Retrieval Interfaces for Video Databases

Andrew S. Gordon and Eric A. Domeshek
The Institute for the Learning Sciences, Northwestern University
1890 Maple Avenue, Evanston, IL 60201
gordon@ils.nwu.edu, domeshek@ils.nwu.edu

Abstract

Rich multimedia databases are useless without the ability to access the right piece of information at the right time. We have been investigating the development of indexing and retrieval systems for databases of stock video. By using a simplified conceptual indexing framework, we have been able to focus our efforts on the creation of semantic networks that organize and catalog stock video indexes and on the development of end-user indexing and retrieval interfaces.

In this paper, we focus on the interface issues. We compare two classes of information retrieval interfaces: *Query and Search* interfaces and *Zoom and Browse* interfaces. Previous researchers have developed knowledge-rich retrieval systems that follow the Query and Search approach. We argue that it is better to offer choices than to play guessing games. Accordingly, we present a version of a Zoom and Browse interface that we have developed for the retrieval of stock video clips based on scene content indexes.

1. Introduction: Indexing for Video Databases

Modern computers can play video, they can store video, they can even ship video around a network. As the multimedia juggernaut gathers momentum, a growing number of users are planning to exploit these capabilities by building large-scale video repositories -- on-line databases intended to make video available for a variety of purposes. But while the underlying coding, storage, and transmission technologies are well advanced, the technology for organizing, indexing, and retrieving appropriate video segments is in its infancy. Of course, piles of any kind of information are useless unless users can get at the bits they need when they need them. Accordingly, we require an indexing scheme suited to video and the purposes for which it is being retrieved, and we need an interactive retrieval system that employs those indexes to help users find clips. In this paper, we briefly sketch our approach to indexing and then focus on design principles for a retrieval interface.

We have chosen to focus on a particular real-world video retrieval problem. Many video production companies are moving towards digital technology and are trying to establish on-line databases of stock footage. Stock video footage includes clips produced or acquired by a company that are general enough to be used in many productions. These collections are often managed by dedicated video librarians who become very familiar with the available stock. Requests for stock video range from the very specific, e.g. "close up of water ripples", to the very abstract, e.g. "symbolic images of people designing the future".[1] Librarians depend not only on their familiarity with a given collection, but also on their basic, commonsense knowledge about the everyday world to help them interpret abstract requests in more concrete terms, or to establish possible search contexts for narrowly specific requests. In moving toward on-line databases where clips are instantly accessible to a variety of end-users, system designers must determine how to replace the intelligence of video librarians in a way that makes effective video retrieval more widely available.

The effectiveness of any retrieval system depends critically on the quality and character of the indexes that are assigned to each item in the database. Video is conceptually rich; it is used not only to show the viewer new places, people, activities, and objects, but also to teach lessons, make points, and evoke emotions. One of the great challenges in video indexing is finding ways to embody these concepts in symbolic descriptions that can be searched for matches to users' queries. Ideally, indexes for video should be composed from a language that is expressive enough to represent the breadth of concepts that can be conveyed by video while being restrictive enough to support realistic retrieval mechanisms. When it comes to symbolically describing video, nothing approaches the completeness and flexibility of natural language. However, natural

[1]Requests from Andersen Telemedia, part of Arthur Andersen & Co., S.C.

language descriptions have a pair of significant drawbacks which cause serious headaches for designers of retrieval algorithms. Natural language descriptions are neither canonical (there are an enormous number of ways to say exactly the same thing) nor unambiguous (a single word or phrase can have an enormous number of meanings).

In our effort to develop intelligent retrieval systems for stock video databases we have selected a conceptual indexing scheme which is a simplified version of those typically employed in case-based reasoning systems (Kolodner, 1993). In our framework, indexes for a case are described as a set of independent concepts, each of which is represented as a single node in a semantic network. For example, a clip that depicts a professor walking around a college campus may simply be indexed as *professor*, *walking*, and *college-campus*, where each of these concepts is a node in the conceptual organizations of people, activities, and places, respectively. The major difference between this style of index representations and those implemented in other conceptual indexing systems is that there is no relational structure between the concepts that make up a case's index. Relational structure is an absolute requirement when completely representing cases for the purpose of drawing analogies, analyzing causal chains, and adapting cases to new situations, but our investigations to date have suggested that unstructured index representations may be sufficient for many retrieval tasks.

