Guest Editorial

Making sense of card sorting data

Sally Fincher[1] and Josh Tenenberg[2]

(1) Computing Laboratory, University of Kent, Canterbury, UK
E-mail: S.A.Fincher@kent.ac.uk
(2) Computing and Software Systems, Institute of Technology, University of Washington, Tacoma, USA
E-mail: jtenenbg@u.washington.edu

Abstract: *Among the knowledge elicitation techniques card sorting is notable for its simplicity of use, its focus on subjects' terminology (rather than that of external experts) and its ability to elicit semi-tacit knowledge. Card sorting involves categorizing a set of pictures, objects or labelled cards into distinct groups using a single criterion. This paper focuses on the challenges associated with analyzing the data that result from card sorts, especially when large data sets are generated. Traditional semantic analysis methods that require direct researcher interpretation of elicited linguistic terms are distinguished from methods that are purely syntactic, and hence can be automated. Each paper within this special issue is summarized and its contribution to card sorting in general, and data analysis in particular, is highlighted. The set of novel computational techniques presented in several of the papers in this issue is examined. The paper concludes by noting that even large-scale data sets can be meaningfully analysed by combining well-known interpretative methods with the new computational approaches presented within this special issue.*

Keywords: card sorting, card sorts, knowledge acquisition, knowledge elicitation

Among the elicitation techniques, card sorting has a number of distinct advantages. Salient among these is the simplicity of its administration for the researcher: items in a stimulus set are given to a participant, who then sorts them into groups. The items can be pictures, physical objects, words or fragments of domain-specific representation. In 'open' card sorts, items are grouped according to the participant's choice, and the participant provides a name for each group and for the criteria by which items are grouped. In repeated card sorts, participants are prompted to re-sort the stimuli into a different set of groups using another criterion, repeating until the subject is no longer able to do so. In 'closed' card sorts, participants are constrained in some way: most often, the criteria are provided so that the participants only have to choose which of the set belongs to any given category.

There are other advantages to the use of card sorts. One is that the simplicity of administration scales, so that defining the participant interview protocol among researchers in large-scale studies is simplified as well. The categorization task places no special cognitive burdens on research subjects, such as time pressure or memory limitations, and hence is suitable for all ranges of domain expertise. Open sorts, since subjects generate their own terms in naming criteria and categories, are subject – rather than researcher – centred. Card sorting can even allow for the comparison of participants who do not share a common language through the use of non-linguistic stimuli. And as Upchurch *et al.* (2001) point out, card sorting can elicit some of the semi-tacit knowledge that traditional interviews and questionnaires fail to access. All the studies reported in this special issue use 'open' card sorts.

Where card sorting presents challenges to the researcher is not in administration but in analysis: how does one make sense of the data? This special issue of *Expert Systems* provides a number of different perspectives and techniques for making sense of card sorting data. Between them, the papers in this issue demonstrate a range of uses to which card sorting has been put to make inferences about participant knowledge. These uses include using card sorts to elicit beliefs about Website similarity (and possible plagiarism), to unearth gender differences among office workers toward women's work attire and to capture conceptual structures about computer programming among student programmers.

What do card sorts represent?

Card sorting is a categorization task. As pointed out in Upchurch *et al.* (2001), 'Card sorts originated in George Kelly's Personal Construct Theory. Personal Construct Theory is based on the belief that different people categorize the world differently, but with enough commonality to let us understand each other but enough differences

to make us individuals' (p. 85). There is evidence to suggest that the way in which participants categorize entities *externally* reflects their *internal*, mental representation of these concepts. And, as Maiden and Hare (1998) discuss, categorization is an active construction of individuals, rather than simply being a reflection of ontological truth-in-the-world. 'Lakoff's (1987) review of categorization research concludes that categories are not "out there in the world", external to people. Rather, mental representations depend on factors specific to each person including experience in the world, perception, imaging capabilities, and motor capabilities' (p. 284).

