SimulSort: Multivariate Data Exploration through an Enhanced Sorting Technique

Inkyoung Hur and Ji Soo Yi

School of Industrial Engineering, Purdue University
315 N. Grant Street, West Lafayette, IN 47907-2023, USA
{ihur,yij}@purdue.edu

Abstract. Sorting is one of the well-understood and widely-used interaction techniques. Sorting has been adopted in many software applications and supports various cognitive tasks. However, when used in analyzing multi-attribute data in a table, sorting appears to be limited. When a table is sorted by a column, it rearranges the whole table, so the insights gained through the previous sorting arrangements of another column are often difficult to retain. Thus, this study proposed an alternative interaction technique, called "SimulSort." By sorting all of the columns simultaneously, SimulSort helps users see an overview of the data at a glance. Additional interaction techniques, such as highlighting and zooming, were also employed to alleviate the drawbacks of SimulSort. A within-subject controlled study with 15 participants was conducted to compare SimulSort and the typical sorting feature. The results showed typical sorting and SimulSort work with comparable efficiency and effectiveness for most of the tasks. Sorting more effectively supports understanding correlation and reading corresponding values, and SimulSort shows the potential to more effectively support tasks that need multi-attribute analyses. The implications of the results and planned future work are discussed as well.

Keywords: Sort, SimulSort, information visualization, multi-attribute data analysis, tabular information, and decision support system.

1 Introduction

Sorting, arranging items in an ordered sequence, is one of the well-understood and universally used interaction techniques in many software applications, such as spreadsheets, word processors, and even email clients (e.g., sorting emails by date). Many user interfaces that contain any kind of list or table of data employ the sorting feature. With sorting, people can accomplish various tasks, such as searching for a particular or extreme value, identifying the general patterns of values, or determining relationships between values in two or more different columns.

However, sorting does not appear to be ideal for supporting multi-attribute data analyses. For example, suppose a consumer tries to select a car. She then collects information about various cars and archives the information on a spreadsheet. While reviewing the cars, she may sort the spreadsheet by the "price" column to find the ten least expensive cars. Then, she might want to pick out the most fuel efficient car among the ten. However, by sorting the information with the "fuel efficiency"

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 684–693, 2009.

column, she loses insights gained through the previously sorted information. To avoid this, she should have marked the ten cars on the spreadsheet or somewhere else. However, even with this extra effort, she may run into problems in analyzing multi-attribute information if the number of considered attributes increases.

To resolve this issue, the present study proposed an interaction technique, called "SimulSort." When SimulSort is used on a table of data, all columns in a table are sorted simultaneously, so that the corresponding attributes of a data point (e.g., the price and fuel efficiency of a car) are no longer shown in the same row. Instead, when a mouse cursor hovers on a cell of the table, all corresponding cells (or different attributes of a record) in different columns are highlighted. This visual representation of tabular data may be perceived as somewhat unfamiliar, but it would help people compare two or more choices by considering multiple attributes simultaneously. An additional visualization technique, such as zooming, was also employed to overcome some drawbacks of SimulSort. To identify the advantages and disadvantages of SimulSort over the typical sorting feature, we also conducted a usage study.

The rest of this paper is organized as follows: Section 2 discusses relevant literature regarding sorting interaction techniques and information visualization. Section 3 introduces SimulSort and interaction techniques. Section 4 lists the research hypotheses used in comparing typical sorting and SimulSort. Section 5 details the design of the experiment. Sections 6 and 7 describe the results of the experiment and their implications. Lastly, section 8 discusses the conclusions of this study and some future work.

2 Background

2.1 Sorting

In spite of the ubiquity of sorting, we found little literature discussing how sorting has been used so far. In 1785, Priestley made an early breakthrough in visual sorting with the first timeline charts, which used individual bars to compare the life spans of persons [1]. In 1901, Hollerith developed a sorting machine, called a "card sorter," that was employed to sort US census data, saving more than two years [2]. In 1967, Bertin emphasized the importance of the correspondence between the act of sorting values and its representation in helping readers gain a meaningful understanding of the data and useful retention [3]. This visualization theory of Bertin became a discipline of sorting visualization [4]. Other than these, most of literature regarding sorting has been devoted to developing and evaluating different sorting algorithms in computer science, which is not the focus of this paper.

