# A Tool for Analyzing Categorical Data Visually with Granular Representation

Kousuke Shiraishi, Kazuo Misue, and Jiro Tanaka

Department of Computer Science, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan
{shira,misue,jiro}@iplab.cs.tsukuba.ac.jp

**Abstract.** Categorical data appears in various places, and dealing with it has been a major concern in analysis fields. However, representing not only global trends but also local trends of data simultaneously by conventional techniques is difficult. We propose a visualization method called “granular representation” for analyzing categorical data visually. Our approach visually represents data as a set of objects and allows intuitive analysis instead of the traditional way with tables of numbers. We developed a tool by integrating granular representation and bar charts. The effectiveness of the tool is demonstrated using real data about media consumption.

**Keywords:** categorical data, visualization, multi-dimensional analysis.

## 1 Introduction

The visualization of categorical data occurs in various fields, such as mass media and marketing research. Visualizing categorical data has two purposes [1]. The first purpose is to present global trends or characteristics of data to the audiences of books, televisions, newspapers, and so on. In this case, the process of visualizing data is simple and facile because results are shown with conventional graphics, such as a bar chart or pie chart. The second purpose is to analyze data in detail to use it in research. For this, analyzing both global and local trends of data in further detail is necessary. In other words, discovering relations among two or more attributes is needed for multi dimensional analysis. Generally, this visualization is performed with spreadsheet software such as Microsoft Excel or statistical tools such as Statistical Package for the Social Sciences. However, analyzing intuitively from tables of numbers and letters is difficult, especially for casual users.

We present a granular representation (GR) technique for analyzing categorical data visually. Our approach represents individual entities in data separately as small circles. Users are able to drill down data and analyze both global and local trends interactively.

## 2 Analyzing Categorical Data

This section describes the general process of analyzing categorical data. An overview of the process is shown in Fig. 1. The process is divided into two parts: “tabulation”

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 342–351, 2009.

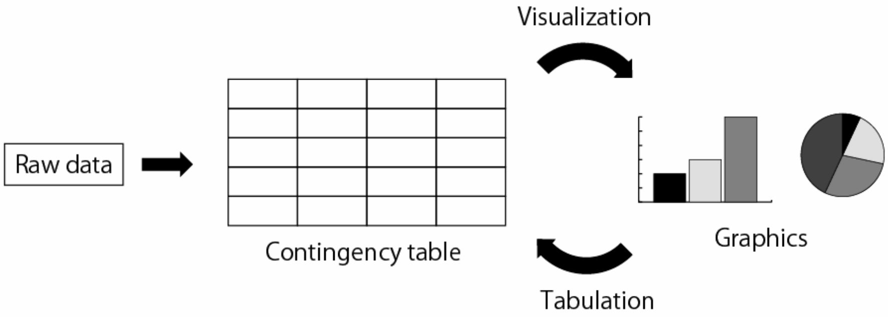


**Fig. 1.** Analyzing categorical data is divided into two parts: tabulation and visualization

and “visualization”. “Tabulation” forms a table consisting of the distribution of categories. These tables are called cross tables or contingency tables. Initially, most of the categorical data is provided as lists in which each record corresponds to an individual entity. This raw data is transformed into a contingency table by counting the frequencies of categories in attributes. This counting of frequencies is performed automatically. For instance, this calculation can be performed in Excel by simply selecting the attributes of interest.

The second part of the process is “visualization”. As the name indicates, in this part, the table made in the preceding “tabulation” part is represented by visualization techniques. There are several types of visual representations available. Graphics such as a bar chart or pie chart are the most common as visual representations.

Analysis is usually a recursive process that repeats “tabulation” and “visualization”. In these processes, if the number of attributes and categories increases, maintaining correspondence among elements in charts and tables is difficult, especially for casual users.

Sometimes, categorical data may include an attribute that is difficult to categorize. An example is comments included in questionnaires and census data. Although knowing the tendency of comments as a whole is possible by reading all the comments in order, the important thing is to analyze the relations between comments and segments, such as “what segments of people make what comments”. However, conventional graphics, such as a bar chart or pie chart, are only able to represent global data trends. Analyzing from both the global and local sides simultaneously is difficult.

