Yaşar Tonta, Indexing in hypertext databases. In: *Studies in Multimedia:State-of-the-Art Solutions in Multimedia and Hypertext.* (p. 21-30) Eds.: Susan Stone and Michael Buckland. Medford, NJ: Learned Information, 1992.
http://yunus.hun.edu.tr/~tonta/yayinlar/indexing.pdf

INDEXING IN HYPERTEXT DATABASES

Yasar Tonta
School of Library and Information Studies,
University of California at Berkeley, CA 94720
tonta@sherlock.berkeley.edu

WHAT IS HYPERTEXT?

As Franklin puts it, "defining hypertext is much like describing beauty: part of it is inherent in the eye of the beholder" (1). Nevertheless, hypertext can be defined as "non-sequential reading and writing." Hypertext systems contain frames of text, pictures, sound, and animation that are organized non-linearly in a network of linked frames. From any frame, users can access a variety of other frames containing text or other media (2). Users follow various sequences of frames and links to retrieve the information they require or add new frames and connections between them. Hypertext systems with non-textual contents are called "hypermedia" to denote the fact that hypertext can not only include text but also non-textual data such as sound, images, animation and audio-visual data in digitized forms. The terms hypertext and multimedia are used throughout this chapter to cover both hypertext and hypermedia systems.

The origin of hypertext goes back to 1940s when Bush outlined his vision of Memex, which was to have some hypertextual features such as linking and associating different parts of documents in a way that would facilitate what he called "associative thinking" (3). Twenty years later, inspired by Bush's ideas, Nelson coined the terms hypertext and hypermedia to describe the associative linking of information into large networks (4).

Today, various hypertext systems, though most of them are experimental, have been developed in different areas such as education and publishing. Systems developed tend to be small in size with not-so-sophisticated information retrieval capabilities, because hypertext systems are a relatively new phenomena; the oldest hypertext system is less than ten years old. Furthermore, there are some teething problems that need to be addressed before hypertext systems can be used in a wide variety of disciplines.

PROBLEMS WITH HYPERTEXT SYSTEMS

Conklin summarizes the problems with hypertext systems under two headings: disorientation and cognitive overhead (5).

Disorientation is the tendency to lose one's sense of location and direction in a non-linear document. Browsing through non-linear documents leaves people with a general feeling of disorientation, being lost, or of losing context. Foss, in working with NoteCards, has identified one of these disorientation problems as "the embedded digression problem" (when users following a chain of cross-references become distracted from the main task and never return) and another as "the art museum phenomenon" (when users, so deluged by information, cannot summarize at session's end what they have seen) (6). Foss has developed support mechanisms for helping users stay oriented, including history graphs and twin cards (7).

Cognitive overhead occurs when users are offered many alternatives when reading hypertext documents. For example, each node may be linked to several other nodes and users may not be able to figure out which link to follow. Therefore an additional effort and concentration is necessary to maintain several tasks or trails at one time.

Marchionini and Shneiderman, and Carando point out the difference between searching for facts and browsing (8, 9). Unlike database systems, hypertext systems are not designed for fast and efficient fact retrieval. Rather, they support unhurried and informal information seeking. These

opposing interaction styles may frustrate users expecting a different kind of support. Marchionini and Shneiderman contrast these two objectives. In the first instance, users searching for a fact want a swift and efficient way to reach their goal; that is, to extract information. Such users may be professional on-line searchers or just in a hurry (for example, software engineers who need a fact quickly). The second type of information seeker may wish to browse through system knowledge, unconcerned with efficient search performance; these may be users with a long-term commitment to the domain of interest, who believe that random knowledge acquired during information examination will be of future use.

Many researchers emphasized that there is a thin line between order and "chaos" in presenting information in hypertext systems (10). Hypertext systems tend to fragmentalize information, which makes managing information difficult. The reality is that currently we are developing what van Dam calls "docu-islands" of knowledge that are incompatible with one another. Just when it seems that compatibility problems of microcomputers have eased somewhat, new, more complex hyperdocument systems will make all those interconnections obsolete (11).

