Social Epistemology 18:333-356, 2004

From Classification to Indexing:

How Automation Transforms the Way We Think

F. Allan Hanson
Department of Anthropology
University of Kansas
1415 Jayhawk Blvd.
Lawrence, KS 66045
USA
hanson@ku.edu

February 13, 2005

Abstract

To classify is to organize the particulars in a body of information according to some meaningful scheme. Difficulty recognizing metaphor, synonyms and homonyms, and levels of generalization renders those applications of artificial intelligence that are currently in widespread use at a loss to deal effectively with classification. Indexing conveys nothing about relationships; it pinpoints information on particular topics without reference to anything else. Keyword searching is a form of indexing, and here artificial intelligence excels. Growing reliance on automated means of accessing information brings an increase in indexing and a corresponding decrease in classification. This brings about a shift from the modernist view of the world as permanently and hierarchically structured to the indeterminacy and contingency associated with postmodernism.

Keywords

Classification
Indexing
Modernism
Postmodernism

Worldview
Automation

Some years ago I wanted to use, as the epigraph for a chapter I was writing on vocational interest testing, a passage by Mark Twain: "A round man cannot be expected to fit a square hole right away. He must have time to modify his shape." My source indicated that the passage occurred in Twain's book *Following the Equator*, but I needed to verify it. I went to the library and pulled the thick volume off the shelf. Unfortunately it had no index, so I began scanning each page. After nearly an hour of eye-breaking tedium, I finally found it. Had the full text of that book been available in electronic form, I could have conducted a keyword search for, say, "round man," and I would have found the passage in a matter of seconds.[1] Essentially what I was doing by scanning the pages, and also what I would have been doing in the keyword search, was constructing an index for the book. A funny kind of index perhaps, having but a single entry, but an index nonetheless.

The point of this little story is that indexing with the aid of artificial intelligence has important advantages over print indexes, which are exclusively the product of human intelligence. But there is an equally important counterpoint: artificial intelligence is inferior to human intelligence when it comes to classification. The following pages explore this point/counterpoint with the objective of establishing the more general point that our growing reliance on automated means of accessing information has brought an increase in indexing and a corresponding decrease in classification. This has significant consequences for how we think and how we view the world. The consequence I am especially concerned to explore is that automation erodes the clear structures, permanence, and profundity characteristic of the modernist worldview and is conducive to a worldview that is compatible with the fluidity, indeterminacy, and lack of depth characteristic of postmodernism.

Please note that the discussion is largely limited to applications of artificial intelligence that are currently in general use. My concern is not with the research and development dimension of artificial intelligence, but with its cultural consequences, and these are discernable only for those applications that are widely used.

Classifying and indexing

The first order of business is to distinguish classifying from indexing. To classify is to organize the particulars in a body of information according to some meaningful scheme. The body of information may be as small as the contents of a single article or as large as the entire corpus of recorded knowledge. The classification scheme involved may be as unique and focused as the table of contents of a book or as general and widespread as the Dewey Decimal System for cataloging all materials in libraries. In all cases, the distinctive feature of classification is that it reflects ideas about meaningful relationships among the components of the body of information being classified: that some of them are more general or more specific than others, that they are related in various ways, and so on.

An index, on the other hand, is a finding device that connects a symbol for a topic (usually in the form of an image or a word) with whatever material is pertinent to that topic in a body of information stored in human memory, in print, or electronically. Indexing conveys nothing about relationships that may exist among different topics. It simply pinpoints information on particular topics. Automated keyword searching, which locates words or phrases in a database that match the query, is an outstanding example of indexing, and it is the one of particular interest in this essay.

In non-automated contexts the distinction between classification and indexing is not particularly prominent because human intelligence conjoins the two functions comfortably, and

in some cases almost imperceptibly. Printed indexes, for example, often combine indexing and classification. Pulling a book randomly from the shelf in front of me, Anthony Standen's *Insect Invaders*, I see the following entries in the index:

Plum curculio, 86, 182
Potato beetle, 13, 137-38, 192
Praying mantis, 187
Predators, 185-88, 190-95
Prickly pear, 97-201
Propolis, 105-06
Prussic acid, 166-67, 168

Each of these entries is an example of pure indexing: a single topic with no identified connection to any other topic. Two topics in the index, however, take a different form. One of them is:

> Pest control, by quarantine, 144-50; by chemicals, 153-70; other methods of, 172-201; eradication and, 205-23.