The simplicity of this framework allows the use of very basic matching algorithms. For each concept in the semantic networks of the system, the set of all the cases indexed by that concept can be compiled and effectively encoded as a one-dimensional bit vector. At retrieval time, such a bit-vector represents a case-set: all the cases indexed by the concept. To determine the set of cases that are indexed by the conjunction of two or more distinct concepts, a simple intersection of each concept's case-set can be performed. In this manner, all of the stock video clips that are indexed by both *professor* and *college-campus* can be determined by intersecting the case-sets associated with the nodes for these two concepts. This mechanism offers an effective method of matching query representations to index representations during the retrieval process.

In simplifying the representation of indexes, we have been able to focus our efforts on two issues: (1) the creation of semantic networks that organize indexes for stock video clips, and (2) the development of end-user indexing and retrieval interfaces. With respect to index concepts, here we just briefly note that we have identified several categories of indexes that are appropriate for stock video clips, including scene content descriptions (place, people, activities, things), points made by the clips, camerawork, narrative functions, production information, and relationships among clips in the database. Our initial efforts have focused on developing and organizing an indexing vocabulary for the first category only: scene content descriptions.

While the proper organization of these index types is critical to the success of our retrieval system, this paper addresses an equally important question: What style of retrieval interface will meet the retrieval needs of stock video database users? In section 2, we discuss different interface styles for information retrieval systems and argue for the one we have chosen to pursue in our research. In section 3, we describe the current state of the stock video retrieval system we are developing.

2. Retrieval interfaces for conceptual databases

In designing intelligent multimedia retrieval interfaces, indexing cases is only half the battle; complete systems must incorporate appropriate end-user retrieval interfaces. Regardless of the quality of the case indexes, no system will be successful unless users can easily locate cases that meet their needs. In general, two classes of retrieval methods have been developed that are applicable to retrieving cases from a case library. The first method, *Query and Search*, has been explored in a number of other intelligent multimedia retrieval systems. The alternative that we have implemented, the *Zoom and Browse* method, is an outgrowth of research in hypermedia browsing systems.

2.1. Query and Search

The first approach to intelligent retrieval is Query and Search, which has been developed out of a long tradition of standard database record retrieval and deductive retrieval databases in AI. In Query and Search, users must construct a request to be parsed by the system into some representational form which is matched against the indexes in the database by a search algorithm. Advanced Query and Search systems embed flexible, knowledge-rich matching into their search algorithms, allowing them to retrieve cases similar or related to the user's query when an exact match cannot be made. Aiming to mimic human librarians, the best Query and Search systems may elaborate on users' queries with extensive inference in order to find items that satisfy their needs.

Two recent examples of intelligent multimedia retrieval systems that use the Query and Search method

are presented in the work of Chakravarthy (1994) and Lenat and Guha (1994). In Chakravarthy's work, user's queries are mapped directly to concepts in a semantic network (organized as sets of synonyms in Wordnet) and matched against indexes for each case. During the matching process, several rules may be used to infer whether a case's index satisfies the user's query based on the semantic relationships between the two. The system described is capable of making many obvious matches, such as returning a picture captioned as "a Dalmatian" in response to the user query "dog". It is also capable of more insightful matches, such as returning "closeup of an arrow hitting the bullseye of target" in response to a query for "shooting". Lenat and Guha take a related approach in their research, utilizing the extensive domain knowledge of the CYC knowledge base to expand both user queries and case indexes to increase the likelihood of successful matches. By drawing reasonable inferences from the captions of video clips, their system is capable of generating the impressive match between the user query "Find images of shirtless young men in good physical condition" and clip captions like "Pablo Morales winning the men's 1992 Olympic 100-meter Butterfly event" and "Three blonde men holding surf boards on the beach".

There are some difficulties associated with the Query and Search method that still need to be addressed. In addition to the obvious computational expense inherent in performing wide-ranging inference, these approaches tend to produce false positives, i.e. they retrieve cases that do not address the needs of the user. While it is reasonable to guess that "Three blonde men holding surf boards on the beach" might include a shot of a shirtless man in good physical condition, it is not necessarily the case. Expanding indexes through inference is somewhat like trying to cleverly service requests for video without actually watching the clips you select; without the means of verifying the legitimacy of inferences, you are bound to make lots of errors. To determine if the level of incorrect inferences is a significant problem, system designers must consider the nature of the retrieval task that the system is supporting. In stock video clip retrieval tasks, video producers consistently argue that more is better; they would prefer to be overloaded with options rather than miss out on a potentially useful clip. But even producers have their limits, and it appears that the retrieval rules used in these systems would generate many candidates that video producers would not accept. For retrieval tasks where precision is important, inference-based systems like these may be more useful as intelligent indexing aids that could suggest additional indexes to be verified by a human indexer while storing a case into the database.

