Multimedia Search: An Authoring Perspective

Stephen W. Smoliar
James D. Baker
Takehiro Nakayama
Lynn Wilcox
FX Palo Alto Laboratory, Inc.
3400 Hillview Avenue, Building 4
Palo Alto, CA 94304
Telephone: +1-415-813-6703 (Smoliar)
Fax: +1-415-813-7081
Email: {smoliar, baker, nakayama, wilcox}@pal.xerox.com

1. The Nature of Authoring

If we are to achieve a world in which hypermedia has become the basis of the documents we exchange, then that world will require that we first achieve the goal of authoring software that makes hypermedia documents as easy to create as text documents are through current word processing environments. However, we can only succeed in inventing authoring technology if we also invent the relation of that technology to the human behaviors of both writing and reading (Baker *et al.*, 1996). We already tend to identify authoring with writing, but it is often overlooked that the role of the reader is just as important to effective communication as that of the writer.

While we may all have learned to read from printed books, our appreciation of other media allows us to generalize those intuitions we have formed about reading. When we talk about "reading" a film, a painting, or even a personality evoked by a stage actor, there may be some argument as to whether or not we are using the verb "read" in a literal or metaphorical sense. We feel it is important that this view of reading *not* be taken as mere metaphor. Rather, adopting the terminology of Ferdinand de Saussure (1986), we wish to view the objects we encounter in printed books, films, paintings, and even the personalities evoked by actors as instances of a single collective called *signs*. Saussure proposed a new science, to be called *semiology*, "which studies the role of signs as part of social life." Semiology is thus concerned with both the nature of signs and how we "read" them as a fundamental aspect of our human behavior.

The attempt to characterize the nature of signs began with those that seemed the most straightforward, the linguistic signs (de Saussure, 1986):

> A linguistic sign is not a link between a thing and a name, but between a concept and a sound pattern. The sound pattern is not actually a sound; for a sound is something physical. A sound pattern is the hearer's psychological impression of a sound, as given to him by the evidence of his senses. This sound pattern may be called a 'material' element only in that it is the representation of our sensory impressions. The sound pattern may thus be distinguished from the other element associated with it in a linguistic sign. This other element is generally of a more abstract kind: the concept.

This approach is a useful one because it can be generalized to signs in other media that must be founded on not only a concept but also some set of sensory impressions that are perceived as a pattern. Saussure then assigned names to designate these two elements of a single sign. The pattern of sensory impressions was called the *signifier*, and the concept associated with that pattern was called the *signified*. Louis Hjelmslev then introduced the idea of "planes" for distinguishing the respective domains of signifiers and signifieds, described by Roland Barthes (1973) as follows: "The plane of the signifiers constitutes the *plane of expression* and that of the signifieds the *plane of content*."

What does this have to do with reading and writing signs? First of all, it is necessary to put aside any view of reading as a relatively passive task: A passive approach to reading is one in which the reader is not paying attention! An alternative approach, which has been introduced by Umberto Eco (1979), is that reading is an *activity* in which the reader, negotiating between signifiers and signifieds, *reconstructs* the material being read, transforming one body of signs into another. This approach is based on the hypothesis that a reader has understood a document when he is capable of recounting its content (and perhaps its expression) back to himself (or to anyone else). This reconstructive process frequently takes place in the reader's head, but if it is translated to some concrete medium, then, for all intents and purposes, it is another writing task. In other words an authoring technology that supports writing in a manner both efficient and unobtrusive may be equally valuable in the support of reading.

2. Where Does Search Fit Into the Processes of Reading and Writing?

Where does search fit into this story? While hypermedia existed before the rise of the World Wide Web, we now tend to think of the Web as the most significant repository of hypermedia documents. More important, however, is that, as that repository grows, it will become the *primary information resource* for authors preparing the documents it contains. In other words the Web is a potential paradigm shift of authors' attitudes towards libraries. After all, most authors still tend to consult a variety of resources as part of the normal course of a writing task. If those resources are now going to be available through Web pages, what will be the effect on the resulting writing behavior?

The nature of writing behavior itself has been an object of study for many research projects that we have reviewed elsewhere (Baker *et al.*, 1996). In an attempt to generalize the results of these projects from text to hypermedia, we have developed our own model of reading and writing illustrated in Fig. 1. For the purposes of this discussion, we shall not dwell on the generation of outlines and the rendition of a hypermedia document, concentrating instead on the role of search in the first two stages of this model.

When we think of a writer researching material in a conventional library, we usually envisage a process that begins by accumulating a pile of books and articles. Then, as each of those items is read, the writer tends to make note of specific *points*, traditionally assigning each point to the proverbial index card (Baker *et al.*, 1996).

