User Interface Design Methodologies for CD-ROM Information Retrieval or "How to Find that Needle in a Haystack"

Greg Ferguson
Hewlett-Packard
Application Support Division
Mountain View, California

"There may be millions of fine thoughts, all encased within stone walls of acceptable architectural form; but if the scholar can get at only one a week by diligent search, his syntheses are not likely to keep up with the current scene."

Vannevar Bush
"As We May Think"
The Atlantic Monthly; July 1945

The preceeding quote describes a phenomenon that exists today, even with the vast amount of intelligence our society possesses. Computer networks, on-line databases, and satellite link-ups are all technologies that are being used to bring information to us on a daily basis. This information influx permits us to formulate decisions in a timely fashion. As we obtain and assimilate more data, our ability to make superior decisions increases, as well as our productivity. Yet, much of our time is still spent diligently searching for the right information.

Now envision 250,000 pages of information on a 120mm disc that can be accessed by a personal computer. CD-ROM is the product of that vision. The dilemma we are constantly struggling with is the transfer of knowledge (or lack thereof); part of the answer is CD-ROM.

What is CD-ROM?

CD-ROM stands for Compact Disc Read Only Memory. It is an optical disc having a diameter of 120mm made from the same material as bullet-proof glass. Information is stored on a CD-ROM in one continous spiral track that is approximately three miles long. The storage capacity of a CD-ROM is greater than 600 megabytes, which translates to 250,000 typed pages, or greater than two million sectors on a HP3000. The CD-ROM drive is connected to a personal computer, making it a neatly packaged solution. The cost factor for large volume pressing of CD-ROMs is also enticing. The latest quotes from mastering plants are anywhere from three to ten American dollars per disc. It should be noted, however, that the first disc, and the data preparation phase, is where most of the costs are incurred.

The CD-ROM disc cannot be written to, making it an ideal media for the collection and distribution of static information. It has a well-defined file structure, formerly labeled High Sierra, that has now become an international standard (ISO 9660). Codes used for the detection and correction of possible errors in the data are supplied with each block of data (2K bytes) that is written to the disc. These error codes are so efficient that only one error will escape uncorrected in 10 quadrillion bits (1 followed by 16 zeroes). Compact Discs have the capability for storing not only audio and text; but graphics, color photographs, full-motion video, animation, or any combination of the aforementioned.

Consider the Shortcomings...

How do we access all that data? There are certain unique characteristics about the hardware and the media that software developers must consider when outlining a strategy for the retrieval of information from a CD-ROM disc.

The first problem is in the physical speed of the drive. The CD-ROM drive rotates at a relatively slow speed when compared to most magnetic media of today, such as a Winchester hard disc. The drive operates at fluctuating speeds depending on where the laser is on the surface of the CD-ROM disc (faster on the inner tracks). This is known as constant linear velocity, or CLV. Most magnetic media of today operate at an unvarying speed or constant angular velocity, CAV. The CLV format is ideal for the storage of large amounts of data, but poor for the retrieval of individual blocks of data. The head movement of the laser must be accompanied by the mechanical process of speeding up or slowing down the disc. This movement accounts for high seek and latency times. Transfer rates of data coming off the disc fall between the rate of a high-speed floppy disc drive and a hard disc drive. The developer must consider these facts when creating and positioning files on the disc and strive to counterbalance the poor seek and latency times of the hardware.

One benefit of the CD-ROM could also pose a problem for some developers when dealing with more volatile types of information: the disc is read-only. Formulating an update scheme could involve writing the information provider's/user's input to another media (eg. hard disc) and displaying the most current information at retrieval time. Where the information comes from is totally transparent to the user. Response times will be visibly slower when multiple media-types are read. An alternative could involve distributing a new CD-ROM. The less static the information is, the more frequently new discs may have to be generated. Webster's Dictionary does not have to be updated as frequently as the market price of Hewlett-Packard stock!

Other Concerns

The types of data that the user will be accessing will play a major part in determining the design of your CD-ROM retrieval system. Numeric data will require certain options within the retrieval software that may not need to be implemented in an all text-based system. Graphics, audio, and video will all require certain elements be built into the interface for their retrieval and use.

