A New Measure Summarising 'Information' Conveyed in Cluster Analysis of Card-Sort Data:

Application to a Neonatal Intensive Care environment.

Gary Ewing,[1] **Robert Logie,**[2] **Jim Hunter**[1], **Neil M**^{c}**Intosh,**[3] **Sue Rudkin**[2], **Yvonne Freer**[3] and **Lindsey Ferguson**[3]

Abstract. The aim of this paper is to describe a novel approach to the analysis of data obtained from card-sorting experiments. The card-sorting experiments were performed as a part of the initial phase of a project,called NEONATE, that has an aim to develop decision support tools for the neonatal intensive care environment.

Physical card-sorts were performed using clinical "action" and patient "descriptor" words. Thirty-two staff (8 junior nurses, 8 senior nurses, 8 junior doctors and 8 senior doctors) participated in the actions cards sorts and 32 staff (of similar classifications) participated in separate descriptors card-sorting experiments. There was considerable overlap of staff in the two groups (of 32), but the actions and descriptors experiments were held several months apart. The card-sorts were replicated for all staff in the descriptor experiments, and for nurses only during the actions card-sorts.

The card-sort data were analysed using conventional cluster analysis to produce tree-diagrams or dendrograms. These showed interesting differences in the way the various classes of staff mentally map clinical concepts, and some of these differences are briefly highlighted here. However, the main aim of this paper is to present a method of summarising the differences in the card-sort data for the various staff categories, and further, to develop a method to quantitatively measure the different information requirements for the different staff classes in order to facilitate user-interface design. This information theoretic based technique may also have application later in the development of a clinical support system as a metric for the appropriateness of particular formats for clinical information presentation.

1 INTRODUCTION

The modern intensive care unit is an environment that requires medical and nursing staff to deal with large amounts of, and many different types of, information in making clinical decisions. It has been shown that just displaying these data in their raw form does not of itself lead to improved patient care. The work reported in this paper formed part of the initial effort in an ongoing project, NEONATE, to develop decision support for clinical staff (doctors and nurses) in a neonatal intensive care environment–Neonatal Intensive Care Unit (NICU) at the Simpson Maternity Hospital, Edinburgh. The component of this initial phase, which is reported here, focussed on developing a concise lexicon of terms used by clinical staff during clinical practice, and further, the use of these lexicons as the basis for card-sorting experiments designed to elucidate the way clinical staff mentally organise those terms. This information was used to design the interface for a software tool to allow a trained observer (research nurse) to record (in a standardised fashion on a computer database) the clinical activities of doctors and nurses as a data gathering exercise for the NEONATE project. (Detailed physiological data such as heart rate, blood pressure, are automatically collected by the NICU's computerised monitoring system.)

The standard way of analysing card-sort experiments is by performing cluster analysis in order to generate tree-diagrams (or dendrograms) as a graphical representation of the relationships between the concepts under study. However, the dendrograms generated by cluster analysis of card-sort data are often very complex and difficult to interpret. Even though, a visual analysis of tree-diagrams may possibly provide useful insight into how people mentally organise concepts, it is tedious and does not provide a quantitative measure of the complexity or order implicit in the dendrogram. The work reported here aims to fill this gap by providing a simple graphical summary, and a quantitative measure of the difference in the information carried by various dendrograms that are being compared. This work is still under development, but the feasibility of the approach is demonstrated here.

2 METHODOLOGY

The scope of this section is to outline the methodology employed in the data gathering, performance and analysis of the card-sorting experiments.

2.1 Clinical Lexicons

In order to elicit lexicons for both patient "descriptors" and clinical "actions" we interviewed medical and nursing staff at different levels to delineate their roles in the unit and the vocabularies they use to categorise neonatal data obtained by observation and physical means. Senior clinical staff subsequently reviewed these lists for consistency, and to remove synonyms and singletons (single words used by only one member of staff). The derived actions lexicon contains 51 terms, while the descriptors lexicon contains 166 terms.

[1] Department of Computing Science, King's College, University of Aberdeen, Aberdeen, UK
[2] Department of Psychology, King's College, University of Aberdeen, Aberdeen, UK
[3] Neonatal Intensive Care Unit, Simpson Maternity Hospital, Edinburgh, UK

2.2 Card-Sorting Experiments

Following the interviews, we carried out card-sorting experiments using the two lexicons as the concepts to be sorted. Card-sorting was used as an elicitation technique because "Concept Sorting" is a well-known technique, and studies in Cognitive Psychology and related fields[2, 6] have shown it to be effective and very efficient.

