Palay, A.J., and Fox, M.S., (1981), "Browsing Through Databases", *Information Retrieval Research*, R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams (Eds.), pp. 310-324, Butterworth & Co. (Publishers) Ltd.

Browsing through databases

Andrew J. Palay and Mark S. Fox

20.1 Introduction

Significant progress has been made over the last decade in the design and construction of database systems that go beyond their specialised applications in providing multi-purpose access to data. Structures such as relational, hierarchic and network databases provide efficient and flexible structuring and storage of data. Concurrently, research has progressed in providing powerful languages for accessing these databases — for example, relational calculus and algebra, QBE (Zloof, 1977) and natural-language interfaces (Mylopolous *et al.*, 1975; Waltz, 1975; Sacerdoti, 1977; Hayes-Roth, Mostow and Fox, 1978).

It is important to note that the majority of the research in the area of database access has focused on the area of *parameterised search* (PS). PS can be characterised as strictly focused, in the sense that the user must specify exactly the set of attributes that the records must contain and/or satisfy. There are a number of problems associated with PS-based access to databases. In particular:

(1) The interface to the database system is often too complex for the untrained user. In order to properly specify the parameters of a search, the user must understand a highly stylised and idiosyncratic interface language. Work in the area of natural-language understanding has been directed towards solving this problem.

(2) The logical structure of the database is normally different from the user's view of the database. If the user does not understand the structure of a database, it is doubtful that he will be able to specify the proper set of parameters and values in order to retrieve the information he desires.

(3) The user is unable to become familiar with the organisational structure of the database. Records are inserted into the database according to some classification schema. Since the user only specifies parameters and never views the system as a whole unit, the user has little chance of learning the classification schema.

(4) There is no quick and easy way to access *related* records. Each access, whether related or independent, requires a new search specification.

The BROWSE system has been developed as an alternative to PS-based database access systems. The BROWSE system relies on *browsing* as the primary mechanism for information retrieval. Browsing can be characterised as a heuristic search in a well-connected space of records. Some of the heuristics most commonly used in browsing library databases are:

(1) *If* book X is interesting, *then* what else has the author of X written?

(2) *If* book X is interesting, *then* what other books are in the same category?

(3) *If* a symposium article is interesting, *then* what else appeared in the same symposium?

(4) *If* the author of an interesting paper is from an institution X, *then* what else has been published at that institution?

(5) *If* there is an interesting paper in a journal, *then* what else appeared in that journal?

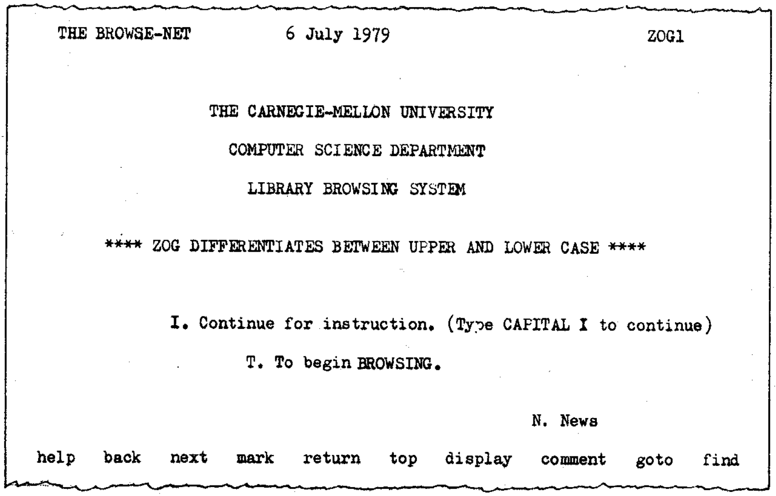
The goal of the BROWSE system is to provide *browsing* access to databases by building the search heuristics directly into the database as quick access paths between related records.

20.2 An example

A Browse system database consists of a set of frames. A frame is a single CRT (TV) screen of information. The purpose of the frame is to provide information to the user and to provide quick access to further related information. Options provide links to related frames. A user moves between the frames by selecting an option, which results in the system's displaying a new frame. The database can be viewed as a set of tokens, with the options providing the relations between the tokens.

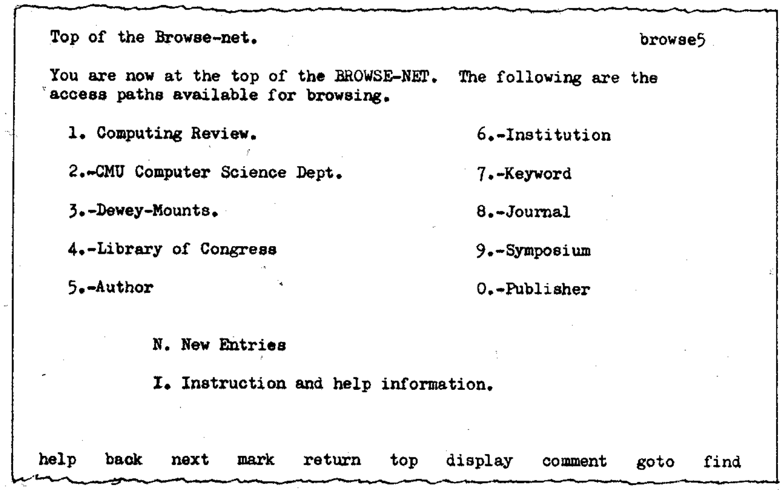
The following example is taken from a BROWSE database developed for an on-line library catalogue system.

