International Workshop on Visual Analytics (2011)
S. Miksch and G. Santucci (Editors)

Contingency Wheel: Visual Analysis of Large Contingency Tables

Bilal Alsallakh[1,2] , Eduard Gröller[1] , Silvia Miksch[1] and Martin Suntinger[2]

[1]Vienna University of Technology, Austria   [2]UC4 Software, Austria

Abstract

We present the Contingency Wheel, a visual method for finding and analyzing associations in a large $n \times m$ contingency table with $m < 100$ and n being two to three orders of magnitude larger than m. The method is demonstrated on a large table from the Book-Crossing dataset, which counts the number of ratings each book received from each country. It enables finding books that received a disproportionately high number of ratings from a specific country. It further allows to visually analyze what these books have in common, and with which countries they are also highly associated. Pairs of similar countries can further be identified (in the sense that many books are associated with both countries). Compared with existing visual methods, our approach enables analyzing and gaining insight into larger tables.

Categories and Subject Descriptors: categorical data analysis, visual representations and interaction techniques.

1. Introduction

Categorical data, such as hair color or people's country of residence, pose a challenge for visualization [BKH05] as well as for several data mining tasks. Contingency tables are one of the most common ways to summarize categorical data as the first step of analysis. Table 1 shows an example contingency table of Snee's hair-and-eye-color dataset [Sne74]. Several methods have been devised to visualize contingency tables. However, their applicability and interactiveness are limited to small tables with few categories. The aim of this work is to develop an interactive visual method for finding and analyzing associations in a large $n \times m$ contingency table, with $m < 100$ and n being two to three orders of magnitude larger than m. The next section presents related work. Sections 3 and 4 present the proposed visual and interaction metaphors designed to identify the desired associations and patterns within the data. Section 5 provides evaluation results and Section 6 concludes the work.

2. Related Work

One of the common approaches to visualize contingency tables are Mosaic Displays [HK81] and their variations. They represent counts in the table as tiles of proportional size, in a treemap-like fashion. Fig. 1(a) shows a mosaic display of the data in Table 1. This method is limited to small tables due to the large skewness in the distribution and to the large number of zero entries, both of which typically exist in large tables [UTH06]. Parallel Sets [BKH05] is a more recent method that adopts the layout of parallel coordinates, but substitutes data points and lines by frequency based representations. The view is combined with several interactions to support data analysis. Fig. 1(b) shows the parallel sets of the data in Table 1. According to the authors, the view becomes overloaded when the number of categories is getting larger than 30. Correspondence Analysis [Ben90] is a statistical technique that enables displaying the table's row- and column- categories as 2D points. A higher association between a row and a column positions their points closer together, and vice versa (in a way similar to multidimensional scaling). The categories are projected on the two most-contributing orthogonal factors of the chi-square statistic (in a way similar to Principal Component Analysis). Fig. 1(c) shows the result of applying correspondence analysis on Table 1. 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. While the above mentioned approaches can handle more than two dimensions, the readability degrades as the number of categories increases, even with two dimensions. Our new approach addresses this limitation for $n \times m$ tables.

Eye \ Hair	Black	Brown	Red	Blond	Total
Brown	68	119	29	7	220
Blue	20	84	17	94	215
Hazel	15	54	14	10	93
Green	5	29	14	14	16
Total	108	286	71	127	592