Larson approaches the issue from an information retrieval point of view. He claims that many hypertext systems lack even rudimentary search capabilities that operate across multiple hypertext networks. Applying the same techniques used in personal information systems to large hypertext systems would end up creating "chaos." Larson is more concerned with the "scale" issue and warns that "large hypertext databases will, of course, face all of the problems of large bibliographic databases. Yet, so far, there are no indications that hypertext researchers have considered such things as name authority or subject vocabulary control. Both of these will eventually need to be addressed if library-sized hypertext networks are to avoid scattering information that should be kept together" (12).

INDEXING IN HYPERTEXT SYSTEMS

Even though most hypertext systems use some kind of indexing for information retrieval, indexing in hypertext systems so far has not been a research topic by itself. Nevertheless, some researchers have briefly mentioned the indexing subsystem in their hypertext systems. Only a few authors (13,14,15) emphasized the importance of indexing in hypertext systems.

Bush, in his vision of Memex, asserted that human mind "operates by association." He claimed that "with one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain" (16). (See also 17.) Bush suggested a new kind of information indexing --associative indexing-- to incorporate this associative characteristic of the human mental process into the selection of information. Associative indexing is such that "whereby any item may be caused at will to select immediately and automatically another." In this way, the author or "associative indexer" of a document creates associative links or trails between items of information. The reader follows or navigates a trail from item to item, reviewing each item rapidly or slowly according to his/her needs. The reader can choose to navigate trails set up by the authors of the information or create a new web of associative trails reflecting the reader's particular point of view about the information store (18).

Weyer's Dynabook (19) is one of the earliest pieces of research on electronic books that also deals with indexing problems. Weyer compares three different electronic books (Book A, B, C) with different search capabilities. He describes a simple dynamic book (Book C), based on a world history textbook, implemented in the Smalltalk programming system and used by students to answer questions. Book B and Book C had some indexing capabilities as well. Students were more successful in searching Book C than Book A and Book B and they liked some of the design features in Book C such as fast access to subject terms, term highlighting and simultaneous access to subject index and text. (See also 20.)

Egan et al. describe SuperBook, which is "a new presentation system designed to improve the

usability of existing machine readable documents" (21). They present a series of behavioral studies that directly compare how people find information in conventional printed documents and the same documents in SuperBook form.

The authors believe that "people fail to find desired information in textual databases" because of indexing policies. They are skeptical of the value of controlled vocabularies and state that indexing performed by subject experts does not necessarily improve users' performance over automatic techniques. Their system, SuperBook, is enhanced with what they call "rich indexing." "Rich indexing is implemented in SuperBook by means of full content indexing in which every occurrence of every word is indexed, and by aliasing. The latter technique permits users to establish their own search synonyms" (22).

Despite their "rich indexing", Egan et al. soon admitted that a thesaurus might improve the performance of their system, even though at the beginning they were skeptical about the value of controlled vocabularies. This is not surprising because full text databases with automatic indexing exhibit much worse problems than databases with controlled vocabularies. Blair and Maron found that the recall rate in a full text document database is shockingly low (23).

Marchionini and Shneiderman think that, after all, links at every word to every word are clearly not desirable from the user's perspective or that of system performance (24). The trade-offs in machine overhead and user cognitive load (in the form of overchoice) must be weighed carefully. Designers should consider the targeted task domains and typical user population in deciding how fine the access points should be and what links among access points should be visible to users.

In SuperBook, in the authors' words, "other, more complex search mechanisms (e.g., full Boolean search) are not supported because previous research suggests that end users rarely use such mechanisms effectively" (25). One needs to be skeptical about the above statement. For full text indexing, some sort of Boolean search capability should be provided in order to reduce the

information overload even though users have some problems with Boolean operators. Boolean search capabilities have become an indispensable part of online catalogs and users are not totally unfamiliar with basic operators such as AND and OR.