The first frame (*Figure 20.1*) that is displayed welcomes the user to the BROWSE system. In the upper right-hand corner of every frame is the name of the frame (in this case ZOG1). At the bottom of the frame is a set of standard options ('help', 'back', 'next', . . ., 'find'). These options (called global pads) will appear in every frame. They provide a set of system functions that are useful throughout the entire network. The first frame consists of the text welcoming the user to the BROWSE system and three options. The first option allows the user to continue receiving instruction on how to use the BROWSE system. The second option allows the user to move directly to the top of the classification hierarchy. The third option allows the user to get news describing changes to the system. To select an option, the user types in the first letter of the option (an I to select the first option, a T for the second or an N for the third). If there is a pointing device (mouse, touch screen, etc.) available, then the user only needs to point to the option in order to select it. The naive user would continue by selecting the I option for more instruction. An experienced user would select the T option, thus displaying the top of the classification hierarchy (*Figure 20.2*). There are 12 options to this frame. Options that contain a dash (-) after the first two characters do not point to any other frames. They are included as pointers to information that one day will be included. An important property of the BROWSE system is that it can support



```
  THE BROWSE-NET            6 July 1979                        ZOG1


                  THE CARNEGIE-MELLON UNIVERSITY

                   COMPUTER SCIENCE DEPARTMENT

                     LIBRARY BROWSING SYSTEM


       **** ZOG DIFFERENTIATES BETWEEN UPPER AND LOWER CASE ****


             I. Continue for instruction. (Type CAPITAL I to continue)

                  T. To begin BROWSING.


                                           N. News

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.1

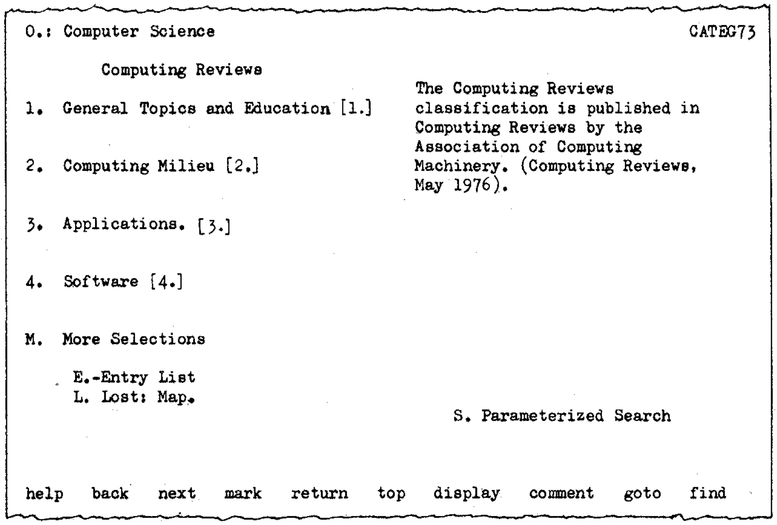


```
 Top of the Browse-net.                                    browse5

 You are now at the top of the BROWSE-NET.  The following are the
 access paths available for browsing.

   1. Computing Review.                       6.-Institution

   2.-CMU Computer Science Dept.              7.-Keyword

   3.-Dewey-Mounts.                           8.-Journal

   4.-Library of Congress                     9.-Symposium

   5.-Author                                  0.-Publisher


             N. New Entries

             I. Instruction and help information.


help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.2

multiple views of the database. At present only the *Computing Reviews* (1976) classification and a new entries list are available.

By selecting option 1, the user moves on to the top of the *Computing Reviews* classification hierarchy (*Figure 20.3*). The frames forming the classification hierarchy each contain a title (0.: Computer Science), a definition section, a list of sub-categories, an entry list option (E), a lost map option (L) and a parameterised search option (S). If there are additional sub-categories, then a more selections option is included (M). Finally, if there is a designated primary supra-category, then an option (P) is included that links the current classification frame to its primary supra-category. The entry list option points to a list of the entries that have been directly classified under the current category. In this case there are no such entries. Finally, the parameterised search option causes the system to begin a parameterised search and would proceed to a set of frames that allows the user to specify the parameters.



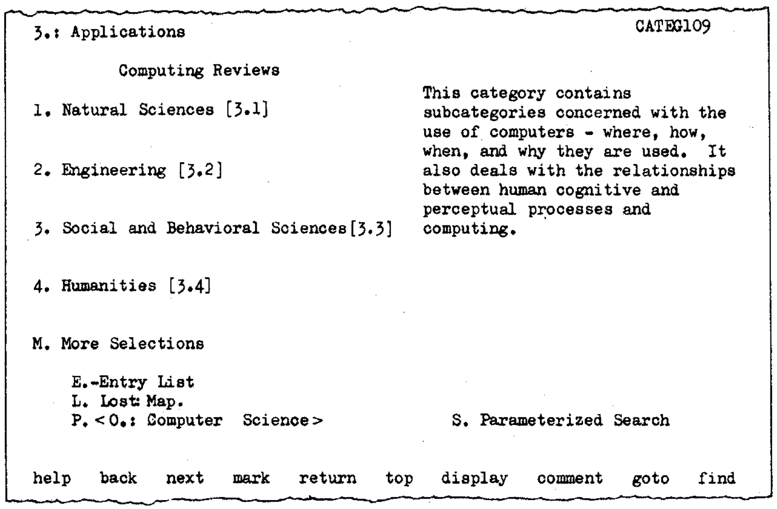
```
 0.: Computer Science                                          CATEG73

       Computing Reviews
                                              The Computing Reviews
 1.  General Topics and Education [1.]        classification is published in
                                              Computing Reviews by the
                                              Association of Computing
 2.  Computing Milieu [2.]                    Machinery. (Computing Reviews,
                                              May 1976).

 3.  Applications. [3.]


 4.  Software [4.]


 M.  More Selections

      E.-Entry List
      L. Lost: Map.
                                                S. Parameterized Search


 help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.3

The initial goal of this sample session is to browse through the database for information related to learning. By selecting the option for Applications, the user learns that the category includes cognitive processes (*Figure 20.4*). By selecting the more selections option (M), the user discovers that Artificial Intelligence is one of the sub-categories of Applications (*Figure 20.5*). If applications was not a useful category, the P option could be selected to go back up the hierarchy, to allow the user to select another search path.

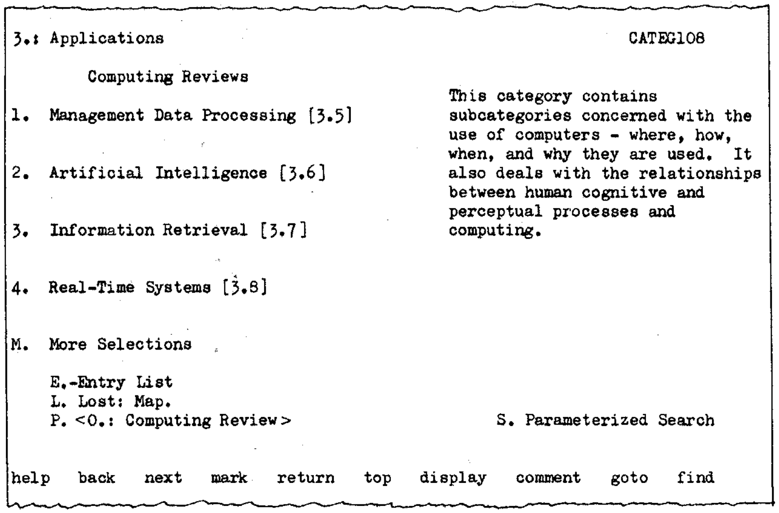
Now by selecting option 2, the user moves onto the frame describing artificial intelligence (*Figure 20.6*) and discovers that learning is one of the sub-categories (under Learning and Adaptive Systems). Note that an entry frame exists for this category. If the user would want to see entries that are directly classified under Artificial Intelligence, he would select option 'E'. With an interest in learning systems, the user now selects option 2 and the goal category frame has been reached (*Figure 20.7*).

There are no sub-categories to Learning and Adaptive Systems. The user has reached a terminal frame in the classification hierarchy. The entry list (*Figure 20.8*) is displayed by selecting the E option. Along with the list of entries, the entry list frame also contains a pointer back to the category frame that leads to the entry list (option R), and options to move back and forth through the entry list (options M and P, which is not shown). This is just one form of indexing provided by BROWSE. In addition, the autogeneration of hierarchic and alphabetical indices has been added.