Security can be a major concern. Due to the abundance of information that can be put on a disc, certain portions may not be accessible to some users. If that is the case, you may wish to encrypt the data on the disc as well as build stipulations for access into the retrieval software.

Graphics and video often require that special hardware boards be installed into the PC to decompress and display the information. A special monitor may be necessary as well. Can your user base fund this expense? If not, the feasibility of having data of that nature on your CD-ROM will be reduced.

The Current Alternatives

Upon deciding that CD-ROM is the answer to your needs, there are a few alternatives to compare in terms of how to construct the user interface.

You may wish to develop the entire interface on your own. This will take a great deal of time and money. Often, that "window of opportunity" of getting a needed product out to your users will pass by while the software is still in the development phase.

Another appoach would be to have an interface built for you by a company that specializes in custom software. This can be extremely expensive.

There are many companies specializing in CD-ROM that offer a full range of services to their customers. This covers everything from the preparation of the data, to a quality assurance check of the CD-ROMs that are returned from a mastering facility. They also offer a variety of user interfaces from which to choose. The intent is to pick one that fits your needs and the nature of your information. These "stock" interfaces can usually be customized to some degree. The issue with service organizations is that generally you receive something that does not quite meet your original expectations, although a product can be delivered within a relatively short period of time.

Finally, a few companies offer a set of software libraries that can be used for the development of a user interface. This concept is often referred to as a toolkit. The underlying search and retrieval software is provided, and the developer is given the ability to construct an interface on top of this basic "search engine". This provides the developer with full customization of the portion that the user accesses, without the total development time that would be required to build an entire system. Royalty fees are often required to license the company's software.

Features of Retrieval Software Packages

All CD-ROM retrieval packages offer the user and developer a wide genre of features from which to choose. This section will outline and briefly explain many of those features.

Keyword Search - Most CD-ROM retrieval packages offer an inverted file management system. In this system, all significant words within textual data are indexed. The concept of the keyword search allows the user to type in a string of keywords, with full boolean constructs (AND's, OR's, NOT's and the use of parantheses), and retrieve a search result set. This set contains pointers to the data that indicate which documents were found that contain keywords and where within each document the keywords were found. A document can be described as the lowest "record" of data found within a CD-ROM database. This is generally one to five screens of textual data.

Expansion or "wildcard" - This feature permits the user to retrieve all indexed terms that begin with a particular base. An example of this would be "COMPUT*", which would retrieve occurrences of such terms as "COMPUTER", "COMPUTATION", etc.

Phrase and Proximity - Use of the proximity operator lets the user limit the area in which keywords are searched for. For example, [word1 word2] indicates that word1 must be found within a supplied distance of word2. This distance is configurable and can be specified in terms of characters or words. The phrase search is specified when a user wishes to look for an exact match on a series of terms within the indexed text.

Thesaurus - Often a thesaurus can be invoked by the user that contains a list of synonym terms that have been specified by the information provider and/or the user. It may or may not be modifiable. By invoking the thesaurus, the user indicates to the search engine to retrieve all synonym terms for the supplied keywords and construct an "OR" condition using those terms. An example of this feature would be to have the user type in "DB", and via the thesaurus, the search engine would return a search result set containing references to the terms "DB", "DATA BASE", "DATABASES", etc.

Topic search - Instead of searching for general keywords, pre-constructed search result sets may be accessed by the user which provide a quick pathway to information relating to a specific topic. The topic "COBOL SORT ROUTINE" may contain pointers to documents such as application notes, software status bulletins, and manuals that all contain information relative to that topic.

Range search - This is a critical feature for designing an interface that will be accessing numeric data. This allows the user to specify a mathematical range when searching for specific data.

Field search - The user may specify certain keywords that will be searched on only within a pre-defined field in the database. An example of such a feature would be "NAME=BILL", where the search engine would look in the pre-defined database field "NAME" for all occurrences of the term "BILL".

Subsearch or "sideways search" - A subsearch is allowing the user to search across a previously retrieved search result set. This lets the user further define the search to work with a manageable set of documents. Searching on "FILE" may retrieve hundreds of documents. Performing a subsearch on that search result, using the keyword "ERROR", will limit the amount of documents that the user has to examine, making a more precise search result set. A "sideways search" involves using a term from a document that has been found via a previous search. Using our previous example, the user may be reading through a document relating to "FILE ERROR CODES", see a reference to "SECURITY", and perform a "sideways" search using "SECURITY" as the search term. A new search result set will be constructed for the user to examine.