Table 1: *An example contingency table [Sne74].*

DOI: 10.2312/PE/EuroVAST/EuroVA11/053-056

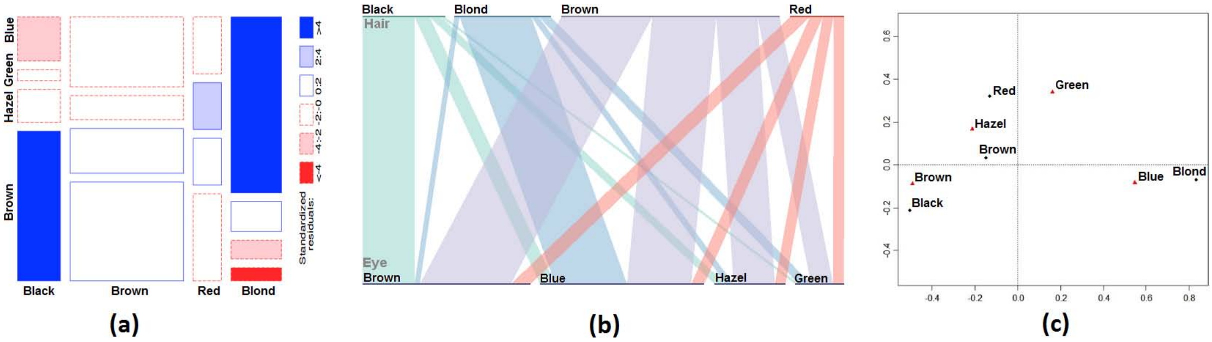


Figure 1: *Three visualizations of the dataset in Table 1: (a) Mosaic Displays. (b) Parallel Sets. (c) Correspondence Analysis.*

3. The Visual Metaphor

Mosaic Displays and Parallel Sets map the table frequencies $(f_{i,j})$ directly to visual elements of linearly proportional size. Our proposed view, the Contingency Wheel, uses an intermediate representation $r_{i,j}$ which denotes the association strength between row category i and column category j:

$$r_{i,j} = \begin{cases} \frac{f_{i,j} - \hat{e}_{i,j}}{f_{i+} - \hat{e}_{i,j}} & f_{i,j} \geq \hat{e}_{i,j} \\ \frac{f_{i,j} - \hat{e}_{i,j}}{\hat{e}_{i,j}} & otherwise \end{cases} \quad , \quad \hat{e}_{i,j} = \frac{f_{i+} \cdot f_{+j}}{f_{++}} \qquad (1)$$

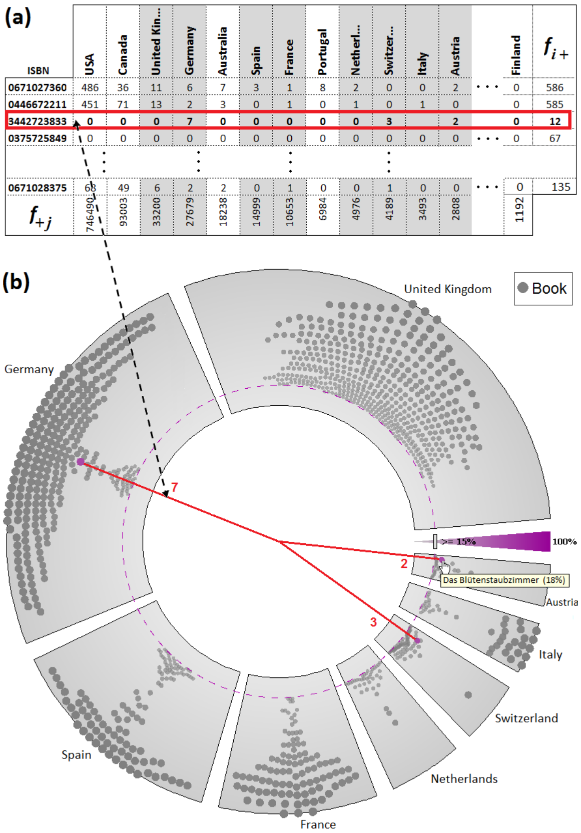
where f_{i+} and f_{+j} are the marginal row and column frequencies, and f_{++} is the sum of all frequencies in the table. Equation 1 performs a piecewise linear interpolation of the association strength for frequencies in the range $[0..f_{i+}]$ to the range $[-1..1]$ where $\hat{e}_{i,j}$, the frequency predicted under the null hypothesis, is mapped to zero (no association). The view is built upon a ring chart whose sectors represent the column categories and are scaled and sorted by their marginal frequencies (f_{+j}). For the i-th row category, a node is created in the j-th sector under these two conditions:

1. $r_{i,j} \geq T_r$ where $T_r > 0$ is a threshold on the association strength to filter out insignificant or negative associations. The threshold is marked with a dashed circle, and can be adjusted interactively via a purple slider (Fig. 2b).
2. $f_{i+} \geq T_s$ where $T_s > 0$. This filters out entire rows that have insignificant appearance in the dataset.