```
3.: Applications                                                CATEG109

      Computing Reviews
                                           This category contains
1. Natural Sciences [3.1]                  subcategories concerned with the
                                           use of computers - where, how,
                                           when, and why they are used.  It
2. Engineering [3.2]                       also deals with the relationships
                                           between human cognitive and
                                           perceptual processes and
3. Social and Behavioral Sciences[3.3]     computing.


4. Humanities [3.4]


M. More Selections

   E.-Entry List
   L. Lost: Map.
   P. < O.: Computer  Science >              S. Parameterized Search


help  back  next  mark  return  top  display  comment  goto  find
```

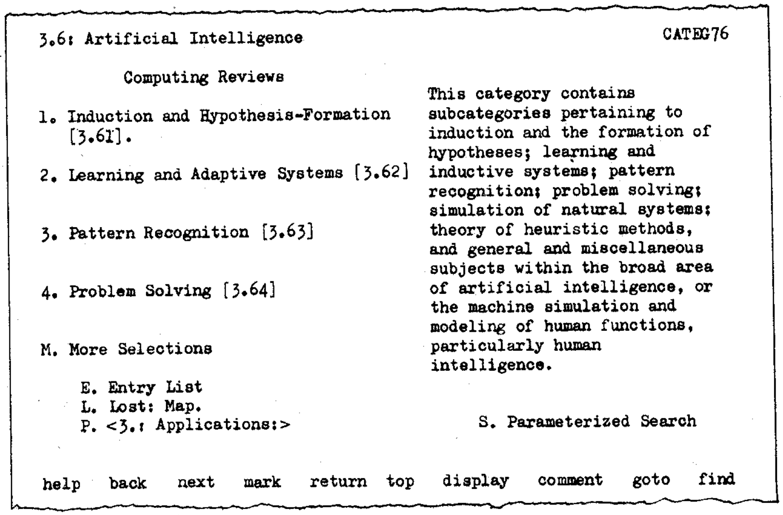
Figure 20.4



```
3.: Applications                                               CATEG108

      Computing Reviews
                                             This category contains
1.  Management Data Processing [3.5]         subcategories concerned with the
                                             use of computers - where, how,
                                             when, and why they are used.  It
2.  Artificial Intelligence [3.6]            also deals with the relationships
                                             between human cognitive and
                                             perceptual processes and
3.  Information Retrieval [3.7]              computing.


4.  Real-Time Systems [3.8]


M.  More Selections

    E.-Entry List
    L. Lost: Map.
    P. <O.: Computing Review>                   S. Parameterized Search

help  back  next  mark  return  top  display  comment  goto  find
```

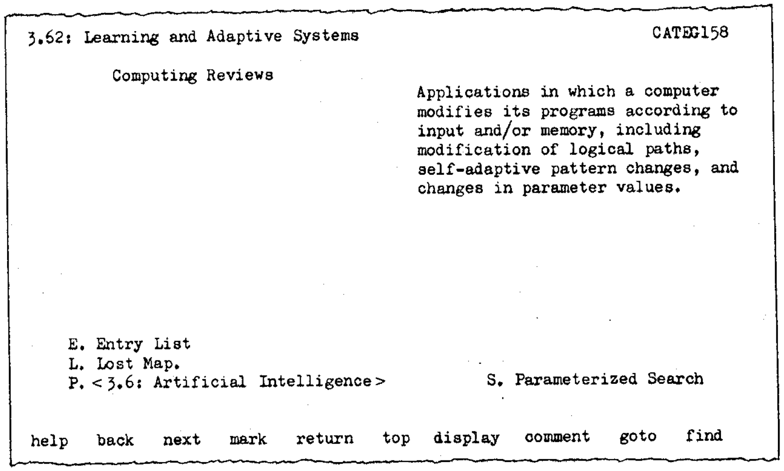
Figure 20.5

The user decides that entry 6 looks interesting. That option having been selected, the frame in *Figure 20.9* is displayed. This frame provides the basic information about the article in question. The user may find additional information about the authors by selecting options A or B. If the user wants information about the Computer Science Department at Carnegie-Mellon University he may do so by selecting option I. In all three cases a list of entries associated with the author or institution will be available. The user can gain additional information about the symposium in which this article appeared by selecting the option F. Along with information about the symposium, a list of all articles will be provided. Finally, there are a set of options that provide additional information about the current entry. Option 1 will provide a list of acronyms and keywords. The keywords are organised as a list of options. If the user wants to see what other entries in the database share a keyword, he can select that keyword. The second option provides the abstract to the paper. The fourth option leads to a list of categories under which the current article has



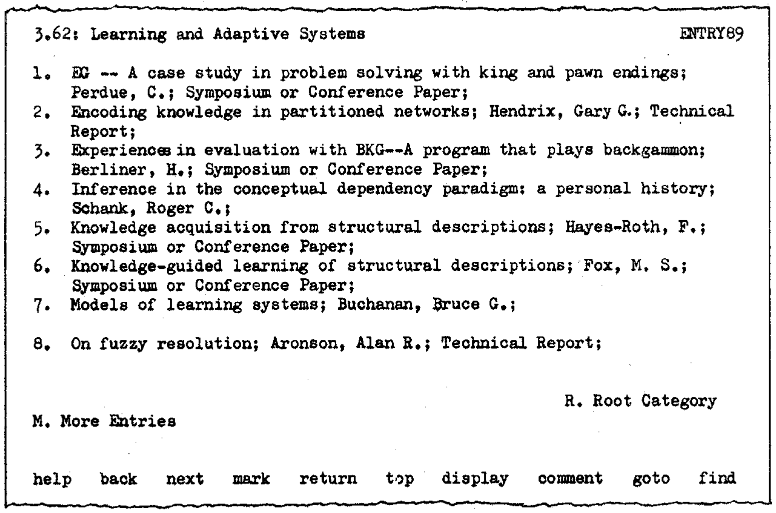
```
3.6: Artificial Intelligence                                    CATEG76

