Frank Nack
CWI, Amsterdam

Kooks, Obsessives, Sturgeon's Law, and the Real Meaning of Search

Jonathan Foote
FX Palo Alto Laboratory

Collecting and organizing is a basic human activity. Watch children at play and notice how much of it is about imposing order on the environment—for example, collecting all the white pebbles in the toy truck, arranging toys by color, or collecting Pokémon cards. In adult life, if you don't collect things yourself, you surely know someone who does. Whether it's Pez dispensers, antique cameras, or ceramic figurines, the urge seems universal. Even if you're personally not an eBay addict or a model train fanatic, you most likely have collections of things that you like to have around, even just the addresses in your address book or your browser's list of favorite URL bookmarks.

Consumer need = collection

Many of the most popular consumer gadgets seem designed to enable collections. By fixing sound into a collectible object—the phonograph was one of the earliest examples—the obsessive record collector has become a staple geek stereotype. Around the turn of the last century, the Kodak Brownie camera took photography from the realm of the specialist to a consumer hobby and gave birth to the family photo album (as well as the ubiquitously annoying shutterbug relative).

The VCR enabled collecting movies and video—it turned television, formerly as ephemeral as audio was before Edison, into something the average consumer could capture and collect. I recall the vast shelves of neatly labeled videotapes collected by a friend's obsessive father. I still wonder what fraction of these **he** ever watched. Lately all this has made the digital transition, with the iPod and TiVo perhaps the latest—and least plausibly capitalized—entrants in the race.

Digital storage allows more things to be collected, and ever more cheaply at that. Photos, addresses, MP3s, even videos barely make a dent in today's capacious hard drives. You might even argue that the blogging phenomenon is a result of the urge to collect the thoughts, links, and musings that you previously considered random. Social bookmarking sites like de.licio.us (http://del.icio.us/) let you manage—and share—your URL collection. Meanwhile, social networking sites like Friendster (http://www.friendster.com/) let you collect … well, people.

Editor's Note

This column is about the human need for collection and the resulting request for retrieval. The conclusions drawn by the author what all this means for media search culminate in a proposition on what research on media search should address.

—Frank Nack

Collection = organization

Now with all this stuff—Pez dispensers, photos, and MP3s—comes the urge to organize. Hence there is iTunes, your shoebox full of newspaper clippings, and the numbers stored in your cell phone. Most people have these basic organizational strategies, but some don't stop there—they want to organize everything. I have a particular sympathy for the listmaking kooks and oddballs who feel compelled to organize the world—for example, Samuel Johnson, who compiled one of the first English dictionaries. And I'm certain similar kooks were behind the first encyclopedias and thesauri. (Although this listmaking tendency is clearly some sort of obsessive disorder, it's arguably benign, and certainly resulted in some useful reference books.)

Continuing in this trend are people like Melvil Dewey,[1] whose eponymous decimal system attempts to numerically categorize the whole of human knowledge. Modern adherents of this creed—or mania—see the computer as the magic tool that will help them encapsulate all knowl-

1070-986X/05/$20.00 © 2005 IEEE

edge. Although it's sometimes difficult to differentiate this from the equally quixotic compulsion to build a thinking machine, the former is often used as a prerequisite to the latter. The idea goes something like "A machine needs an internal model of the world to understand it." From this we get projects like Open Mind Commonsense (http://commonsense.media.mit.edu/cgi-bin/random.cgi) and CYC (http://www.cyc.com/cyc). These projects aim to teach a reasoning system common sense by encoding real-world knowledge into machine-readable form.

Organization = strangulation

In the end, although I admire the impulse, I'm not sure the effort spent on these artificial ontologies will ever amount to much (although I would like to be proven wrong.) The problem is that while they may be useful for particular narrow domains, they're fragile in the grand Noosphere of what people nowadays call reality. Get one step outside of your domain, and your ontology doesn't fit any more.

Here's a fun little game: For any existing classification scheme, think up counterexamples that break it. For example, if your commonsense reasoning system knows about the real world, it should have the concept of a "chair," and somehow attached to that, facts like "a chair is for humans to sit upon," and "a chair has legs."

Let's play the game and think about a miniature dollhouse chair. It's indisputably a chair—it has legs, yet you can't sit in it. And what about those inflatable exercise balls—you can easily sit on them, but do they have legs? What about the legless antigravity chairs yet to be invented? Or car seats? It looks like we'll have to adjust our concept of chairs (or legs) to account for these special cases.

However, you should beware. Just as the adherents of Ptolemaic astronomy needed complex epicycles and eccentrics to fit a mistaken geocentric model to real observations, the hacks and special cases needed to make an ontology work in messy reality should be a warning sign that the underlying system has problems.

So this is the drawback of ontologies: You need to classify things in advance, and even as you do that, the special cases and exceptions to those can proliferate infinitely into a recursive fractal nightmare. Not only this, but when you classify things you make an assumption of what questions will be asked by the scheme that you create. However nobody, and nothing, can generate sufficient metadata to anticipate every information need. Such a scheme would consume more storage than the original object, and still be useless when the Martians come and insist on searching your databases for things with high fnorny content. (What, you didn't include fnorny in your metadata fields? To the disintegration unit with you, short-sighted Earthling!)

Search ≠ organize

In some sense, all these admirable attempts at organization are solving the wrong problem. As long as you can ultimately find the things you want, organization is irrelevant. I'm often amazed when a colleague can retrieve, in seconds, a particular document from a desk that resembles a recycling truck disaster. If you have a good card catalog (remember those?), it doesn't matter what particular shelf the book is actually on. This is why Dewey's system is still useful—the fact that related books may be shelved nearby is nice but certainly not critical. This whole argument is a roundabout way of motivating search as opposed to classification. If ontologies are compile time, search is run time and on demand. With a good search engine, the object itself is its own descriptor! (At this point, you're probably wondering why I haven't mentioned Google. Well, let me get that out of the way: Google, Google, Google. There.)