In spite of a dearth of literature regarding sorting, the universal adoption of the sorting feature by numerous software applications proves that it is a well-understood and useful interaction technique. We found that sorting could be helpful in various tasks, such as finding extreme values or outliers, finding a certain value, verifying the existence of a certain value, detecting patterns of data, understanding the relationships of data, and organizing or clustering data.

However, we also noticed when multiple attributes must be analyzed simultaneously, sorting appears to be limited. As described in the previous scenario, when a tabular data set is sorted by an attribute, the whole table is rearranged, making previous arrangement disappears, which limits the utility of the feature when multiple attributes should be considered together.

One might argue that the multi-column sorting feature, found in Microsoft Excel®, would help overcome this limitation because it sorts multiple columns at the same time. However, this sorting feature is only useful as a tie breaker. For example, if data are sorted by Columns A and B, the data are sorted by Column A first, and sorting by Column B only affects on the rows that have the same values in Column A. Other rows that have different values in Column A will stay the same. Thus, a different approach is necessary to deal with this issue.

2.2 Multivariate Information Visualization

The limitation of the currently available sorting has been remedied in various information visualization techniques. One would be parallel coordinates [5, 6], which layouts multiple axes in parallel and represents data points as multiple lines connecting these axes as shown in Fig. 1. By presenting multiple axes in parallel, a user can explore multiple attributes at a glance without changing sorting orders of different columns. Instead, lines connecting those axes represent a data point. Though various implementations of parallel coordinates have subtle differences in detailed interaction techniques, all of them share the same visual representation technique.

Another approach is parallel bargrams, which is more similar to tabular view. Each row represents an attribute and values in each row are sorted as shown in Fig. 2. Parallel bargrams has been implemented in different systems, such as MultiNAV [7], FOCUS [8] (which evolved into InfoZoom [9]) and EZChooser [10]. Again, different implementations have subtle differences in interaction techniques, such as how filtering works, but they essentially rely on the same visual representation.

As easily noticeable from Fig. 1 and 2, both visualization techniques are designed to deal with multi-attribute data. They help users understand general trends and interesting patterns at a glance without changing the arrangement of information. In other words, the data points are already sorted. In parallel coordinates,

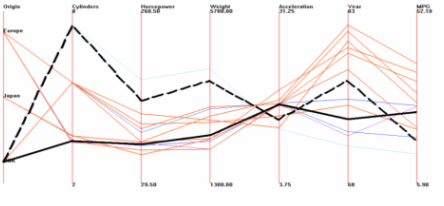

Fig. 1. Parallel coordinates (Parvis [6])

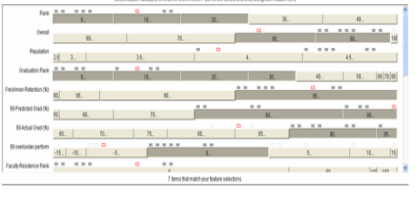

Fig. 2. Parallel bargrams (EZChooser [11])

each axis represents a sorted column. In parallel bargrams, each row is sorted horizontally.

However, they also have limitations. First, the novelty of techniques may cause them to be relatively difficult to learn and use. Both visual representations are quite different from a familiar tabular view, so, for novice users, these tools could be challenging to use. Second, visualizing a large data often hides the details of information. As shown in Fig. 1 and 2, detailed information about each individual data is generally hidden. In other implementations, detailed information could be shown upon requests using tooltip features, but we found that these could be generally problematic since users cannot see these details readily.

Table Lens [12], which was designed to handle a large table with a zooming interface, is inspiring. Though Table Lens does not specifically resolve the limitation of sorting, it is easy to learn and use since it is based on a traditional tabular view and shows the overview of large data and details using the zooming techniques. Table Lens also uses a horizontal bar graph on top of a numerical value in each cell to help a user quickly identify the trends in the data.