Relevance - There are some search engines that will rank your retrieved results based upon certain algorithms; such as the number of times specified keyword(s) occurred within each document. A weighting factor is assigned to each document, and the search result set is sorted according to this factor, prior to being displayed to the user.

Search history - This feature permits the user to save a query string with all related information about the search (ie. number of documents retrieved, date, etc.). The user can employ the search log to re-construct a previous query into the database for fast retrieval to information.

Browsing - Browsing allows the user to see the hierarchy of information within a database and follow a path to specific knowledge. Paths may be pre-defined by the information provider, established by the user, or a combination of both. Browsing can be presented in different ways. The discussion on hypermedia presented in this paper talks about some of those ways. Often times a "topic explosion" approach is taken. The user is presented with a set of themes (ie. magazine articles). By selecting a theme, the user may now be presented with the specific text relating to that theme (ie. the article itself).

Book marks - Through this feature, the user may establish a pointer directly to information that is of specific interest. Use of this tool gives the user faster access when needing to reference particular information again.

Annotation - This will let the user customize the information with certain facts that make the data more relevant. These personalized facts may be referenced upon viewing the annotated document via a keyword search or by other means.

The LaserROM Strategy

LaserROM was developed to aid customers in their use of existing Hewlett-Packard documentation. This tool uses CD-ROM technology as it's "delivery vehicle", taking advantage of the enormous storage capacity of the CD-ROM. The use of a MicroSoft (C) Windows-based graphical interface in conjunction with CD-ROM makes LaserROM an attractive, integrated, PC-based solution. Monthly updates provide customers with current information that is immediately at their disposal within the framework of the PC workstation environment. Full keyword indexing and hierarchically structured browsing supply the user with direct pathways to the primary feature of LaserROM: the information (reference manuals, product catalogs, known problem reports, etc.). These pathways give LaserROM an added value above the printed copy. Additional benefits are described in the following sections. LaserROM is an entirely new way to look at documentation.

Why Windows?

MicroSoft (C) Windows was chosen as the primary operating enviroment for LaserROM for a variety of reasons. Windows provides the user with a heuristic, easy-to-use interface. The Hewlett-Packard commitment to this environment has been conveyed through the introduction of New Wave. The PC industry as a whole is now committed with the announcement of the next operating system, Presentation Manager, which is being developed under Windows. Windows is fast becoming a standard amongst the users of personal computers.

The New Wave Environment

The LaserROM development team was able to work with the the New Wave Human Factors engineering group within Hewlett-Packard, to help make the LaserROM interface conform to guidelines that had been established for development within the New Wave and Windows environments. This effectively gives LaserROM the same "look and feel" as other Hewlett-Packard applications developed within this environment. This approach benefits customers, allowing them to spend more time using the product and less time in the training phase.

The New Wave environment also gives LaserROM an additional advantage above the alternatives we previously discussed. New Wave permits the user to retrieve valuable information without leaving their normal working environment. LaserROM becomes part of a larger, totally-integrated, object-oriented office environment and not just a stand-alone application. The New Wave architecture will be compatible within the DOS, OS/2, and UNIX environments.

In summation, four key advantages were derived through the use of New Wave and Windows in developing the LaserROM user interface:

1) LaserROM was developed under an industry standard.
2) LaserROM is heuristic and easy to use, with a common Hewlett-Packard New Wave user interface.
3) The information is readily and quickly accessible.
4) It allows the user to retrieve information within the framework of their normal operating environment from a well-integrated workstation.

The Features

A key strategic point in the development of LaserROM was to concentrate on the content of the information. When a user is given tools, such as keyword searching, which allow for direct access immediately to relevant data, errors within the data become much more visible. For that reason, an extensive quality check was done to the data that was placed on the CD-ROM. Another primary design strategy was to give the user additional capabilities only if they were easy to use, not to implement features based on a

technology-driven focus. Additional features will be added once they are deemed necessary from a customer standpoint, provided they fit into the overall LaserROM development strategy.

