IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4197

A Task Model For the Retrieval of Statistical Data in Decision Making

Peter N. Hyland
Department of Information Systems, University of Wollongong, Tel: 61 2 42213759, Fax: 61 2 42214474, peter_hyland@uow.edu.au

INTRODUCTION

When faced with complex decisions, business and community leaders often need to use statistical data, either from within their own organisation or from external sources. These sources include government departments, national statistical authorities, international agencies and other businesses. Although such statistical data has been traditionally provided in hard copy, as summary tables and charts, statistical data (SD) is now widely made available via Statistical Information Systems (SIS). SIS have employed technologies like spreadsheets, GIS, statistical databases and statistical packages, such as SPSS, SAS etc, and, more recently, Online Analytical Processing (OLAP).

Given the importance of SD in corporate and community decision-making, it is surprising that the task of retrieving and analysing SD has not been the subject of significant research. To address this shortcoming, this paper describes the usability testing of *Abacus*, an SIS prototype based on a combination of OLAP, Multi-Dimensional Databases (MDDB) and the World Wide Web. During the testing of the prototype, data was gathered about the way users actually work with SD. This data was used to develop a detailed task model for the retrieval and analysis of SD.

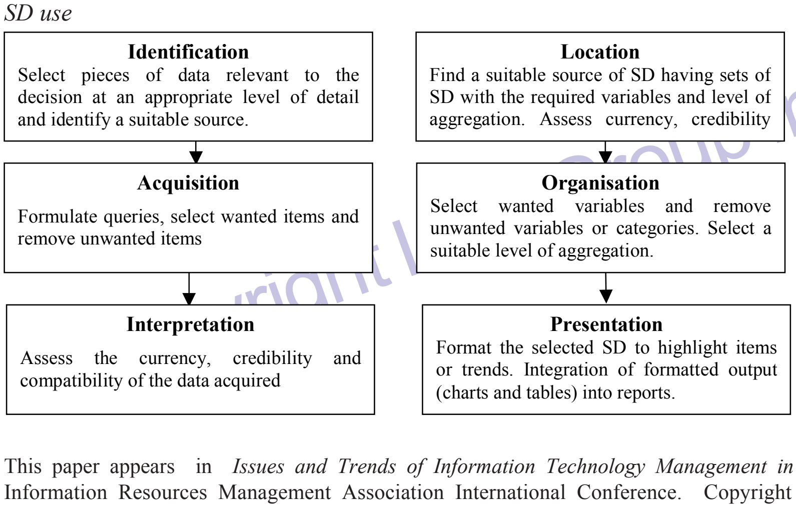
DEVELOPING AN INITIAL MODEL

As mentioned above, the task of retrieving and analysing SD is not well represented in the literature. However, an initial model can be developed from the literature on two related topics. The first concerns the use of information in decision making and the second is the literature on SIS.

The use of SD is frequently associated with decision making and one model of information gathering to support managerial decision making is described by Goodhue (1998). She suggests that the task of information gathering consists of three sub-tasks, shown on the left of Figure 1. A similar model for the use of SD is suggested by the description of an SIS provided by Malmborg and Sundgren (1994). In their description, they associate SIS use with the three sub-tasks shown on the right of Figure 1.

The usefulness of this initial model was confirmed by data gathered in two pilot studies (Hyland and Hasan, 1997; Hyland and Gould, 1998). In addition, examination of several existing SIS suggest that the functions provided by these SIS are consistent with the initial model for SD use. The model was further extended by analysis of data gathered during the usability testing of the *Abacus* prototype.

Figure 1: Comparing a model for information gathering and an initial model for SD use



USABILITY TESTING OF THE *ABACUS* PROTOTYPE

Abacus provides an easy-to-use interface that allows users to create charts or multi-level tables of values by applying the common OLAP tools, such as slicing and dicing, filtering, drilling down and rolling up. The resulting tables could have up to six variables such as Age, Sex, Income, Location etc.

The data used in the prototype was a 1% sample of the 1991 Australian census, having over 20 variables and up to 160000 records. The census data concerns dwellings, families and people. For each of these primary entities, *Abacus* had three separate data sets e.g. People A, People B and People C. Although these data sets described the **same** population, the population in each set was described in terms of **different** variables e.g. People B used the variables Occupation and Location while People C used the variables Sex and Age. *Abacus* was tested by 36 volunteers - university students, academics or professional staff from the local government. Although most volunteers were relatively inexperienced with manipulating SD, there were also a number of quite experienced users of SD. This sample is comparable to those in studies described by Nielsen and Levy (1994).

