MULTIPLE ACCESS METHODS TO HIERARCHICALLY STRUCTURED FULLTEXT DATABASES *

M.A. SHEPHERD, C.R. WATTERS, E.W. GRUNDKE, P. BODORIK
SCHOOL OF COMPUTER SCIENCE
TECHNICAL UNIVERSITY OF NOVA SCOTIA
HALIFAX, CANADA B3J 2X4

ABSTRACT

Many fulltext documents, such as manuals, guides and reference materials, corporate regulations, and legal statutes, are logically and physically arranged in an hierarchical structure. This hierarchical structure may manifest itself as chapter, subchapter, section, subsection, paragraph, subparagraph, etc.

Automated information retrieval systems can exploit this inherent hierarchical structure to integrate a number of different access methods in order to provide highly sophisticated fulltext passage retrieval. These access methods include Boolean, menu, direct access to specific passages, and access through a subject index. Passages are retrieved and displayed within the context of logically adjacent passages, even though the passages may be physically distant. A prototype system has been developed to demonstrate most of these features.

* This project was funded in part by Bell-Northern Research.

HIERARCHICALLY STRUCTURED TEXT

INTRODUCTION

This paper outlines how the hierarchical structure found in some fulltext databases can be exploited to integrate a number of different access methods to the fulltext. Such integration of access methods allows the user to switch freely among the different access methods in order to develop the most appropriate search strategy for the query. The use of the hierarchical structure also allows passages to be presented within a context; i.e., passages which are logically adjacent but physically distant may be identified and presented to the user.

ACCESS TO FULLTEXT

Bibliographic records have an easily identified linear structure consisting of variable and fixed length fields such as author, title, publisher, subject descriptor, etc. A similar linear structure has been superimposed on fulltext databases. The structure, when superimposed on fulltext databases, consists of variable length fields such as title, abstract, text passages, citations, figures, tables, etc.

Many fulltext databases, however, have an inherent hierarchical structure such as the structure found in legal statutes and engineering standards. The superimposed structure described above is not hierarchical, rather it is linear in that no field is recognized as having logically superordinate or subordinate fields. This results in the loss of any hierarchical structure in the fulltext.

The Boolean query is the most widely used access method to fulltext databases, although some systems do provide subject access through manually prepared subject indices (Sprowl, 1981; Wells, 1982) or through manually prepared subject codes that have been assigned to the text passages (Tousignaut, 1984). This is probably due to the fact that the database is easily prepared for Boolean access without human intervention. The Boolean access method, when applied to fulltext retrieval, generally requires the satisfaction of the Boolean conditions within a field of the document, usually a passage of text. This passage is then retrieved for the perusal of the user. Even if the user finds that the field contains pertinent information, the user must browse forwards and backwards within the document in order to place the retrieved information into a context.

The need for alternate access methods has been demonstrated by Kaske and Sanders (1980) and by Geller and Lesk (1983). Kaske and Sanders found that different classes of users of online library catalogues desired different access methods. Geller and Lesk, in comparing menu versus keyword access to online library catalogues and to news wire service stories, found

that menus are preferred when the user does not know what is available while keywords are preferred when the user already has some knowledge of what he is looking for.

Blair and Maron (1985) and Shepherd (1981) also support this need for alternate access methods to fulltext. In the evaluation of a fulltext retrieval system in a litigation support situation, Blair and Maron found that Boolean retrieval of fulltext documents produced precision measures of 79% but recall measures of only 20%. Shepherd found that the Boolean retrieval of paragraphs from the fulltext of scientific journal articles produced precision measures of 73% and recall measures of only 9%. Such results indicate that alternate access methods must be made available to the user to ensure comprehensive retrieval.

Alternate access methods to hierarchically structured fulltext can be readily implemented by exploiting the hierarchical structure. In 1979, Shepherd and Watters exploited this structure in legal statutes and regulations and engineering standards to integrate access to text passages through the Boolean combination of keywords, direct access to previously identified passages of text, and through a manually prepared subject index.

Bramwell (1984) has developed a system which can be used to view structured text through the use of menus. Although the system is primarily menu based, keyword access is available also.

Watters et al. (1985) developed a prototype system that provides integrated access through the Boolean combination of keywords, direct access to previously identified passages of text, and also through menus that were generated automatically from the hierarchical structure of the fulltext. This system permits the user to switch freely among the different access methods in order to develop the most appropriate search strategy for the query.