## 3 Related Work

Many techniques have been developed to efficiently visualize categorical data. There are two approaches: frequency-based and quantification.

Frequency-based techniques represent frequency values in contingency tables proportional to the size or length of the figures. Mosaic Display [2] represents the frequency values with rectangles proportional to the area and lays out the rectangles like tiles recursively. Cattrees [3] is also frequency-based visualization and uses Treemap [4]. These techniques are effective for visualizing a couple of attributes. However, more than three attributes leads to complicated visualizations.

Quantification techniques transform categorical data into graphical representations of quantitative data. Rosario et al. [7] proposed a quantification technique for categorical data and presented an example of visualization using parallel coordinates. Johansson et al. [8] suggested a quantification process for mixed data that contains categorical and continuous variables.

Some techniques use parallel coordinates. Hammock Plot [5] corresponds frequency values with line widths using parallel coordinates. Parallel Sets [6] combines frequency-based techniques and parallel coordinates and offers interactive analysis.

Dust&Magnet [9] is a visualization technique of multivariate analysis for quantitative data. Dust&Magnet represents each data point as a dot and uses a magnet metaphor for intuitive analysis. The magnitudes of the magnets correspond to the attribute values of each data point. Attracting data points with magnets spatially separates positions of each data point depending on attribute value. Similar to Dust&Magnet, our proposed GR technique represents individual records as visual elements. However, Dust&Magnet represents the differences of attribute values by the distances between data points. In contrast, GR represents the quantitative differences among collectives. Thus, several interactions and the interfaces of the tools are different.

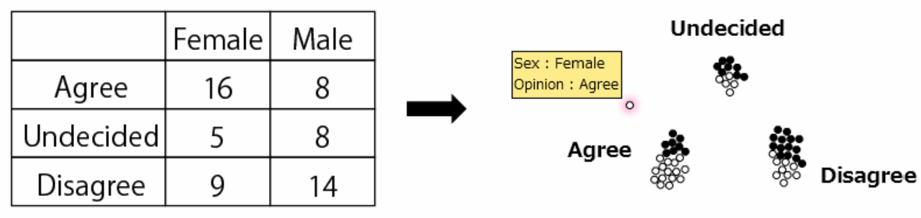
## 4 Granular Representation

Granular representation is the visualization technique we propose. Categorical data, due to its nature, can be thought of as sets of objects. For instance, questionnaires or census data consist of people's opinions. The basic idea of our approach is to effectively represent that nature visually instead of as tables and to enables users to intuitively tabulate and visualize. Consider an example of a contingency table that consists of two attributes, "sex" and "opinion." The "sex" attribute has two categories, "male" and "female," and the opinion attribute has three categories, "agree," "disagree," and "undecided." An example of visualizing this contingency table with GR is shown in Fig. 2. Small white or gray circles represent individual entities, that is, each circle can be thought of as a person in this example. These circles are called elements. Text labels represent the categories of an attribute, e.g., "opinion." In this way, GR displays the frequency value of each cell as small circles like grains.

Users are able to see information about each individual entity (element) by hovering the cursor over the element. See the rectangle on the right side of Fig. 2. This contains information about the element on the left far from the others. It means that this element is female and her opinion is "agree."

We tend to perceive spatially close objects as groups. This has been theorized as the law of proximity by psychologists. If we look at the GR on the right side of Fig. 2, we immediately see that there are three collectives of elements, and the labels reveal their categories. In this figure, the collective at the top is people who have an "undecided" opinion, the left is those who "agree", and the right is those who "disagree".

Color is also good for representing categories of elements. Elements can be easily seen as groups if they have the same properties, such as color, shape, or texture. This perception is called the law of similarity. The elements in Fig. 2 are colored by their "sex" attributes; white represents females and black represents males. With color, the "sex" attribute is identified without a label. In this way, GR represents categories of elements by label and color.



|  | Female | Male |
|---|---|---|
| Agree | 16 | 8 |
| Undecided | 5 | 8 |
| Disagree | 9 | 14 |

**Fig. 2.** Contingency table (left) represented by granular representation (right)

## 5 Interactions

We explained the basis of GR in the previous section. Although the representations are static images, the formation of the element layout is animated. In this section, we introduce various interaction techniques used to control elements. Users can drill down interactively using these techniques.