All volunteers were given about 30 minutes training in the use of *Abacus* and then asked to answer 18 questions of the form:

How many married couples, living in the city of Sydney, earned over $80000 p.a.

A major goal of the research was to identify the strategies used by the participants to answer questions and to move from one question to the next. To promote the evolution of such strategies by participants, the training materials intentionally did **not** demonstrate the use of particular strategies and the questions were intentionally organised so as to prompt participants to develop new strategies.

THE LOCATION SUB-TASK

The Location sub-task in the proposed task model involves not only finding one or more sets of SD but also judging the suitability of those sets of data. In a business situation, a user must decide if a particular source of SD is reliable, if the data is sufficiently current and if the data is accurate. In the model proposed by Goodhue (1998), these decisions are dealt with in a separate sub-task at the end of the information gathering task. When dealing with SD, it would appear that decisions about suitability should be made earlier because the organisation of SD is a complex activity and it would be inefficient to spend time and effort organising SD that was not current, accurate or sufficiently reliable.

Putting aside the issue of data quality, the Location sub-task is primarily concerned with finding a set of data that contains the required variables and the required categories of those variables e.g. does the data have information about Income and is the grouping “$50000 to $80000” included. In some SIS, it may be possible to create new variables or categories, in which case, a data set that did

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

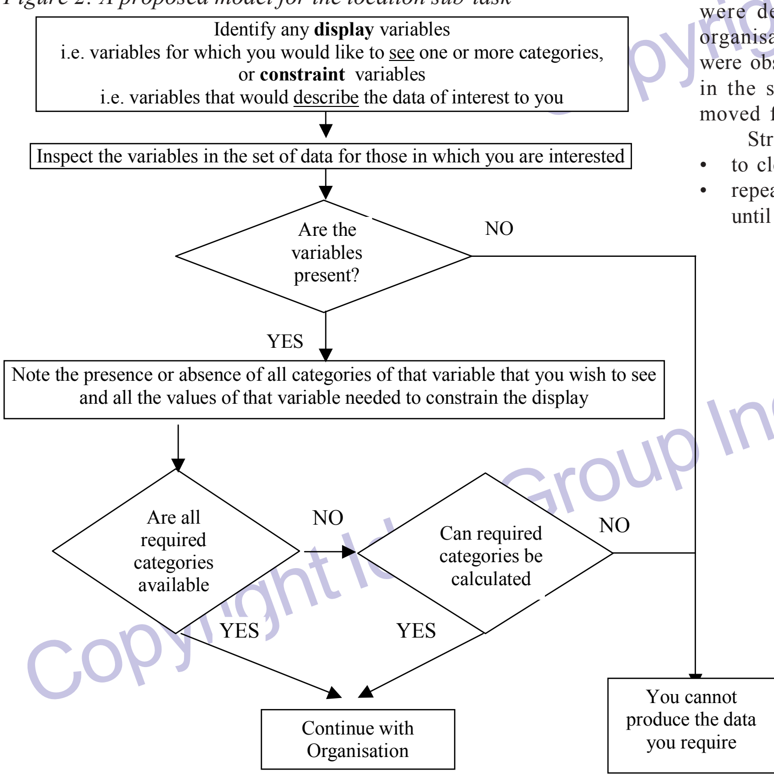
not appear to have the required variables or categories might be modified to produce them and thus satisfy the user's needs. The user would need to take this possibility into account when locating a set of data.

To explore the Location sub-task, 2 of the 18 questions did not contain the name of the required data set. For these two questions, participants were only told which entity the question concerned i.e Dwellings or People, and which variables and categories were needed to answer the question. The amount of time taken by participants to locate the correct data set and the number of errors made while doing so were quite surprising.