There is an important question concerning the Query and Search method that must be addressed: Why are rich inferential search mechanisms necessary in the first place? The answer to this question is that in any system that allows users to formulate arbitrary queries, users will often produce requests for cases that are not in the case library. For stock video clips, where the space of possible descriptions is absolutely enormous, retrieval systems may almost never find cases that exactly matches a complex and specific user query. Rather than returning "No match" in response each time, we must be able to say "No exact match, but here are some close matches". Determining what can be considered close to an arbitrary user query requires the kind of inference that the best of these systems incorporate. There is, however, an alternative to the Query and Search method. In the next section we introduce the Zoom and Browse method, an approach that removes the need for complex inference by prohibiting users from making arbitrary requests for cases that are not in the case library.

2.2. Zoom and Browse

The Zoom and Browse approach takes the position that it is better to offer choices than to play guessing games. That is, rather than having users formulate queries for cases that may not be in the case library, retrieval systems should allow users to browse through the choices that the system has to offer. In this way users can make their own decisions about how closely the system's offerings suit their individual needs. The Zoom and Browse approach is exemplified by Ask Systems (Bareiss & Osgood, 1993), a type of hypermedia navigation application that organizes information, typically stories, in a format based on the flow of human conversations. In Ask Systems all information is linked together according to the questions answered and questions raised by each individual story. When the user is presented with a story, they are shown an organized list of follow-up questions they can ask that will lead them to other stories in the system. Zooming is used here to refer to the beginning of this conversation process, where users are directed towards a particular starting story in a section of the network that is likely to contain the sorts of stories they will find interesting. Browsing refers to the subsequent process where the users navigate through the stories by following the conversational links provided by the system.

There is a substantial difference between conversational hypermedia and information retrieval, but the style of Zooming and Browsing that exists in Ask Systems can be adapted for the task of retrieving

cases from multimedia databases. For this purpose, we introduce a new twist on the previous uses of the Zooming and Browsing interface: rather than stepping through an organization of cases (e.g. stories in an Ask System) we propose that users step through an organization of *case indexes*. In this context, the Zooming process guides the user to some conceptual index that is related to the types of things that they are looking for. During the Browsing process, users navigate through the semantic links that organize the conceptual indexes, consider the options that the system offers, and select those that most effectively meet their retrieval needs. For example, a video producer looking for shots related to the use of computer technology in job training could, within a few mouse clicks, enter the semantic network in the areas of computers or job training to see what video clips the library has available on these topics.

One advantage to this approach to index selection is that it allows users to incrementally discriminate between cases in the database. In large databases, there may be a very large number of cases associated with any single index, so users may have to choose multiple indexes to reduce the number of retrieved cases. To facilitate the selection of multiple indexes, but to avoid forcing the user to guess which index combinations will result in matches, a system must dynamically control the selectability of indexes. Initially, all index concepts should be selectable (assuming that all the concepts are used in indexing some clip). But when an index concept is selected, many of the other index concepts must be disabled to prohibit the user from selecting a conjunction of concepts that would not result in the selection of video clips. Fortunately, only a small subset of all possible index terms are available at any given moment, and it is simple to dynamically determine the selectability of those concepts on the fly; using the notion of case-sets introduced earlier, an index is selectable if the intersection of its case-set and all of the case-sets of the already selected concepts is not empty.

The Zoom and Browse method of index selection can also be used as an interface for end-user indexing of new cases. As stated in section 1, we are currently using a very simple representation for storing cases in our systems, namely an unstructured set of individual conceptual indexes. If the indexes already in the conceptual organizations are sufficient to describe a new case, then by disabling the dynamic index selectability mechanism it is possible to use a Zoom and Browse interface to select all of the indexes for a new case and update those concepts' case-sets accordingly. Further extensions to the Zoom and Browse interface can also allow end-users to modify the conceptual vocabulary and organization when new indexes must be created to properly index a case. To support the data-driven development of an indexing vocabulary, we believe that end-users must be provided with a well-organized initial vocabulary and an intuitive indexing and editing environment.