It should be noted that the case study for SuperBook required the users to come up with answers to factual questions. The kind of questions that come up in actual information retrieval could be quite different from factual questions. Authors' evaluation of SuperBook is based on how successful users were in finding relevant sections in the SuperBook (26). This approach does not require index-using skills that are more sophisticated than the skills needed to use back-of-the-book indexing, which can hardly be compared with the indexing in large databases.

More recently, attempts have been made to build hypertext databases that are larger than electronic books or encyclopedias (27, 28, 29). Research is underway in several institutions to investigate the challenges and opportunities that large hypertext databases will offer. It will be interesting to see, from the indexing point of view, how it is that the indexing techniques developed for electronic books will "scale up" in larger hypertext databases.

The University of California (UC) at Berkeley has a wide variety of multimedia information sources scattered all over the campus. Some multimedia collections are housed in libraries (e.g., the Geography Department's Map Library and the Architectural Slide Library) while others are held by different departments in the campus (e.g., University Art Museum and Lowie Museum of Anthropology) .

The UC Berkeley Image Database Project aims to provide access to some of the image collections in the campus by means of an online catalog with tools for visually browsing surrogate images. Terms to index images are taken from the Art & Architecture Thesaurus (AAT) (30). The software developed for the prototype system is called ImageQuery. ImageQuery allows users, as in traditional online catalogs, to enter search terms taken from the controlled vocabulary (AAT) and

then browse visually the resultant surrogate images associated with the search term. Besser points out that combining visual browsing capabilities with controlled vocabulary terms helps users find relevant images more quickly and efficiently, whereas using index terms only tends to retrieve many images some of which are not quite relevant (31). (See also (32).)

The National Archives of Canada developed an optical disc imaging system called ArchiVISTA to provide access to 20,000 editorial cartoons and caricatures in its holdings. Images of cartoons and caricatures are accessible via subject, artist, publication, place, date and unique item numbers. In addition, the description of archival records for images can also be retrieved from the bibliographic database by means of a minicomputer-supported database management system (MINISIS) (33). (See also the chapter by Stone in this volume).

The Film Repository at the NASA's Johnson Space Center (JSC) "houses a collection of more than 1 million negatives and transparencies, as well as around 10,000 motion picture and audio reels, documenting all aspects of the manned space flight program in this country since 1958" (34). JSC has recently decided to establish intellectual control over the storage and retrieval of images in this collection.

Researchers at JSC have first digitized some images and then concentrated on developing a "visual thesaurus" to streamline access to images in the database. The terms included in the visual thesaurus mainly came from the card catalog for the image collection. The NASA Thesaurus was used for further enriching the visual thesaurus. Hierarchical relationships between the terms were also identified (e.g. broader, narrower, related terms).

The visual thesaurus provides browsing facilities for each term and brings up the corresponding image(s). Personal Librarian was chosen as the data retrieval engine for this project. This software offers relatively advanced retrieval capabilities and is not based on Boolean logic. It weights the terms and ranks the images in the order of their relevance to the search query. The retrieval performance of the Personal Librarian proved to be satisfactory during the preliminary

experiments (35). Marchionini and Shneiderman argue: "Present systems may support browsing strategies attractive to end users but inefficient for fact retrieval. To compensate, cumbersome analytical strategies that take advantage of indexing to improve retrieval may be supported; however, the overall design may become complex. Analytical strategies include consulting thesauri before search, using Boolean connectives, and systematically iterating queries" (36).

It is interesting to note that users prefer using indexes when searching in hypertext databases. Allinson and Hammond found that the index in their Hitch-Hiker's Guide was used by 79% of student users (37). Index use was found to be much higher in the Marchionini and Shneiderman study (38). When subjects were asked to perform efficiently in searching for specific factual information in a Hyperties database, the predominant strategy (14 of 16 subjects) was to use the alphabetical index.