### 5.1 Categorization by Label

Users divide, i.e., “categorize”, elements into collectives to drill down data. In Fig. 2, elements are categorized into three collectives. Categorization of elements is performed using a label. Elements from the contingency table in Fig. 2 are represented by GR on the left side of Fig. 3. The elements are all collected together, so the female and male elements are mixed. To separate the male elements from this collective, the user drags the “male” label, and the elements having the “male” category follow because they are attracted towards the label (right side of Fig. 3). After the dragging stops, the elements are categorized into two collectives; the rightmost one is male and the other is female.

Dust&Magnet supports a similar interaction. It uses this feature as the attraction of magnets. In contrast, GR uses this as the categorization of elements.

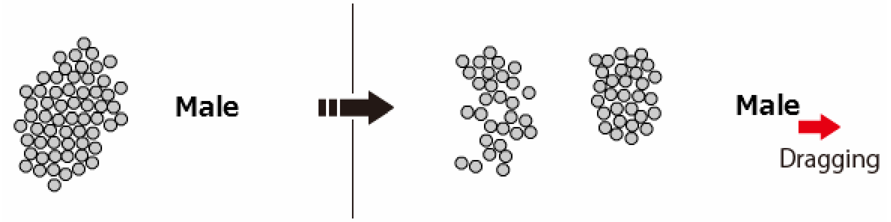


**Fig. 3.** When the “male” label is dragged, elements having the “male” category are attracted towards the label

### 5.2 Controlling Attraction of Label

Dragging a category label attracts all the elements that have that category. However, categorizing elements by the relation of two or more dimensions with this method can be inconvenient. For instance, consider three collectives divided by the “opinion”

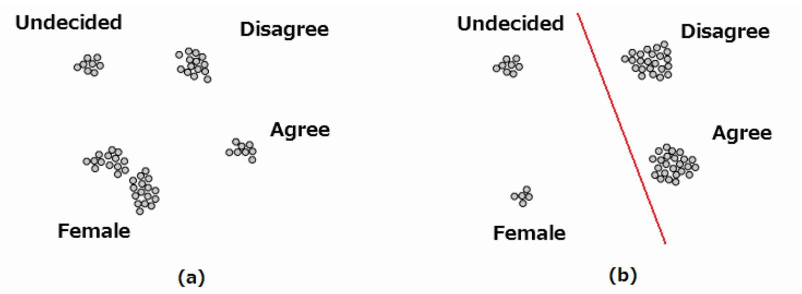


**Fig. 4.** (a) Attraction of a label without a line. (b) Attraction of a label using a line.

attribute, such as in Fig. 2. In this case, if the user drags the “female” label, elements that have the “female” category will be attracted from each of the three collectives, and then the attracted elements will collect together (Fig. 4(a)). To control the attraction of a label, the user draws a line to separate the collectives he or she does not want to be affected by attraction. In Fig. 4(b), the line separating the two collectives on the right side from the label ensures that attraction does not affect those collectives. A label attracts elements from only collectives on the left side of the line. In this way, a line allows users to categorize the relations of elements by more than two attributes to drill down.

### 5.3 Clusters

Attributes usually contain many categories. Categorizing by labels is difficult because each category label must be dragged separately, due to the features of labels. Clustering automatically divides elements that have the same categories by the selected attributes, so the labels do not have to be manually controlled. Clustering by two attributes is shown in Fig. 5. The elements are divided into six clusters. A glance at the figure will reveal the differences in the number of elements at once. Clustering supports users in discovering data trends in advance of categorizing by label.

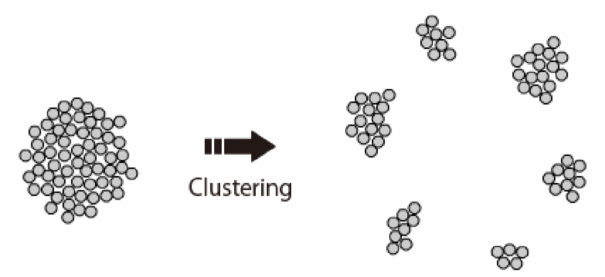


**Fig. 5.** Elements automatically clustered into six collectives by their categories

### 5.4 Merging

Due to our approach representing each individual entity, visualizations are more complicated as the number of data entities is much larger than in other approaches. To deal with data containing large amounts of individuals, GR has a “merging” interaction that represents multiple entities as one element proportional in size to the number of entities. An example of a merging interaction is shown in Fig. 6. The 60 elements on the left side merge into 30 elements (the center) and finally merge into 15 elements. That is, in each merging interaction, two elements merge into one. GR can deal with data containing many records by reducing the number of elements as a whole by merging.

