Editor: Nevenka Dimitrova
Phillips Research

Multimedia Technology in Context

Sharon Flank
DataStrategy

Media management relies on a range of techniques, some of them at the cutting edge. As we move beyond the tolerant early adopters into a user community with a new set of expectations, we can expect new demands. Although early adopters might be lured by the excitement of new technology and willing, even eager, to change the way they do business, the far larger community of media professionals demands that any new technology fit smoothly into their existing processes.

As with any potentially disruptive business process, the key starting point for multimedia application development is the users: Who are they, and what do they want to do? In this article, I examine usability in digital media management, with special attention to searches, which most users employ repeatedly. My insights are drawn from almost a decade of experience with a business-to-business multimedia search application, which has evolved into the photography portal PictureQuest (http://www.PictureQuest.com, see Figure 1). My work with both end users and corporate media archivists has led me to conclude that a focus on user context is the next big hurdle we face to ensure the creative community will embrace our technologies. Planning system design and search strategies from the users' viewpoint increases the likelihood that their needs will be served, both when the system is new and as they come to accept it as part of the creative process.

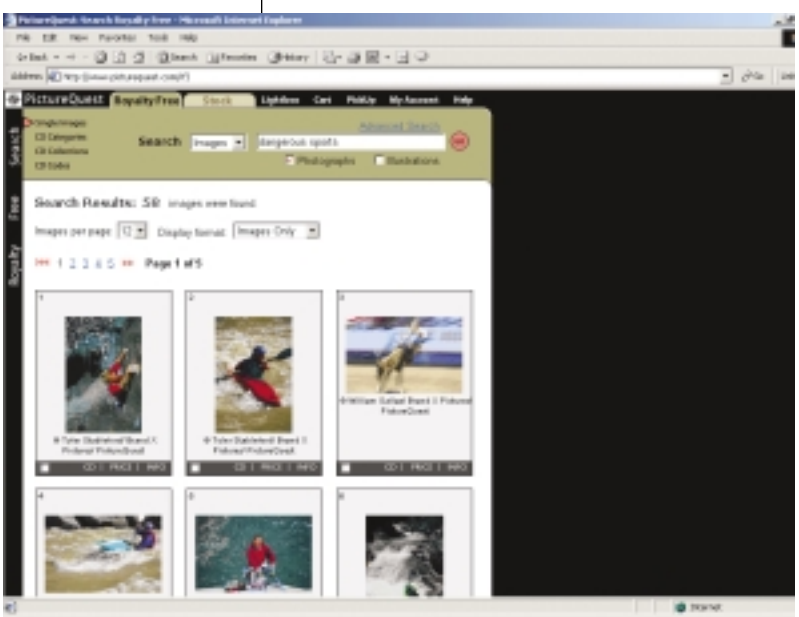

Figure 1. PictureQuest processes tens of thousands of image searches every day.

User profiles

A lot of different kinds of people access multimedia, but only some of them are important in a business context. In general, we face a cultural divide. The key business users are artistic and visually oriented, the kind of audience that Apple has appealed to historically. Creative professionals search media repositories, either internal to their companies or in outside libraries, to find material for advertising, publishing, feature films, and Web sites. On the other hand, Web surfers in general look for pictures and videos on the Web so they can plan their vacation, see what a monarch butterfly chrysalis looks like, and so on. Among the most popular media searched for are images of celebrities, including sports icons and movies stars. Pornography is also popular and often drives technological innovation because its consumers are willing to pay for the multimedia experience. Electronic commerce is significant as well, either in conjunction with the uses I've just described, for use in catalog sales, or even simply for purchasing a copy of the media object itself.

To build a successful application, the software developer needs to know not only who the users are, but what they want and how the rest of their business operates. A software solution should be crafted around the user's defined, documented business needs, using technology appropriate to the task.

A business multimedia application must meet these criteria:

- *Speed*. One to two second response times are essential because time is money.

1070-986X/02/$17.00 © 2002 IEEE

- *Usability*. The application must be comprehensible at a glance and usable for hours on end. These two criteria can suggest different approaches and melding them can be a challenge. Figures 2 and 3 show some examples of how MediaPartner, the core software underlying PictureQuest, has streamlined a digital media management user interface to balance those requirements.

- *Adequacy*. Get the right answer often enough to be useful. This will vary by domain, although in my experience customers don't like to be told that their requirements are less stringent, even when they are. It might be that certain retrieval techniques are accurate enough to find a particular replay in a sports domain but not fine enough to pick out the right segment from a library of news videos.