3 SimulSort

Thus, we propose an alternative visualization technique, called "SimulSort," which combines the strengths of the reviewed visualization techniques but still retains and leverages the well-understood sorting feature in a tabular view. Fig. 3 is a screenshot of SimulSort, and contains a data set of 70 used cars as an example. As shown in the figure, every column is sorted. The attributes of a car are no longer shown in a single row. Instead, in order to retrieve values of a certain car, the highlight feature should be used. When a mouse cursor hovers over the cell, Car No. 11, the corresponding cells are highlighted in yellow as shown in Fig. 3. This visualization supports the similar tasks that parallel coordinates and parallel bargrams support. Comparing Car No. 11 (highlighted in yellow) and Car No. 12 (highlighted in light blue) over multiple attributes becomes much easier. A user does not need to sort different columns multiple times to compare these two cars.

One might notice that some values are not shown in Fig. 3. For example, the mileage of Car No. 11 is not shown in the figure. Thus, a zooming feature is provided, so that a user can zoom out to show all of the rows at a glance as shown in Fig. 4 by selecting a radio button on the top left portion of the screen. Of course, due to the limitation of screen real estate, in the zoomed-out view, the detailed number for each cell cannot be shown. Instead, bar graphs provide information to help a user compare between values. When the user wants to have details, the screen can be changed back to the zoomed-in view.

SimulSort is implemented using Adobe Flex and flare, so that it can be interactive and accessible through most of the web-browsers in the market. The data source used in SimulSort is XML, so that any type of tabular data that is convertible to XML used by SimulSort.