The other is "Insects." It has twenty-two divisions, including "classes of," "bodies of," "forest," "imported," "harmful to man and animals," and "good." These two entries are internally organized as general topics divided into categories. They represent a hybrid approach called, not surprisingly, classified indexing. The same can be said of much more complex classification schemes, such as the Dewey Decimal System when it is used to locate items in a library. These too are classified indexes, or, if you will, classifications that also serve the function of finding devices, or indexes.

A sliding, overlapping relation between classification and indexing may also be glimpsed in the history of encyclopedias. The encyclopedic movement began in the late eighteenth century with the lofty goal of providing a systematic compilation of all knowledge. With the accent on "systematic," early encyclopedias were primarily classifications. In the twentieth century, however, encyclopedias began to concentrate on presenting rapidly and easily accessible

information on a large variety of particular topics, arranged in alphabetical order (Dolby 1979:167-8). That is, their primary function shifted from classification to indexing.

When it comes automated information management, however, the distinction between indexing and classifying becomes more pronounced, because those applications of artificial intelligence in general use are very good at the one and very bad at the other. Classifying, as I have said, is based on meaningful relationships. Artificial intelligence is poor at classifying because it can deal only with meanings that are expressed in utterly explicit, unambiguous terms. It is at a loss to deal with the many meanings that are couched in metaphor, satire, or double entendre, or that depend on context or delicate nuances. Indexing, on the other hand, operates by locating matches for particular topics or queries. This is where artificial intelligence excels. It is fast: contemporary search engines search millions of documents in less than a second. It is general: most printed documents lack indexes, but *any* digitalized text or database is subject to electronic searching. And it is customized. Everyone has had the frustrating experience of finding print indexes too general and/or being unable to divine what terms the indexer selected for the topic of one's interest. With automation users create their own *ad hoc* indexes with topics of their own choosing (the keywords they enter), specific to their particular interests (Harrington 1984-85:546, Bowker and Star 1999:292, Bolter 1991:22, Richmond 1965:5). This transforms the index from one-size-fits-all to a more powerful, highly customized tool for information retrieval.

The next three sections discuss why artificial intelligence is so good at indexing and so bad at classifying. That will put us in a position to address the cultural consequences of the difference.

How automated indexing works

When computers were first coming into common use, automated systems engineers found it impracticable to design information retrieval on the basis of subject classification or any other method except free text retrieval of words, commonly known as keyword searching (Bintliff 1996:346-7). When it comes to text management, what artificial intelligence does exceptionally well is to find matches. Especially when combined with Boolean operators (AND, OR, NOT) and proximity controls (instructions that the keywords must not be separated by more than a specified number of words), keyword searching becomes a powerful tool for locating occurrences of particular words or combinations of words in one or more documents. That is to say, it indexes the documents for those words or word combinations.

Automated indexing works differently depending on the kind of search program used and the number of documents to be searched. The simplest is the "find" function of a word processor such as WordPerfect or Microsoft Word, which is limited to searching single documents. The user enters a string of characters (part of a word, a single word, a sequence of words, numbers) and the word processor scans through the document to locate matches with the search string. These are presented one at a time, in the order they appear in the document, it being necessary to click the "find next" button to move from one match to another. Less than exact matches can be found with "wildcards." "?" represents any single character, so a search for "th?s" will return both "this" and "thus." "*" is used for any number and combination of characters, so a search for "th*s" produces "otherwise" and "the purpose," among many others. Microsoft Word also has a "sounds like" option for searching. It seems, however, only to return results with the same initial sound. Thus a search for "thing" produces "think" but not "bring," while a search for "bring" in the same document produces "brings" and "Berring" but not "thing."

Indexing multiple documents is more complicated. Especially interesting are Internet search engines. These have the daunting task of indexing the staggeringly numerous and utterly unorganized sites on the World Wide Web according to users' search queries (Maze, Moxley and Smith 1997:16-17). A typical way to accomplish that is by means of discovery programs called by various terms such as robots, crawlers, or spiders. These are programs that locate web sites. They accomplish this by visiting all the hyperlinks found in one site, all the hyperlinks in each of the linked sites, and so on through an indefinite number of generations of links until millions of sites have been located. The discovery robot alphabetizes the addresses (URLs) of the sites it has found and eliminates duplicates. A harvester robot then visits each of those sites and develops a list of the words (except "stop words" such as "and," "the," "of" and so on) that appear in it. A few search engines, such as Altavista, include all the non-stop words in each site, but most limit their lists in one way or another, such as to title, abstract, or first one hundred words in the site. For each word on the list, the harvester robot also records information about how many times it appears in the site, where it appears, and so on. A database consisting of a master list of words is created, with each web site entered under each word found in it. When someone enters a query in a search engine, the words in the query are compared with those in the database. The results of the search are all of the web sites that have been listed under the words in the database that match the words in the search query (Maze, Moxley and Smith 1997:14-26, 96; Kustron 1997).