      Computing Reviews
                                             This category contains
1. Induction and Hypothesis-Formation        subcategories pertaining to
   [3.61].                                   induction and the formation of
                                             hypotheses; learning and
2. Learning and Adaptive Systems [3.62]      inductive systems; pattern
                                             recognition; problem solving;
                                             simulation of natural systems;
3. Pattern Recognition [3.63]                theory of heuristic methods,
                                             and general and miscellaneous
                                             subjects within the broad area
4. Problem Solving [3.64]                    of artificial intelligence, or
                                             the machine simulation and
                                             modeling of human functions,
M. More Selections                           particularly human
                                             intelligence.
   E. Entry List
   L. Lost: Map.
   P. <3.: Applications:>                       S. Parameterized Search

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.6



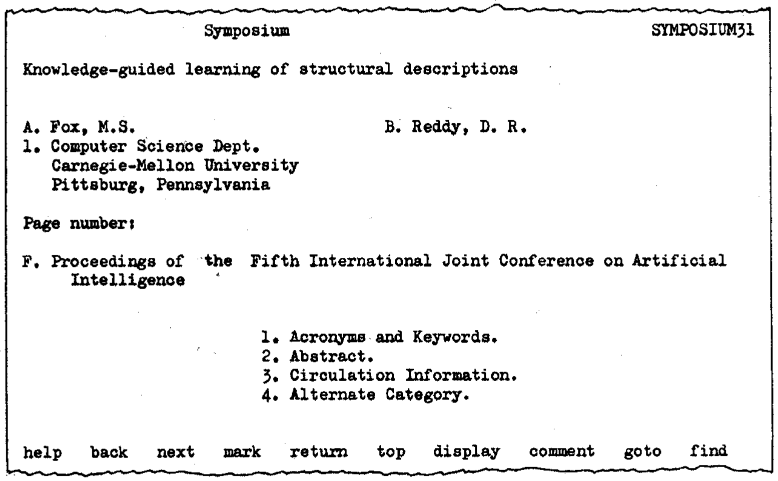
```
3.62: Learning and Adaptive Systems                            CATEG158

      Computing Reviews
                                             Applications in which a computer
                                             modifies its programs according to
                                             input and/or memory, including
                                             modification of logical paths,
                                             self-adaptive pattern changes, and
                                             changes in parameter values.







   E. Entry List
   L. Lost Map.
   P. < 3.6: Artificial Intelligence >         S. Parameterized Search

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.7



```
3.62: Learning and Adaptive Systems                              ENTRY89

1.  EG -- A case study in problem solving with king and pawn endings;
    Perdue, C.; Symposium or Conference Paper;
2.  Encoding knowledge in partitioned networks; Hendrix, Gary G.; Technical
    Report;
3.  Experiences in evaluation with BKG--A program that plays backgammon;
    Berliner, H.; Symposium or Conference Paper;
4.  Inference in the conceptual dependency paradigm: a personal history;
    Schank, Roger C.;
5.  Knowledge acquisition from structural descriptions; Hayes-Roth, F.;
    Symposium or Conference Paper;
6.  Knowledge-guided learning of structural descriptions; Fox, M. S.;
    Symposium or Conference Paper;
7.  Models of learning systems; Buchanan, Bruce G.;