It is known that experts organize information differently from novices: they form abstractions based on semantic characteristics, fundamental to the domain of their expertise, rather than on syntactic or domain-irrelevant characteristics (such as alphabetical organization or grouping by shape or colour). Allwood (1986) reporting the study of McKeithen *et al*. (1981) noted that 'novices used general memory strategies while experts used a more specific strategy' and although 'no reliable evidence was found that novices have less organized knowledge than experts . . . the novices showed large variation in their organization of the investigated concepts' suggesting that they did not share the commonly held domain knowledge of the experts.

This characterization of novice behaviour in sorting tasks (i.e. the variability of categorizations between subjects) aligns with findings of Weiser and Shertz (1983), as reported in Allwood, who 'instructed [them] to sort programming tasks into as many categories as they liked taking as much time as they wanted. The results showed that the novices used significantly less time and had a greater within-group variability with respect to how they clustered the problems compared with the experts.'

Eliciting the structures (representation) of knowledge (in our case via card sorts) is a more reliable indicator of expertise than quantities of facts, as demonstrated by a series of investigations by Chi *et al*. (1981) and Chi and Koeske (1983) where 4-year-old children were quizzed about features of two sets of dinosaurs, one set familiar and one not. Their responses were plotted using nodes and links and the familiar set showed tighter and stronger links than the unfamiliar set, even though the *number* of facts known about each set of dinosaurs was the same. Second, Chi gave physics students and educators a selection of representations of standard physics problems and asked them to group them. The educators used underlying principles (e.g. friction or gravity) as the basis for organization, whilst the students used features of the diagrams (pulleys or inclined planes).

It is thus plausible to take as an assumption that individuals construct meaningful internal categories to reflect their understanding of distinctions in the world. Card sorts serve as a 'contrived technique' (McGeorge & Rugg, 1992) that can be effective in eliciting our individual, and often semi-tacit, understanding about objects in the world and their relationships to one another.

Types of analysis

Taken together, the papers in this collection demonstrate a range of different analysis methods that can be used to make sense of card sort responses – from statistical counts of categories per sort, to content analyses, to similarity metrics between subjects. Additionally, these papers highlight a dimension that has not yet been addressed in the literature. This is the importance of *scale* in choosing appropriate analysis methods, in particular when card sorts are used in large-scale studies that contain hundreds or thousands of sorts.

Traditional analyses of card sort data use *semantic* methods, those methods that rely upon interpretative judgements by individual researchers on the meanings of the respondents' utterances. These methods can provide rich insights but require correspondingly high investment of time and scrutiny. For example, the kinds of inferences that Gerrard and Dickinson make about gender differences in office workers' attitudes about working women are of the kind that can only be uncovered through a close, semantic analysis of this type. Other analysis methods that can be brought to bear on card sort data are *syntactic*. These methods rely on statistical characteristics of the data set that can be automated (or semi-automated) and it is in the interpretation of these results that researchers seek insight.

Purely semantic methods are daunting to a researcher having several thousand category names to analyse, and in the arena of closed, or constrained, card sorts, several tools and techniques have been developed to help with syntactic analysis. Typically used in information architecture, tools such as EZSort and WebSort allow researchers to gather data from large populations.[1] The results of the card sorts are then subjected to cluster analysis and the results are viewed as a dendrogram. These tools are predicated on research questions which relate to aggregated aspects of the corpus – e.g. what Web-page structure would the population expect – and the clustering that the tools perform is directly on the articles of interest, i.e. the organization of the stimuli presented to the participants is the object of the investigation.

By contrast, in open card sorts, the organization of the stimuli by participants is often taken as a representation or characteristic of something else; neither categories nor criteria are specified in advance; and rarely is an aggregate analysis of interest. Rather it is the characterization of an individual within the corpus and the comparison to other

[1] *EZSort was a freely available tool from IBM: it was archived on the 25 January 2005. WebSort is a pay-for-use service run by Larry Wood from http://websort.net.*

participants that is sought. This means that open card sort data are not amenable to syntactic analysis tools of this kind, and until now there have been no tools to assist the researcher using open card sorts. However, both Deibel, Anderson and Anderson and Fossum and Haller present tools in this volume.