Yet, even Google works only on an index or abstraction of the data, collected in advance. For text, this works pretty well, but for media? That's where we experience problems—similar problems to those we've seen for ontologies. In short, if we index or process the data ahead of time, we can only search the data that we've indexed. Google's index treats punctuation like white space (at least this week), so if you care about the difference between braces and parentheses, you're significantly out of luck. The lesson we can learn here is that every time you build a search engine, you make assumptions about what the user will ultimately want. And that can be wrong, especially for media that isn't text.

Search media = ?

Let's talk about media retrieval. Let's start several decades ago, way back in the dark ages, before, yes, even Clippy (Microsoft Office's paper clip icon) was around to help—back when information meant text. Based on Cyril Cleverdon's 1974 paper,[2] the information retrieval community settled on what has come to be known as the

"Cranfield" model of information retrieval. In the Cranfield approach, a user has an information need that can be expressed as a search query that will allow an automated system to retrieve documents. Ideally, the retrieved documents are relevant to the information need and the user goes away satisfied. The folks at certain double-vowelled search enterprises are now rather rich for doing this rather well.

This approach is so pervasive that it's pretty much ruined most attempts at media information retrieval (that is, for media besides text). Why? Because the assumptions that are good for text may be all wrong for media. What are documents? What's a query? And, Clippy help us, what's an information need? The assumptions that most of us make can be pretty awful. For example, take the whole area of content-based image retrieval. Although I've spent years looking for a counterexample, it's my sad conclusion that nobody ever needs to search an image collection based on low-level features. The reader is invited to prove me wrong.

Take the concept of a media document. While text exists in nicely parseable chunks like files, paragraphs, and sentences, media doesn't and likely never will. What is a media document? An entire DVD of a Hollywood movie? An MP3 file of your favorite song? A broadcast news story? What about your favorite radio station? Things like cross-fades and L-edits (where the soundtrack and video are spliced at different times) make even video shot boundaries problematic. I would even wager that automatic shot-boundary detection is approaching the human limit; that is, humans would disagree with each other at about the same rate they disagree with an automatic shot boundary determination.

Next, look at the notion of a query for multimedia. A text query is useful only if you have sufficient text in your metadata to allow Cranfield-like (or Google-like) methods. This, of course, assumes those methods are appropriate (for example, you don't need to search your news broadcast database by, say, the number of faces—or you have a magic artificial intelligence robot that can distill the semantics of a given media clip into text.)

Finally, and most pertinently, what does a user's "information need" mean in the context of media? Okay, there's a tiny fraction of possible searches where text approaches are relevant—for example the TRECVID[3] news broadcast retrieval. But this is only a microscopic subset of the space of possible (let alone useful) applications.

A proposition for search research

Let me propose a new alternative. Let's start with Sturgeon's law (http://www.jargon.net/jargonfile/s/SturgeonsLaw.html). Sturgeon's law is as universal as Murphy's law, and even briefer. Though cruder variants exist, Sturgeon himself put it like this:

> Ninety percent of everything is crud.

I presume you're a consumer of media, so I trust you will agree that Sturgeon's law holds supreme in that domain as well. No matter if your taste runs to boy bands, televangelists, or cat magazines, I'm sure you find a small fraction of those to be superior to the rest. So here's a great job for media retrieval: extract the 10 percent that's good stuff from the dross. Call it Sturgeon's Razor:

> Life is short.

Consider the mathematician of anecdote, who, estimating that he had roughly 10,000 days left in his life, created a 100 × 100 grid, and makes a daily ritual of checking off another square. Consider the Buddhists who routinely meditate on death. Consider that more than 400 books get published every day, and that's just in the US. You will never be able to read more than a fraction. What are the chances that you missed a good one? How much time did you waste with boring ones, let alone bad TV shows?

So I would like to add a corollary to Sturgeon's law:

> Life is too short for crud.

So let's extend our media filter beyond Sturgeon's razor. Of all the possible media files in the world, find me the best one, given I have *N* minutes to spend. And of course, media files include Web pages, books, sunset images, and (why not) actual, real sunsets. And let's broaden our definition of *best*—What film will inspire the best conversation at next week's dinner party? What information will make me the best-informed voter? What book will most inspire me to change my life for the better? What media will best serve me at this moment?

How's that for an information need? So, yes, media retrieval is important if you look at it this way. We in the business have an opportunity to improve the lot of a lot of humanity. What a responsibility! I have a fear that on my deathbed I will look back upon my life and see ... an endless sequence of vapid sitcoms. Sturgeon's razor can save me, and possibly you as well. That's *media impact* for you, and someone has to build it—so let's get started. **MM**

References

1. M. Dewey, *A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library*; http://www.gutenberg.org/etext/12513.
2. C.W. Cleverdon, "User Evaluation of Information Retrieval Systems, *J. of Documentation*, vol. 30, no. 2, 1974, p.170.
3. A.F. Smeaton, P. Over, and W. Kraaij, "TRECVID: Evaluating the Effectiveness of Information Retrieval Tasks on Digital Video," *Proc. 12th Ann. ACM Int'l Conf. Multimedia* (ACM MM 04), ACM Press, 2004, pp. 652-655; http://doi.acm.org/10.1145/1027527.1027678.

Readers may contact Jonathan Foote at foote@fxpal.com.

Contact Media Impact editor Frank Nack at Frank.Nack@cwi.nl.