- *Accuracy*. The user's experienced accuracy is more important than a scientific measure. It makes sense to focus on search precision over recall, letting in some false negatives so as to screen out the false positives. Why? Users' confidence is eroded if they see junk, so precision must be high. Few users know what is in the collection, so recall is of less importance. When we fine-tuned PictureQuest, we were amazed that users were so unconcerned about recall (although photographers who remembered each picture they had submitted were much more interested). Furthermore, what is displayed on the first retrieval screen is critical because users are lazy, and if they don't like what they see, they may become discouraged and drift off. Given the combination of preferences for precision and the first screen, the classic information retrieval test of "Precision at 20" may be the best measure of practical accuracy. This test refers to the idea that the precision of the top 20 images returned is both easier to measure and more reflective of the user's perception of system accuracy.

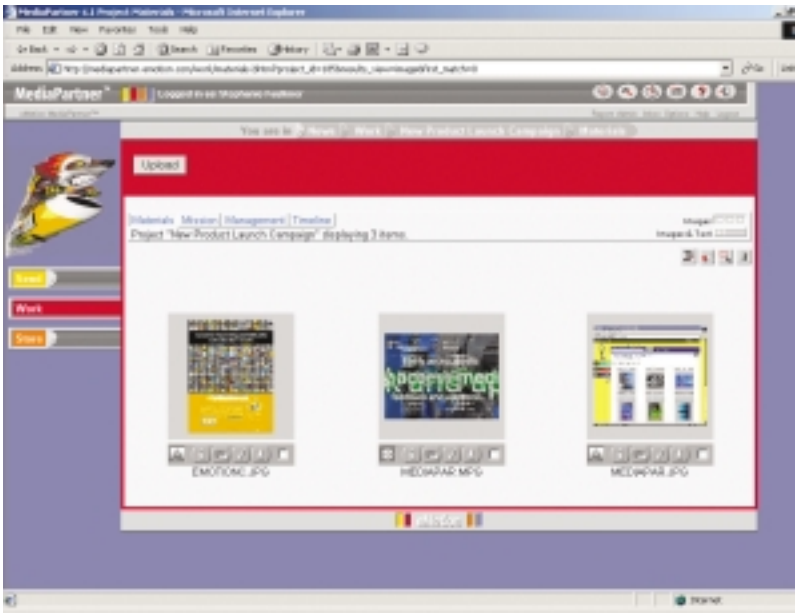

Figure 2. A spare, white background lets the user focus on the images themselves.

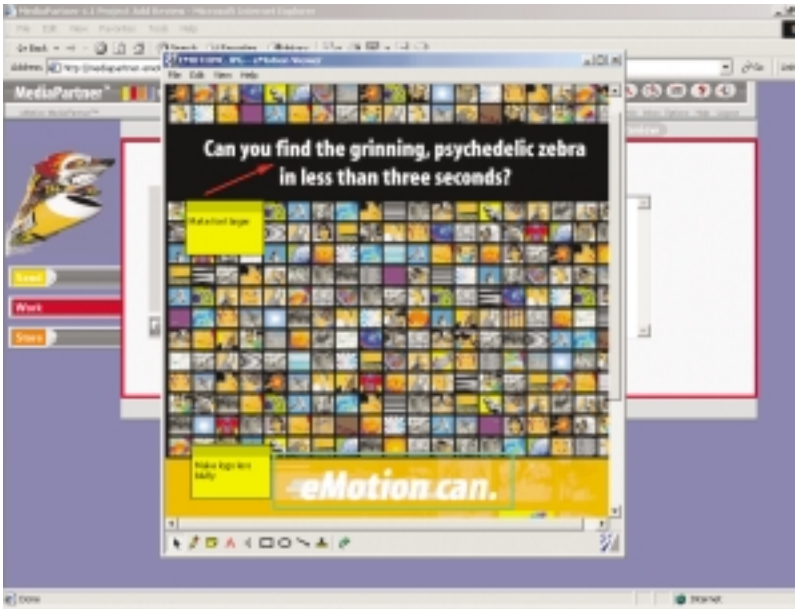

Figure 3. Annotation tools facilitate creative collaboration.