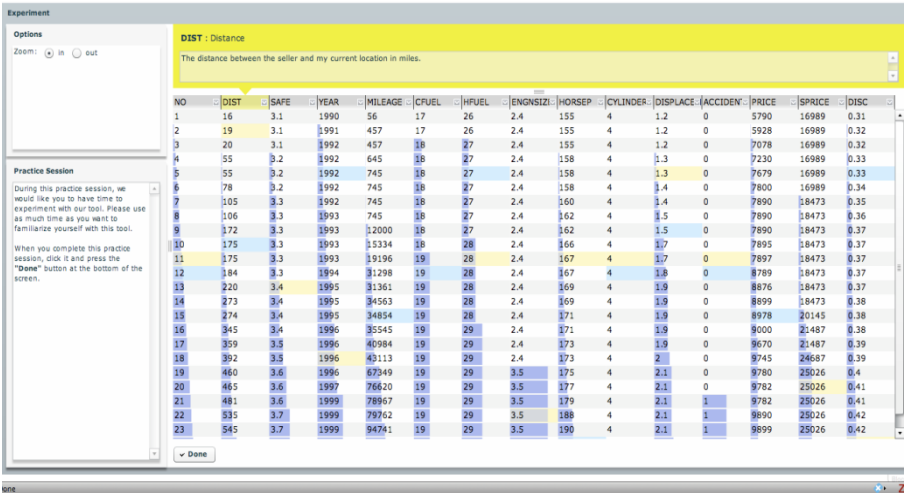

Fig. 3. A screenshot of a table using SimulSort with a zoomed in view

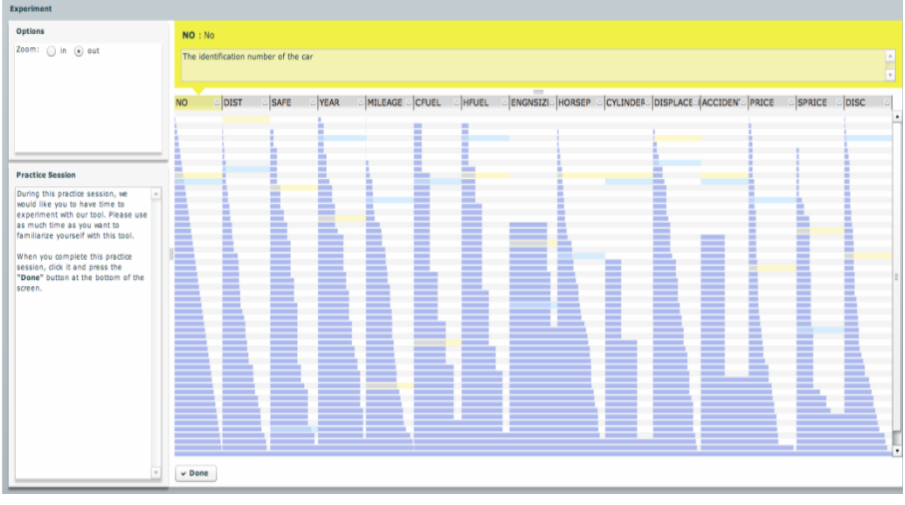

Fig. 4. A screenshot of a table using SimulSort with a zoomed out view

4 Hypotheses

After pilot studies with three participants, we found potential in SimulSort to support various tasks. To test this speculation, two sets of hypotheses were constructed. The first set of hypotheses is comparing the effects of two sorting techniques (typical sorting vs. SimulSort) on performance measures, such as a response time and accuracy, of various tasks. A sub set of tasks were selected from the low-level analytic activities for multivariate data surveyed by Amar et al. [13].

- H1-1: Participants retrieve values more quickly and correctly while using typical sorting than while using SimulSort.
- H1-2: Participants filter values more quickly and correctly while using typical sorting than while using SimulSort.
- H1-3: Participants sort values more quickly and correctly while using typical sorting than while using SimulSort.
- H1-4: Participants find a correlation more quickly and correctly while using typical sorting than while using SimulSort.

In addition, we would like to investigate the effect of different sorting techniques on the decision performances, such as response time and decision quality. We also hypothesized that the discrepancy will increase as the number of attributes to consider increases from three to seven.

- H2-1: Participants make a decision with the smaller number (three) of criteria more quickly and confidently while using SimulSort than while using typical sorting.
- H2-2: Participants make a decision with the larger number (seven) of criteria more quickly and confidently while using SimulSort than while using typical sorting.

5 Methods

5.1 Participants

Fifteen participants (7 males and 8 females; age: 21 – 31) at Purdue University were recruited for the experiment. The completed education level of these participants varied from high school to master's degree, but all had general, high-level computer experience. On average, they have used a computer about 13 years. Fourteen out of 15 participants reported that they are "comfortable" or "very comfortable" in using computers. At the end of experiment, they were compensated at the rate of eight US dollars per hour.

5.2 Datasets

Three artificial datasets of used cars were generated for the experiment. Each dataset has 70 used cars and 14 attributes. Random numbers were generated to make three datasets different enough to avoid any learning effects. At the same, to make complexities of the three datasets equivalent, the average of inter-attributes correlations were adjusted around zero (i.e., 0.0001, -0.0002, and 0.0003). This blocked the potential effects of inter-attributes correlations on decision quality [14].

5.3 Procedure

When a participant arrived at the lab, the pre-task survey was conducted to collect subject's demographic information, including education levels and computer literacy. After the pre-task survey, participants were asked to answer questions on paper in one baseline setting (i.e., without any sorting feature) and two experimental settings (i.e., with typical sorting and with SimulSort). The orders of the two experimental settings

and used datasets were randomized. Whenever a new setting was introduced, an experimenter described how to use the interface, and participants had a chance to explore the interface without time limitation. After completion of each setting, the subjects were asked to answer questions regarding decision qualities. At the end, participants were interviewed before being compensated.

5.4 Measures

Response time, accuracy, and decision quality were used as performance measures. The response time was derived by taking the difference of the recorded times each task started and completed. Task accuracy is binary in conformity to the answer is correct or incorrect. Decision qualities of each decision were captured through the post-task questionnaire in terms of confidence of their decisions, using a five-level Likert scale:

$$Decision\ Quality = \frac{\text{WeightedAdditiveValue Choice} - \text{WeightedAdditiveValue Worst}}{\text{WeightedAdditiveValue Best} - \text{WeightedAdditiveValue Worst}} \quad (1)$$

6 Results

While analyzing data, t-tests, Mann-Whitney test and two proportions tests were employed for the quantitative measures (e.g., response times and decision qualities) and binary measures (e.g., accuracy), respectively. Table 1 summarizes the uncovered findings, the effects upon the tests gives the means, standard deviations (S.D.), and *t* value, W value, and *p* values for each question and hypothesis.