While a simple listing of the occurrences of a keyword in order of appearance may be adequate for searching single texts, sheer bulk renders this insufficient when it comes to searching the Internet. It was estimated that in the year 2000 there were about a billion web pages in existence and that their number was doubling every three to six months, with some 600

new ones being added every minute (Arnold and Colson 2000:44). On May 24, 2002, the search engine Google claimed to search 2,073,418,204 web pages, a number that increased to 3,307,998,701 by September 11, 2003 and more than doubled to 8,058,044,651 as of February 12, 2005. In the face of magnitudes this large, it is clear that even the special powers of artificial intelligence are hard pressed to keep pace. In 2000 Richard Belew wrote that the robots of most search engines manage to visit and analyze only about a third of all web sites (p. 296). The number of sites located with most search terms is staggering; it is not uncommon for a search engine to return hundreds of thousands of sites that satisfy the query. The user can diminish the results by refining the search query, but search engines themselves also cope with massive numbers of finds by automatically evaluating their relevance to the search query and presenting them in rank order. This represents another advance that artificial intelligence achieves in indexing. Manual indexing only lists places where certain topics are addressed; usually it does not specify which of those places contain the richest information on the topic.

Still, the way that search engines achieve their ranking is related to how human beings use print indexes. Searching for a topic in the index of a book, one is likely to go first to places where several successive pages are indexed for it, because an extended discussion of the topic is more likely to satisfy the user's need than just a mere mention. Or, looking in an index to periodical literature, one is likely to go first to those periodicals that have many articles on the subject rather than just one or two, because publications in periodicals with a particular interest in a topic probably contain the most important articles on it. Similarly, search engines rank sites automatically by frequency and location of keywords. That is, the more often terms in the search query occur in a site, and the more they occur in strategic places (the title, section headings), the higher the ranking of that site (Kustron 1997). Another automatic method, used by Direct Hit, is

to rank sites on the basis of the number of visits they receive. A third, pioneered by Google, ranks sites on the basis of the number of other sites that link to them, taking into account the rank of those other sites on the basis of sites that link to *them*, and so on (Arnold and Colson 2000:44, Jacso 1999).

Attempts to automate classification

For all their ingenious ranking methods, search engines still return findings as a disorganized mass that includes many documents irrelevant to what the user is looking for, and that may not include the most relevant ones. This problem might be solved, as one dream has it, if artificial intelligence could be programmed to classify as human intelligence does. The result would be an optimal system of information storage and retrieval in the automated mode of thinking that combines the computer's capacity to conduct rapid searches with the librarian's ability to organize it.[2]

I do not think this is likely to happen because, as I have said, artificial intelligence is not very good at classifying. This may appear to fly in the face of the facts, because subject classifications are prominently displayed and available for use in online library catalogs, Internet search engines, and other automated systems. This is entirely true, but they do not really automate classification. The subject function of an online library catalog simply provides automated access to classifications that have been produced by human catalogers in accordance with the Library of Congress or Dewey Decimal classification systems. The web sites retrieved through the classification systems offered by search engines such as Yahoo!, the Open Directory Project and many others have also been selected and categorized by human rather than artificial intelligence.

To develop a truly automated classification of, say, the World Wide Web would be highly desirable because it is unrealistic to imagine that human classifiers could control the entire and growing avalanche of information on all topics available there. And the effort clearly has been and is being made—although, as the following review will demonstrate, with very limited success. At this point we enter into an area of technical research and development in artificial intelligence. I will try to touch on some of the high points, although much current work will be left aside and surely some new schemes will have been introduced by the time what I am writing now reaches print. It is important to remember that the concern here is not with the technical aspects of cutting edge research projects but with the cultural consequences of information technology. Hence the focus is on the main features of programs in general use.