8.  On fuzzy resolution; Aronson, Alan R.; Technical Report;


                                                       R. Root Category
M. More Entries


help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.8

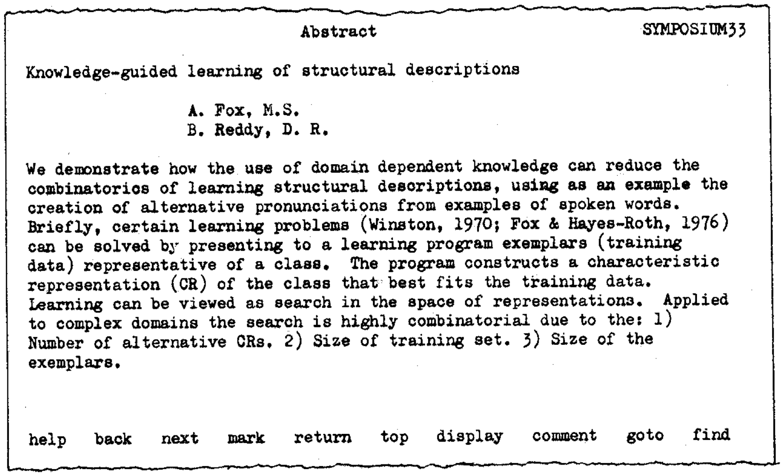


```
                      Symposium                               SYMPOSIUM31

Knowledge-guided learning of structural descriptions


A. Fox, M.S.                          B. Reddy, D. R.
1. Computer Science Dept.
   Carnegie-Mellon University
   Pittsburg, Pennsylvania

Page number:

F. Proceedings of  the  Fifth International Joint Conference on Artificial
     Intelligence

                              1. Acronyms and Keywords.
                              2. Abstract.
                              3. Circulation Information.
                              4. Alternate Category.

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.9

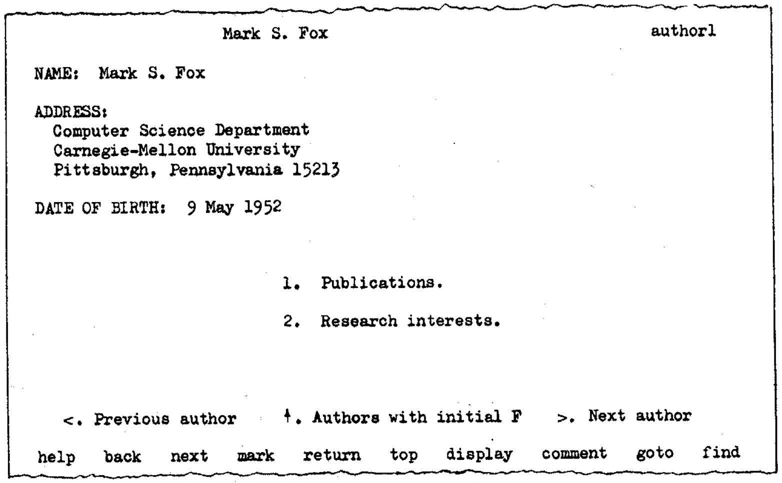
been classified. These categories are options that point back into the classification hierarchy.

By selecting option 2 the user can view the abstract of the paper (*Figure 20.10*). The user decides that he wants to see additional information on M. S. Fox. By selecting option A in *Figure 20.10*, he moves to a frame giving some information about the author (*Figure 20.11*). Options are provided for linearly moving through the author list (<, >), and going to the author index frame (↑). Selecting option 1, the user is led to a frame listing of all M. S. Fox articles that



```
                              Abstract                        SYMPOSIUM33

Knowledge-guided learning of structural descriptions

                 A. Fox, M.S.
                 B. Reddy, D. R.

We demonstrate how the use of domain dependent knowledge can reduce the
combinatorics of learning structural descriptions, using as an example the
creation of alternative pronunciations from examples of spoken words.
Briefly, certain learning problems (Winston, 1970; Fox & Hayes-Roth, 1976)
can be solved by presenting to a learning program exemplars (training
data) representative of a class.  The program constructs a characteristic
representation (CR) of the class that best fits the training data.
Learning can be viewed as search in the space of representations.  Applied
to complex domains the search is highly combinatorial due to the: 1)
Number of alternative CRs. 2) Size of training set. 3) Size of the
exemplars.


help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.10

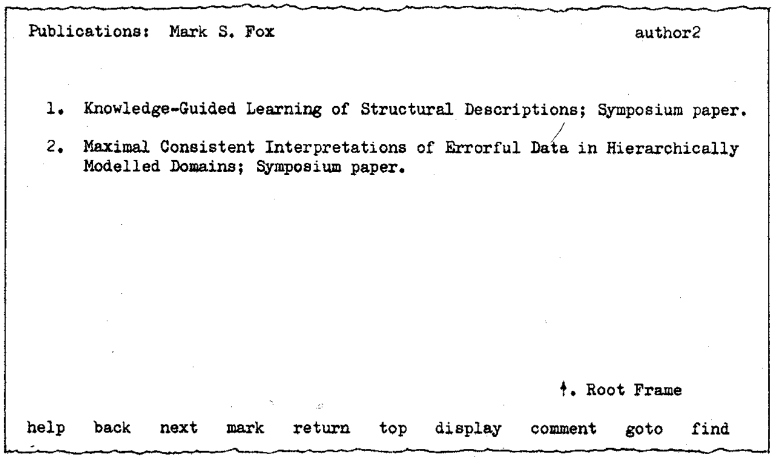


```
                     Mark S. Fox                               author1

NAME:  Mark S. Fox

ADDRESS:
  Computer Science Department
  Carnegie-Mellon University
  Pittsburgh, Pennsylvania 15213

DATE OF BIRTH:  9 May 1952


                           1.  Publications.

                           2.  Research interests.


  <. Previous author    ↑. Authors with initial F    >. Next author

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.11

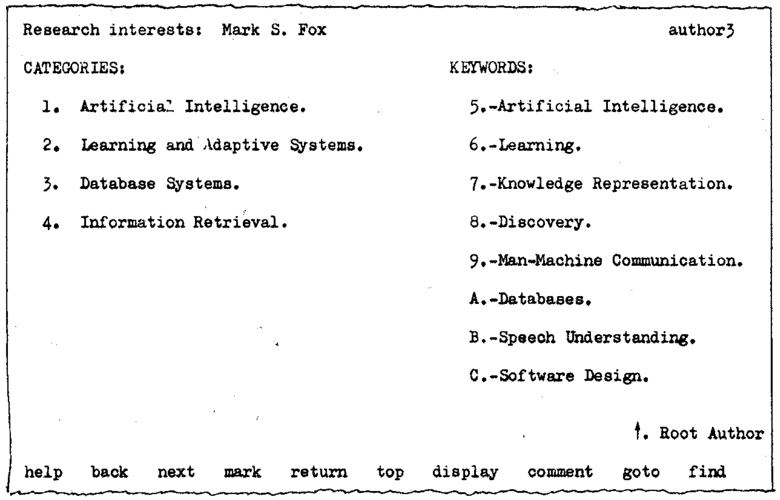
are currently in the database (*Figure 20.12*). Seeing nothing of additional interest the user now decides to see M. S. Fox's other areas of interest (*Figure 20.13*).

Also interested in Information Retrieval, the user selects option 4. This puts him back into the classification hierarchy (*Figure 20.14*). He may now continue browsing in that area. If the user has managed to get lost, there is the lost map option (L). By selecting that option the user can get a global view of the classification hierarchy that surrounds the current category (*Figure 20.15*). Surrounding categories can be reached by selecting any of the options.



