
clerical/a...	31623.0
scientist	22951.0
customer s...	21850.0
lawyer	20563.0
unemployed	14904.0
tradesman	12086.0
homemaker	11345.0
farmer	2706.0

☑ Show Selected Entities

max assoc.	support	Title	genre	Rele
0.425	2,223	Godfather, The (1972)	3 Action | Adventure	1,97
0.445	1,692	Godfather: Part II, The (1974)	3 Action | Adventure	1,97
0.362	774	Godfather: Part III, The (1990)	3 Action | Adventure	1,99

3 items (1 highlighted).

Movie ID \...	Sum 1000209	K-12 student (23290)	self-employed (46021)	scientist (22951)	executive (105425)	writer (60397)	homemaker (11345)	academic/educator (85351)	programmer (57214)	technician/engineer (72816)	other (130499)	clerical/admin (31623)	sales/marketing (49109)	college/grad student (131032)	lawyer (20563)	farmer (2706)	unemployed (14904)	artist (50068)	tradesman (12086)	customer service (21850)	retired (13754)	doctor/health care (37205)
333	410	20	13	5	39	23	4	27	25	15	60	9	27	79	5	1	7	21	3	14	1	12
339	673	19	33	11	52	32	22	77	38	38	81	26	35	98	11	3	7	34	6	8	9	33
2316	236	5	10	6	25	14	2	33	14	20	21	12	10	29	1	1	5	9	5	2		12
1371	792	11	43	27	81	38	5	69	80	85	100	16	37	85	14	2	10	28	8	19	12	22
2174	1495	33	64	41	169	78	16	107	94	127	186	46	71	214	33	5	25	68	23	40	10	45
1374	1448	22	71	48	174	68	8	119	128	158	167	41	70	148	31	6	17	57	15	35	18	47
1375	927	17	46	33	97	43	5	75	86	102	118	25	37	98	18	5	13	37	9	24	8	31
915	365	9	12	5	31	20	12	42	20	20	54	15	18	42	7	1	5	22	2	5	8	15
1376	1126	21	47	36	119	58	10	89	104	131	135	27	45	138	23	7	14	43	9	29	8	33
1377	1031	31	37	22	103	48	10	72	60	90	139	26	64	167	15	4	16	49	13	28	4	33
916	425	11	15	4	38	24	16	47	29	24	57	18	17	51	10		4	27	6	5	9	13
918	189	7	9	4	13	16	2	20	6	13	21	17	3	17	4		4	13		3	9	8
1012	301	13	17	9	29	17	10	26	8	25	43	12	7	35	6	1	2	14	3	5	6	13
2826	750	22	34	18	65	31	3	36	53	75	104	17	43	118	10	4	15	31	17	20	7	27
3629	275	4	19	1	29	29	4	27	13	12	36	7	8	26	13		2	16	5	6	4	14
1015	234	26	10	3	14	16	5	9	6	18	30	5	12	46	3	1	6	6	1	4	4	9
1016	156	3	10	2	12	14	4	15	8	7	26	5	3	16	3	1	3	7	5	4	3	5
1019	575	11	36	15	70	31	5	55	36	53	80	16	21	46	8	3	8	21	7	10	14	29
3489	722	31	28	19	75	33	6	50	43	59	88	21	37	110	12	2	11	36	13	19	5	24
344	766	26	35	17	75	35	10	57	40	57	102	21	35	134	13	3	10	34	11	18	7	26
275	80	3	1		5	6		8	1	2	15	3	3	21	2	1	4	2		1		2
2252	186	2	13	3	12	16	3	19	8	12	25	8	11	23	5	1	2	9	2	3	3	6
277	170	7	8	1	14	11	3	15	9	10	18	3	8	28	6		2	10	2	3	1	11
923	1116	20	48	25	108	75	12	124	72	67	137	41	49	131	33	4	18	62	8	17	20	45
924	1716	34	72	58	194	100	13	164	132	153	199	48	76	172	38	4	25	91	18	30	36	59
3701	504	1	26	14	70	25	4	36	26	60	72	9	29	45	8	1	6	21	7	15	6	23
3703	1034	10	50	29	123	56	2	69	80	112	127	30	57	106	20	5	15	61	14	29	10	29
926	403	7	16	9	40	30	6	55	18	22	58	18	10	29	10		8	28	5	7	12	15
3704	722	6	36	20	80	33	5	44	50	81	99	21	37	68	16	6	11	42	12	22	8	25
1020	392	24	16	3	34	[illegible]	7	37	20	23	44	13	14	71	11	1	4	19	5	8	4	16
784	385	14	19	4	36	[illegible]	2	23	21	20	62	10	21	72	4		4	15	7	14	3	16
858	2223	40	98	66	278	116	26	196	133	172	252	62	110	275	65	6	28	102	28	42	40	88

3706 rows

☑ Show oringial data items behind the cell clicked in the c

ID	User.ID	Movie.ID	rating
883831	5336	2023	5
853590	5120	2023	5
794931	4755	2023	5
684343	4089	2023	5
665355	4005	2023	5
606127	3681	2023	5
567791	3483	2023	5
516122	3185	2023	5
42620	292	2023	5
474236	2909	2023	5

101 items (0 highlighted).