Nearly all projects to automate classification have tried to make artificial intelligence mimic human intelligence, the ultimate objective being computer decisions that are indistinguishable from those made by human beings. Human classification consists of two tasks. One is to design the classification scheme itself, and the other is to apply an existing scheme by allocating particular items to one or another of its categories. The actual work of most human classifiers is limited to the second task: they slot particular items in the categories of existing classificatory schemes. This, for example, is what the cataloguers who classify library materials in line with the Dewey Decimal or Library of Congress systems do. The most prominent line of research and development of automated classification has been the attempt to design computer applications to do that too.

As of July 26, 1999 a website describing itself as "a clearinghouse of projects, research, products and services that are investigating or which demonstrate the automated categorization, classification or organization of Web resources" listed thirty-nine such endeavors.[3] According to

Koch, Day et al. (1997:34), “the most important project in the area of automatic classification is OCLC's research project Scorpion.”

Of the two tasks involved in classification—to devise a classification scheme and to allocate particular items to the proper categories in the scheme—Scorpion is concerned only with the second. It adopts as its scheme the human-devised Dewey Decimal system and aims to classify websites and other electronic documents automatically according to its categories. It works by comparing “input documents” with “concept documents.” Concept documents are electronic records containing words and phrases describing the subject for each category in the Dewey classification scheme. Input documents are websites or other electronic documents to be classified. Scorpion compares the words and phrases (weighted according to their richness of content, frequency, and position) in an input document with all concept documents, and classifies it in the category or categories to which it bears a preset degree of similarity (Subramanian and Shafer 1997). Other experimental systems such as KeyConcept also work by comparing input and concept documents (Madrid and Gauch 2002, Gauch, Chaffee and Pretschner 2002).

At first glance this looks like a recapitulation of what human classifiers do when they compare the content of a particular document with the descriptions of categories in the classification scheme and place it in the category where it fits best. But actually it is a form of indexing, the difference being precisely the one that distinguishes classifying from indexing as they were defined at the start of this essay. The human classifier assimilates the specific to the general by comparing the specifics of the document to be classified in question with the general categories of the scheme. This is the essence of classification. Programs like Scorpion, on the other hand, compare one set of specific things (the contents of the input document) with other sets of specific things (the contents of the concept documents) and make their decisions on the

basis of matches between particular terms. They remain entirely with the relation between particulars. This is the essence of indexing.

Scorpion's results have been good but not perfect, and human review is needed to weed out misclassifications (Hickey and Vizine-Goetz 1999). The reason, I believe, lies in the differences between programs such as Scorpion as elaborations of indexing rather true mirrors of human classification procedures. Keith Shafer, designer of Scorpion, wrote: "While Scorpion cannot replace human cataloging, Scorpion can produce tools that help reduce the cost of traditional cataloging by automating subject assignment when items are available electronically. For instance, a list of potential subjects could be presented by Scorpion to a human cataloger who could then choose the most appropriate subject" (Shafer 1997).

Scorpion has remained an experimental project. Being unavailable to the public, one cannot evaluate how it works in practice. Northern Light (www.northernlight.com), however, seems to work on similar principles. This is a once-popular search engine that has fallen on hard times.[4] The following refers to Northern Light as it was in late 2002, when anyone could search it. As with other search engines, Northern Light retrieves documents on the basis of matches with the terms in users' queries, presented in order of relevance to the query. Northern Light's distinctive feature is that it then sorts the results into subject categories and presents them in "Custom Search Folders."

The folders represent a human-constructed classification scheme of over 200,000 categories derived from the Dewey Decimal system and the Library of Congress Subject Headings, with sixteen top-level subjects heading taxonomic hierarchies varying from seven to nine levels deep (Notess 1998, Ward 1999). Northern Light automatically sorts the results of each search by comparing the words (weighted for importance in various ways) contained in

each document returned by the search query with terms from controlled vocabularies established for each of the subject categories (Ward 1999, see also Notess 1998, www.northernlight.com/docs/search_help_folders.html, visited March 20, 2002). [5] Although many of the technical details are proprietary and not publicly divulged, this seems to be the same as Scorpion's procedure of comparing input documents with concept documents.

Results of a search are returned in up to eighty folders. They bear the names of categories in the Northern Light classification scheme, and they are intended to enable searchers to identify by subject the located documents that are most likely to satisfy their needs. Clicking a subject folder produces the search results classified under that subject plus a new set of folders that represent subtopics of that subject. For example, clicking a folder named "psychology" produces a screen with folders labeled "social psychology" and other divisions of psychology. Opening one of them produces the results classified under that subtopic together with a new set of folders at the next lower level. One can continue that process until the hierarchical levels are exhausted.