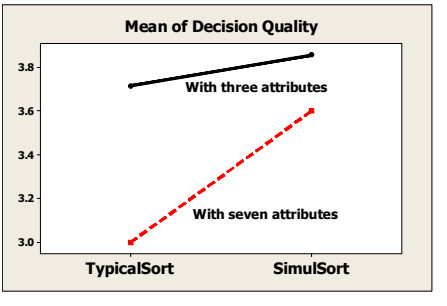

Fig. 5. Means of the decision quality for different settings and number of attributes

Participants while using SimulSort did not demonstrated significantly different performances in most of the activities (H1-1 and H1-2 were falsified) from while using typical sorting features except for two activities. For example, participants sorted values more quickly (H1-3 was partially supported) and found a correlation more correctly (H1-4 was partially supported) while using typical sorting than while using SimulSort, as shown in Table 1.

When considering the number of criteria in the decision making activity, we found slight differences in the mean scores of decision quality, as shown in Fig. 5. SimulSort and typical sorting features were not judged to be different on the response time and decision quality with three criteria (H2-1 was falsified). The mean of decision quality of SimulSort with seven attributes was 3.6, which was higher than the mean of decision quality of typical sorting features, 3.0, but the difference was not statistically significant at the error level of 0.05, though (H2-2 was falsified).

Table 1. Results of t-test, Mann-Whitney test, and two proportions test of using typical sorting and SimulSort

Question/ Artifact	Response time (seconds)		Accuracy (Decision Quality)	
	Mean (S.D.)	**t(df)/W , p values**	**Mean (S.D)**	**t(df)/W , p values**
Question 1	What is the government-tested safety rating of car #21? (H1-1)			
Sort	40 (35)	t(25) = -0.28	100%	Z = 0
SimulSort	43 (26)	p = 0.778	100%	p = 1
Question 2	Which used cars are made in 2008? (H1-2)			
Sort	21 (14)	t(21) = -1.41	100%	Z = 1.04
SimulSort	32 (27)	p = 0.173	93%	p = 0.301
Question 3	List the ten most closely located cars. (H1-3)			
Sort	**57 (34)**	**t(22) = -3.30**	93%	Z = 0
SimulSort	**113 (56)**	**p = 0.003***	93%	p = 1
Question 4	What attribute has the highest correlation with the year attribute? (H1-4)			
Sort	117 (111)	t(24) = -0.66	**73%**	**Z = 3.46**
SimulSort	140 (77)	p = 0.514	**20%**	**p = 0.001***
Question 5	Which cars are the two best if you have the following criteria (lower price, more recently manufactured, and lower mileage)? (H2-1)			
Sort	158 (108)	t(25) = -1.66	3.7(0.9)	W=196
SimulSort	227 (110)	p = 0.109	3.9(0.8)	p = 0.7336
Question 6	Which cars are the two best if you have the following more comprehensive criteria (Lower price, More recently manufactured, lower mileage, high city/highway fuel efficiency, high government tested safety rating, lower number of accidents, more closely located)? (H2-2)			
Sort	183 (128)	t(24) = 0.13	3.0(1.2)	W=199.5
SimulSort	176 (161)	p = 0.901	3.6(1.0)	p = 0.1569

*Statistically significant differences are found at the error level of 0.05.

7 Discussion

These results show that typical sorting features and SimulSort are generally comparable, but they potentially have different advantages and disadvantages related to the accomplishment of different tasks. These two sorting techniques provide comparable effectiveness in supporting most of the tasks except for sorting and finding correlations tasks, but SimulSort might potentially support decision making with large number of criteria.

It took a significantly longer time to find ten minimum values when using SimulSort than when using typical sorting (H1-3). Based on our observation, this result was

due to the fact that SimulSort required scrolling a bar to look up the values in highlighted cells that were scattered, while typical sorting required seeing only one row in which related values existed. Finding correlations was challenging with use of SimulSort (H1-4) because the positions of many highlighted cells must be remembered, which caused errors. In the other hand, with typical sorting features, participants simply sorted one column and compare other columns.