Although the individual information of merged elements cannot be seen when the cursor is hovered over them due to the merged elements containing several elements, users are able to divide merged elements into their former states. Merging and dividing interactions can be frequently switched. Therefore, users can merge when analyzing the global side and divide to see individual information.

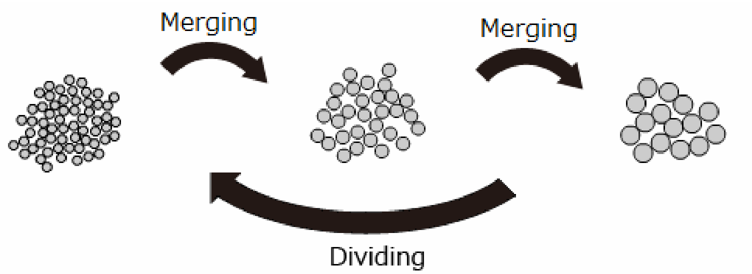


**Fig. 6.** Two elements merge into one element of proportional size

## 6 Construction and Implementation of Tool

In this section, we talk about the interface of the tool developed with our approach. In section 2, we explained the process of analyzing categorical data as being roughly divided into two processes, “tabulation” and “visualization”. Categorization by label or clustering corresponds to the “tabulation” process, and comparing the number of elements among collectives corresponds to the “visualization” process. Although GR supports both tabulation and visualization, a bar chart is more suitable for comparing values. We designed an interface for the tool by integrating GR and a bar chart.

A screenshot of the tool, which is divided into two views, is shown in Fig. 7. On the left is a settings and chart view, where users can control settings and where data are represented by bar charts. On the right is the main view, where data are represented by GR. The user mainly works in this view.

To integrate these two views, linking and brushing techniques are applied in the tool. Linking connects the common elements between two different visualizations. Between the left and right views, the bar charts and colors of elements are linked. Brushing summarizes a part of all the elements. Users can make bar charts of a part of whole elements by selecting the desired part.