Numerous trials that I ran showed that documents with high relevance ratings contained in the first folders presented tended to be quite germane to the query. However, major differences from how a human classifier would proceed become increasingly evident as one plunges deeper in the list of folders. For example, the twenty-first folder presented in a search for "Social anthropology" on December 14, 2002 was "British pound." That folder contained fourteen items, most of which are reviews of anthropological books. They are classified in the folder "British pound" because those reviews all specify the cost of the books in pounds sterling. Clearly, Northern Light's automated classifier does not have the capacity to weed out similarities between documents that would be deemed irrelevant by a human classifier.

On the other hand, Northern Light often fails to classify items in categories that human classifiers would certainly select. A search for "lemurs" on December 11, 2002 produced 1175 items. One of the folders generated by that search is "lemurs." This and other trials suggest than when a search query is identical with a category in the Northern Light classification scheme, one of the folders produced will be for that category. But the "lemurs" folder generated by the search query "lemurs" contains only nineteen items, under 2% of the total found in the search. Obviously the vast majority of items returned for the search query "lemurs" do not qualify for inclusion in the folder of the same name. Indeed, of the first twenty items listed as most relevant to the search query, only three of them are found in the folder "lemurs" generated by that search. The omission of five others is understandable because they concern the New York software company Lemur Networks, but the other twelve obviously pertain to the animal and would unquestionably be included in the category "lemurs" by a human classifier. These trials demonstrate that, in practice, Northern Light's automated classification function is severely limited when measured by human standards.

The reason, as with Scorpion, is that artificial intelligence is limited to the indexing procedures of specific matches and cannot recapitulate the judgments relating specifics to generalities that are the stuff of human classification. Northern Light inappropriately delivers "British pound" as a category of "social anthropology" because the terms that appear in a certain number of documents returned in a search for "social anthropology" match terms in the controlled vocabulary or concept document for "British pound" in Northern Light's classification scheme. It fails to include many documents appropriately found in a search for the keyword "lemurs" in the folder named "lemurs" because the number or location of the terms in those

documents do not sufficiently match the terms in the controlled vocabulary or concept document for “lemurs” in the classificatory scheme.

Another problem is that artificial intelligence has difficulty dealing with synonyms, homonyms, metaphor, and meanings carried more by context than particular words. To computers, words such as “plant” and “tree” are simply strings of symbols while to human beings they are meaningful signs. For humans, “plant” has several distinct meanings, one being “factory,” another being a certain category of biological organisms, and a third being a verb meaning to place a seed in the ground (or, by metaphorical extension, an idea in someone’s mind, etc.). “Tree” too has multiple meanings, one of them being a subcategory of “plant” in the second sense. Numerous techniques for sorting out such differences in automated contexts have been devised, such as examining other nearby words to determine the probability of which meaning of a homonym is intended. For instance, if “plant” occurs in close textual proximity with “industrial” it is likely that it means “factory,” while proximity with “tropical” indicates that its meanings as a biological organism is in play. This is not a sure bet, however, for some texts speak of the shrubs, trees and other plants in industrial parks, while others discuss the special challenges of heating, ventilation and air conditioning in (industrial) plants located in the tropics.

If, however, computers could be “taught” the meanings of words and phrases and ways to know when a term is used metaphorically or which homonym is intended, then they could manipulate them more effectively. Such is the intention of certain projects now in the research and development stage. The Cyc Project is engaged in building a knowledge base of millions of common sense propositions, initially articulated by human analysts and then stored so as to be accessible by computers (Reed and Lenat 2002). Among those propositions are many dealing

with the taxonomic relations among things, such as that carrots and beans are kinds of vegetables, jellybeans are a kind of candy (and not a kind of vegetable), and so on.[6]