Thirty-two subjects consisting of, 8 junior nurses (JN), 8 senior nurses (SN), 8 junior doctors (JD) and 8 senior doctors (SD), participated in the actions card sorts and thirty-two staff (of similar levels) participated in separate descriptors card-sorting experiments. There was considerable overlap[4] of staff in the two groups (of 32), but the actions and descriptors experiments were held several months apart.At least one week after the initial sessions, the card-sorts were replicated for all staff in the descriptor experiments, and for nurses only in the actions experiments. Each actions card-sorting session took about an hour to complete on average, while each descriptors card-sort session required about 1.5 hours to complete. The level at which staff were categorised, whether junior or senior, was decided upon by the senior clinical staff involved in the project.

The actual card-sorting procedure we asked subjects to perform is illustrated in figure 1. During a session each subject was presented with a physical pile of cards, each with a term from the appropriate lexicon. These (level-0) cards were marked on the back with a bar-code (3 of 9 code), containing the term on the front of the card, and a unique identifing alpha-numeric code. We also produced cards that were blank on one side and bar-coded on the other, again with a unique identifing alpha-numeric code, to be used as indicated in the next paragraph.

Each subject was asked to sort the cards into piles of "similar" cards, without any prompting as to how many piles to create or what attributes to use to sort the cards. Once satisfied, the subject was asked to name each group (without restriction) and these names were written by the experimenter on to (level-1) blank cards. The experimenter then entered the names and codes of the cards, within their sorted groups, into a computer database by means of a bar-code scanner. This saved considerable time and minimised errors in data entry. The subject was then asked to sort the cards containing the names of the groups that had just been created into higher groups; i.e. groups of groups. Again, the names of the groups were written by the experimenter on to (level-2) blank cards. This process continued until the subject was happy with the result.

2.3 Analysis of the Card-Sort Experiments

Within the context of the work described in this paper, cluster analysis calculates the strength (distance) of the perceived relationships between card-pairs, and displays these relationships graphically (dendrograms). We performed hierarchical agglomerative cluster analysis of the card-sort data using a free software package called EZCalc[3], which was designed to be used with its companion software, EZSort, that facilitates computer based card-sorting experiments. As we did physical sorts, the data files needed to be formatted[5] and pre-processed to allow use by EZCalc. The pre-processing also checked for consistancy within the card-sort data files. As EZCalc used arbitrary weighting for second level groups (the highest level of hierarchical grouping it could handle) in the card-sort, we only used first level sort data for our analysis. We employed the *average linkage* method of cluster analysis, as this provides a good compromise between the extremes of other methods[1].

[4] Many of the staff who participated in the card-sorts were also involved in the interviews to establish the lexicons. However, there was several months gap between the interviews and the card-sorts.

[5] Into the format generated by EZSort.

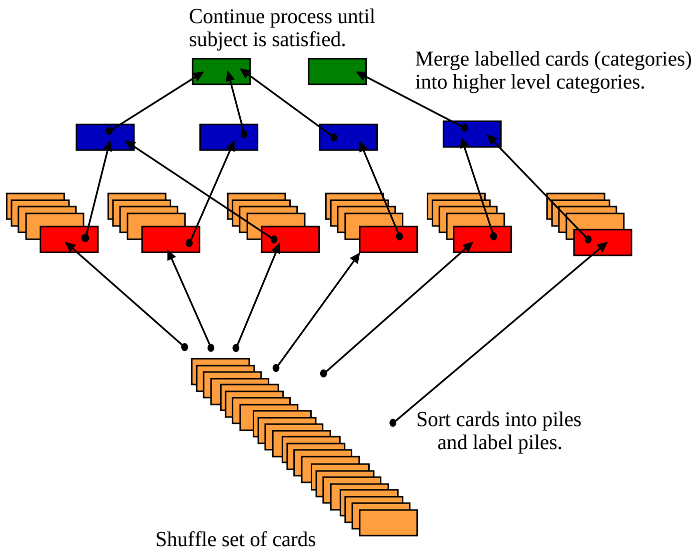


Figure 1. Card-sorting procedure.

Using in-house software we carried out further analysis of the card-sort data. This software produced *distance matrices*, which quantified how often all the possible pairs of cards were grouped together by a group of subjects. Based on the distance matrices further graphical and analytical measures were produced, and are described in section 3.2, while the outcomes of applying these measures are presented in section 4.

3 DEVELOPMENT OF A QUANTITATIVE INTERPRETATION OF CLUSTER TREE-DIAGRAMS

In this section we will develop a quantitative measure of the amount of "information" mismatch, in terms of the entropy (defined by equation 3 in section 3.1), or amount of structure, displayed by different dendrograms under comparison. (In our case the dendrograms derived from the card-sorts for the different staff categories in the NICU.)

