

## BROWSING IN LARGE DATA BASES

Douglas D. Dankel II  
Assistant Professor  
Computer and Information Sciences Department  
University of Florida  
Gainesville, FL 32611

The formation of data bases containing information on mechanical systems for troubleshooting purposes has become increasingly popular and important. Examination of these data bases by humans can be very costly. A system called BROWSER was developed to heuristically search a data base containing information on U.S. Navy aircraft with little or no human intervention. BROWSER searches the data base guided by models and heuristics looking for interesting patterns or configurations. The user is then notified of the existence of these patterns.

### 1. INTRODUCTION

Consider the tasks of trying to find similarities between the causes of the crashes of several aircraft of a common type, the causes or early warning signs for various diseases, the characteristics of stocks with high growth potential or advanced warning signs of severe weather conditions. All of these tasks require a large amount of data and valuable time spent sifting through the data. Browsing or troubleshooting computer systems of the future will eventually perform these tasks. BROWSER is designed to be the prototype of such systems.

BROWSER is an automated system for troubleshooting with data bases. It is based on Douglas Lenat's automated discovery in mathematics [1] [2] and is designed with a modified production system architecture [3] [4]. BROWSER explores a data base using models describing the data and a large collection of data-dependent and data-independent heuristics.

An example of a BROWSER data base is a data base containing information on maintenance actions performed on U.S. Navy aircraft [5]. The data base is explored through the definition of data subsets and their examination using simple statistical techniques. Within a data subset BROWSER looks for regularities such as recurring sequences of data or unusually high occurrence rates of a particular datum. BROWSER attempts to find

sequences of maintenance actions which precede a particular maintenance action. Such an action might be the failure of a component in a system. If such a pattern can be found, it could be used to give warning of the possible failure of that component. When one of the preceding events occurs the part which is due to fail could be repaired or replaced eliminating possible adverse conditions which might later occur due to the failure of the component (e.g., the aircraft crashing).

When comparing two data subsets BROWSER looks for statistically significant differences in the occurrence rates of a particular datum. BROWSER can compare aircraft with very good maintenance histories to aircraft with very poor maintenance histories to find differences. Discovery of these differences might lead to modifications on the aircraft with poor histories which will improve their system performance.

All discoveries or facts about the data are reported to the user. The user could be given the option of creating alerters [6] [7], which would monitor for the occurrence of these facts in new data entering the data base. (Alerters are not currently implemented.)

### 2. IMPORTANT BROWSER FEATURES

BROWSER contains three significant features. First is the representation of facts and knowledge known about the data base. This knowledge is contained in models and data-dependent heuristics. Second is the controlled execution of tasks within the system. For this, BROWSER uses an agenda, an ordered list of all tasks which are to be executed, data-

-----  
This work was supported by the Office of Naval Research under Contract N00014-75-C-0612 while the author was at the University of Illinois at Urbana-Champaign.

independent heuristics, and a controller, which executes the agenda tasks. The final significant feature is the representation of new knowledge. New knowledge is represented and stored in concept trees. Each concept represents a subset of the data and contains several facets which describe the subset.

## 2.1 Data Base Knowledge

Knowledge of the data base is stored in models and data-specific heuristics. Five general types of models exist in BROWSER. The first four model types provide detailed knowledge about the data in the data base while the fifth provides general knowledge on data structures useful for examining the data.

The lowest level models are the Data Base Models. One Data Base Model exists for each type of data file in the data base. Each model contains a description of the data fields which compose the data file, including the field's retrieval name (used in the data base query language), an equivalent English name, the starting position of the field in a data record, the length of the field, and the type of value contained in the field (numeric or character). The Data Base Models are used mainly by the data base management system for retrieving data from the data base.

The Data Specific Models provide information on how the various data fields are related to each other. These models show the semantics of the data fields. Again, one Data Specific Model exists for each type of data file in the data base.

The Specific Models are of two varieties. The first shows how the data records, described semantically in the Data Specific Models, combine to form larger units called Conceptual Units. For the aircraft data base an example Conceptual Unit is a maintenance action performed on an aircraft. Each performed maintenance consists of several actions, such as the removal of an aircraft part, the repair of a part or the installation of a part, where each action is described by a single data record in the data base. The second type of Data Specific Models describes the objects from which the data was gathered. For the aircraft data base these models are of the F-4 and A-7 aircraft.

Typical Models are the most general data specific models. They describe generalizations of the Specific Models. Generalized knowledge which applies across Specific Models is stored in the Typical Models. BROWSER's Typical Models include a model of a typical aircraft and the model of an aircraft's history. The typical

aircraft model represents common knowledge known about all aircraft. The aircraft's history model shows how the Conceptual Units of performed maintenance combine to form the overall maintenance history of an aircraft.

The final model type is that of the General Models. They provide information on the typical types of data structures which exist within the data. Included are models of a network, a tree, and a list.

Data-dependent heuristics are contained within the models. These heuristics provide knowledge about the data values in the data fields and other knowledge not represented by the structure of the model. Interconnections between models are provided so this heuristic knowledge can be shared.