The created node has the following polar coordinates:

- The radius is mapped linearly from the association strength $(r_{i,j})$ where 0 is mapped to the inner radius, and 1 is mapped to the outer radius of the view.
- The angle bears no semantics and is computed by a layout algorithm to minimize both node overlapping inside each sector, and the node distance to its center. Layouting a sector with n nodes requires $O(n \cdot log(n))$ operations.

Fig. 2(b) shows the wheel view of the Book-Crossing dataset [ZMKL05] which breaks down ratings on 270,170 books (the rows) by country (the columns). To simplify the view, only eight countries are included (Fig. 2a). The thresholds are set to $T_r = 0.15$ and $T_s = 12$. The nodes are scaled by $r_{i,j}$ to emphasize books that are highly associated with specific countries.



ISBN	USA	Canada	United Kin...	Germany	Australia	Spain	France	Portugal	Netherl...	Switzer...	Italy	Austria		Finland	f_{i+}
0671027360	486	36	11	6	7	3	1	8	2	0	0	2	···	0	586
0446672211	451	71	13	2	3	0	1	0	1	0	1	0		0	585
3442723833	**0**	**0**	**0**	**7**	**0**	**0**	**0**	**0**	**0**	**3**		**2**		**0**	**12**
0375725849	0	0	0	0	0	0	0	0	0	0	0	0	···	0	67
	⋮			⋮			⋮			⋮					
0671028375	68	49	6	2	2	0	1	0	0	1	0	0	···	0	135
f_{+j}	746490	93003	33200	27679	18238	14999	10653	6984	4976	4189	3493	2808		1192	

Figure 2: *(a) A large table from the Book-Crossing dataset with eight selected countries. (b) The Contingency Wheel showing how the books are associated with the selected countries. A node in a sector close to the outer radius represents a book highly associated with the sector category. One book is selected (marked red), and all nodes representing it are highlighted and connected (the red lines).*

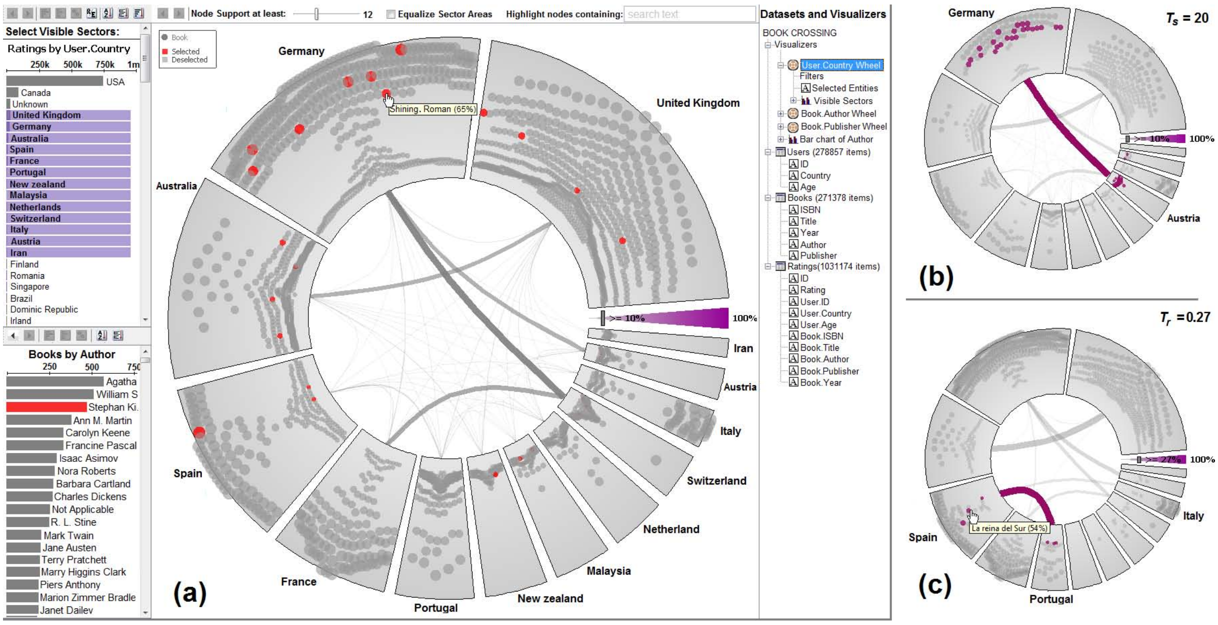


Figure 3: *(a) The main screen: from the top-left bar chart, the visible wheel sectors are selected. From the bottom-left bar chart, nodes that represent books by "Stephan King" are highlighted. (b) The node-support threshold is set to 20. The inter-sector links remain almost unchanged. (c) The association-strength threshold is set to $T_r = 0.27$. New inter-sector links are now visible.*