Search options

We can search multimedia files using a variety of search techniques or by combining two or more techniques. There is nothing natural about a keyboard or a mouse-and-menu interface. In real life, if you want help finding something, you explain what it is in words, you point to it, or if you can draw, you draw a picture of it. In the long run, we should aim to take advantage of human intelligence where we can and gather input as elaborate as we can. If, in the process, we end up asking users to do things that are more natural, such as explaining, pointing, and drawing, then we are combining the best processes. The following options aren't currently robust search option possibilities, but we can expect

Creating fielded data is the obvious choice for characterizing objects, and multimedia is no exception.

them to supplant the artificial interaction modes once they become effective.

Voice input

Speech-recognition technology makes it possible for computers to understand the words in a voice query. Speech-to-text takes the spoken signal and turns it into written words. The initial work in speech recognition focused exclusively on signal processing, ignoring context completely. We've now reached the point where additional advances will rely on linguistic insights about what the user is actually saying.

Advances in speech recognition will require moving beyond the signal and a vocabulary list to include syntax (word-ordering information) and even discourse analysis (pronoun reference and other cohesion devices). This development parallels your intuition—it's easier to write down what you hear if you understand the language you're hearing. On the other hand, it's not so easy to take dictation in Hungarian or Korean if you don't understand what is being said.

The use of smaller devices, with smaller keyboards or odd styluses, is also a spur for implementing voice-based search. The cultural issue is significant here as well because creative media professionals might be more likely to use voice input because they aren't necessarily eager to write, spell, or type.

Drawing

The software engineers who build multimedia systems might find it difficult to imagine using a drawing as input, but the creative professionals who use the system to find material for their artwork are generally quite comfortable with drawing.

The techniques used to match a drawing will rely on shape and perform template matching to select the elements in the database that most closely resemble the sketched input. Drawing is a complex search mode because it operates in a multidimensional feature space, forcing the search engine to consider many parameters at once.

Fielded metadata

Creating fielded data is the obvious choice for characterizing objects, and multimedia is no exception. Certain temptations, however, lead system designers astray. If we have information available, we want to keep it, and if it's neatly categorizable into separate fields, that seems, from a database standpoint, to be the right thing to do. But there are pitfalls in creating a separate field for everything. Some poor person is going to have to look at those fields, and someone is going to have to fill them in and remember what goes where. Fields are good for noncontent information, such as date created and artist name. Fields aren't good for content information, like "contains minorities" or "over 40 years old." This information is best folded into a more open-ended description. The conflict comes from trying to take an essentially infinite set of description possibilities and turn it into a finite data checklist. Either information will be lost in the transition, or the resulting structure will be horribly unwieldy.

Along the same lines, beware of the temptation to insert dozens of searchable fields. In my experience, users don't like to fill out big forms. PictureQuest began with an attractive interface with a dozen fields to fill in. We logged every query and gradually realized that no one was using our beautiful interface—everyone used a single search field. Even when they should have used other fields, they didn't. Our photo research department was constantly fielding calls from users who wanted only vertical pictures but had not bothered to select that field and from users who wanted only model-released images but hadn't clicked on that either.

Nowadays, users' experience with Web search engines has led them to expect even more dogmatically that one search field will be sufficient. Think about your own Web search experience. Even though you're an expert, Do you use the advanced search features on the big form, or do you try to type a few words into a single field first, hoping that the engine will be smart enough to find what you want?

Go ahead and include an advanced search option if you must (and many of our customers have demanded it). Figure 4 gives an example of the advanced search option in MediaPartner 4.1.

Do try to convince your customer that advanced search should be an option rather than the default choice because most users will select painless and instant over tunable and highly accurate.

Touch screen

With the advent of image-recognition techniques that use information about a specific region of the image, we should expect that users will choose to identify a region of an existing file and ask for matches (such as color, shape, and texture). Although this selection can be accomplished with a mouse or a pen-based input device, some people will surely just want to touch the screen and point to the part they want.

Natural language

When users attempt to search by matching descriptive metadata, they ought to be able to use normal English, with natural-language processing providing the supporting intelligence. Users shouldn't have to navigate an unfamiliar hierarchy or know the exact words used in cataloging. They simply type in what they want to see, in plain English (or another natural language), and the best matches pop right up at the top of the list. The ability to find files in one pass eliminates the need to traverse unfamiliar categories or search again within an imprecise search. The following sections describe techniques that comprise natural-language search.

Part-of-speech tagging. One simple, effective way to understand what exactly a user is looking for, and whether a particular description matches it, is to use part-of-speech tagging. This technique, by now quite mature in natural-language processing, distinguishes nouns, verbs, adjectives, adverbs, and other, less significant parts of speech. It can also help determine when a word has a particular meaning—for example, crane (your neck) or crane (a bird or piece of equipment).