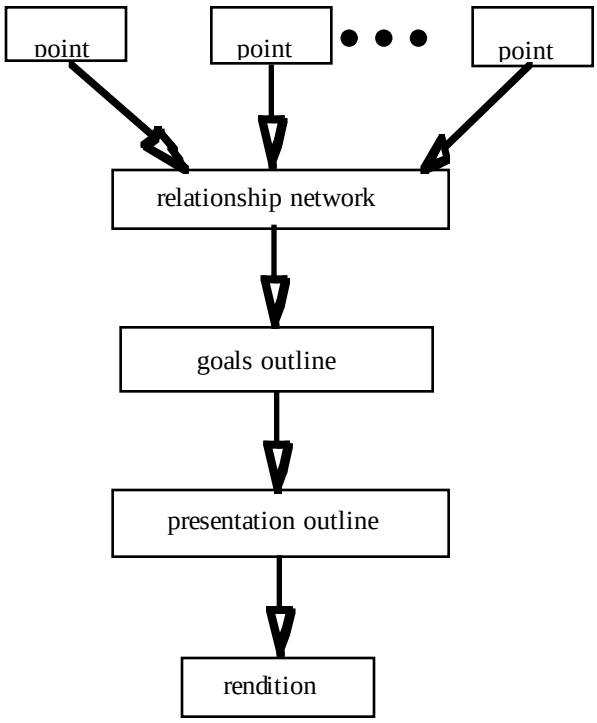

Fig. 1. Model of reading and writing

This need to *extract* such points from "the literature" will be just as important if that literature is on the Web as it would be in any other library. Sometimes, however, the writer has a specific point in mind that needs to be validated by consulting available sources. In this case that point becomes an *object of search*: The writer needs to associate it with a source document in order to establish a confirming citation or to help find additional material relating to the point being developed.

However, a writer does more than cull points from large quantities of reading matter. The actual *synthetic* side of writing begins with the hypothesizing of *relationships* among those points. One of the most important of these relationships is the organizing of points into groups, and such grouping leads to another approach to search. Once a group has been defined, one often wishes to add further points to it, seeking out additional confirming evidence. Thus, one point may serve to drive the search for other points, provided the nature of the grouping relationship can be suitably articulated.

Furthermore, extracting points and searching for related points can be just as important to reading as they are to writing. Eco's reconstructive approach to reading comprehension is often driven by the reader's ability to relate the content of a new

document to already familiar material. Thus, the reader may be as likely to undertake a "search-by-relationship" on a point identified while reading as a writer would be while engaged in a literature search. From this point of view, search may be one of the most fundamental activities for both writing and reading. If its significance tends to be overlooked, it is because so much of it frequently takes place in the subconscious of both the writer and the reader.

3. Implications for Different Media

The whole point of the model in Fig. 1, however, is that "points" need not necessarily be restricted to simple sentences of declarative text. Art historians see such "points" as particular regions of a canvas, musicologists hear them in melodic, harmonic, rhythmic, or even timbred gestures, and film theorists may make a point out of an entire *mise en scène*. We are currently working on a hypermedia authoring environment that supports the collection and management of points in any of these media, as well as the definition of relationships among those points (Baker *et al.*, 1996). However, this environment must support not only the accumulation of those points but also search processes through which they may be encountered and extracted.

From a semiological point of view, the target of a search is always a *sign*. However, the terminology of semiology allows us to be more specific. A search operation is usually confined to only one of Hjelmslev's planes, and explicitly identifying that plane may facilitate expressing the search target more clearly. This distinction may be found in the data model for the Visual Information Management System (VIMSYS) (Gupta, Weymouth, and Jain, 1991) developed under the InfoScope project at the University of Michigan. That model partitioned information into planes of its own (not to be confused with Hjelmslev's planes) that distinguished data derived from *images* (Hjelmslev's plane of expression) from those derived from *domain objects* being represented by those images (Hjelmslev's plane of content). Bearing this distinction in mind, let us now examine the nature of search with respect to the media of text, images, audio, and video.

3.1 Text

When one has a specific pointer (URL) to the target text, the World Wide Web works very well. However, this is not always the case. When one doesn't have a pointer but specific text in mind, he might want to use the current keyword search technology. There are some search engines available on the Web (e.g., http://www.altavista.digital.com/); but there is no guarantee that their search space includes the target text. They can search only a limited part of the Web, because it is impossible to take a snapshot of a structure that may be changing at any place at any moment. Furthermore, it is often a difficult task to compose a set of keywords that selectively represents the target of the user's search.