Indexing in existing hypertext systems, however, is not satisfactory and more sophisticated indexing techniques should be tried. Nielsen reports: "It is important that big hypertexts will be the most useful ones and it is therefore important to address the issues of overview in large information spaces. We are currently working on methods for assigning relevance to links based on an information retrieval measure of similarity between the two linked nodes as well as an estimate of the user's current interests. However, it is yet too early to judge the usability of such methods for real work. If such metrics can be tuned to correspond to users' actual intentions it would be possible to filter out links with ratings below some cutoff point" (39).

Marchionini and Shneiderman summarize the search features that should be implemented in hypertext systems:

Search features like Boolean connectives, string search, proximity limits, and truncation facilitate rapid access to information, but cause additional cognitive load on the part of the user and substantial preprocessing of the database itself.

Systems that provide only browsing features allow casual, low cognitive load

exploration, but are typically inefficient for directed search tasks or fact retrieval. Defining a hybrid system that guides discovery seems an appropriate compromise, but involves a number of trade-off decisions. How deeply the database is indexed, whether some automatic controlled vocabulary is included, and how feedback is summarized and even formatted on the screen affect the strategies users will apply. If every word is indexed, the possibility of information overload increases. Therefore, features for filtering such as frequency of occurrence per node or support for NOT operators must be enhanced. If a controlled vocabulary is included, automatic thresholds must be established, or the user must be prompted to apply the controlled vocabulary or be alerted to its effects. For example, in an encyclopedia, a query that retrieved more than 50 articles could automatically trigger a narrowing function (40). (See also 41, 42.)

Lynch discussed some of the potential problems in large hypertext databases (43). As database size grows, the number of links between terms will increase tremendously. This will in turn increase the number of links to follow, thereby creating potential "disorientation" and "information overload" problems. Lynch does not see automatic indexing as a solution for large hypertext databases, since there will be many terms and many links assigned to a given occurrence of a term. He also finds the tree structure that small hypertext databases have inadequate to locate information in large hypertext databases "both because the access points may not have a natural hierarchical structure and because the number of tree nodes that it may be necessary to transverse to reach a useful starting may be large" (44). Lynch offers the following techniques to tackle retrieval problems in large hypertext databases and to assist users in identifying relevant material from large result sets: keyword and phrase searching, Boolean queries, proximity searching, term weighting and result ranking.

Salton and Buckley investigated the applicability of generating content links in hypertext databases automatically (45). They tried to generate automatic content links between terms based on the global term and phrase matches and tested the usefulness of this method. They suggested that further research is needed in this area, although the preliminary results they obtained are somewhat promising.

Regarding indexing and classification, Farmer warns that the library profession, as the prime collectors and organizers of information, needs to look at how hypermedia will affect traditional cataloging and classification: "In the future, cataloging codes may include sections on how to set up webs of information; what types of relationships should be expressed; how to guide users through a hypermedia database; and how to create their own links and trails. New cataloguing standards, which define the structures and relationships that can be imposed on a hypermedia document, may need to be developed" (46).

ISSUES AND FUTURE DEVELOPMENTS

Indexing for small size hypertext databases such as electronic books and encyclopedias is quite different from that of large hypertext systems. Existing hypertext systems have indexes similar to back-of-the-book indexes, which do not "scale up" to large hypertext systems. Larson describes this "scaling up" issue as the "order of magnitude" problem (47). If the size of the hypertext database gets bigger, the database needs to be rearranged. Indexing principles used in small databases simply do not work for large databases.

Larson points out that additional problems are faced by hypertext systems that attempt to support very large databases of multimedia information: "The problems of cataloging and indexing non-textual materials are well-known, and simply putting a collection of such materials into digital form does not address the problem" (48).

Hypertext databases should support a multitude of data types such as text, sound, images,

animation and audio-visual data. New data types such as point data, raster data, vector data, and spatial and temporal data should also be supported in hypertext databases. For instance, many satellite and remote sensing instruments produce a regular array of point measurements. Similarly, data on topographical maps are represented as vectors. It is suggested that one approach to representing this type of data in hypertext databases will be to use existing thesauri of geographical regions and place names that include the cartographic coordinates of places (49).

