# Short Talks

## USER MODELS AND BEHAVIOR

## IDENTIFYING HIGH-LEVEL UNIX TASKS

**Russell J. Branaghan, James E. McDonald &
Roger W. Schvaneveldt
New Mexico State University**

### INTRODUCTION

Task analysis (TA) in systems design has a number of goals, including identifying information needed by the user, knowledge and skills that the user needs, and potential errors that the user might make. However, a necessary precursor to TA is to determine what the important tasks are in the first place. This research is aimed at developing a methodology for identifying tasks in command-based human-computer systems. The following paragraphs delineate the motivation for the approach, as well as the general procedures in this methodology.

The common denominator of all TAs is the description of jobs in terms of identifiable units of activities (McCormick, 1979). This involves selecting some level at which the tasks will be described. However, it is not clear what this level is, and there are no formal methods for finding out. Currently, identifying these meaningful units is done on the basis of intuition, which is problematical.

The present approach is based on the following assumptions: (l) a task is a psychologically meaningful unit of work behavior which is performed for its own sake, and is not dependent on other job actions for its meaningfulness (Phillips, Bashinski, Ammerman, & Fligg, 1988), (2) in a command-based system, a task is a sequence of commands issued to achieve some goal, and (3) identifying tasks involves discovering command sequences which occur much more often than chance.

### METHOD

This work used data which was previously collected by Anderson (1988). Protocols were obtained from nine experienced UNIX users. The user's *.login* files were altered so that histories were collected of each interaction the users had with the system and were automatically mailed to the experimenter. The protocols contained all of the commands that the users issued during their interactions with the system.

All command sequences occurring at least three times in the protocols were identified. These were then ranked according to their scores on the following measures:

Co-occurrence frequency: The frequency with which the command sequence of interest occurred in the protocols.

Conditional probability: The probability of command y following a sequence of commands x, where x can be a sequence up to five long.

Mutual information: A measure of the observed probability of a sequence in a protocol divided by the expected probability of the sequence in the protocol. The expected probability is found by multiplying the component independent probabilities. This ratio is taken to the log base 2.

Intersection over union: The frequency of a sequence of commands occurring together divided by the sum of the independent command frequencies minus the intersection.

These scores were computed for all sequences from length 2 to 6. A composite of these scores yield a number of promising sequences. An exhaustive list is beyond the purview of this paper, but a small sample of sequences identified as good by the composite is shown below.

**Sample Results:**

| | | | | |
|---|---|---|---|---|
| refer | tbl | pic | ficpic | eqn |
| emacs | soelim | refer | tbl | pic |
| ls | (pipe) | more | | |
| tbl | (pipe) | eqn | (pipe) | troff |
| set | biff | set | | |
| biff | set | mail | | |
| tbl | eqn | troff | | |
| cd | pwd | ls | | |

### POTENTIAL BENEFITS

This research is preliminary in nature. However, if valid, it offers a number of advantages over approaches involving interview, observation, and intuition. First, it is replicable, i.e. different analysts should reach the same conclusions given the same data. Second, it is likely to be faster than traditional methods, since it abbreviates time-consuming interview processes. Finally, it can be almost completely computerized.

We are currently conducting studies to determine the validity of methods such as these as well as to determine which measure identifies the best tasks. Further research might be aimed at what information is sacrificed when forsaking interview and observational techniques.

### REFERENCES

Anderson, M. P. (1988). *The transfer and declarative representation of procedural knowledge.* Unpublished doctoral dissertation, New Mexico State University.

McCormick, E. J. (1979). *Job Analysis: Methods and Applications.* New York: AMACOM.

Phillips, M. D., Bashinski, H. S., Ammerman, H. L., & Fligg, C. M. (1989). A task analytic approach to dialogue design. In M. Helander (Ed.) *Handbook of Human-Computer Interaction.* New York: North Holland.

### CONTACT INFORMATION

Russell J. Branaghan, James E. McDonald &
Roger W. Schvaneveldt
Department of Psychology
New Mexico State University
Box 3452 Las Cruces, NM 88003
(505) 646-6206

## UNDERUTILIZATION OF ARCHIVAL FACILITIES IN A UNIX ENVIRONMENT: A RESOURCE ALLOCATION PROBLEM

**M. Elliott Familant**
**Bellcore**

Ideally, information should be distributed among storage devices according to how frequently it is accessed. Rarely accessed information should be stored on the cheapest (and slowest) devices, frequently accessed data should be store on the fastest (and more expensive) devices.

End-users can move information between different secondary storage devices, remove information they create, compress information on disk, and create copies of information (for the purpose of backing it up). All these actions result in a pattern of information distribution that can be evaluated in terms of optimization. This study asked whether users in a computing environment in which there are few restrictions will allocate information between storage devices in a way that optimizes their use.

This paper focused on information that is stored for archival purposes. A file was defined as a candidate for archival based on the date when the file was last accessed. File access occurs when a user applies some operation to the contents of a file. Operations include: editing, moving, copying, compressing, executing, and listing.

Users are provided with a default media on which they can store information. In the Bellcore UNIX computing environment, this is a portion of a fixed disk. Bellcore also has two archival systems: a system for transferring data stored on disks to tape and a method for compressing files. To optimize their storage, users need to engage in the extra step of moving the information to tape, or transform it into a less accessible form by compressing it.

The study presented here analyzed the files in the directories of 79 users of UNIX minicomputer systems. The dates when each file was last accessed were computed. The analysis looked at the number of files that have been unaccessed for various time intervals (at least 6 months, at least a year, etc.). Additional analyses were done to determined the proportion of files that were compressed and the average and range of file name lengths.

Data on 52,705 files was obtained. The initial analysis looked at the number of files that were unaccessed as a function of time. These analyses provide evidence that significant proportions of storage space on the disk packs were being consumed by files that have not been accessed for long periods of time. For example, 25,656 files or 48.7 percent of the files surveyed had not been accessed for 6 months or longer when the sample was taken. Over 32 percent of the files surveyed had not been accessed in the past year. A small percentage of users account for a large percentage of the disk storage used. We found 20 percent of the users consumed 65 percent of the storage space in this survey.

An examination of existing archival resources showed that they were being underutilized. Over 79 percent of the sample had never archived a file on tape. An almost equal proportion of the sample had never compressed a file. A number of causes were identified for this situation. First, archival facilities are generally not well advertised. Therefore, many users may simply not know that these facilities exist. Second, an analysis of the user-interface of the tape archival system revealed an awkward and hard to use procedure for archiving information on tape. Third, underutilization of the other archival system, file compression must be attributed, in part, to a lack of incentives by the users to archive information.

The file compression facility is both easy to use and is readily accessible. Although some of its underuse can be attributed to users not knowing about the facility, users who had compressed files in the past also underuse it. A likely reason for the underutilization of this system is that its use rarely brings a user closer to accomplishing the task goal towards which the user is working. Optimization of storage is usually not one of the goals that users pursue. In fact without coercion, deliberately moving information to cheaper media is something of an altruistic act. The user receives no immediate benefit from performing this action and must consume scarce resources (e.g. time) in order to do it.

File storage is a shared resource. The amount of storage available generally exceeds the needs of any one user. Therefore, no one user needs to reduce their personal consumption of file storage resources in order to accomplish the task on which he or she is working. Because they receive no benefit from conserving, users in general will use the most expensive and most easily accessible file storage devices. However, in situations in which there are many users, this will result in a significant waste of storage resources.