Fig. 10. Visual exploration of the Godfather trilogy: (a) search box, (b) search result, (c, d) star plot of associations of selected item to user occupations, (e) row of selected movie in contingency table, (f) raw data of user ratings, (g-i) rating histograms for the three movies of the trilogy.

5 State of the Art

Heat maps can be used as a generic method for visualizing large matrix data as a colored image [5]. While they can provide an overview of the data distribution, they are limited in terms of exploring associations. Methods dedicated for visualizing contingency tables are usually designed to handle a small number of categories. Based on what the visual representations depict, they can be classified into three types:

Frequency representations These methods map the table frequencies f_{ij} to visual elements of proportional size. Mosaic Displays [14] and their variations use tiles to represent the frequencies (similar to Treemaps [35]). Parallel Sets [2] and their variations, such as Circos [25], represent frequencies as stripes or ribbons between visual elements that depict the categories. These approaches offer an intuitive visual representation that can be divided further to accommodate additional dimensions. However, they can handle only a relatively small number of categories (≤ 30 for Parallel Sets). With larger tables, the clutter increases in Parallel Sets, and the increased skewness and number of zeros in the table values make it difficult to identify and compare the tiles in Mosaic Displays [40].

Deviation representations Association Plots [26] use bar charts to show the deviations between the actual frequencies f_{ij} and the expected frequencies $\hat{e}_{ij}$ (Eq. 1). Sieve Diagrams [8] plot f_{ij} as sieves in Mosaic Displays of $\hat{e}_{ij}$ to show how both deviate from each other. Sieves with smaller holes represent higher associations. A recent approach was proposed for exploring proportions in multivariate categorical data [27]. It adopts the layout of Parallel Sets, but depicts the proportionality of relationships between the categories instead of f_{ij}.

Intermediate representations Correspondence Analysis [3] (CA) projects the categories to points in a 2D space, spanned by the two most contributing factors of the χ^2 statistic, in a way similar to Principal Component Analysis [19]. A higher association between categories of the same class positions their points closer together, in a way similar to multidimensional scaling [24]. The approach can also accommodate additional categorical dimensions [9]. However, with a growing number of categories, the plot becomes more difficult to read. It lacks an intuitive structure as its axes bear no interpretable semantics. Johansson et al. [18] and Rosario et al. [33] proposed methods for quantifying categorical data based on CA. The quantified data can then be visualized using scatter plots or parallel coordinates. However, the latent numerical variables used for the quantification are not always easy to interpret.

The dot-based Contingency Wheel uses deviation representations for the cells as dots along the radial dimension. Like many approaches for dealing with large data [6, 29, 38] it uses data reduction to handle tables having a large number of rows. Also, it employs alpha blending to reveal overlapping, as done by other approaches for dealing with similar issues [7, 17, 22]. In contrast, Contingency Wheel++ employs a frequency-based approach to abstract large data, as used by many other techniques [15, 21, 32]. Also, it makes use of interactive visual analytics techniques to enable the exploration of individual data items. As the use case illustrates, asymmetrically-sized tables with a small number of columns (≤ 30) and thousands of rows can be handled efficiently by Contingency Wheel++ without filtering the data.

6 Conclusion

Contingency Wheel++ employs novel visual analytics methods that address the major shortcomings of the original dot-based wheel for visualizing and discovering patterns in large categorical data. It improves on the computational component by introducing an association measure based on Pearson's residuals to alleviate the bias in the association measure originally used. It eliminates the scalability and readability limitations caused by overlapping dots, by using a frequency-based abstraction that shows distributions rather than individual entities. Finally, it offers a multi-level overview+detail interface to explore individual entities that are aggregated in the visualization or in the table along with their attributes. The use case demonstrates how these methods can be used to find nontrivial patterns in large categorical data, and how further attributes can be analyzed in separate views or by coloring the histograms in the wheel visualization.

Future work aims to conduct comparative user studies to assess the effectiveness and efficiency of Contingency Wheel++, and to apply it to different real-world domains. Also, we are exploring further measures of associations and column similarities. Finally, we are investigating the applicability of our approach to other problems having similar data structures, such as point-set memberships or the class probabilities computed by a fuzzy classifier for a large number of samples.

Acknowledgement We thank Peter Filzmoser for statistical help, Margit Pohl, Simone Kriglstein, and Florian Scholz for insightful ideas and evaluation, Georg Beischlager and Josef Tiefenbacher for implementation help, and Hannes Obweger and Martin Suntinger from UC4 Software for reviews. This work was supported by the Austrian Federal Ministry of Economy, Family and Youth via CVAST, a Laura Bassi Centre of Excellence (No. 822746), and by the Austrian Science Fund (FWF) through the ViMaL project (No. P21695).

References

[1] B. Alsallakh, E. Gröller, S. Miksch, and M. Suntinger. Contingency Wheel: Visual Analysis of Large Contingency Tables. In *EuroVA 2011: International Workshop on Visual Analytics*, pages 53–56, Bergen, Norway, 2011. Eurographics Association.