This section is divided into two parts. In the first section we discuss a classical information communications system and define key concepts, while in the second section we derive a notional information system based on the card-sort data, and derive properties analogous to those we defined for the classical system. Using the information system analogy is compelling, since we are interested in how the different staff groups use and think about clinical information, and how they communicate this to each other.

3.1 Information Theoretic Model

We will briefly describe the classical information theoretic model used to define information transfer in an information channel[8, 7, 5]. Shown in figure 2 is a simple diagram of the basic components of a communications system. This system consists of an information source with a symbol set, $\{a_i\}$, (alphabet) containing N symbols each with a probability of occurrence of $P(a_i)$; an information receiver or user with a symbol set, $\{b_i\}$, containing M symbols each with a probability of occurrence of $P(b_i)$; and an information channel that represents the interactions between the source and the user. Note, that

$$\sum_{i=1}^{N} P(a_i) = 1 = \sum_{j=1}^{M} P(b_j).$$

The information channel, Q, is characterised by an $M \times N$ matrix whose elements represent the conditional probability of the user receiving (interpreting) a particular symbol given that a particular symbol was "sent" by the source; i.e.

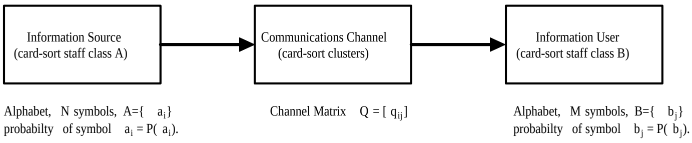


Figure 2. A Basic Information System

$$Q = \begin{bmatrix} P(b_1|a_1) & P(b_2|a_2) & \dots & P(b_1|a_N) \\ P(b_2|a_1) & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ P(b_M|a_1) & P(b_M|a_2) & \dots & P(b_M|a_N) \end{bmatrix}. \quad (1)$$

The source and user probabilities may be represented as vectors thus:

$$\mathbf{s} = \begin{bmatrix} P(a_1) \\ P(a_2) \\ \cdot \\ \cdot \\ \cdot \\ P(a_N) \end{bmatrix}, \text{ and } \mathbf{u} = \begin{bmatrix} P(b_1) \\ P(b_2) \\ \cdot \\ \cdot \\ \cdot \\ P(b_M) \end{bmatrix}.$$

The information system can then be completely described by the equation[8, 7]:

$$\mathbf{u} = Q\mathbf{s}. \quad (2)$$

The *entropy* of the source is defined as:

$$H(\mathbf{s}) = -\sum_{i=1}^{N} P(a_i) log P(a_i), \quad (3)$$

where $H(\mathbf{s})$ quantifies the average amount of information per symbol obtained by observing a single source output. The units of entropy are defined in r-ary units, where r is the base of the logarithm used. An increase in entropy represents an increase in the amount of disorder or uncertainty in the information, and thus represents an increase in the average information per symbol (since it is harder for the user to "guess" its identity). Likewise the entropy of the information user is:

$$H(\mathbf{u}) = -\sum_{j=1}^{M} P(b_j) log P(b_j). \quad (4)$$

The *equivocation* of $\mathbf{s}$ with respect to $\mathbf{u}$ is the average information associated with one source symbol, assuming the observation of the output symbol that resulted from its occurrence. This is defined as:

$$H(\mathbf{s}|\mathbf{u}) = -\sum_{i=1}^{N}\sum_{j=1}^{M} P(a_i, b_j) log P(a_i|b_j), \quad (5)$$

were $P(a_i, b_j)$ is the probability of a_i being "sent" and b_j being "received"(i.e. joint probability of a_i and b_j). The difference between $H(\mathbf{s})$ and $H(\mathbf{s}|\mathbf{u})$ is called the *mutual information* of $\mathbf{s}$ and $\mathbf{u}$ and gives the average rate of information transfer per symbol (in r-ary units); i.e.

$$I(\mathbf{s}, \mathbf{u}) = H(\mathbf{s}) - H(\mathbf{s}|\mathbf{u}). \quad (6)$$

Equation 6 can transformed into a form suitable for calculation of $I(\mathbf{s}, \mathbf{u})$, by substituting equation 3 and equation 5 into equation 6 to produce:

$$I(\mathbf{s}, \mathbf{u}) = \sum_{i=1}^{N}\sum_{j=1}^{M} P(a_i) q_{ji} log \frac{q_{ji}}{\sum_{k=1}^{N} P(a_k) q_{jk}}. \quad (7)$$

The minimum value possible for $I(\mathbf{s}, \mathbf{u})$ is zero, which occurs when the input and output symbols are statistically independent (i.e. $P(a_i, b_j) = P(a_i) \cdot P(b_j)$), since $H(\mathbf{s}) = H(\mathbf{s}|\mathbf{u})$. The maximum value of $I(\mathbf{s}, \mathbf{u})$ is called the channel *capacity* (C), and is maximised for a particular source distribution, $\{P_{max}(a_i)\}$, over all possible choices; i.e.

$$C = \max I(\mathbf{s}, \mathbf{u}). \quad (8)$$

The channel capacity quantifies the maximum rate of information transfer per symbol that the channel can reliably transmit.