4. User Interaction

Fig. 3(a) shows a screenshot of our prototype. From a bar chart of the column categories, the user can select which sectors to show in the wheel. By clicking on a node, it will be selected along with all visible nodes of its row category. Connections similar to star coordinates show the frequencies $f_{i,j}$ for this category (Fig. 2b). Multiple nodes can be selected individually, by polar ranges, or by clicking on a sector. Other nodes of the same categories as of the selected nodes will also be selected. To reveal if books associated with one sector are also associated with other sectors the user can manually select the sector's nodes, or choose to augment the wheel with inter-sector links. Between each pair of sectors, a link is drawn whose thickness and opacity are proportional to the strength of the association between these two sectors, denoted by rc_{j_1,j_2}. This value is larger, when the two sectors share more books, and when these books have higher associations with both sectors:

$$rc_{j_1,j_2} = \frac{1}{f_{+j_1}+f_{+j_2}} \cdot \sum_{\substack{(0<i\leq n)\wedge(T_s\leq f_{i+})\\ \wedge T_r\leq r_{i,j_1}\wedge T_r\leq r_{i,j_2}}} (r_{i,j_1}+r_{i,j_2}) \qquad (2)$$

In Fig. 3(a) one can observe large associations between several pairs of countries. These associations remain consistent over different values of T_s, the node-support threshold, which confirms that they are not due to outliers (Fig. 3b). By changing T_r, the association-strength threshold, new pairs of associated countries become visible (Fig. 3c). By clicking on a link between two sectors, the nodes representing shared books will be selected. By examining them one can verify that large associations between countries are due to their common languages. In case the rows represent entities, an attribute of them can be mapped to node color. This helps further in analyzing what the nodes associated with a certain column category do have in common (Fig. 4). Additionally, several interactive queries can be performed on the nodes. In Fig. 3(a), books by the selected author in the bar chart of authors are highlighted in the wheel. Also, instant text search can be used to select or highlight the desired nodes.

5. Scenario-based Evaluation

We presented the proposed view to representatives of different departments of UC4 Software and to faculty members (eight users in total), and asked them to perform tasks on the Book-Crossing dataset. The dataset records 1,031,175 ratings given by 278,856 users on 270,170 books, along with the author and publisher of each book as well as the age and country of each user. Examples of these task are:

1. Find out which publishers are associated to each other, and which geographical markets they reach.
2. Which authors write children's literature? Which other authors are popular among the fans of Sandra Brown?

Fig. 4 shows snapshots of the views that can be interactively used to answer these questions. They visualize two contingency tables that break down ratings given by users (row categories) by the book's author (resp. publisher). We observed an initial hurdle in the translation of the tasks into