Although the results were not statistically significant, SimulSort can be regarded as helping users develop high confidence in the decisions made. In the post-task interviews, many subjects reported that SimulSort was helpful when multiple attributes were considered at the same time, though subjects are no more likely to use SimulSort over typical sorting features for decision makings with the small number of criteria. Although SimulSort could be unfamiliar to the participants, some subjects quickly made a strategy for better decision making. One strategy was to find the item that had more factors of positive influence in the lower position and more factors of negative influence in the upper position, when data are in ascending order. We observed that, although subjects did not seem to know the accurate values, they did know what the strengths and weaknesses were. In addition, using zooming-out feature and SimulSort together in making decisions indicated that bar graphs in one screen delivered visual messages that allow users to understand data and possibly mental tradeoff.

8 Conclusions

This paper proposed a new multidimensional sorting technique, which was designed to help users make decisions based on multiple-attribute information. We empirically compared SimulSort with the typical sorting feature, which showed that these two interaction techniques are generally comparable in all except the two tasks tested.

One limitation of this study is the lack of participants. Some of non-significant differences could be due to the limited number of participants. In two decision tasks, some interesting patterns were found, but it could not be proven that these differences were statistically significant.

However, it was still encouraging to find that SimulSort was quickly understood and used to support various tasks effectively. Due to the unfamiliarity of SimulSort, we expected that some participants had troubles of understanding it. However, most of the participants used it properly and came up with some heuristics, amplifying the advantages of SimulSort. Further investigation will provide more clear evidence of the effectiveness of SimulSort.

However, many other questions remain unanswered. Is it possible to combine typical sorting and SimulSort? Would the performance results be different if SimulSort had more flexibility to choose increasing order or decreasing order in each column? At what minimum number of attributes would SimulSort shows any statistically significant improvement of performances? How and where do users eyes move when using SimulSort? What is the subtle cognitive mechanism for sorting visualization regarding decision making? Those questions will be investigated in further research.

References

1. Priestley, J.: A description of a chart of biography; with a catalogue of all the names inserted in it, and the dates annexed to them. In: Eighteenth century collections online, London (1785)
2. Knuth, D.E.: The art of computer programming. Addison-Wiley, Reading (1968-1973)
3. Bertin, J.: Semiology of graphics: diagrams, networks, maps. University of Wisconsin Press, Madison (1983)
4. Mazza, R.: Introduction to Information Visualization (2004)
5. Hauser, H., Ledermann, F., Doleisch, H.: Angular Brushing of Extended Parallel Coordinates. In: IEEE Symposium on Information Visualization (2002)
6. Parvis, http://home.subnet.at/flo/mv/parvis/
7. Lanning, T., Wittenburg, K., Heinrichs, M., Fyock, C., Li, G.: Multidimensional Information Visualization through Sliding Rods. In: Proceedings of the working conference on Advanced visual interfaces table of contents, Palermo, Italy, pp. 173–180 (2000)
8. Spenke, M., Beilken, C., Berlage, T.: FOCUS: the interactive table for product comparison and selection. In: 9th annual ACM symposium on User interface software and technology, pp. 41–50. ACM, Seattle (1996)
9. Spenke, M., Beilken, C.: InfoZoom-Analysing Formula One racing results with an interactive data mining and visualisation tool. In: International Conference on Data Mining, Cambridge University, Cambridge (2000)
10. Wittenburg, K., Lanning, T., Heinrichs, M., Stanton, M.: Parallel bargrams for consumer-based information exploration and choice. In: 14th annual ACM symposium on User interface software and technology, pp. 51–60. ACM, Orlando (2001)
11. Verizon myEzchooser, http://brisa.merl.com:8080/myezchooser/index2.htm
12. Rao, R., Card, S.K.: The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In: Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, pp. 318–322. ACM Press, New York (1994)
13. Amar, R., Eagan, J., Stasko, J.: Low-Level Components of Analytic Activity in Information Visualization. In: Proceedings of the IEEE Symposium on Information Visualization, pp. 111–117 (2005)
14. Lurie, N.H.: Decision Making in Information-Rich Environments: The Role of Information Structure. J. Consumer Research. 30, 473–486 (2004)