```
Publications: Mark S. Fox                                 author2


  1. Knowledge-Guided Learning of Structural Descriptions; Symposium paper.

  2. Maximal Consistent Interpretations of Errorful Data in Hierarchically
     Modelled Domains; Symposium paper.









                                                  ↑. Root Frame

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.12



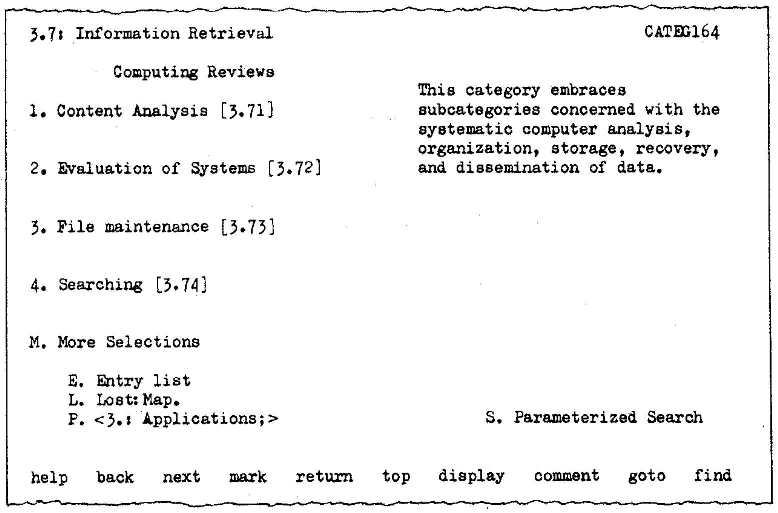
```
Research interests: Mark S. Fox                               author3

CATEGORIES:                            KEYWORDS:

  1. Artificial Intelligence.            5.-Artificial Intelligence.

  2. Learning and Adaptive Systems.      6.-Learning.

  3. Database Systems.                   7.-Knowledge Representation.

  4. Information Retrieval.              8.-Discovery.

                                         9.-Man-Machine Communication.

                                         A.-Databases.

                                         B.-Speech Understanding.

                                         C.-Software Design.

                                                        ↑. Root Author

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.13

20.3 Browsing and searching

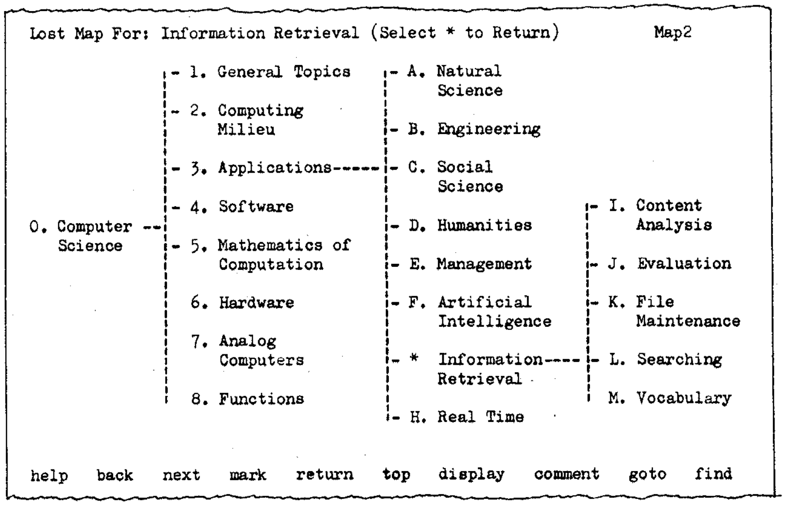
The BROWSE system relies on browsing as its primary method of database access. However, there are many times when searching is desired. The user may already know exactly for what he is looking and should not have to move through the network of frames to get there.

The BROWSE system includes the ability to specify searches. As was seen in the example (*Figure 20.2*), each category frame in the current system has an option that allows the user to specify a parameterised search. The parameterised search differs from searches in normal PS-based systems in the following:



```
3.7: Information Retrieval                                    CATEG164

      Computing Reviews
                                      This category embraces
1. Content Analysis [3.71]            subcategories concerned with the
                                      systematic computer analysis,
                                      organization, storage, recovery,
2. Evaluation of Systems [3.72]       and dissemination of data.


3. File maintenance [3.73]


4. Searching [3.74]


M. More Selections

   E. Entry list
   L. Lost: Map.
   P. <3.: Applications;>                   S. Parameterized Search

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.14



```
Lost Map For: Information Retrieval (Select * to Return)        Map2

                |- 1. General Topics   |- A. Natural
                |                      |     Science
                |- 2. Computing        |
                |     Milieu           |- B. Engineering
                |                      |
                |- 3. Applications-----|- C. Social
                |                      |     Science
                |- 4. Software         |                     |- I. Content
0. Computer   --|                      |- D. Humanities      |      Analysis
   Science      |- 5. Mathematics of   |                     |
                |     Computation      |- E. Management      |- J. Evaluation
                |                      |                     |
                |  6. Hardware         |- F. Artificial      |- K. File
                |                      |      Intelligence   |      Maintenance
                |  7. Analog           |                     |
                |     Computers        |- *  Information-----|- L. Searching
                |                      |     Retrieval       |
                |  8. Functions        |                     |  M. Vocabulary
                                       |- H. Real Time

help  back  next  mark  return  top  display  comment  goto  find
```

Figure 20.15

(1) The search that is initiated is context dependent. The user may browse through the classification hierarchy until he finds the area in which he is interested. When he initiates a search at that point, only entries that are

classified under that category will be searched. Thus, if the user selected the parameterised search option at the frame for Artificial Intelligence, the search procedure would only search entries directly classified under Artificial Intelligence and entries classified under all of the sub-categories to Artificial Intelligence.

(2) When the results of the search are returned to the user, he is able to browse through the list of satisfied entries. He may also browse outside the list of satisfied entries. The current search procedure returns a list of frames in the system that satisfy the search parameters. An additional set of options are provided that allow the user to move through the list. The list can be viewed as an additional classification structure. Thus, the user can make full use of the browsing capacity of the system.