3.2 Development of a Measure of Relative Information in Dendrograms

We now develop an approach to measure the relative information associated with particular dendrograms using the basic model discussed in section 3.1, and by utilising analogies between the two concepts, which at first sight seem quite different.

As described in section 2.3, one of the products of the card-sorting experimental data were distance matrices, which are symmetrical (about the major diagonal) and the elements have values between zero and unity. Since the matrices were symmetrical we converted them to triangular matrices thus:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N-1} & d_{1,N} \\ 0 & d_{2,2} & \dots & d_{2,N-1} & d_{2,N} \\ 0 & 0 & d_{3,3} & \dots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \dots & d_{N-1,N-1} & \cdot \\ 0 & 0 & \dots & 0 & d_{N,N} \end{bmatrix} \quad (9)$$

Since the elements of the distance matrix have values within the range zero to one, (i.e. $0.0 \leq d_{ij} \leq 1.0$) they can be considered as random variables with associated probability values[6]. However, since

$$\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij} \neq 1, \quad (10)$$

the elements of the distance matrix as a whole do not represent a probability distribution. To overcome this difficulty we developed an analogue to a cumulative probability distribution by the following means.

Each element d_{ij} of the distance matrix D represents the joint probability $P(a_i, b_j)$ of two cards appearing together in the same pile[7] in a card-sort. We applied a test to each d_{ij} of:

$$f_k = \begin{cases} f_k + 1 & d_{ij} < \tau \\ f_k & d_{ij} \geq \tau \end{cases} \forall\, d_{ij} \in D, \quad (11)$$

[6] The d_{ij} have also been generated from frequency data.
[7] Deemed similar in some sense by experimental subjects.

where τ is a monotonically increasing threshold value such that $0 \leq \tau \leq 1$, while τ was increased in discrete steps of $\delta : \delta < \min \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{p=1}^{N} \sum_{q=1}^{N} |d_{ij} - d_{pq}|$, and f_k is frequency of the test being passed for each d_{ij} at the k^{th} increment in τ.

The f_k were normalised to a range $0.0 \leq \hat{f}_k \leq 1.0$ by dividing by $\frac{N!}{2(N-2)!} = \frac{N(N-1)}{2}$, the number of possible combinations of different pairs of cards from a stack of N cards. A stylised plot of $\hat{f}_k$ is shown in figure 3 (actual plots are shown in section 4.2). This graph is analagous to the plot of a cumulative probability distribution (F), as the y-axis represents $F_\tau(d_{ij}) = P(d_{ij} \leq \tau)$ (i.e. the probability that the distance between any card pair is less than some distance value $d = \tau$), and the x-axis is the distance (d). A property of any cumulative probability distribution of a discrete random variable $X \in \{x_1, x_2, \ldots, x_n\}$ is:

$$P(x_i) = F(x_i) - F(x_{i-1}). \tag{12}$$

Figure 3 serves as a summary of a distance matrix, and in fact the dendrogram(s) produced from it. The straight line graph in depicted in figure 3 represents the the plot expected when the distance matrix is completely random; i.e. the entropy is maximum. The other curves in figure 3 represent different degrees of order (entropy) in two hypothetical distance matrices.

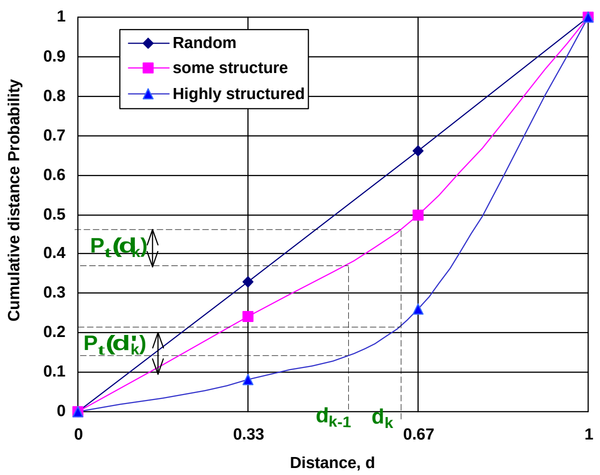


Figure 3. A stylised plot of the proportion of card-pairs (cumulative probability) that are separated by a distance less than d