The main feature of LaserROM is its keyword search capability. The utilization of full boolean constructs, and a thesaurus created by the information provider, supplies the user with the ability to "zoom in" on data of high points of interest. Proximity searching, phrase searching, and the use of an expansion operator are all supported by the software.

Another way to access information is through the LaserROM browse facility. Browsing allows the user to look at how the information is layered through a "structure explosion" capability. For example, the user selects a database and is then presented with a list of all possible manuals within that database. Selecting a manual then generates an "electronic table of contents", and so on. At any time, the user may reverse the process by examining a previous level.

Because approximately fifty percent of Hewlett-Packard manuals contain graphics, being able to display that data type became imperative. Activating an icon within the text provides the user with the alternative of viewing the figure only if deemed necessary. It also allows the user to scroll through a document on the screen quickly, without the burden of watching a graphic imbedded within the text slowly re-paint. The figure is displayed within its own window, giving the user the flexibility of moving or sizing the window. This window may be left on the screen for the user to refer back to, while reading different sections of the text.

A complete, thorough help system was also fully integrated into LaserROM. It gives the user a generic index from which to access topics discussed within, and a context sensitive mode that can be used to see specific text relating to menu items or individual windows with their associated commands.

The Help subsystem, a computer-based training package that comes standard with LaserROM, and the heuristic nature of Windows all contribute to the fact that LaserROM does not require use of a manual. As a matter of fact, one was not printed for those specific reasons. The best training for LaserROM is to use it.

The Future in CD-ROM Information Retrieval

Clearly, CD-ROM is a technology whose time has come. The addition of interactive media and graphical user interfaces has made information assimilation exciting, rather than mundane and tedious as it often has been. It is human nature to take more interest in the things we enjoy doing, and to ignore or prolong the tasks we despise.

Many new technologies have been introduced that try to enhance the nature or content of information, as well as bring to the user an element of excitement and intrigue when using them.

Hypermedia

The theory of hypertext and hypermedia goes back to the days of the quote found at the beginning of this paper. In 1945, Vannevar Bush synthesized the idea of a "memex" - a complete information "storeroom" at the disposal of each individual that could be accessed and manipulated within the framework of the user's working environment. Information was intertwined with cross-linkages and was stored on microfiche (later updated to being on a computer). Theodore Nelson (circa 1965) has been credited with coining the term hypertext, and along with Douglas Engelbart, have become founders of the ideas that lie behind the technology. The simple concept of hypermedia is to allow users to establish linkages between all data types so that these linkages form a natural "knowledge path" that may be followed. This perception embraces the idea of non-sequential reading vs. normal, sequential reading. Linkages are generally not carried with the data, but rather stored in a separate database. The information provider may also establish linkages within the data, allowing the user to follow many different paths, if so desired.

Hypermedia can be thought of as a tool that allows users to "teach" the computer the way they themselves think. Eventually, software may be smart enough to emulate the knowledge processing traits of the end user and apply these rules to new information when it is added into the system, thereby automatically establishing the necessary linkages.

As you can imagine, the problem of updating these linkages and incorporating hypermedia into a changing environment is a question that software developers will be (and have been) wrestling with for quite a while. Two such tools currently on the market that deal with this technology are Guide and Idex(C) by Owl International and Hypercard(C), recently introduced by Apple.

Multimedia

With the use of technolgies such as CD-I or DV-I (Compact Disc- or Digital Video-Interactive), full-motion video on a CD-ROM disc is now possible. You can imagine the uses in the fields of home entertainment and especially interactive training. Being able to look up and read the textual portion of how to fix a disc drive head crash, invoke a detailed graphic of the drive internals, view a full-motion video sequence of someone fixing the drive (freezing the video when necessary), and having each step narrated-all on your personal computer-is not a concept that falls only within the realms of science fiction anymore.

Conclusion

The world of CD-ROM is growing. Products are being introduced that give the user more power in accessing information then was ever deemed possible before. The door is open to both information providers and software developers to coordinate their efforts and utilize the tools I have outlined. Through their efforts lies not just the future of information distribution and retrieval, but the future of computing in our society as well.