As a result, part-of-speech tagging limits spurious choices and bad synonyms. It's approximately 99 percent accurate. Primarily statistical in nature, it operates quickly and reliably.

Stemming and morphology. Users don't want to have to type in every variant of a word, and even wild cards can be frustrating for the novice user. Natural-language search uses the smarter cousin of stemming, called *morphology*, which understands more about what constitutes an ending and how to figure out which root word to match against, for example,

- run-s;
- run-ning, not runn-ing; and
- ran yields run.

Pattern matching. Recognition of names and noun phrases can make a retrieval system much more capable. We can accomplish both using a pattern matcher. For names, the natural-language engine incorporates a data file of first name variants, for English and other languages—for example, Bill, William, Billy, Will, Willy, and Guillaume. It also uses a pattern for the ways that we express names in English, paying special attention to which elements can be omitted or added and still refer to the same person, such as a middle name, middle initial, or Jr. Thus, we can refer to George W. Bush as George Bush (while still preferring the exact match that does contain the middle initial, so as not to confuse him with his father) but not as Jeb W. Bush, even though two out of three elements still match. The patterns for other languages are different. For example, in Spanish, we can include the mother's surname at the end or not, so that Juan Veracruz Lopez is the same person as Juan Veracruz but not the same as Juan Lopez.

Recognizing modifier–noun groupings makes more precise retrieval possible. Adjectives and

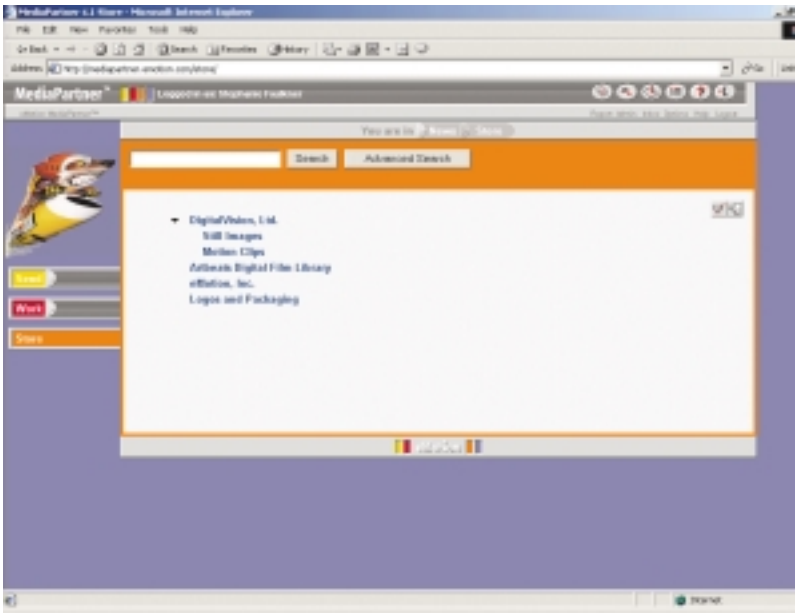

Figure 4. MediaPartner's standard interface hides advanced search options because few users employ them.

Making search effective isn't simply a technological question. It requires a clear understanding of what the intended use is.

nouns that serve as modifiers are less central to a match than the head noun and should be ranked accordingly. Thus, for the query "fire truck," it wouldn't be appropriate to rank the matches with "truck fire" as 100 percent correct. Furthermore, if you are searching for tiger, then tiger shark is hardly an optimal match.

Semantic net. Searchers don't want to play a vocabulary guessing game. They don't want to guess which words were used to catalog a file before they can find it. Similarly, cataloguers save time and money when they can simply describe files using normal English, without having to select particular words from a controlled vocabulary or maintain an ever-changing thesaurus. A *semantic net* can support natural-language search, freeing both searchers and cataloguers from fruitless games.

What exactly is a semantic net? It moves beyond a thesaurus to include other relationships besides synonyms, including hierarchical terms, part terms, and other relations. Optimally, it should be possible to share the task of creating and maintaining the semantic net across all the software applications that have a need for it, so that specialists are responsible for its creation and all users can contribute to its improvement.

The best model, then, is to start with a resource and then tailor it, rather than starting with nothing and having each organization reinvent the wheel. One recommended starting point for English is WordNet, a resource created at Princeton University with US government funding. WordNet contains more than 100,000 terms and is available on the Web at http://www.cogsci.princeton.edu/~wn.