Papers in this issue

In providing a brief discussion of each paper, we highlight the manner in which each set of researchers couple their analysis methods to their research questions and the constraints of their study setting.

Rugg and McGeorge's paper, reprinted from 1997, is a 'must read' for those new to card sorts, and useful to review even for those with card sort experience. In this paper the authors situate card sorts among the elicitation techniques, state the kinds of study settings in which it can be fruitfully employed, provide a step-by-step tutorial on how to run card sorts in experimental settings and suggest a number of analysis methods. Among these are a count of the number of criteria by each participant to estimate amount of domain categorization knowledge; examining the type of criteria to determine if they are observable, subjective or extrinsic; looking for commonality of criteria and their distribution across the study population; and examining the categories for commonalities, characteristics (e.g. abstract versus concrete), lapses of knowledge and significant absences.

Gerrard and Dickinson use card sorts to investigate office workers' perceptions of women's work attire. These researchers are interested in comparing responses of men and women when viewing visually-rich stimuli – pictures of women in different work attire – without imposing any prior researcher model on the terminology subjects use to describe their categories. Gerrard and Dickinson employ a superordinate analysis, where criteria or category names are grouped into higher-level constructs based on similarity of meaning. In doing so, they unearth assumptions by a significant subset of only the male subjects about the marital status of the women depicted in the stimulus set. The authors also report on gender differences in the number of *dichotomous* sorts (those sorts having exactly two categories) and relate this to previous research findings on novice/expert differences. This is one of the very early mentions of this phenomenon in the literature and we are pleased to be able to include this previously unpublished work here.

The paper by Martine and Rugg shows how card sorts can be used to generate a similarity metric for Web pages or other visual stimuli. This metric is obtained by analysing a subject's *co-occurrence matrix* that encodes the frequency with which pairs of cards are placed into the same category group. Martine and Rugg point out several advantages of this approach versus more semantic-based methods for developing a similarity metric. Not only are co-occurrence matrices much less resource-intensive for the researcher to derive, they also allow comparison of responses without any intermediate coding or interpretation by the researcher, thus reducing a significant source of bias. In addition, this particular analysis permits the comparison of subjects who do not share a common language or culture. Finally, because this metric relies only on grouping choices and not the names of groups, it can uncover *semi-tacit* knowledge that is embedded in the subject's categorizations but which the subject might be unable to articulate.

The remaining papers all share a common origin, and this is the study setting described in Sanders *et al.* This study was part of a National Science Foundation funded project, Bootstrapping Research in Computer Science Education. The editors of this special issue (with Marian Petre, Open University, UK) were organizers of this study. The Bootstrapping study was a multinational, multi-institutional card sort investigation of the conceptual structures of students of initial programming. The focal questions concerned what meanings student programmers have for individual programming concepts, and for the relationships they make between these concepts. Sanders *et al.* describe the specific use of card sorts, which resulted in a collection of over 1000 sorts and 5000 category names. Given this scale of study and the number of researchers involved, it was necessary to use syntactic analysis methods whenever possible. There were few difficulties in obtaining statistical measures on the aggregated data (e.g. number of sorts, categories per sort) or with respect to any identifiable subpopulation (e.g. men and women, high- and low-performing students based on previous course grades, students with knowledge of specific programming languages). Yet there was difficulty in addressing the study's focal questions from this analysis.