Following the model shown in Fig. 1, writers need to try to relate as many relevant combinations of text as possible when they want to accumulate points and organize them into a relationship network. Fortunately, because the text on the Web is machine-readable, we can use it not only as targets of search but also for assistance in formulating queries. In other words one uses information from the plane of

expression of one document as cues for searching the plane of expression of other documents; and, as we learn more about which features on the plane of expression tend to yield the best search results, we may use those features as a basis for indexing, rather than the conventional keyword-based approach. However, because we are searching on the plane of expression, rather than the plane of content, we must be satisfied with only a list of potential candidates, many of which may have nothing to do with the target the user had in mind. Nevertheless, by beginning the search on the plane of expression, analysis on the plane of content, which is far less susceptible to automation, may be concentrated on the candidate set, rather than the entire search space.

3.2 Images

The better part of this decade has seen considerable effort towards advancing technology in "content-based image retrieval" (Furht, Smoliar, and Zhang, 1995). Unfortunately, none of this work really has anything to do with content in Hjelmslev's sense of the word. A humbler awareness of history and a more generous acknowledgment of Hjelmslev's perspective would rightfully call this technology *expression*-based retrieval.

From a Saussurean point of view, expression-based retrieval would be driven by sound patterns. However, if we are interested in images rather than language, we may assume that signifiers consist in patterns of visual, rather than auditory, stimuli. If such patterns are to drive our abilities to perform search operations, then we need to understand their nature before we understand how we may exploit them. What, after all, *are* patterns of visual stimuli?

One fruitful way to approach this question is to consider the research of Gerald Edelman (1987). For over a decade Edelman has been investigating the construction of automata that not only are capable of general visual perception but also have designs that reflect our knowledge of the neural architectures that support *biological* perception. Edelman's model takes a two-stage approach to the perception of visual patterns. The first stage is the recognition and classification of *image features*, while the second stage identifies patterns in terms of relationships among those features.

From this point of view, even expression-based retrieval still comes up short, because almost all technological advances have addressed the detection and representation of image features. It *is* true that this technology can now provide mathematical models for a variety of these features, including color, texture, object shape, and edge relationships of the sort one might encounter in a hand-drawn sketch (Furht, Smoliar, and Zhang, 1995). However, while the technology for representing an object's shape is promising, the segmentation technology that identifies an object in the first place is far weaker. As Edelman (1987) has demonstrated, that technology requires building beyond the recognition of features to the more sophisticated identification of actual patterns. Consequently, even the representation of features such as color and texture still cannot be readily associated with objects, but only with either the entire image or some isolated region of that image (determined *a priori*).

This means that, if we wish to include the search for image signs in the course of either writing or reading, we shall have to reduce our expectations of what such a

search is likely to retrieve. First of all, we need to provide support in expressing our query in a form that has only to do with image features, rather than patterns of those features, let alone any properties on Hjelmslev's plane of content. Thus, if we wish to formulate such a query based on a specific source image, we must first reduce that image to its features and then decide which of those features will be incorporated in the query (and how they will be incorporated). Finally, we have to accept that searching on such a query can only be viable if it yields a set of candidates, some (if not many) of which may have nothing to do with the content we had in mind and many of which may only be valuable to the extent that they can help us formulate a more accurate query (Furht, Smoliar, and Zhang, 1995). Clearly, this is not the best of situations; but, on the basis of some experiences with such feature-based retrieval systems (Seybold, 1994), it *does* seem to be better than nothing.

3.3 Audio

In some cases, the author of a hypermedia document may want to create textual links to audio. For example, in an educational document on various types of musical instruments, the author may want to provide links from the instrument names to their pictures as well as to sound samples produced by each instrument. While it is reasonable to do this for a small number of examples, it is not feasible to label manually large collections of music by the instrument being played. However, using techniques similar to speaker identification and segmentation (Wilcox *et al.*, 1994), it may be possible to search arbitrary collections of music for examples of a particular instrument. This capability would not only provide the author with an easier means to locate musical examples, but also allow the reader to enrich the document by locating examples on his own.

Also, in a hypermedia document analyzing a particular piece of music, the author may want to point out to the reader a particular melodic theme. While the author might highlight the melody line in the musical score, it would also be informative to highlight the melody in the music itself. This requires first that the melody be found within the context of the entire musical piece, and second that the melodic line be segmented from the accompaniment. Locating the melody within the piece could be performed using techniques for wordspotting in speech (Wilcox and Bush, 1992). Segmentation of the melodic line, however, requires partial transcription of music.

Finally, it is important to recognize that there is more to audio than speech and music. There is a broad variety of other sound objects that may be targets of search operations. Classifying these sounds is no easy matter (Smoliar, 1993). Various descriptive ontologies have been proposed, but none have been particularly effective. Indeed, in the case of film and video, the evidence strongly suggests that the classification of a sound is tightly coupled to what the audience sees while the sound is heard (Metz, 1985). Thus, while feature-based search may be the only viable approach, there is a significant danger that its performance for audio will be far weaker than is currently achieved with images.