We can always integrate additional resources, but starting from scratch is a bad idea because a large manually created resource might only reach 30,000 terms, while still requiring enormous maintenance overhead.

WordNet incorporates multiple relationships, including

- synonymy (cougar to puma to mountain lion),
- hyponymy (hierarchy terms, such as bird to owl), and
- meronymy (part–whole relationships, such as snow to snowflake, beach to sand, and car to brakes).

Some people believe that reasoning is appropriate for metadata management. They believe that when they encounter, for example, crimson robe, they should invoke a reasoning module and a full-scale artificial intelligence knowledge base. With this approach, they can determine that crimson is a kind of red, robe is a kind of garment, and garment is a kind of physical object, and therefore it has the characteristic "has_color." This is considerable overkill, especially when WordNet will tell you that crimson is a kind of red, red is a color, and robe is a garment. The fact the crimson modifies robe tells you everything else you need to know to support intelligent search and retrieval, with far less processing overhead.

Discourse-based strategies. Although commercial information retrieval hasn't yet embraced discourse strategies, eventually they will become important. A user will point to an image and request one that is "like this but with dark hair." Such a request is complex because it assumes both a model of an existing query or file and knowledge of where the substitution should be made. Research in natural-language processing continues to improve discourse processing.

Another way to exploit discourse is by combining image search and textual information. If an image caption refers to someone in the picture as "left, seated," then face detection can be used to find the person who is at left and seated so that the name can be attached to the right person.[1]

The importance of discourse in time-based media is still a research issue. In a video segment, two concepts might appear close together and be tightly linked, or they might appear close together but in two different topic segments and there-

fore have no link. Discourse analysis will help us determine when to link nearby concepts together.

Usability issues

Making search effective isn't simply a technological question. It requires a clear understanding of what the intended use is. When creating the design for metadata (descriptive information about a file, separate from its content), ask

- How are people going to search for this?
- What are the reuse options?

Too often, software engineers ignore the end user, often a creative professional, and approach the metadata question as if they were designing a database to be read by a machine rather than a system people will use. Questions that aren't helpful (but often lead software developers astray) include

- What categories of information can I attach?
- What will people need to know once they have found what they want?

We can give users information once they've successfully found a useful file—for example, this image was shot with a particular type of camera, it can't be used after 2003 or in Canada, and so on—but those are supplemental data, not searchable data.

To build a system that enhances business productivity, first model the types of users and determine what functions they use. What do they do most often? For how long at a stretch will they use a particular functionality? Do they use the system often enough to remember shortcuts?

Beware of technology for the sake of technology. Instead, ask what the technology buys you. Does it make the application more effective, scalable, likely to survive 10 years? The second and third waves of users will be less interested in innovation and more concerned with preserving their existing business processes.

Also beware of jumping on a bandwagon before it makes business sense. Is face recognition mature enough to be helpful? Object recognition? If you can't distinguish a pencil from a flagpole or a deer from a carousel horse, will your customer still want to pay for the technology? If face recognition works only when the person looks directly at the camera in good lighting with a neutral expression from not too far away, is that good enough for you? In what constrained environments could these advanced technologies be useful? Businesses used voice recognition when all it could manage was "Say or press 1." If your application can limit the scope sufficiently so that a nascent technology is still worthwhile, then its limitations might not matter to you.

Natural-language search might soon be part of basic operating systems. Will there then be a need for specialized natural-language applications? In the end, users will realize that optimized search engines can be useful for searching different kinds of text and that media poses particular genre challenges that won't be handled by the standard OS-based searches.

Conclusion

Media management relies on a range of techniques, some just emerging. Planning system design and search strategies from the users' viewpoint increases the likelihood that you and your systems will serve their needs, both when the system is new and as it ages. Open-ended data should be handled with open-ended methods, like natural-language processing, and unnatural interactions such as keyboards and menus should be minimized. Only then will media applications truly serve their business context. MM

References

1. R.K. Srihari, "Automatic Indexing and Content-Based Retrieval of Captioned Images," *Computer*, vol. 28, no. 9, Sept. 1995, pp. 49-56.

Readers may contact Sharon Flank at sflank@post.harvard.edu.

Contact Visions and Views editor Nevenka Dimitrova at Phillips Research, 345 Scarborough Rd., Briarcliff Manor, NY 10510, email nevenka.dimitrova@philips.com.