The wording of the first of these two questions indicated that participants needed to find a data set concerning Dwellings. The data set that was required to answer the question was Dwellings A, so it should only have been necessary for participants to inspect the first of three sets of data about Dwellings. Many participants found the task very simple; the minimum time taken was only 4 seconds and 30% (N= 11) of participants completed the task in 10 seconds or less. However, 27% (N= 10) of the participants took between 30 and 90 seconds to locate the correct data set. Moreover, the behaviour of participants was often quite illogical. For example, some participants ignored the reference to Dwellings in the question and began by inspecting the data sets about People. Other participants **repeatedly** inspected the data sets named Dwellings B and Dwellings C but did not inspect Dwellings A. The second Location sub-task also seemed to cause participants significant problems, although somewhat less than the first.

It should be remembered that the actions being described are only the simplest aspects of the entire Location sub-task and only require the participant to confirm the presence of relevant **variables**. The Location sub-task would normally require the participant to confirm the presence of appropriate **categories** of those variables as well. For example, a user who wanted to know the number of plumbers in a region would not be satisfied with a data set containing an Occupation variable if the lowest level of aggregation for Occupation were Tradesmen. Given the apparent difficulty in completing the relatively simple Location sub-task in this study, it can only be assumed that the more complex activity of matching categories as well as variables would be problematic for many users.

Figure 2: A proposed model for the location sub-task



Based on the data gathered during usability testing and reflection on the development of the prototype, the following normative model of the Location sub-task appears to be viable and useful. As with Goodhue's (1998) model for information gathering, there might be overlaps between the Location sub- task shown below and Organisation sub-task.

THE ORGANISATIONS SUB-TASK

The pilot studies suggest that the Organisation sub-task involves four steps that could be carried out serially or in parallel. Firstly, the user must identify the required variables and include those variables, and only those variables, in the display (i.e. table, chart, map). This could involve the addition of variables to the display or the removal of unwanted variables from the display. It may be necessary at this time to calculate new variables (e.g. convert an existing variable to a percentage, combine two or more variables etc). The user must then assess the suitability of the default level of aggregation and either zoom in or zoom out on one or more categories of the chosen variables. Once again, the user may need to calculate new categories or group categories together. The final step is the placement of variables and categories in such a way as to highlight any features of the data that the user thinks are important. In the case of a table, for example, this might involve placing one or more variables as rows while other variables are placed as columns. In cases where this resulted in multi-level tables i.e. having multiple rows or multiple columns, the placement of variables in that hierarchy would also be significant. These appear to be the major steps in the Organisation sub-task in a variety of SIS.

STRATEGIES USED IN THE ORGANISATION SUB-TASK

One of the objectives of user testing was to find out how participants would use the various OLAP functions to find required data. The structure of both the training materials and the questions themselves were deliberately designed to encourage participants to evolve organisation strategies. Only two fundamentally different strategies were observed. However, there were a number of interesting variations in the second of these strategies and also in the way participants moved from one question to the next.

Strategy 1 was:

- to clear any existing table and then
- repeatedly apply **slices**, one for each variable used in the question, until the answer is shown.

This approach does not result in a true table being formed and so participants who used this method were actually in violation of some of the instructions on their question sheets. This strategy was used at least once by 27.8% of participants (N=10). This approach is very fast and many of the participants who used it completed the questions very quickly as a result. Unfortunately, it is also quite limited in its applicability because it cannot be used when there are two conditions for a single variable e.g. finding the number of people who are either widowed or divorced. Despite its limitations it is a very effective strategy and should be used wherever possible. However, users need to understand its limitations and the use of other strategies to overcome those limitations.

Strategy 2 was:

- to clear the existing table, then
- repeatedly add a variable as a row or column and
- zoom in to those variables to the required level of aggregation.

This strategy was used at least once by 88.8% (N=32) of the participants. It often resulted in a very large table that participants found difficult to use. Faced with such large tables, participants often had to scroll repeatedly backwards and forwards through the table, keeping track of their position by putting their finger on the monitor screen. Despite this difficulty, many users never evolved any strategy for controlling table size. There were, however, many variations on this method that did solve the problem of table size by combining this strategy with other OLAP functions, as described below.

Given that this strategy often results in very large tables, it is not very effective by itself. It may be useful for simple tables containing only one row and one column, but even here, table size could still become a problem if one or both of the variables on the table had a large number of categories. In such cases, or when dealing with more complex tables, some variant of this strategy would be needed.

There were four successful variants of Strategy 2.