To use the Zoom and Browse method for the purpose of index selection, the Zooming process must be effective enough to place the user into a section of the conceptual organization that is close to the concepts they are looking for. Likewise, the browsing process requires that the conceptual organization of indexes can be navigated by people unfamiliar with semantic networks as used in AI research. Determining the feasibility of these requirements for stock video retrieval has been one of the central focuses of our research. We are currently in the process of developing a large-scale video retrieval system and accompanying indexing tools based on the Zoom and Browse method. These systems, which are in the early stages of development, are described in section 3.

3. Retrieval systems for stock video

In September of 1994 we began our project to construct a stock video indexing and retrieval system for Andersen Telemedia, a medium-sized video production facility which is part of Arthur Andersen & Co., S. C. We started by collecting sample video clips, consulting with producers, and reviewing the video production literature. Design work focused on analyzing these sources to uncover an initial set of high-level conceptual categories for indexing, and settling on an interaction style and accompanying interface. Over time, we collected a larger representative corpus of clips, and extended and refined selected parts of the indexing scheme. As of September 1995, we have a second implementation of the interface, the underlying storage and retrieval architecture, and an initial corpus of 1000 indexed video clips. The system is constructed in Delphi running on a Windows PC, and makes use of a Paradox database to manage both the library contents and the conceptual indexing scheme.

As mentioned earlier, our initial forays into index development have focused on scene content description: we aim to record the most salient aspects of such concrete aspects of a scene as the place in which a clip was shot, the activities going on, and the people and things present in the scene. To organize the rich conceptual space of scene content indexes, we capitalize on the situated nature of most human activities, and on users' expectations about how activities will typically unfold. People expect activities

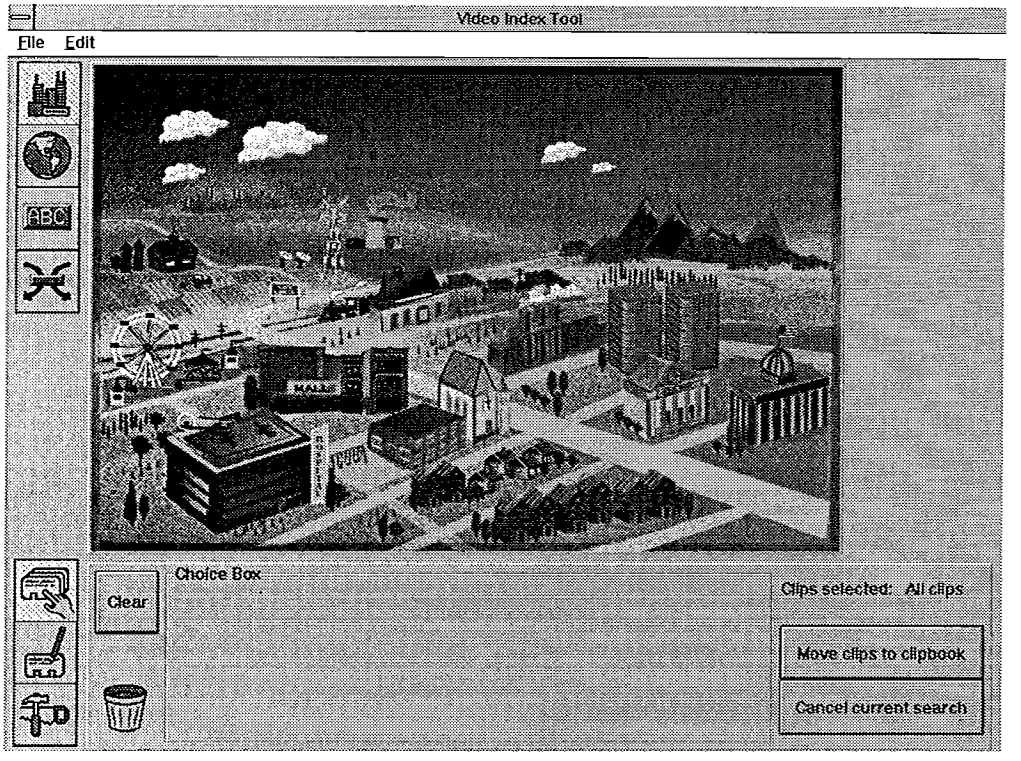

Figure 1: The stock video retrieval system showing the first picture of a zooming interface for types of places in the world.