3.4 Video

For all intents and purposes all problems concerned with searching video are subsumed by problems concerned with searching both images and audio. Thus, as is the case with images, we cannot really expect video search to rise above the level of stimuli features, let alone accommodate any technique that is truly content-based. The best we can hope for is that a repository of video source material be indexed with respect to one or more *abstractions* of that content.

The simplest technique is to formulate abstractions in text and reduce the problem to search in the text domain. While useful results based on free text are feasible, search may be facilitated by constraining the text according to different "styles of discourse." One such style is stratification, where text is constrained to account for a time table of events (Aguierre Smith, 1992). Another style is that of object-oriented frames (Smoliar, Zhang, and Wu, 1994). In this case the "text" is basically constrained to forms similar to that of a propositional calculus, and search may require supplementing conventional matching with inference on those propositional forms. Yet another style is that of the database record, and a variety of different approaches to data modeling for video databases have been proposed (Furht, Smoliar, and Zhang, 1995).

Alternatives to text styles have not yet been explored very extensively, but several have been investigated (Furht, Smoliar, and Zhang, 1995). One possibility is to reduce the problem strictly to one of image search. This is achieved by first reducing the video to a collection of static images. If these images are *key frames* extracted from the individual camera shots, this approach becomes feasible with respect to the amount of data that must be stored and subjected to search. However, the method also inherits all of the shortcomings of the current state of the art of image search. Another possibility is to abstract a video strictly into its camera movements. These movements may be regarded as features at the video level, rather than the image level. They are characteristically unique to video and film data. The problem is that it tends to be difficult to formulate useful queries based on these features. Finally, there are features concerned with movement in the entire image frame, rather than just movement due to the camera. These features (or some approximation of them) are often explicitly detected and recorded as part of a compressed representation of a video source. If the video is stored in that compressed form, then searching based on those features does not require that the images first be reconstructed. Again, however, the key problem is one of being able to formulate useful queries in terms of those features.

4. Bibliography

Aguierre Smith, T. G., If You Could See What I Mean … Descriptions of Video in an Anthropologist's Video Notebook, Master's thesis, Massachusetts Institute of Technology, Cambridge, September 1992.

Baker, J. D. *et al.* Reinventing Reading and Writing in the Context of Hypermedia, submitted to the International Working Conference on Integration of Enterprise Information and Processes (November 1996).

Barthes, R. *Elements of Semiology*, translated by A. Lavers and C. Smith, New York: Noonday, 1973.

Eco, U. Introduction: The Role of the Reader, *The Role of the Reader: Explorations in the Semiotics of Texts*, Bloomington: Indiana University Press, 1979, pp. 3–43.

Edelman, G. M. *Neural Darwinism: The Theory of Neuronal Group Selection*, New York: Basic Books, 1987.

Furht, B., Smoliar, S. W., and Zhang, H.-J. *Video and Image Processing in Multimedia Systems*, Boston: Kluwer Academic Publishers, 1995.

Gupta, A., Weymouth, T., and Jain, R. Semantic Queries with Pictures: The VIMSYS Model, Proceedings of the 17th International Conference on Very Large Data Bases (September 1991), pp. 69–79.

Metz, C. Aural Objects, *Film Sound: Theory and Practice*, edited by E. Weis and J. Belton, New York: Columbia University Press, 1985, pp. 154–161.

de Saussure, F. *Course in General Linguistics*, edited by C. Bally and A. Sechehaye with the collaboration of A. Riedlinger, translated and annotated by R. Harris, La Salle: Open Court, 1986.

Seybold, IBM Unleashes QBIC Image-Content Search, Seybold Report on Desktop Publishing, Vol. 9, No. 1 (September 1994).

Smoliar, S. W. Classifying Everyday Sounds in Video Annotation, *Multimedia Modeling*, edited by T.-S. Chua and T. L. Kunii, Singapore: World Scientific, 1993, pp. 309–313.

Smoliar, S. W., Zhang, H.-J., and Wu, J.-H. Using Frame Technology to Manage Video, Proceedings: Second Singapore International Conference on Intelligent Systems (November 1994), pp. B189–B194.

Wilcox, L., and Bush, M. Training and Search Algorithms for an Interactive Wordspotting System, Proceedings: International Conference on Acoustics, Speech and Signal Processing (March 1992), pp. 97–100.

Wilcox, L. *et al.* Segmentation of Speech Using Speaker Identification, Proceedings: International Conference on Acoustics, Speech and Signal Processing (April 1994), pp. 161–164.