The volume of multimedia data that is to be fed by satellites, the Earth Observing System (EOS), and sensors is mind-boggling. It is expected that 1950 x 109 bytes (terabyte) of raw data will be received every day from these stations. Furthermore, the amount of multimedia data will reach 1.0 x 1016 byte (pedabyte) bytes in 15 years! Providing access to such large volumes of multimedia data requires sophisticated indexing techniques. For instance, in order to find "all the images of Lake Tahoe taken by, say, Orbit Platform #1 for the period June-September 1986," spatial data and time need to be indexed. Furthermore, performing this search requires indexes on the result of a function and not the raw value. "Indexing functions for images and text often return a collection of values for which efficient access is desired" (50, 51).

Automatic indexing of images, unlike text, may not be possible for some time to come. Satisfactory image recognition software has yet to be developed. This will exacerbate the indexing problems in large-scale multimedia databases. Voice recognition software gets better. It might be possible to automatically index recognized words (52).

Currently it appears that supporting visual browsing capabilities with controlled vocabularies such as AAT improves retrieval effectiveness in image databases. Developing specialized "visual thesauri" for other disciplines might help solve indexing and retrieval problems in other multimedia databases as well. In conclusion, as van Dam put it, designing hypertext systems needs people "who...can think about classification and indexing." It is librarians and information workers who could provide safer journeys in tomorrow's colossal information spaces.

ACKNOWLEDGMENTS

I am grateful to Professors Ray R. Larson and Michael K. Buckland for their comments on an earlier draft of this chapter.

NOTES

(1) Carl Franklin, "Hypertext Defined and Applied," Online 13 (May 1989) 37-49.

(2) Carolyn L. Foss, "Tools for Reading and Browsing Hypertext," Information Processing & Management 25 (1989) 407-418.

(3) Vannevar Bush, "As We May Think," Atlantic Monthly 176 (July 1945) 101-108.

(4) Ray R. Larson, "Hypertext and Information Retrieval: Towards the Next Generation of Information Systems," ASIS '88 Proceedings of the 51st ASIS Annual Meeting Atlanta, Georgia October 23-27, 1988. Ed. by Christine L. Borgman and Edward Y.H. Pai. (Medford, New Jersey: ASIS, 1988). Vol 25, 195-199.

(5) J. Conklin, "Hypertext: An Introduction and Survey," IEEE Computer 20 (1987) 17-41.

(6) Foss, "Tools for Reading."

(7) Patricia Carando, "Shadow - Fusing Hypertext with AI," IEEE Expert 4 (Winter 1989) 65-78.

(8) Gary Marchionini and Ben Shneiderman, "Finding Facts versus Browsing Knowledge in Hypertext Systems," IEEE Computer 12 (January 1988) 70-80.

(9) Carando, "Shadow."

(10) Torrey Bayles, "A Context for Hypertext: Some Suggested Elements of Style," Wilson Library Bulletin 63 (November 1988) 60-62.

(11) Ann F. Bevilacqua, "Hypertext: Behind the Hype," American Libraries 20 (February 1988) 158-162.

(12) Larson, "Hypertext."

(13) Ibid.

(14) Clifford A. Lynch, "Hypertext, Large Databases, and Relational Database Management Systems," in National Online Meeting Proceedings. (Medford, NJ: Learned Information, 1989), 265-270.

(15) Linda Farmer, "Hypertext: Links, Nodes and Associations," Canadian Library Journal 46 (August 1989) 235-238.

(16) Bush, "As We May Think."

(17) Lauren B. Doyle, "Indexing and Abstracting by Association," American Documentation 13 (October 1962) 378-390.

(18) Farmer, "Hypertext."

(19) Stephen A. Weyer, "The Design of a Dynamic Book for Information Research," International Journal of Man-Machine Studies 17 (July 1982) 87-107.

(20) Stephen A. Weyer and Alan H. Borning. "A Prototype Electronic Encyclopedia," ACM Transactions on Office Information Systems 3 (January 1985) 63-88.