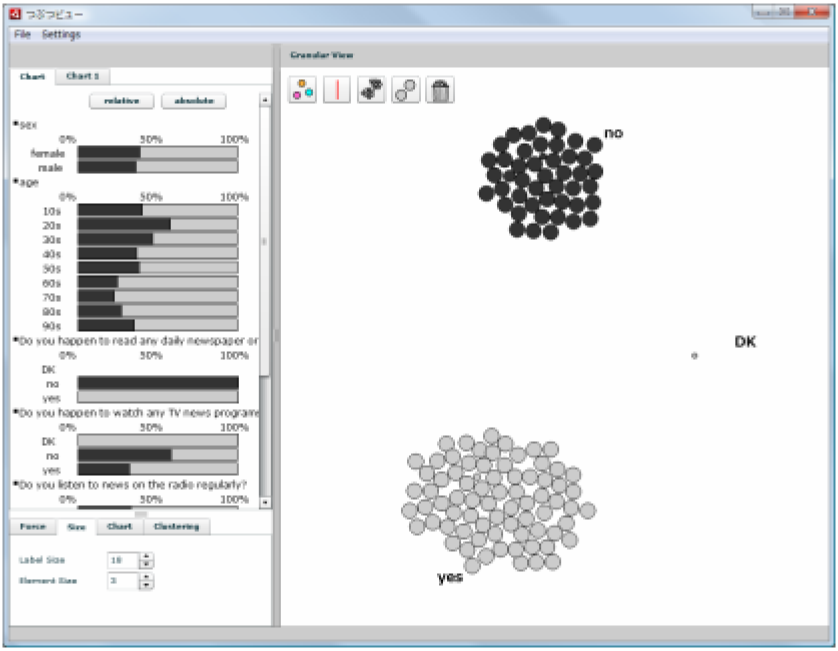
### 6.1 Drawing and Layout of Elements

The layout of the elements is manipulated through a simple physical model, and they are frequently animated. There are two forces: a repulsive force among elements to avoid occlusion and an attractive force to attract elements by labels. After the elements are dragged by their labels, they spread out in a circle due to the effects of their repulsive forces. Seeing elements as a large clustered circle helps users to compare numbers. For example, Fig. 7 shows two clusters, the lower one larger than the upper one. The attractive force of a label changes according to the distance between the label and elements. Thus, elements are attracted by a strong force when the distance is long.

## 7 Walkthrough

Here we show an example of analysis using the tool. The data set used in this section is a media consumption survey for April to May, 2006 [10]. We chose eight attributes from the data set: two basic attributes, “sex” and “age”; four questionnaire attributes about the consumption of media, such as television, radio and the Internet; and two attributes with unique variables being comments about liking or disliking newspapers. In this example, we analyze “the difference between people who read newspapers and those who do not.” As the data set is read, all the individual records are represented as gray circles in the main view. By default, there are no bar charts on the right.

Next, the user divides the elements into two collectives: one reads newspapers and the other does not. To divide the elements, the user chooses the “yes” and “no” labels of the “Do you happen to read any daily newspaper or newspapers regularly?” attribute from the left view. The “yes” and “no” labels then appear in the main view.



Label Size 10
Element Size 3

**Fig. 7.** Screenshot of tool

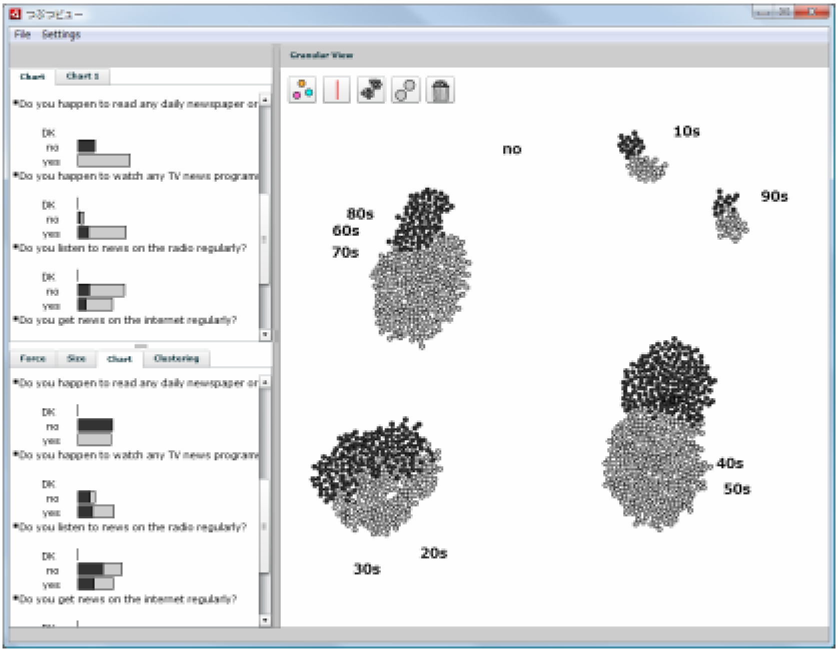

File
Settings
Chart
Chart 1
*Do you happen to read any daily newspaper or
DK
no
yes
*Do you happen to watch any TV news program
DK
no
yes
*Do you listen to news on the radio regularly?
DK
no
yes
*Do you get news on the internet regularly?
Force
Size
Chart
Clustering
*Do you happen to read any daily newspaper or
DK
no
yes
*Do you happen to watch any TV news program
DK
no
yes
*Do you listen to news on the radio regularly?
DK
no
yes
*Do you get news on the internet regularly?
Granular View
no
10s
90s
80s
60s
70s
40s
50s
20s
30s

**Fig. 8.** Elements are divided by ages. Upper bar charts represent values of 60s, 70s and 80s collective, and lower ones represent values of 20s and 30s collective.