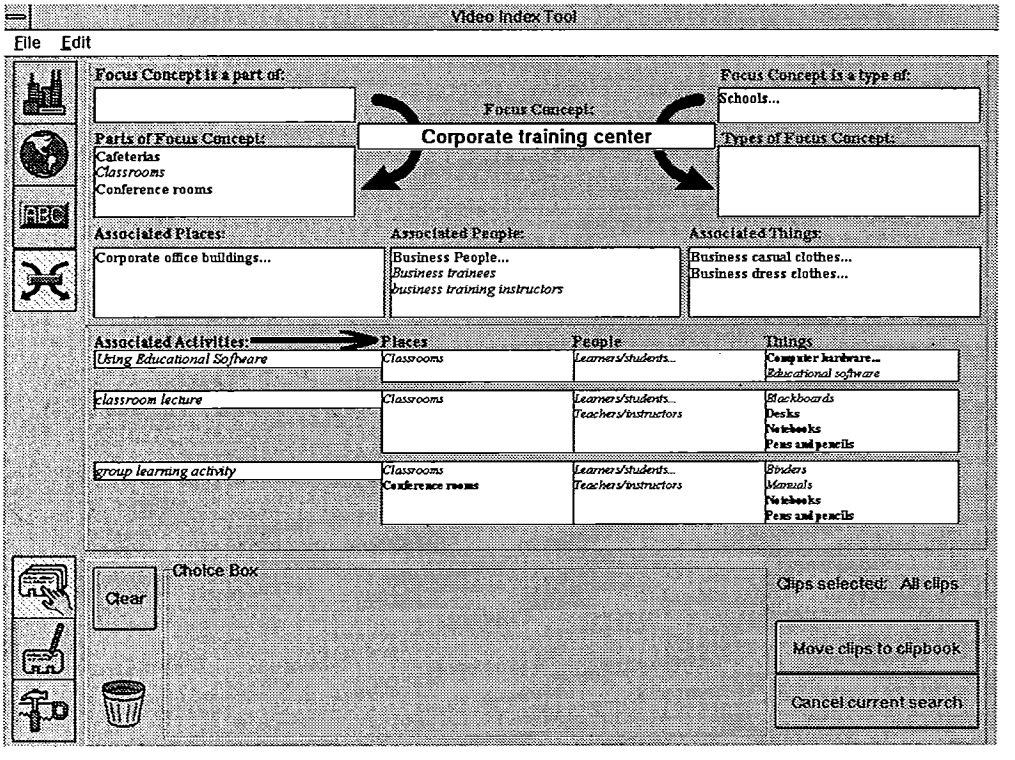

Figure 2: The stock video retrieval system showing the browsing interface with Corporate training center as the central focus concept.

to follow stereotypical sequences and to involve certain kinds of actors and props; they expect certain activities, people and things to be found in certain known places. As an example, baseball games, professional baseball players, and baseballs can all typically be found in baseball stadiums. We have developed a hierarchy of places which includes places with proper names as well as concepts for general places at various levels of abstraction. We use this organization as the backbone of a larger conceptual network that includes indexes for activities, types of people and objects. Much of our recent efforts have focused on elaborating the system's knowledge about activities, capitalizing on the progress that has been made in representing actions and activities in previous AI research (Schank, 1982).

The result is a set of networks organizing places, activities, people, and things that capture relationships both within categories (primarily taxonomic and partonomic relationships) as well as expectations about relationships across categories (primarily expectations about locations and activity participation). These networks directly support various facilities for zooming and browsing. We have developed several zooming and browsing views that offer alternate ways of exploring the space of index terms, all integrated into a standard screen layout as pictured in Figure 1. Initially, a user would be offered a zooming view such as the cartoon map shown dominating the top of the screen in Figure 1; starting with a picture offering "all the places in the world", a couple of clicks will establish a focus that can begin to support richer contextual browsing. For example, a user that is looking for clips about the use of technology in job training might begin by selecting the graphic for *Educational service places*. This will bring the user to a second zooming view with graphics for various types of educational places, including *Corporate training centers* and *Elementary and secondary schools*. If the user clicks on one of these graphics the system leaves the zooming view and displays the browsing view with the selected item as the center of focus.