[2] F. Bendix, R. Kosara, and H. Hauser. Parallel sets: visual analysis of categorical data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 133–140, 2005.

[3] J. P. Benzécri. *Correspondence Analysis Handbook*. Marcel Dekker, New York, 1990.

[4] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel. KNIME: The Konstanz information miner. In *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 319–326. Springer Berlin Heidelberg, 2008.

[5] J. Bertin. *Semiology of graphics : diagrams, networks, maps*. University of Wisconsin Press, Madison, Wisconsin, USA., 1983.

[6] A. Dix and G. Ellis. By chance: enhancing interaction with large data sets through statistical sampling. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '02, pages 167–176, New York, NY, USA, 2002. ACM.

[7] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization, 2002.*, pages 117–124, 2002.

[8] M. Friendly. Graphical methods for categorical data. In *SAS User Group International Conference Proceeding*, volume 17, pages 190–200, 1992.

[9] M. J. Greenacre and J. Blasius. *Multiple correspondence analysis and related methods*. Chapman & Hall/CRC, 2006.

[10] GroupLens. MovieLens data sets. http://www.grouplens.org/node/73. Accessed: August 2012.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Nov. 2009.

[12] R. L. Harris. *Information Graphics: A Comprehensive Illustrated Reference*. Oxford University Press, Inc., New York, NY, USA, 1999.

[13] M. Harrower and C. Brewer. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal*, pages 27–37, June 2003.

[14] J. A. Hartigan and B. Kleiner. Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pages 268–273. Springer-Verlag, 1981.

[15] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *IEEE Symposium on Information Visualization*, pages 127 – 130, 2002.

[16] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, Mar. 2004.

[17] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure in visualizations of dense 2d and 3d parallel coordinates. *Information Visualization*, 5(2):125–136, June 2006.

[18] S. Johansson, M. Jern, and J. Johansson. Interactive quantification of categorical variables in mixed data sets. In *Proceedings of the 12th International Conference on Information Visualisation*, pages 3–10, Washington, DC, USA, 2008. IEEE Computer Society.

[19] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, Oct. 2002.

[20] D. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining*, volume 4404 of *Lecture Notes in Computer Science*, pages 76–90. Springer Berlin / Heidelberg, 2008.

[21] R. Kosara, F. Bendix, and H. Hauser. Timehistograms for large, time-dependent data. In O. Deussen, C. Hansen, D. Keim, and D. Saupe, editors, *Symposium on Visualization (VisSym)*, pages 45–54, 340. Eurographics Association, 2004.

[22] R. Kosara, S. Miksch, and H. Hauser. Focus+context taken literally. *IEEE Computer Graphics and Applications*, 22:22–29, 2002.

[23] S. Kriglstein, F. Scholz, M. Pohl, B. Alsallakh, and S. Miksch. Contingency wheel evaluation: Results from an interview study. Technical Report CVAST-2012-2, Vienna University of Technology, Vienna, Austria, March 2012.

[24] J. B. Kruskal and M. Wish. Multidimensional scaling. *Methods*, 116(2):463–504, 1978.

[25] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[26] D. Meyer, A. Zeileis, and K. Hornik. Visualizing independence using extended association plots. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria*, 2003.

[27] H. Piringer and M. Buchetics. Exploring proportions: Comparative visualization of categorical data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 295 –296, 2011.

[28] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009.

[29] D. Rafiei and S. Curial. Effectively visualizing large networks through sampling. *Visualization Conference, IEEE*, pages 375 – 382, 2005.

[30] J. N. K. Rao and A. J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *The Journal of the American Statistical Association*, 76:221–230, 1981.

[31] J. O. Robinson. *The Psychology of Visual Illusion*. Dover Publications, Inc., 1998.

[32] J. Rodrigues, J.F., A. Traina, and J. Traina, C. Frequency plot and relevance plot to enhance visual data exploration. In *Proceedings of Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2003)*, pages 117–124, Oct. 2003.

[33] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang. Mapping nominal values to numbers for effective visualization. *Information Visualization*, 3(2):80–95, June 2004.

[34] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, Dec. 1979.

[35] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics (TOG)*, 11(1):92–99, Jan. 1992.

[36] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings of IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[37] J. S. Simonoff. *Analyzing Categorical Data*. Springer-Verlag, New York, USA, 2nd edition, 2003.

[38] M. C. Stone, K. Fishkin, and E. A. Bier. The movable filter as a user interface tool. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94, pages 306–312, New York, NY, USA, 1994. ACM.

[39] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, 2005.

[40] A. Unwin, M. Theus, and H. Hofmann. *Graphics of Large Datasets: Visualizing a Million*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[41] M. Wertheimer. Laws of organization in perceptual forms. In W. D. Ellis, editor, *A sourcebook of Gestalt psychology*, pages 71–88. Routledge and Kegan Paul, 1938.

[42] T. Xiong, S. Wang, A. Mayers, and E. Monga. A new MCA-based divisive hierarchical algorithm for clustering categorical data. In *Proceedings of IEEE International Conference on Data Mining*, pages 1058–1063. IEEE Computer Society, 2009.