We now wish to draw an analogy with the information system discussed in section 3.1. Let us consider the information source, $\mathbf{s}_c$, to be a distance matrix of a particular class of clinical staff $c \in$ {JN,SN,JD,SD}. A distance matrix holds the information required to produce graphical illustrations (dendrograms) of the way clinical staff construct mental schemas related to clinical actions or patient descriptors, and can be viewed as an encoding of this information. The source "symbol" set, in this case, is the set of card pairs that are identified by the coordinates for the d_{ij} in D. However, the symbol probabilities are not simply the d_{ij} since they do not sum to unity as shown in equation 10. A method to convert the d_{ij} to probabilities (or analogues thereof) is required. To this end consider the following.

Convert the triangular distance matrix into a single vector by concatenating the rows;

$$\begin{bmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,N-1} & d_{1,N} \\ 0 & d_{2,2} & \ldots & d_{2,N-1} & d_{2,N} \\ 0 & 0 & d_{3,3} & \ldots & . \\ . & . & . & . & . \\ . & . & \ldots & d_{N-1,N-1} & . \\ 0 & 0 & \ldots & 0 & d_{N,N} \end{bmatrix} = \begin{bmatrix} d_{1,1} \\ . \\ . \\ d_{1,N} \\ d_{2,2} \\ . \\ . \\ d_{2,N} \\ d_{3,3} \\ . \\ . \\ . \\ d_{N,N} \end{bmatrix} = \mathbf{d_v}, \tag{13}$$

where $\forall\ i, j\ i \geq j$. Let the elements of vector $\mathbf{d_v}$ be indexed as d_k, such that ascending k implies increasing magnitude of d_k; i.e. the d_k are ranked.

A method to convert the d_k to probabilities is demonstrated in figure 3, where a d_k value on the x-axis corresponds, via a specific curve, to a value of $F_\tau(d_k)$. From equation 12, $P(d_k) = F_\tau(d_k) - F_\tau(d_{k-1})$. Likewise, the properties of an information user, $\mathbf{u}_c$, can be derived in a similar manner, but utilising the distance matrix of another staff class.

The information source and user symbol probabilities can then be calculated from the appropriate distance matrices and may be represented by the vectors:

$$\mathbf{s_c} = \begin{bmatrix} P_c(a_1) \\ P_c(a_2) \\ . \\ . \\ . \\ P_c(a_M) \end{bmatrix}, \text{ and } \mathbf{u_c} = \begin{bmatrix} P_c(b_1) \\ P_c(b_2) \\ . \\ . \\ . \\ P_c(b_M) \end{bmatrix},$$

where $M = \frac{N(N-1)}{2}$, and is the length of a vector produced from a distance matrix as defined in equation 13. Therefore by analogy with equation 2, we may write the equation to describe our notional information system as:

$$\mathbf{u_c} = Q_c \mathbf{s_c}. \tag{14}$$

We propose to use the mutual information, $I(\mathbf{s}, \mathbf{u})$, (see equation 7) as a metric for the amount of mismatch in the information that different categories of staff use in structuring the concepts they use in clinical practice (within the limitations of the card-sorting experiments). Since, once given the card-sorting results, we have fixed source and user alphabets and probability distributions. For each pair of staff classes (one source, the other user, defining a channel), there exists only one value of $I(\mathbf{s}, \mathbf{u})$, and that is equivalent to the channel capacity (as defined in equation 8). This requires that we define a channel matrix, Q_C, of the form shown in equation 1, which would completely determine the channel capacity according to the system defined by equation 14. Further, this requires the definition of the conditional probabilities of $P_c(b_j)|P_c(a_i)$, which are the elements of the matrix Q_c in equation 14, and are analogous to those in equation 1.

Consider two distance matrices obtained respectively from the card-sorts of staff category A and staff category B. For a given value of the distance threshold τ, the derived cumulative probability distribution (as in figure 3) for A is $F_\tau(d_{a_i})$ and $P_c(a_i) = F_\tau(d_{a_i}) - F_\tau(d_{a_{i-1}})$, where d_{a_i} is d_k (at a_i), and $d_{a_{i-1}}$ is d_k

(at a_{i-1}). Similarly for the same τ, B has $F_\tau(d_{b_j})$ and $P_\tau(b_j) = F_\tau(d_{b_j}) - F_\tau(d_{b_{j-1}})$. Therefore, we can argue that the notional information source has sent a "message" that the symbol (card pair, a_i) represents the distance of $\Delta d = d_{a_i} - d_{a_{i-1}}$, and has probability $P_c(a_i)$ of occurring. The information receiver "interprets" that message to correspond to the symbol b_j, which has probability $P_c(b_j)$ of occurring. Hence, the probability of b_j occurring given that a_i has occurred, $P_c(b_j)|P_c(a_i)$, is the probability $P_c(b_j)$ that corresponds to the same τ_{a_i} and $\tau_{a_{i-1}}$ that are associated with symbol a_i with probability $P_c(a_i)$. By this approach all the elements of the matrix Q_c can be determined.