The ability to combine browsing and searching, it is hoped, reduces the problem with parameterised searches of not finding all the entries in which the user is interested. Often closely related entries are not returned by the search. With the user being allowed to browse throughout the network, he can often find the entries related to the list of returned entries.

20.3.1 Future directions in searching

At the present time the search procedure used by the BROWSE system is independent of the rest of the system. The search is initiated and returns a simple list of entries. In the future an integrated version of searching will be included in the system. Instead of producing a list of entries, the new search procedure will be used to restrict the view of the database. For example, if the user is again at Artificial Intelligence, and he decides he only wants to see information related to Games, then by specifying that parameter the user will only be shown entries in the system that do relate to Games. For example, in *Figure 20.8* only options 1 and 3 would be included in the entries list. Instead of actually doing the search, the parameters would be used to remove, from the view of the user, all unrelated entries. At any time the user may refine or expand the scope of his search.

20.4 Building BROWSE systems

The BROWSE system is composed of two separate systems. There is a display system, called ZOG, and a system that is used to create the frame-structured database, called BROWSE.

20.4.1 The ZOG system

The BROWSE system was designed to use the ZOG system, developed at Carnegie-Mellon University by Robertson, McCracken and Newell (1979), as its display system. ZOG has its roots in the University of Vermont's PROMIS system (Schutz, Cantrill and Morgan, 1971). To understand the philosophy of BROWSE, one must understand the philosophy of ZOG.

ZOG is a rapid-response, large-network, menu-selection system used for man–machine communication. A ZOG user sits in front of a terminal on which a frame is displayed. The frame consists of some text and a set of options. At the discretion of the user an option is selected and almost 'instantaneously' a new frame is displayed. The process then starts again.

ZOG was designed on the basis of a philosophy developed at the PROMIS Laboratory of the University of Vermont. ZOG's basic features are:

(1) *Rapid response.* When a user selects an option that leads to another frame, the next frame should appear fast enough so that the user does not feel that he is waiting for the system. A user must feel free to explore surrounding frames without concerning himself with the time it takes to display each frame.

(2) *Simple selecting.* The act of selection should be a simple unitary gesture. At the present time there are two forms of selection available. These are (1) single-character input from the keyboard and (2) a touch screen.

(3) *Large network.* The network should be large enough to provide most of the information needed by the user.

(4) *Frame simplicity.* The frame display should be kept simple. The user should be able quickly to assimilate the information contained in the frame. The idea of frame simplicity has led to the development of frames that contain a small amount of text and up to five or six options. This is not necessarily true of frames developed for the BROWSE system, where the simplicity comes from the very structured nature of the information. Although a large amount of information may be displayed on a single frame, the information desired by the user can still be assimilated quickly.

(5) *Transparency.* The user should be able to understand exactly what the system is doing and what he needs to do to gain additional information. At no point should the user feel that he has lost control of the system.

(6) *Communication agent.* ZOG has been designed to act as a communication agent between a user and another system. As a communication agent, ZOG presents commands to the user in a simple format, as well as an explanation of what the command will do. When the user makes a selection, ZOG sends the more complex set of commands to the other system for processing. This facility is used by the BROWSE system for the parameterised search interface.

(7) *External definition.* Unlike many menu-selection systems, ZOG nets are databases which exist independent of and external to the ZOG system. Frames can be defined external to and/or within the ZOG system.

The basic philosophy of ZOG is that a menu-selection system can be an effective communication system if the user can move around in the system quickly and if there is a large network available to meet the user's needs.

A menu-selection system allows the user to have almost complete knowledge of what is occurring in the system whenever he selects an option. It also allows for related information to be located nearby (by placing a link between related frames). Menu-selection systems normally have a disadvantage in the time it takes to move from one frame to another. This is solved by the fast response of the ZOG system. Another characteristic of many menu-selection systems is that the same information is provided to all users, regardless of their needs. An expert user is presented with the same information as the naive user. In ZOG different paths would be provided for each level of user. The naive user would get more information about what he is doing, while the expert user would be presented with just the frames needed to perform the task.

20.4.2 The BROWSE system software

A major problem with menu-based systems is the creation of the menus. Experiments with the ZOG system have shown that the average rate of frame creation is approximately 5–30 frames per hour, depending on the task (McCracken and Robertson, 1979). Given a database that contains in excess of 1 million frames, it becomes clear that some machine-aided creation mechanism is needed.

The BROWSE system solution to that problem was to create a software package external to the display system (ZOG), to create the database network. The software is designed to take a standard database and translate it into a frame-structured database. The reasons for using an external database and a translation system are:

(1) To allow for the modification of the frame formats. The design of frame formats and network structures may change, which would require the frame database to be re-created. ZOG does not provide facilities for the automatic modification of frames and structure.

(2) To allow for the creation of different networks for different display types. At the present time the BROWSE system runs only on a standard 24 by 80 character display terminal. A completely different frame structure would be used for a high-resolution display terminal.

(3) To allow for parameterised searches. It would be cumbersome to carry out the parameterised search within the frame network. Thus, the search is carried out by a separate program, which accesses the original database.

The BROWSE system software maintains a separation between the information to be displayed and the form in which it is to be displayed. The actual information is held in the external relational database. A separate frame-format database (FFD) contains the information on how to display the information. The FFD contains frame formats for each type of entry in the system. For example, for a library system a separate frame format would exist for a book, technical report, subject classification, etc. The frame format describes:

(1) The information to be placed in the frame.
(2) The placement of the information.
(3) The information necessary to actually create the frame.
(4) The options to appear in the frame.
(5) The type of frame to which each option may lead.

The frame format consists of (listed in increasing order of complexity): (1) window descriptions, (2) option descriptions, (3) fill descriptions, (4) group descriptions and (5) index descriptions.

The window description (WD) is the basic unit of the frame format. The WD includes information describing a two-dimensional area into which text will be placed. The WD contains a starting co-ordinate for the window. This position is relative to the group that accesses the window. The WD also contains a minimum and a maximum length and width of the window.