(21) Dennis E. Egan et al., "Formative Design Evaluation of SuperBook," ACM Transactions on Information Systems 7 (January 1989) 30-57.

(22) Ibid.

(23) David C. Blair and M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System," Communications of the ACM 28 (March 1985) 289-299.

(24) Marchionini and Shneiderman, "Finding Facts," 78.

(25) Egan et al., "Formative Design," 37.

(26) Ibid., 40-41.

(27) I. Ritchie, "Hypertext - Moving Towards Large Volumes," The Computer Journal 32 (December 1989) 516-523.

(28) Peter Cooke and Ian Williams, "Design Issues in Large Hypertext Systems for Technical

Documentation" in Hypertext: Theory into Practice. Ed. by Ray Mcaleese. (Norwood, NJ: Ablex, 1989), 93-104.

(29) Kenneth Utting and Nikole Yankelovich, "Context and Orientation in Hypermedia Networks," ACM Transactions on Information Systems 7 (January 1989) 58-84.

(30) Toni Petersen, "Developing a New Thesaurus for Art and Architecture," Library Trends 38 (Spring 1990) 644-658.

(31) Howard Besser, "Visual Access to Visual Images: The UC Berkeley Image Database Project," Library Trends 38 (Spring 1990) 787-798.

(32) Howard Besser and Maryly Snow, "Access to Diverse Collections in University Settings: The Berkeley Dilemma," in: Beyond the Book: Extending MARC for Subject Access. Ed. by Toni Petersen and Pat Molholt. (Boston, MA: G.K. Hall, 1990), 203-225.

(33) Gerald Stone and Philip Sylvain, "ArchiVISTA: a New Horizon in Providing Access to Visual Records of the National Archives of Canada," Library Trends 38 (Spring 1990) 737-750.

(34) Gary A. Seloff, "Automated Access to the NASA-JSC Image Archives," Library Trends 38 (Spring 1990) 682-696.

(35) Ibid.

(36) Marchionini and Shneiderman, "Finding Facts," 71.

(37) Lesley Allison and Nick Hammond, "A Learning Support Environment: the Hitch-Hiker's Guide," in Ray Mcaleese, ed. Hypertext: Theory into Practice. (Norwood, NJ: Ablex, 1989), 62-74.

(38) Marchionini and Shneiderman, "Finding Facts."

(39) Jakob Nielsen, "The Art of Navigating Hypertext," Communications of the ACM 33 (March 1990) 300.

(40) Marchionini and Shneiderman, "Finding Facts," 78-79.

(41) Ben Shneiderman, "User Interface Design for the Hyperties Electronic Encyclopedia," in

Proceedings of Hypertext '87 (Nov 1987) 199-204.

(42) Xianhua Wang and Peter Liebscher, "Information Seeking in Hypertext: Effects of Physical Format and Search Strategy," ASIS '88 Proceedings of the 51st ASIS Annual Meeting Atlanta, Georgia October 23-27, 1988. Ed. by Christine L. Borgman and Edward Y.H. Pai. (Medford, New Jersey: ASIS, 1988). Vol 25, 200-204.

(43) Lynch, "Hypertext."

(44) Ibid., 267.

(45) Gerard Salton and Chris Buckley, On the Automatic Generation of Content Links in Hypertext. Technical Report TR 89-993. (Cornell University, Department of Computer Science, 1989).

(46) Farmer, "Hypertext," 238.

(47) Ray R. Larson, "Indexing and Intellectual Access in Multimedia Databases," Seminar presented at UC Berkeley, 12 October 1990.

(48) Larson, "Hypertext," 197.

(49) Ibid.

(50) Clifford Lynch and Michael Stonebraker, "Extended User-Defined Indexing with Application to Textual Databases," in: Proceedings of the Very Large Databases Conference. Los Angeles, CA, September 1988.

(51) Information presented in this paragraph comes from Professor Michael Stonebraker's seminar series at UC Berkeley (Spring 1991), and from an unpublished proposal.

(52) Larson, "Indexing."