The value for the mutual information is obtained by applying equation 7 in the system defined by equation 14. This in effect measures the mismatch of information represented by the dendrograms for the different staff categories.

In the next section, actual data derived from the card-sorting experiments for both action and descriptor words is presented.

4 RESULTS and DISCUSSION

Due to space constraints, only the results of the actions card-sorting experiments are discussed. This does not limit the intended scope of this paper, as we are demonstrating the feasibilty of the proposed information measure, and not interpreting in detail the results of the card-sorting experiments. Firstly, the results of the cluster analysis are presented, secondly, the cumulative probability graphs are displayed, and finally, the results of the calculations of the information channel capacities, for the information "channels" between the four staff classes under consideration, are yielded and discussed.

As indicated in section 2 the card-sorts were replicated for for nurses in the actions experiments. The initial and replicated vectorised distance matrices (see equation 13) were statistically compared using the both Wilcoxon Signed Ranks Test (as a nonparametric alternative to the t-test, since plots of the data indicated they were not normally distributed).

The results of the statistical tests for the actions data are shown in table 1 to table 2, and show with very high confidence ($p < 0.00001$), for both junior and senior nurse data, that initial and replicate data come from the same probability distribution; e.g. the results between separate experiments for the same staff class are consistent. This is also reinforced by correlation (Pearson) tests: junior nurses ($r = 0.91$, $p < 0.0001$), senior nurses ($r = 0.92$, $p < 0.0001$).

Table 1. Wilcoxon Signed Ranks Test for sort1 and sort2 (replicate) of the card-sorts for junior nurses.

		N	Mean Rank	Sum of Ranks
JN-S2	Negative Ranks	679(a)	392.92	266795.50
JN-S1	Positive Ranks	107(b)	397.15	42495.50
	Ties	439(c)		
	Total	1225		
(a) JN-S2 $<$ JN-S1 (b) JN-S2 $>$ JN-S1 (c) JN-S1 = JN-S2				

	JN-S2 - JN-S1
Z	-17.865(a)
Asymp. Sig. (2-tailed)	0.0000

Table 2. Wilcoxon Signed Ranks Test for sort1 and sort2 (replicate) of the card-sorts for senior nurses.

		N	Mean Rank	Sum of Ranks
SN-S2	Negative Ranks	392(a)	272.84	106955.00
SN-S1	Positive Ranks	151(b)	269.81	40741.00
	Ties	682(c)		
	Total	1225		
(a) SN-S2 $<$ SN-S1 (b) SN-S2 $>$ SN-S1 (c) SN-S1 = SN-S2				

	SN-S2 - SN-S1
Z	-9.184(a)
Asymp. Sig. (2-tailed)	0.0000

4.1 Dendrograms

Cluster analysis was performed on the processed card-sorting data and dendrograms have been produced for the various classes of staff. This has yielded some interesting results. However, only the major findings are briefly discussed here as a detailed discussion is beyond the scope of this paper.

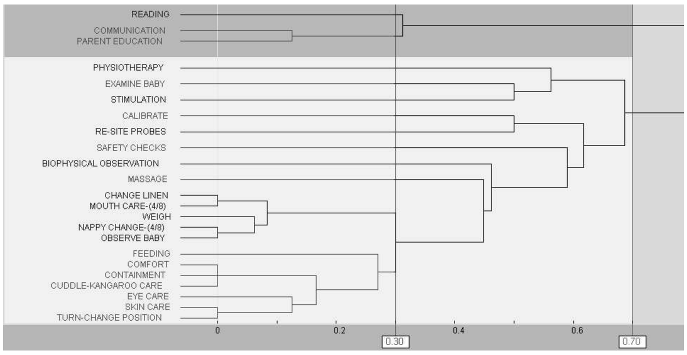
Shown in figure 4(a) is a section of the dendrogram derived from the actions card-sorts for junior nurses, while figure 4(b) displays a section of the dendrogram derived from the actions card-sorts for senior doctors. Both of these sections of dendrograms show some common terms of interest from the actions lexicon, and are now briefly discussed.