The option description (OD) contains information that relates to options in the ZOG frame. Included in that information is the selection character for the option, the touch area of the option (in case a pointing device is available), text to be included as part of the option text, the frame to which the option leads and a WD that is to be used to hold the text of the option.

The fill description (FD) describes the information from the relational database that is to be used to fill a window. The FD also points to a WD. In evaluating an FD, the system retrieves the contents of the fields in the current record. The FD contains information on what to do if the contents of a field are null, as well as the text to place before and after the text found. All the text specified is concatenated together and is then placed in the window. Also included in the FD is information describing the size, type and typeface of font to be used in displaying the text (for use with a high-resolution display), as well as commands describing how to display the text in the window (centred, flushed right, flushed left, filled).

The group description (GD) is one of the two major units of the frame format. The GD points to a set of FDs. If the GD is describing an option, then the GD will also point to an OD. The GD also contains an absolute anchor point for the group. An anchor defines the root position of the group on the frame. The starting positions provided by the WDs indirectly referenced by the GD are relative to that anchor point. The anchor point of the group can be set relative to other groups (for example, place the anchor point of group X one line below the last line used by group Y).

The other major unit of the frame format is the index description (ID). Associated with every category frame is an option to a list of entries. That list is a linear index, ordered alphabetically by title, to all the entries in that category. One of the major problems with an evolving system is that the indexes to the system must constantly be rebuilt. The ID has been included in the frame format to allow the system mechanically to create indexes. The system is capable of creating a variety of index types (linear, hierarchical, alphabetical). The ID points to two sets of GDs. The first set refers to groups that are used as titles to the index. The second set is used for actually creating the index. The ID also contains information on how to form the list of entries to be used in creating the index.

20.5 Conclusion

With the proliferation of Viewdata-type systems (Fedida, 1976) in the next few years, public access to databases will be greatly increased. With a large influx of computer-naive users, new methods for database access must be provided. The BROWSE system represents such a method. It is hoped that building into the database system a set of browsing heuristics will enable the naive user easily to locate the information he desires.

Acknowledgements

This research was supported by the Defense Advanced Research Projects Agency (DOD), Arpa Order No. 3597, monitored by the Air Force Avionics Laboratory under contract F33615-78-C-1551 and the Office of Naval Research under contract N00014-76-0874.

We wish to thank the people involved in the ZOG project for providing the important display component of the BROWSE system. Also, our special thanks to Earl Mounts, head of the Computer Science Browsing Room at Carnegie-Mellon University, for his help in developing the BROWSE system.

References

Computing Reviews (1976). 'Categories of the computer sciences', *Computing Reviews*, May, 172–198

FEDIDA, S. (1976). 'Viewdata — an interactive information medium for the general public using the telephone network', International Broadcasting Convention, *IEE Conference Publication*, No. 145

HAYES-ROTH, F., MOSTOW, D. J. and FOX, M. S. (1978). 'Understanding speech in the Hearsay-II system', in BOLC, L. (Ed.), *Speech Communications with Computers*, Springer-Verlag, Berlin

McCRACKEN, D. and ROBERTSON, G. (1979). 'Editing tools for ZOG, a highly interactive man–machine interface', in *Proceedings of the 1979 International Conference on Communications, IEEE*

MYLOPOLOUS, J., BORGIDA, A., COHEN, P., ROUSSOPOULOS, N., TSOTSOS, J. and WONG, H. (1975). 'TORUS: natural language understanding system for data management', *Proceedings of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, USSR*, pp. 414–421

ROBERTSON, G., McCRACKEN, D. and NEWELL, A. (1979). 'The ZOG approach to man–machine communication', Computer Science Department Technical Report, Carnegie-Mellon University

SACERDOTI, E. D. (1977). 'Language access to distributed data with error recovery', *Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Mass.*

SCHULTZ, J., CANTRILL, S. and MORGAN, K. (1971). 'An initial operational problem oriented medical record system — for storage, manipulation and retrieval of medical data', *AFIPS Proceedings*, **38**, 765–777

WALTZ, D. (1975). 'Natural language access to large data bases: an engineering approach', *Proceedings of the Fourth International Joint Conference on Artificial Intelligence, Tbilisi, USSR*, pp. 868–872

ZLOOF, M. M. (1977). 'Query-by-example: a data base language', *IBM Systems Journal*, No. 4

21

A probabilistic algorithm for nearest neighbour searching

Stephen F. Weiss

21.1 Introduction

In this chapter we examine a probabilistic approach to information retrieval. We assume that documents and queries are represented as binary vectors. The value in a particular position of a vector indicates the presence (1) or absence (0) of the concept associated with that position. For a given query q, documents are ranked for retrieval on the basis of a similarity or distance measure calculated from the document and query vectors*.

This retrieval model is really just a special case of the nearest-neighbour problem. That is, given a set of N points in n space, and a distinguished point q, find the m points that lie nearest q according to some distance measure. In our retrieval model the documents are the points, the query is q, n is the total number of concepts in the system and 'nearest' is synonymous with 'greatest similarity'. In the discussion that follows we shall concentrate on the case where $m = 1$. Cases where $m > 1$ are straightforward generalisations.

The standard way of doing a nearest-neighbour search is to examine all the documents, calculate the similarity measure for each and then select the m best. This requires $O(N)$ time for a collection of N documents, which may be prohibitively expensive and time-consuming for large N, especially when interactive response is required. The optimal nearest-neighbour algorithm (Friedman, Bentley and Finkel, 1977) requires only $O(\log N)$ time but is unusable if the dimensionality of the space is high. Specifically, the optimal algorithm has a multiplicative constant of approximately 1.6^n, where n is the dimension of the space. Information retrieval systems typically have hundreds or even thousands of concepts, and in such situations $1.6^n \log N$ is much larger than N, even when N is very large.

In this chapter we begin by presenting briefly a deterministic nearest-neighbour search algorithm that is faster than the $O(N)$ search and achieves the same results as a full search, but is not crippled in high-dimensional spaces. We then present a modification to the basic algorithm that allows the user to specify a maximum tolerable level of error (which may be zero). This tolerance

* We shall assume throughout this chapter that the similarity measure used has range [0,1], where 1 indicates maximal similarity and 0 indicates minimal similarity.