For example, it would be reasonable to be interested in whether the students in the subject population shared common meanings for the terms they themselves use. Using the methods of Upchurch *et al.* (2001), this would involve performing a verbatim analysis of all 6000 category and criteria labels (i.e. labels that are identical), performing a *gist* analysis on all 6000 labels (e.g. interpreting 'difficult' as meaning the same thing as 'hard') and then performing a superordinate analysis on the 'gisted' categories and criteria names. But, even given sufficient time and capacity necessary for this intensive approach, researchers would then have to determine the meanings that different subjects provide for labels interpreted as semantically similar (e.g. does the word 'difficult' mean the same thing to subject X as it does to subject Y?). Rugg and McGeorge suggest that *laddering* might provide insights into the subject's meaning of terms elicited during card sorting. But, for a data set of this size, this only presents an equally intractable problem of how to compare card sort *and* laddering data among several hundred respondents and several thousand utterances.

Faced with this problem, the remaining papers describe and apply new computational techniques for making sense of open card sort data; and although they are essential for large-scale studies such as the Bootstrapping study, they are equally appropriate for use on small studies as well.

The paper by Deibel *et al.* provides a key insight in making sense of large card sorting data sets – develop a measure of similarity between two different sorts (whether or not from the same person) that depends strictly on the card groupings into which the individual cards are placed. In this regard, it bears a striking similarity to Martine and Rugg's analysis, as both are centrally concerned with the distribution of cards into groups. But the differences between these two methods are important: whereas Martine and Rugg develop a similarity metric on the *stimuli*, Deibel *et al.* develop a similarity metric on *sorts*. The similarity metric is simply the *edit distance* between the two sorts, i.e. the number of 'moves' (taking a card from one category group and placing it into another) required to turn one sort into another.

Deibel *et al.* use this to show that participants who might be considered a subpopulation, e.g. who all provide category labels concerning programming difficulty (e.g. 'difficult', 'hard'), may mean quite different things by these, as there is little similarity between the cards that they place into these categories. Because this metric depends only on the grouping choices and not the terms that subjects use to describe the groupings, the technique is applicable to comparing specific card sorts within any knowledge domain. The authors conclude by demonstrating the use of edit distance to determine neighbourhoods of similar sorts, and how researchers can use these neighbourhoods to gain insight into specific semantic hypotheses ('what might subject X mean by "words I hate"?') and also to focus semantic analysis on large neighbourhoods of similar sorts.

However, whilst successful, this method is heavily computational (cubic time in the size of the input), and relies upon specialized knowledge of graph theory. It might therefore remain unusable except to a small subset of researchers with the requisite knowledge to program these algorithms. Anticipating this difficulty, Deibel *et al.* have made their computer programs available to the research community for exploratory analysis of their own data sets.[2]

Fossum and Haller return to an analysis method described in Rugg and McGeorge: using the number of a participant's sorts as an estimate of the amount of domain knowledge. 'If all the respondents use large numbers, then there is considerable knowledge involved.' Fossum and Haller point out that this assumes that all subjects are not simply repeating the same sorts, or trivial variants. What they seek instead is a basis for combining a subject's number of sorts with the *orthogonality* or aggregate difference between their sorts. In essence, they measure the volume of conceptual space occupied by a subject's sorts; highly similar sorts, even if there are lots of them, occupy only a narrow slice of this space, while highly differentiated sorts occupy a larger amount of space.

Fossum and Haller measure orthogonality by summing the edges in a minimum spanning tree on Deibel's edit distance between pairs of a subject's sorts encoded as edge weights in a complete graph. They provide considerable empirical evidence (using the Bootstrapping data and the data of McCauley *et al.* (2005)) that those subjects whose sorts exhibit low orthogonality can be taken as having less domain categorization knowledge than those subjects with high orthogonality. They conclude with a caution that there are limits to using orthogonality as a measure of domain knowledge, by showing that randomly generated sorts have very high orthogonality compared to those of human subjects.

The paper by McCauley *et al.* describes a follow-up to the Bootstrapping study, whose participants were 65 graduating students who performed 291 sorts using the Bootstrapping stimulus set. The authors undertake a superordinate analysis to generate a set of 16 higher-level characterizations of the students' category and criteria labels, what they call *content analysis groups* (CAGs). These CAGs include Abstract/Concrete, Parts of a program and Language paradigm.