- There is a difference in structure of the dendograms across the different staff groups, and an extreme difference in structure between junior nurses and senior doctors. In the case of the latter, there is evident a much richer structure with more groups (more discerning) than is evident for the former, who formed large groupings with little discernment;
- Junior nurses did not group "Biophysical Observation" with"Examine Baby" (whereas the other three staff groups (SN, JN, SD) grouped these closely together) nor do they seem to be helped much by the monitor-"Observe Baby" was not grouped with "Biophysical Observation", implying that we need 'attention getting' mechanisms.

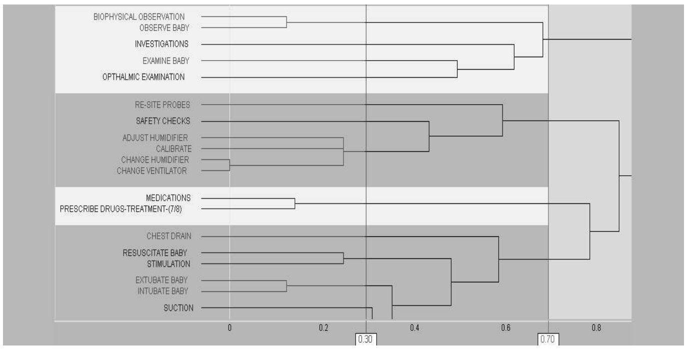
It is clear from the dendrograms that the various groups of staff within the NICU interpret and categorise data differently. The grouping of data appeared to be associated with particular professional practice. For example the actions related to artificial ventilation were, in general, grouped the same for senior and junior nurses; senior doctors had a similar group but omitted the management of the ventilator humidifier - this was grouped with issues of equipment safety. Like senior doctors, junior doctors similarly clustered actions related to artificial ventilation but in their minds, humidifier management and equipment safety were associated with routine nursing care. These variations in representation may correspond to differences in knowledge and or professional role and responsibilities. The dendrograms of the junior nurses and of the senior doctors displayed the greatest difference of the 4 staff groups.

4.2 Cumulative Probability Graphs

We produced graphs of the form shown in figure 3 from the distance matrices derived from the card-sorting experiments. These graphs provide a summary of the structure of the dendrograms that are



(a) Section of the "Actions" dendrogram for junior nurses.



(b) Section of the "Actions" dendrogram for senior doctors.

Figure 4. Example sections of dendrograms.

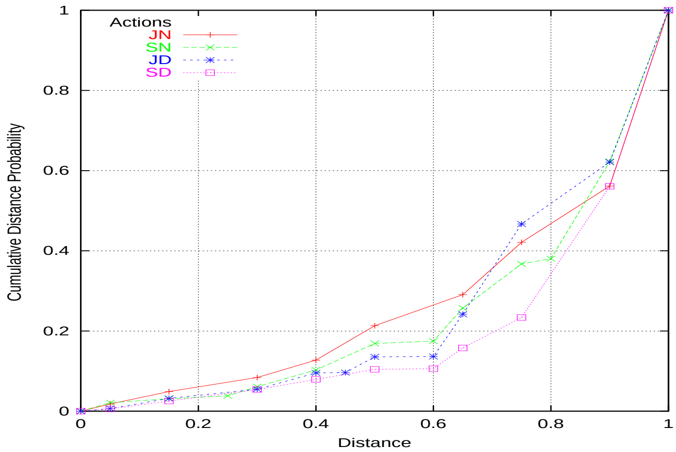


Figure 5. Plot of the cumulative probability that any "Actions" card pair are separated by a distance less than d

produced from cluster analysis of the card-sort data. As the loci of the graphs approach a straight line from (0,0) to (1,1), (become less concave) the associated dendrogram becomes less structured, and a straight line graph would represent a structureless (maximum entropy), meaningless 'dendrogram'. (This was verified experimentally.) Thus the area under the curve is related to the degree of organisation of the associated dendrogram. Also, the actual shape of a curve indicates the level of branching of the associated dendrogram for various distance values; e.g. the trajectory of a curve remaining shallow and then rapidly increasing after a distance, d, would indicate that most branching (in the associated dendrogram) is ocurring at a distance greater than d.

Consider figure 5, which shows the graphs of the proportion of card-pairs with distances greater than values of distance versus that distance. Visual inspection of the graphs indicate that the degree of structure in the dendrograms increases for staff class in the order:- junior nurses, senior nurses, junior doctors, senior doctors, at least for values of $d < 0.6$. This agrees with what is seen on the actual dendrograms, and intuitively makes sense in light of the actual roles of the various staff classes (A deeper discussion of staff roles is beyond the scope of this paper, and will be discussed further in a forthcoming paper[4].). At this stage we have not calculated any confidence limits for the graph, and so no definitive quantitative statements can be made about differences in graph attributes.