## 2.2 Controlled Execution of Tasks

The second significant feature of BROWSER is the use of an agenda to control the execution of tasks within the system. BROWSER's basic operation is to define new subsets of data and explore them. At any one time BROWSER might have 50 or more subsets of data being considered for examination, many additional subsets currently being examined, and some subsets which have already been fully examined. The system uses an agenda to keep track of all of the tasks associated with the subsets which must be performed. These tasks are of two types: those which fill in the facets of a concept and those which explore and examine the data subsets associated with the concept.

The tasks on the agenda are ordered. This ordering is determined by several factors. Certain tasks must be executed before others: e.g., the definition of the data contained in a subset must exist before that subset can be retrieved from the data base, and the subset must be retrieved before it can be examined for regularities. Additionally, some tasks appear to be more interesting to perform than others because of the results of previous tasks. Each task has a list of reasons why it should be executed, each reason has a value and the sum of all of the values, the task priority, gives the position of the task within the agenda.

The basic procedure of BROWSER is as follows. All tasks are sorted according to their priorities. The highest priority task is removed to be executed. All heuristics, both data-dependent and data-independent, which are relevant to the execution of this task are gathered. These heuristics are then executed one by one. When all of the relevant heuristics have been executed the cycle is repeated. This

cycle is continually repeated until all tasks have been executed.

### 2.3 Representation of New Knowledge

The final significant feature of BROWSER is the representation and storage of new knowledge. New knowledge consists of facts gathered by BROWSER from the data base and is stored as faceted concepts on concept trees. Initially the data base is described by a small set of concepts: ALL-PLANES (the entire data base), CRASH (aircraft which have crashed), NON-CRASH (aircraft which have not crashed), HI-MAIN (aircraft requiring excessive maintenance) and LO-MAIN (aircraft with good maintenance histories). BROWSER creates more subsets and their associated concepts by examining existing subsets of data and offering ideas on how the existing subsets might be restricted to form smaller subsets or, possible, generalized to form larger subsets. Each new subset is described by a concept in a concept tree with its facets holding the important and relevant information desired about the subset. Note that the subsets of data described by the concepts can intersect in their coverage of the data base, and that new subsets are always created from existing subsets.

The concepts hold the relevant information needed about the data subsets. They represent the subsets and show their relationship to the other subsets. Each concept consists of a number of facets or characteristic knowledge about the subset. Typical facets include: NAME - the name used to refer to the subset; DEFN - the definition of the data contained within the subset; INTEREST - a numeric measure of how interesting the subset of data is; GENERALIZATIONS - pointers to more general (larger, containing) subsets; SPECIALIZATIONS - pointers to more restricted (smaller, contained) subsets.

### 3. SUMMARY OF RESULTS

The initial implementation of BROWSER was designed to test the feasibility of automatic browsing. It was implemented in MACLISP and ran on a KI DEC-10 computer, occupying a total of 400 pages of memory (this includes the data base management system). The data base consisted of one year of maintenance authorization records for six aircraft. For this initial implementation two of the data related models were implemented: the Data Base and Data Specific Models for the particular data files used.

The initial testing occupied 24.5 CPU hours. The data base management system used 54% of this time retrieving the data. Approximately one half of the remaining time was used examining the data returned. The examination suggested the creation of 224 data subsets. I found 45 of the 115 subsets actually created to be of interest. Of the subsets suggested, 162 were suggested because data values occurred at statistically significant rates. I judged one half (82) of these subsets to be worthwhile to create. Several interesting differences were found between the data subsets, differences too complex to summarize here [8].

While the initial implementation of BROWSER was limited in its scope, it has provided some important ideas and insights for developing more complex systems. It is hoped that in the near future such systems will be able to provide valuable information to users, information which currently is not available or is very expensive to obtain.

### References

- [1] Lenat, D.B., "AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search", SAIL-AIM-286, Stanford University, July 1976.
- [2] Lenat, D. B., "Automated Theory Formation in Mathematics", in Proc. IJCAI-77, MIT, Cambridge, Mass., Aug. 1977, pp. 833-842.
- [3] Davis, R. and J. King, "An Overview of Production Systems", STAN-CS-75-524, Computer Science Department, Stanford University, Oct. 1975.
- [4] Lenat, D. B. and J. McDermott, "Less than General Production System Architectures", in Proc. IJCAI-77, MIT, Cambridge, Mass., Aug. 1977, pp. 928-932.
- [5] Tennant, H., "The PLANES Database", Working Paper 14, Advanced Automation Group, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, June 1978.
- [6] Buneman, O. P. and H. L. Morgan, "Alerting in Database Systems: Concepts and Techniques", The Wharton School, University of Pennsylvania, Dec. 1975.
- [7] Cohen, S. F., Alerters in Network Databases", The Wharton School, University of Pennsylvania, Feb. 1977.
- [8] Dankel, D. D. II, "Browsing in Large Data Bases", Ph. D. Thesis, Department of Computer Science, University of Illinois at Urbana-Champaign, to be completed Fall 1979.