HIERARCHICAL STRUCTURE AND INTEGRATION

Many fulltext databases have an hierarchical structure that is inherent in the logical organization of the text itself. This structure may manifest itself in the form of units such as chapters, subchapters, sections, subsections, paragraphs, subparagraphs, etc. Fulltext databases exhibiting this hierarchical structure include legal statutes and regulations, corporate manuals, textbooks, and journal articles as well as many others. Figure 1 is an example drawn from engineering standards (American Society for Testing and Materials, 1983).

HIERARCHICALLY STRUCTURED TEXT

5. Front-Surface Diamond Polishing

5.1 Apparatus:

5.1.1 Polishing Machine -- Oscillating-tub polisher or other similar small laboratory-scale polishing machine capable of providing randomized motion of the silicon specimen over the polishing pad.

.
.
.

5.1.4 Rigid Plate, of glass or other similar hard material compatible with the chosen polishing machine and capable of providing a flat support for the polishing pad during polishing.

5.2 Reagents and Materials:

5.2.1 Diamond Slurry -- Synthetic or natural diamond with grain size in the range 0.5 to 3 um, inclusive, suspended in a nonaqueous liquid or paste carrier.

5.2.2 Solvent -- Suitable nonaqueous solvent for removing diamond slurry subsequent to polishing.

5.2.3 Wax -- Glycol phthalate or other similar wax having a melting temperature of less than 150C.

5.2.4 Dry Air -- Source of clean, dry air suitable for drying the speciment.

Figure 1. Extract of Hierarchically Structured Text.

The text in such a database may be considered on both a physical level and a logical level: physical in the sequential order of passages of text in the database, as might occur on the printed page, and logical as in the hierarchical order of superordinate and subordinate passage. In the example of Figure 1, the passage of text at 5.2 is physically adjacent to the passage at 5.1.4 and at 5.2.1, but is logically adjacent to the passage at 5 and the passage of text at 5.2.3. Even though the logically adjacent passages may occur several physical pages apart, they may be brought together for retrieval and display purposes, as in Figure 2.

The hierarchical structure may be exploited to provide Boolean access within the logical structure as opposed to the physical structure. It may also be exploited to provide menu access, subject access, and direct access to the same passages of text.

5. Front-Surface Diamond Polishing

5.2 Reagents and Materials:

5.2.3 Wax -- Glycol phthalate or other similar wax having a melting temperature of less than 150C.

Figure 2. Retrieved passages showing appropriate superordinate passages for context.
Result of direct access request, "5.2.3".
Result of Boolean query, "POLISHING AND TEMPERATURE".

Direct Access to Text Passages

A user may directly access a passage of text if its identification code is known. This code may be determined from previous access, "see" references in retrieved passages, menus, or from the original document. Figure 2 is an example of the retrieval results in response to the direct request for "5.2.3". Note that the logically superordinate passages are retrieved in addition to the passage resident at the identified node in the hierarchy. If the request was for "5", all passages in section 5 would be retrieved. If the request was for "5.2", the superordinate passage at 5 and all of section 5.2 would be retrieved.

Boolean Access

The evaluation of the Boolean query takes place within the hierarchical structure of the document. The query, "POLISHING AND TEMPERATURE" may be satisfied by the term, "POLISHING," in 5, and by the term, "TEMPERATURE," in 5.2.3, even though the logical passage at 5.2 intervenes and the passages themselves are physically distant. This query would retrieve all passages on the hierarchical path as shown in example in Figure 2. See Shepherd and Watters (1979) for a full description of implementing the Boolean AND, OR, and NOT operators within the hierarchical structure.

Menu Access

Menu retrieval is similar to retrieval using the table of contents at the front of a book. The user is shown the titles of

each main section in the document. Upon selection of a section, the titles of the subsections of that section are displayed. This search through the hierarchy is continued until the user terminates the menu access. At the selection of each subordinate level, the physical unit or header at the subordinate level is displayed with the menu of its subordinate level. This provides additional context for the the choices on that menu. Menu access can start at any point in the hiearchy by selecting the menu option with the appropriate menu code. The menu codes corresponds directly to the section identification.

The identification code of each menu is the same as the code of the physical unit of text at that level. Figures 3a through 3d show the sequence of appropriately labeled menus leading to the retrieval of the physical unit, 5.2.3.

DATABASE 674

ROOT MENU

Standard Practice for Preparing High-Resistivity n-Type Silicon for Spreading Resistance Measurements

1. Scope
2. Summary of Practice
3. Significance and Use
4. Interferences
5. Front-Surface Diamond Polishing
6. Diamond Bevel Polishing

Select from menu or enter (M+any menu code), (D+any direct code), (B for Boolean query), (S to stop):

Figure 3a. Initial Menu for this Document

MENU 5

Front-Surface Diamond Polishing

1. Apparatus
2. Reagents and Materials
3. Procedure

Select from menu or enter (M+any menu code), (D+any direct code), (B for Boolean query), (S to stop):

Figure 3b. Menu afer selecting "5" in Figure 3a.