1. Limit the size of the table by **filtering** out unwanted categories
   This is an effective and versatile variation that will allow the user to generate a compact table that contains only the desired data.
2. **Slice** one or more variables not on the table
   This is a quick and effective variation that allows the user to generate a relatively compact table. In most cases it will not allow the user to show only the desired data because unwanted categories of some variables will still be displayed. It allows the inclusion of only a single condition for each variable not on the table. It is far less time consuming than defining a rule and, under ideal circumstances, the information provided on the screen is highly communicative about the extent of the data being displayed. It is also very easy to view different slices sequentially e.g. slice first on Age = 40, reset the slicer and then slice on Age = 55.
3. Use a combination of slicing, filtering and zooming
   This is by far the most flexible variation, combining the versatility and compactness of variation 1 with the speed and simplicity of variation 3. It's only drawback is that the user must be proficient with all the main OLAP functions and must be able to develop quite a complex plan of action. Using such a complex strategy, a user could easily become confused and make an error of logic.
4. Change rows to columns or vice versa
   This was not an important variation in the testing but it may be useful in the workplace.

Having described the strategies and the variations on those strategies it is timely to make two comments. Firstly, the majority of participants always used strategy 2 without any variation. In fact, the variations on this strategy were only used in 25% of the situations when they would have been appropriate. Nonetheless, it is quite impressive that significantly complex strategies were developed and used in the participants' first encounter with *Abacus*.

Secondly, in assessing the significance of these variations, it must be remembered that the training materials only showed participants how to use the OLAP functions, not how to combine them with one another. Consequently, the evolution of the initial strategy and these 4 variants on that strategy is an indication of the ability of users to integrate OLAP functions to produce desired results.

A second group of variations was observed in the way participants moved from one question to the next. When moving from one question to the next, most participants would clear the existing table completely. The generation of a new table at the start of a question may be appealing to participants because it returns the participant to a stable point from which he or she can follow the same strategy every time. It also ensures that the participant does not have any residual settings in place from a previous question. However, the questions themselves were organised in such a way that it was, at times, very inefficient to start a new table. Instead, it was often possible to use some of the OLAP functions to simplify the movement from one question to the next and preserve much of the current table.

The 3 variations on this process included:

- adding new variables to an existing table,
- removing an unwanted category from an existing table and
- slicing a new variable rather than adding it to the table.

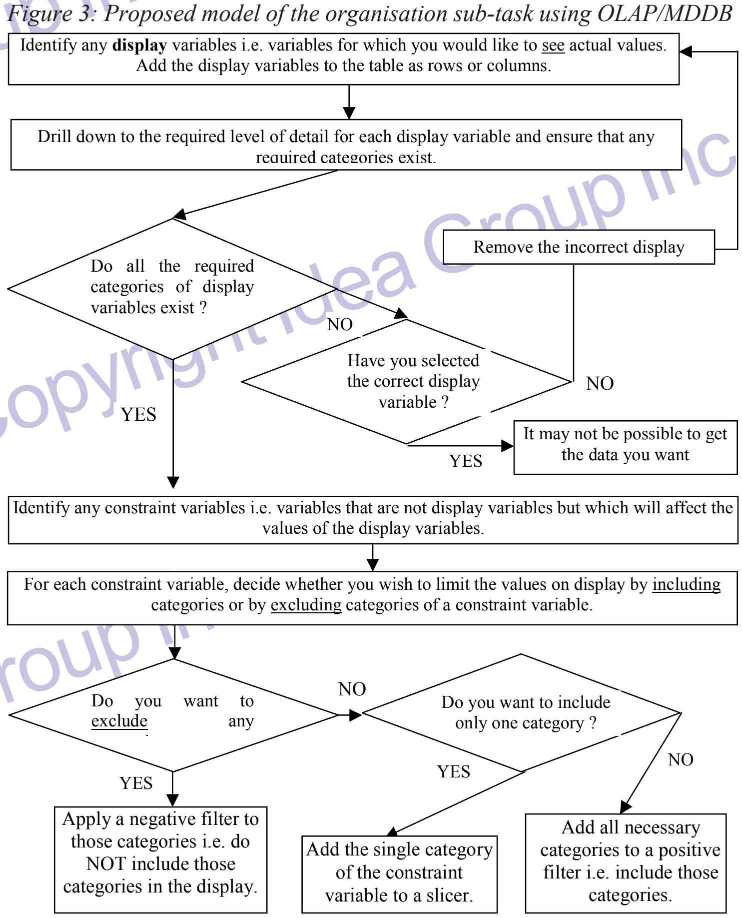
These variations did not make it any easier for participants to answer the questions, they only reduced the time taken to complete all of the questions. Although the questions were deliberately organised to test the usefulness of these variations it was expected that participants would be too involved with getting the answers correct to bother about such efficiencies. Judging from the number of individuals who used these variations, this is clearly not the case. For example, one variation was used by 24% (N= 7) of the participants.