They use Fossum and Haller's orthogonality metric to validate the degree of structural similarity within all of the sorts for each CAG and Deibel *et al.*'s edit distance metric to generate *exemplar sorts* for each CAG, where each exemplar is at the centre of the conceptual space occupied by all of the sorts in a CAG, i.e. the sort having minimal aggregate distance to all of the other sorts in the CAG.

The particular card groupings of this exemplar can then be examined, providing additional insight into the meaning of the CAG as a whole, regardless of the actual terminology used by the different subjects who generated the constituent sorts. What this paper demonstrates is that considerable insights into subjects' conceptual structures, even within a large study, can be obtained using card sorts and a combination of computational and semantic analysis methods.

Conclusion

This special issue brings together a collection of papers describing the use and analysis of open card sorts. We reproduce a landmark paper encapsulating a useful tutorial and include papers that report on early studies of considerable interest. We include papers that report on newer applications of open card sorts and the development of tools for syntactic analysis. Both small- and large-scale studies are represented and each paper describes the use of (at least one) analysis technique.

[2] *The UW Card Sort Analyzer is available from http://www.cs.washington.edu/homes/deibel/CardSorts/.*

We anticipate that the papers combined in this volume will provide a rich resource for future researchers interested in this elicitation technique and, in closing, we draw attention to a further commonality between them. That is, what each of the papers exposes is how the use of syntactic and semantic methods in combination can provide deeper insights into subjects' knowledge structures than either type of method alone.

Acknowledgements

Part of this material is based upon work supported by the National Science Foundation under Grants DUE-0122560 and DUE-0243242. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

ALLWOOD, C.M. (1986) Novices on the computer: a review of the literature, *International Journal of Man–Machine Studies*, **25**, 633–658.

CHI, M.T.H. and R. KOESKE (1983) Network representation of a child's dinosaur knowledge, *Developmental Psychology*, **19**, 29–39.

CHI, M.T.H., P. FELTOVICH and R. GLASER (1981) Categorization and representation of physics problems by experts and novices, *Cognitive Science*, **5**, 121–152.

LAKOFF, G. (1987) *Women, Fire and Dangerous Things*: *What Categories Reveal about the Mind*, Chicago, IL: University of Chicago Press.

MAIDEN, N. and M. HARE (1998) Problem domain categories in requirements engineering, *International Journal of Human–Computer Studies*, **49**, 281–304.

MCGEORGE, P. and G. RUGG (1992) The uses of 'contrived' knowledge elicitation techniques, *Expert Systems*, **9** (3), 149–154.

MCKEITHEN, K., J.S. REITMAN, H.H. REUTER and S.C. HIRTLE (1981) Knowledge organisation and skill differences in computer programmers, *Cognitive Psychology*, **13**, 307–325.

UPCHURCH, L., G. RUGG and B. KITCHENHAM (2001) Using card sorts to elicit Web page quality attributes, *IEEE Software*, 84–89.

WEISER, M. and J. SHERTZ (1983) Programming problem representation in novice and expert programmers, *International Journal of Man–Machine Studies*, **19**, 391–398.

The authors

Sally Fincher

Sally Fincher is a lecturer in the Computing Laboratory at the University of Kent where she leads the Computing Education Research Group. She holds a BA in philosophy and computer science (University of Kent, UK) and an MA in English (Georgetown University, Washington, DC). She is Editor of the journal *Computer Science Education*, jointly with Renée McCauley. Her principal research areas are computer science education and patterns and pattern languages, especially patterns for interaction design.

Josh Tenenberg

Josh Tenenberg is an associate professor in the Computing and Software Systems program in the Institute of Technology at the University of Washington, Tacoma. He holds a BM in music performance (San Francisco State University, USA) and an MS and PhD in computer science (University of Rochester, USA), where his primary research was in artificial intelligence. His research areas have included automated planning, knowledge representation and reasoning, reinforcement learning, temporal logic, and cognitive modelling of computer programming. Most recently, his research is in computer science education, where he is investigating student software design and metacognition.