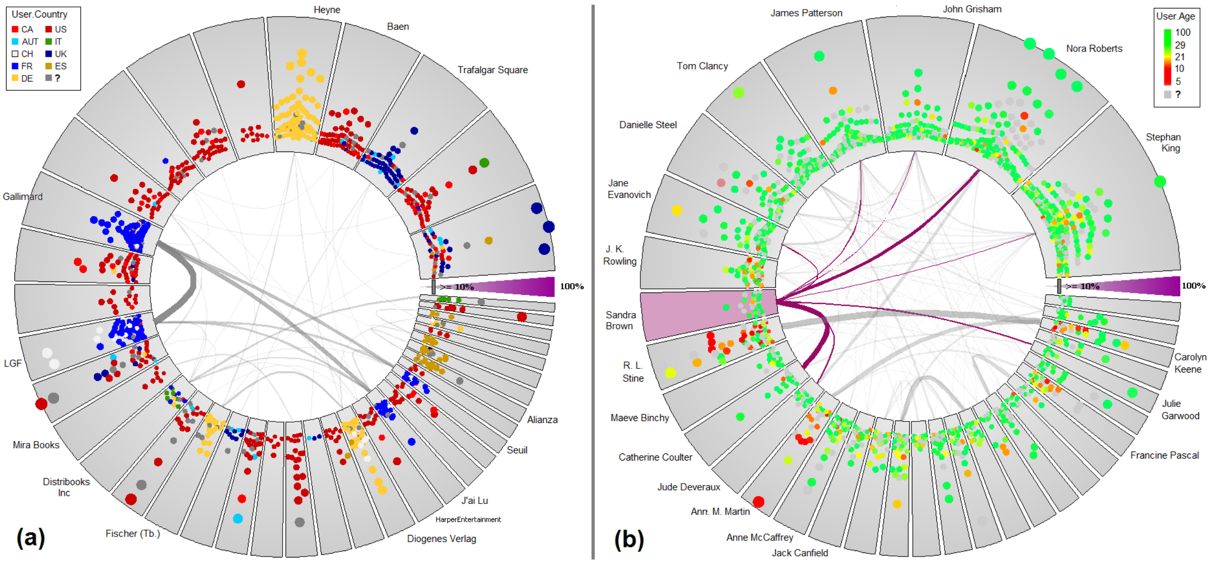


Figure 4: *Using Color for finding patterns: (a) Visualization of user ratings broken down by publisher. The nodes represent users and are colored by country. (b) User ratings broken down by author. The nodes represent users and are colored by age.*

the correct configuration of the view. For task 1 users intuitively decided to map the attribute "publisher" to sectors, whereas for task 2 users initially chose a different attribute and then switched as they did not receive viable results. Once the appropriate configuration for answering task 2 was reached (Fig. 4b) it was straightforward to identify authors of children's literature by coloring the nodes by the user's age. Clicking on the sector representing "Sandra Brown" enabled to identify related authors via inter-sector links. In fact, these authors happen to share the same writing genres.

The visual metaphor of sectors and associations was perceived as intuitive. In particular, the possibility to interactively filter the nodes based on association strength helped in better understanding the metaphor. Coupled with the inter-sector links, it helped to identify fans that are strongly associated with two authors. Moreover, these links helped in accidentally finding irregularities in the data, such as different spellings for the same author, resulting in thick links between the repeated categories. The node-placement metaphor enabled revealing interesting patterns in the data. For example, many users noticed in the wheel of Fig. 2 that more nodes in non-English speaking countries tend to be highly associated with these countries, whereas the opposite holds for English-speaking countries. The unused space left by the placement algorithm was sometimes confusing, possibly because the angular position is not interpretable. Also, the distribution of data attributes within sectors was in some cases hard to understand despite the use of node color and size. The view was able to handle large contingency tables with 100 columns, and hundreds of thousands of rows. Finally, the association measures in equations (1) and (2) suffer from biases due to different sector sizes.

6. Conclusion

We presented the Contingency Wheel, an interactive visualization technique for analyzing associations in large 2-way contingency tables. User feedback highlighted the value of the presented visual and interaction metaphors in discovering and analyzing associations in the data, and the effectiveness of linked views in brushing and filtering the data in the wheel. Future work will focus on exploring different placement algorithms and association measures, as well as providing interactive drill-down in the sectors and edge bundling.

Acknowledgement This work was supported by the Austrian Federal Ministry of Economy, Family and Youth via CVAST, a Laura Bassi Centre of Excellence, and by UC4 Software - Austria.

References

[Ben90] BENZÉCRI J. P.: *Correspondence Analysis Handbook*. Marcel Dekker, New York, 1990. 1

[BKH05] BENDIX F., KOSARA R., HAUSER H.: Parallel sets: visual analysis of categorical data. In *Proceedings of the IEEE Symposium on Information Visualization* (2005), pp. 133–140. 1

[HK81] HARTIGAN, KLEINER: Mosaics for contingency tables. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (1981), Springer-Verlag, pp. 268–273. 1

[Sne74] SNEE R. D.: Graphical display of two-way contingency tables. *The American Statistician* (1974), 9–12. 1

[UTH06] UNWIN A., THEUS M., HOFMANN H.: *Graphics of Large Datasets: Visualizing a Million*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 1

[ZMKL05] ZIEGLER C.-N., MCNEE S. M., KONSTAN J. A., LAUSEN G.: Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web* (2005), ACM Press, pp. 22–32. 2