As the user drags these labels, the elements are attracted towards their relevant labels and divide into two collectives.

To analyze the difference of these two collectives, the user colors each collective differently and then selects each collective to make bar charts (Fig. 7).

The chart view in Fig. 7 shows there could be a difference in opinion by "age". Older people read newspapers more often than younger people, except people under 20 and those in their 90s. To analyze this trend in detail, the user further divides the elements by age. Here, the user obtains four collectives divided by age.

Figure 8 shows that the elements are divided into four collectives and all the merged elements are released. The upper chart view shows the frequencies of the 20s and 30s collective and the lower one shows the frequencies of the 60s, 70s, and 80s collective. From the bar charts on the left, many differences between the two collectives can be seen. For instance, take the attribute "Do you happen to watch TV news programs regularly?" The "yes" rates of the 60s, 70s, and 80s collective are higher than those of the 20s and 30s collective. In other words, senior people get news from the TV more frequently than the young.  In addition, it is possible to analyze the local trends between these two collectives by seeing the comments for each collective.

## 8 Discussion

We believe that GR has three advantages, as we have analyzed several data with the tool using GR.

The first advantage is the simultaneous representability of both the quantitative and qualitative sides of data. The quantitative side means the global trends of the data, i.e., analyzing the frequency distribution by comparing the numbers of elements among the collectives. The qualitative side means the local trends. The user can analyze this by hovering the cursor over elements and seeing individual information. Traditional visualization techniques, such as a bar chart or pie chart, are only able to represent the global side of data. For instance, in the previous walkthrough, the elements were divided into five collectives by the age attribute. Seeing the individual information for each collective uncovers trends between segments and unique variables such as comments.

The second advantage is the ability to represent absolute values. Frequency-based techniques represent the rates of two or more variables by comparing shape size or length of the representative figure. Although these techniques are suitable for comparing quantities, they cannot represent the absolute values of numbers. In contrast, GR represents individual entities as individual, independent elements; therefore, absolute values can be compared.

The third advantage is intuitiveness. Most traditional analysis tools need special knowledge to use. In particular, casual users require a lot of time to understand the tools. GR visually represents the structure of data, and the interactive control of animated elements enables users to analyze the data intuitively.

## 9 Conclusion and Future Work

We proposed granular representation (GR), a visualization method for analyzing categorical data, and several interactions that fit our approach. We designed an interface for a tool by integrating charts and GR to improve the effectiveness of both representations. An analysis example using real data was presented.

Although we discussed the advantages of GR in the previous section, the validity of these advantages has not yet been evaluated. For our next work, we plan to evaluate the effectiveness and validity of the advantages.

## References

1. Friendly, M.: Visualizing Categorical Data. Sas Inst. (2000)
2. Friendly, M.: Mosaic displays for multi-way contingency tables. American Statistical Association 89(425), 190–200 (1994)
3. Kolatchm, E., Weinstein, B.C.: Dynamic visualization of categorical data using treemaps (2001), `http://www.cs.umd.edu/class/spring2001/cmsc838b/Project/Kolatch_Weinstein/index.html`
4. Johnson, B., Shneiderman, B.: Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In: Proceedings of IEEE Information Visualization 1991, pp. 275–282 (1991)
5. Schonlau, M.: Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. In: Proceedings of the Section on Statistical Graphics, American Statistical Association (2003)

6. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann Pub., San Francisco (1999)
7. Yi, J.S., Ponder, R.M., Stasko, J., Jacko, J.: Dust & Magnet: multivariate information visualization using a magnet metaphor. Information Visualization 4, 239–256 (2005)
8. Johansson, S., Jern, M., Johansson, J.: Interactive Quantification of Categorical Variables in Mixed Data Sets. In: Proceedings of IEEE International Conference on Information Visualization (IV 2008), pp. 3–10 (2008)
9. Rosario, G.E., Rundensteiner, E.A., Brown, D.C., Ward, M.O., Huang, S.: Mapping nominal values to numbers for effective visualization. In: Proceedings of the IEEE Symposium on Information Visualization 2003 (INFOVIS 2003), pp. 80–95 (2003)
10. Biennial Media Consumption 2006, Pew Research Center Data Archive (2006), `http://people-press.org/dataarchive/`