Extending the model of the Organisation sub-task

Based on the data gathered during usability testing and critical reflection on the development of the prototype, a model of the Organisation sub-task is proposed. Since this particular analysis does not concern itself with the Location sub-task, the model assumes that all of the required variables and all of the required categories of those variables are available. Figure 3 shows a normative model for the sub-task i.e. a set of recommended steps for users to follow. Although it is based on a substantial body of empirical evidence, it does not attempt to model all the strategies observed. Instead it represents the results of critical reflection on various strategies and proposes one strategy that appears to produce the required results in a very efficient manner.

Unlike the proposed model of the Location sub-task, which would be appropriate for a number of types of SIS, the model of the Organisation sub-task is only appropriate for SIS having analytical functions like those in OLAP/MDDB applications. The sorts of activities carried out with other types of SIS are extremely varied. Given the enormous differences in functionality provided by the full range of

Figure 3: Proposed model of the organisation sub-task using OLAP/MDDB



available SIS, it does not appear feasible to produce a single task description for them all.

THE PRESENTATION SUB-TASK

The Presentation sub-task consists of two parts – selecting the display type and the display style. The **display type** might be a table, chart or map. The decision of which of these types of display to use would depend on the amount of data, the level of detail that is necessary and the audience and purpose for which the display is intended. Charts and graphs may be more helpful in showing trends while tables provide more detail. Obviously maps are only appropriate where a geographical variable is important. The **display style** would include fonts, colours, shading, cross-hatching etc. These can be important factors in determining the usefulness of a display of SD. Although the *Abacus* prototype supports both tables and charts, the questions completed in usability testing did not involve the use of different display types or the incorporation of such displays into reports or other documents. Consequently, no details about the Presentation sub-task can be determined from the current study.

CONCLUSIONS

The data gathered during usability testing have allowed us to identify some of the strategies that can be employed when retrieving and analysing SD. Knowledge of these strategies has allowed us to confirm the usefulness of the model proposed in Figure 1 and to provide additional detail of some of the sub-tasks suggested in that model.

Although the results described above are interesting and useful, they are subject to limitations. For example, the strategies described have been evolved to cope with a specific set of questions, which may have elicited only a limited set of strategies or pre-disposed users to one strategy more than another. In addition, the model for the Organisation sub-task is somewhat dependent on the OLAP technologies used in the prototype. A different set of strategies would be found if the prototype had used GIS, for example. Having said this, it is difficult to imagine how users could arrive at the desired result without having some mechanism to control the addition of variables, the paring of unwanted categories or the aggregation or disaggregation of the data.

The current research has provided an initial model of the task of SD use. This model could be tested in a number of ways. A formal experiment could be conducted to measure differences in performance between two groups of users, one having access to the model and the other not. These groups could be required to carry out tasks similar to those in the current study using *Abacus* or some similar tool.

References

Goodhue, D.L. 1998, 'Development and Measurement Validity of a Task-Technology Fit Instrument for User Evaluations of Information Systems', *Decision Sciences*, vol. 29, no. 1, pp. 105-138.

Hyland, P.N. & Hasan, H. 1997, 'The Use of External Statistical Data in Local Government', *Urban Policy and Research Journal*, vol. 15, no. 4, pp. 279-289.

Hyland P.N. & Gould E. 1998, 'External Statistical Data: Understanding Users and Improving Access', *International Journal of Human Computer Interaction*, vol. 10, no. 1, pp. 71 – 83.

Malmborg E. & Sundgren B. 1994, 'Integration of Statistical Information Systems - Theory and Practice', *Scientific and Statistical Database Management - Proceedings of the International Working Conference, 1994*, pp. 80–89.

Nielsen, J. & Levy, J. (1994) Measuring Usability: Preference Vs. Performance, *Communications of the ACM*, **37**(4), pp. 66-75.

Related Content