MENU 5.2

Reagents and Materials

1. Diamond Slurry
2. Solvent
3. Wax
4. Dry Air

Select from menu or enter (M+any menu code), (D+any direct code), (B for Boolean query), (S to stop):

Figure 3c. Menu after selecting "2" in Figure 3b.

5.2.3 Wax -- Glycol phthalate or other similar wax having a melting temperature of less than 150C.

Figure 3d. Text Retrieved after selecting "3" in Figure 3c.

Subject Access

Subject indices to the fulltext, similar to the alphabetic index at the back of a book, may be created either manually or automatically. In either instance, the alphabetical index provides access to appropriate passages of text. In Figure 4, the index word is the truncated, "POLISH*", and the subheadings are the menu titles of the passages in which instances of the term appear. The appropriate text passages could be accessed either by selecting one of the subheadings in a menu-type of selection, or by requesting the appropriate passages through the direct access mode of retrieval.

POLISH*

5	Front-Surface Diamond Polishing
5.1.1	Polishing Machine
5.1.4	Rigid Plate
5.2.2	Solvent

Figure 4. Subject Access to Truncated Term, "POLISH*"

DISCUSSION

The above example permits access to fulltext passages by direct access, Boolean access, menu access, and through a subject index. Not all hierarchically structured databases lend themselves to menu access or to the automatic creation of subject indices due to the lack of appropriate headings. All

hierarchical databases do, however, lend themselves to direct access, Boolean access within the hierarchy, and to the presentation of passages of text within a logical framework. Should a manually prepared subject index be available, it too can be used as a method to access the text.

While there has been no experimental testing of the effect of the integration of access modes on the user or on search results, it has been shown that alternate access methods to a database are desirable, and that the access method selected is a function of the type of query and of the knowledge and experience of the user (Kaske and Sanders, 1980; Geller and Lesk, 1983). The exploitation of the hierarchical structure, as described above, does provide alternate access methods to the database. The user can freely switch among the available access modes to develop the most appropriate search strategy. The switching among modes is easily accommodated as the implementation of the modes is based on the hierarchical structure inherent in the database.

REFERENCES

AMERICAN SOCIETY FOR TESTING AND MATERIALS. "Standard Practice for Preparing High-Resistivity n-type Silicon for Spreading Resistance Measurements", in Annual Book of ASTM Standards. Vol. 10.05, pp. 583-587. Philadelphia. 1983.

BLAIR, D.C. and MARON, M.E. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System", in Communication of the ACM. Vol. 28, No. 3, (March 1985), pp. 289-299.

BRAMWELL, B. "Browsing Around a Manual", in Proceedings of the Canadian Information Processing Society. (May 1984), pp. 438-442.

GELLER, V.J. and LESK, M.E. "User Interfaces to Information Systems: Choice vs. Commands", in Proceedings of the Sixth Annual Internatinal ACM SIGIR Conference on Research and Development in Information Retrieval. (June 1983), pp. 130-135.

KASKE, N.K. and SANDERS, N.P. "Evaluating the Effectiveness of Subject Access: The View of the Library Patron", in Proceedings of the 43rd ASIS Annual Meeting. Vol. 17, (October 1980), pp. 323-325.

SHEPHERD, M.A. "Text Passage Retrieval Based on Colon Classification: Retrieval Performance", in Journal of Documentation. Vol. 37, No. 1, (March 1981), pp. 25-35.

SHEPHERD, M.A. and WATTERS, C.R. "Hierarchical Retrieval from Structured Text", in Proceedings of the Third International Study Conference on Classification Research. ed. A. Neelameghan. (1979), pp. 466-472.

SPROWL, J.A. "WESTLAW vs LEXIS: Computer-Assisted Legal Research Comes of Age", in Program. (G.B.). Vol. 15, No. 3 (July 1981), pp. 132-141.

TOUSIGNAUT, D.R. "DIF: A Drug Monograph Fulltext Online File", in Proceedings of the Fifth National Online Meeting. (April 1984), pp. 389-392.

WATTERS, C.R., SHEPHERD, M.A., GRUNDKE, E.W., and BODORIK, P. "Integration of Menu Retrieval and Boolean Retrieval from a Fulltext Database", Technical Report, School of Computer Science, Technical University of Nova Scotia, 1985.

WELLS, W.W. Jr. "LEXIS and WESTLAW: The Strengths and Weaknesses", in Legal Reference Service Quarterly. Vol. 2, No. 2 (1982), pp. 51-61.