4.3 Information Measures

We have developed an information theoretic model, quantifying the relationships of the distance matrices for different staff classes, as outlined in section 3.2, and defined by equation 14. We then implemented equation 7 (logarithm to base e was used), using the parameters derived from the distance matrices for the different staff classes, and defining separate information systems, where the source and user for each system was defined for all the possible combinations of staff classes. This will become clear by inspection of table 3, which quantifies the mutual information (in Nepers, log base e, per symbol [card-pair]) for all the possible notional information systems for the four classes of staff. For example, if we define an information system were the source is derived from the junior nurse's distance matrix, and the user is dervied from the senior nurse's distance matrix, the mutual information is 0.32 information units per symbol (Neper).

When the source and user are identical, the mutual information is zero, as there is no 'surprise' or uncertainty in determining the user symbol given the source symbol. The larger the number implies a greater information mismatch between source and user. Perusal of table 3 indicates that there is the greatest information mismatch between senior doctors and the other three groups, whereas the information mismatch between the junior nurses, senior nurses and junior doctors is about constant at 0.3 Neper. This reflects what was shown in the dendrograms and the graphs described earlier. However, a numerical value is given that summarises the information differences in the dendrograms, and as such may be used by computer algorithms as a metric of information mismatch.

Table 3. Mutual information measures, as a summary of the difference in information content of dendrograms for the various staff classes. The units are Nepers (natural logarithm) per symbol.

Staff Class	JN	SN	JD	SD
JN	0.00	0.32	0.31	0.47
SN		0.00	0.29	0.48
JD			0.00	0.43
SD				0.00

5 SUMMARY and CONCLUSION

We have described card-sorting experiments designed to elicit knowledge, from domain experts in a Neonatal Intensive Care Unit, about how they mentally map clinical concepts. These experiments produce data on how the subjects group concepts based on some notion of similarity. Often, this card-sort data is not processed further[3], but if processed further, the usual mode of analysis is cluster analysis, which is applied to the data to produce dendrograms. These dendrograms give useful incite into how people mentally organise concepts, but are often complex and tedious to analyse, and are not easily amenable to inclusion in computer algorithms.

We have shown here, the feasibility of using both a graphical method and a numerical method to summarise the information implicit to dendrograms. The graphical method, which is not fully explored here, allows direct visual comparison of different (but related) dendrograms. This facilitates a quick analysis of the difference in structure of dendrograms under consideration. The numerical method uses an analogy to an information/communications system model to produce measures of the information mismatch reflecting the difference in structure of dendrograms (of the same domain). These information metrics are suitable for direct use in computer decision algorithms.

Note, this paper has only discussed the feasibility of the approaches developed within, and further development is needed, in particular, the analysis of confidence intervals for the methods discussed.

ACKNOWLEDGEMENTS

This work was undertaken as part of the NEONATE project; we are grateful to both the UK ESRC and EPSRC for providing funding under the People at the Centre of Communication and Information Technologies (PACCIT) program. We also acknowledge the input of our co-workers on that project: Peter Badger and Alan Maxwell.

We would like to thank the referees for their valuable comments which helped improve this paper.

REFERENCES

[1] Mark S. Aldenderfer and Roger K. Blashfield, *Cluster Analysis*, Sage Publications Inc., Newbury Park, California, USA, seventh edn., 1990.

[2] A. M. Burton, N. R. Shadbolt, G. Rugg, and A. P. Hedgecock, 'The efficacy of knowledge acquisition techniques: A comparison across domains and levels of expertise', *Knowledge Acquisition*, **2**, 167–178, (1990).

[3] Jianming M. Dong, Martin Shirley, and Paul Waldo. A user input and analysis tool for information architecture. http://www-3.ibm.com/ibm/easy/eou_ext.nsf/Publish/649, June 1999.

[4] Y. Freer, N. McIntosh, L. Ferguson, G. Ewing, R. Logie, J. Hunter, and S. Rudkin, 'Similarities in 'concepts' between staff working in a NICU', Utrecht, The Netherlands, (4 - 7, September 2002). Annual Scientific Meeting of the European Society for Paediatric Research.

[5] R.W. Hamming, *Coding and information theory.*, Prentice-Hall, Englewood Cliffs New Jersey, 1980.

[6] P. Johnson, *Human-Computer Interaction: Psychology, Task Analysis & Software Engineering*, McGraw-Hill, 1992.

[7] A.E. Karbowiak, *Theory of Communication*, Oliver & Boyd, Edinburgh, 1969.

[8] Claude E. Shannon and Warren Weaver, *The Mathematical Theory of Communication*, Urbana: The University of Illinois Press., 1949.