Figure 2 shows a browsing view focused on the concept *Corporate training centers* as a result of the zooming process described above. This rather complex browser layout is intended to offer the user a coherent set of index terms forming a conceptual neighborhood around the designated focus. The box at the top and center displays the focus item (*Corporate training centers*). To its right are a pair of boxes showing abstractions and specializations of the focus (*Corporate training centers* are a type of *Schools*). To its left are a pair of boxes showing containers and parts of the focus (*Corporate training centers* have *Cafeterias, Classrooms*, and *Conference rooms* as parts). In the next row down the user is offered sets of directly related concepts drawn from the categories places, people and things. For instance *Corporate training centers* are directly associated with *Corporate office buildings*, *Business people*, *Business trainees*, *Business Training instructors*, *Business casual clothes*, and *Business dress clothes*. The bottom of the browser is devoted to an activity table in which each row offers index choices associated with some activity that is associated with the focus concept. For instance the first row of that table lists *Using educational software* as an activity that takes place in *Corporate training centers*. The boxes next to *Using educational software* indicate that this activity also takes place in *Classrooms* and is undertaken by all types of *Learners/students* using *Computer hardware* and *Educational software*. The central focus concept can be changed by double-clicking on any of the items on this browsing screen. When the focus changes to the selected item, all of the boxes are redrawn to reflect the links between the new focus and its associated concepts.

The zoom and browse views that dominate the screen layout allow users to explore the space of available index concepts so they can assemble descriptions of clips they want to see. At any point, a user has the option of selecting an index visible on the top of the screen and dragging it down to the Choice Box at the bottom of the screen. The system then informs the user how many clips it has that contain the chosen descriptor. If there are several indexes in the Choice Box at once, the system only looks for clips that are labeled with the conjunction of those terms. These calculations are performed using the clip-set intersection algorithm described in Section 1. To prevent the user from forming a request for multiple indexes for which there are no applicable cases, the selectability of each available index is calculated as described in Section 2. As the user moves indexes in and out of the Choice Box and moves from view to view, the system constantly provides visual indication of which concepts may be dragged into the Choice Box to usefully refine its current contents.

As this retrieval system continues to develop we intend to improve both the organization and graphics of the shallow place hierarchy to find the simplest possible means of Zooming to specific places. As our conceptual index organizations develop and our case library grows, we intend to expand this approach to all of the categories of indexes the we believe are applicable to stock video clips. We believe that the current version of our retrieval system provides an effective framework for the development of large-scale

systems that meet the real-world needs of users of stock video databases.

4. Conclusions

If computers are to be effective tools for multimedia and video production, they must be better able to manage large volumes of on-line video data. The challenges in accomplishing this do not end with low level coding, storage, and transmission technologies. We must also discover how to impose some useful organization on the accumulating data, and we must provide effective mechanisms for getting at relevant items on demand. Video is both conceptually dense (it can communicate many different concepts, and many different kinds of concepts), and computationally opaque (computers cannot extract the relevant concepts automatically). Accordingly, we believe it is necessary to develop rich conceptual indexing schemes that are simple enough to be practical for high-volume hand-coding of indexes.

With humans in the loop, the user interface for the system will be just as critical as the indexing scheme itself. We have illustrated how a rich but simple conceptual indexing scheme can be paired with a Zoom and Browse concept exploration system to yield an end-user retrieval, indexing, and conceptual editing tool. The result is a system that shows promise for effectively managing a large library of stock video clips. Further development of indexing vocabularies and scaling up the library of clips managed by the system will reveal how that promise holds up.

Acknowledgments

Thanks to Lon Goldstein, Smadar Kedar, Anil Kulrestha, Eric Lannert, Andre van Meulebrock, Jacob Mey, Craig Persiko, Kiku Steinfeld, Mindy Wallis, Linda Wood, and Raul Zaritsky, who all contributed to this research. Special thanks to Andersen Telemedia for providing information and materials for this work. The Institute for the Learning Sciences was established in 1989 with the support of Andersen Consulting. The Institute receives additional support from Ameritech and North West Water, Institute Partners.

References

Bareiss, R. & Osgood, R. (1993) Applying AI models to the design of exploratory hypermedia systems. In *The Proceedings of the Fifth ACM Conference on Hypertext.* Seattle, WA: ACM Press (pp. 94-105).

Chakravarthy, A. (1994) Toward Semantic Retrieval of Pictures and Video. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 12-18).

Kolodner, J. (1993) *Case-Based Reasoning.* San Mateo, CA: Morgan Kaufman.

Lenat, D., & Guha, R. (1994) Strongly Semantic Information Retrieval. *AAAI-94 Workshop Program for Indexing and Reuse in Multimedia Systems* (pp. 58-68).

Schank, R. (1982) *Dynamic Memory: A theory of reminding and learning in computers and people.* Cambridge, England: Cambridge University Press.