In a related area of development, the Semantic Web aims to enable computers to recognize differences between all kinds of things and concepts by labeling each with its unique "Uniform Resource Indicator" (URI). The familiar Uniform Resource Locators (URLs) for web sites and email addresses are kinds of URIs, so the URI for the University of Kansas can be www.ku.edu and the URI for the author of this article can be hanson@ku.edu. Everything else would have its own URI: one for "plant" in the sense of "factory," another for "plant" in the sense of a kind of biological organism, one for "tree" in the sense of a hierarchical, taxonomic structure, and so on. URIs would also be given to relationship terms such as "is larger than," "is a kind of," and "is married to." Then expressions called Resource Description Framework (RDF) "triples" could be devised to specify relationships among URIs, much as subjects and predicates do in ordinary sentences. When written in an appropriate machine-readable language such as XML (eXtensible Markup Language), expressions such as "The Middle Ages preceded the Scientific Revolution," "Russia is larger than Liechtenstein," or "a tree is a kind of plant" could be processed by and communicated between computers. It would also be possible for computers to conduct inferences based on such expressions. For example, from the propositions "an oak is a kind of tree" and "a tree is a kind of plant", the computer could automatically generate the further proposition "an oak is a kind of plant" (Berners-Lee, Hendler and Lassila 2002, Schwartz 2002). If they can be perfected and brought into general use, developments such as the Cyc Project and the Semantic Web may eventually enable computers to recognize synonyms, homonyms, metaphor, and relationships between specific and general much as human beings do in ordinary discourse. This would represent a major step toward the capacity of

artificial intelligence to simulate human classifying procedures. At the present time, however, these procedures are not in general use and thus have no discernable social consequences.

Can artificial intelligence create classificatory schemes?

Classification, as noted above, involves two distinct tasks. One is to design or devise classificatory schemes, and the other is to allocate particular items to the categories of such schemes after they have been devised. In their efforts to mirror human intelligence, applications such as Scorpion and Northern Light attempt to automate the second task but not the first. Two other approaches initially seem to automate both tasks. What is especially interesting about them is that they do so while holding fast to the unique capabilities of artificial intelligence and make no attempt to emulate human intelligence.

HITS (Hyperlink-Induced Topic Search) operates specifically with web sites. This ingenious technique was devised by Jon Kleinberg to avoid ambiguity in automatic web searches. To explain how it works by example, a web search for the keyword "jaguar" produces, among others, many sites about a brand of automobile, others about a kind of feline mammal, and still others about an Apple Mac operating system.[7] As with Google's technique for ranking the results of a web search, HITS keys off of hyperlinks between sites. The reasoning behind HITS is that sites about the computer operating system are more likely to link to each other than they are to sites about the car or the cat. Automatically assessing all the links among web sites located with the search query "jaguar" reveals three groupings defined by frequency of reciprocal citations, corresponding to the these three meanings of the term. The groupings or categories generated by analyzing hyperlinks can be sorted at several levels of similarity, thus distinguishing between sites dealing with entirely different topics, broadly similar topics, or closely related topics, such as pro-life and pro-choice persuasions in web sites dealing with

abortion ("Hits and Misses" 1998). Thus HITS seems capable of generating a hierarchical, taxonomic classification of web sites returned by a search query, although I will argue in a moment that this is not exactly the case. The basic idea behind HITS is being developed and enhanced in the Clever Project (Chakrabarti *et al.* 1998).

Another procedure, called "clustering," also aims to identify more or less closely related groupings of documents. It achieves this by analyzing the contents of documents themselves rather than the hyperlinks between them. How clustering works becomes clear from a comparison with Scorpion. As explained already, Scorpion compares the text of an "input document" with the texts of each of the "concept documents" that have been prepared in advance for every category in a classification system, and classifies the input document in the category or categories to which it bears the greatest similarity. In clustering, no canonical concept documents are used to serve as standards for comparison. Instead, all the documents in a set, such as those located by a keyword search, are compared for similarities with each other, two by two. This divides the set into clusters of similar documents (Subramanian and Shafer 1997).[8] Because the degree of similarity between document pairs is expressed quantitatively, one can identify a few broad clusters on the basis of general similarity or, within them, narrower clusters defined by higher degrees of similarity. One can also map the clusters, at any level of generality, according to their distance from each other. The result looks very much like a hierarchical, taxonomic classification of documents.

Clustering is in common use, being an important feature of the popular metasearch engines Vivisimo, Excite, Dogpile, MetaCrawler, and WebCrawler. These submit the user's keyword query to several search engines (such as Google, Lycos, Looksmart, AltaVista) and present the top results from each. Using the "preferences" function, the user can set Vivisimo to

return a total of about 100, 200, or 500 results, the default setting being about 200. As of June 2004 searches on the other metasearch engines returned about 100 results. In addition to being presented in the standard search engine format as a relevance-weighted list, the findings are automatically grouped into categories on the basis of the clustering technology just described. Vivisimo, the only one to describe its procedure, generates its clusters using the two to three line summaries of each site provided by the several search engines it consults to locate them.[9] There is some hierarchy, but not particularly deep. In most cases the clusters contain web sites with no subcategories. Some clusters are further divided at one level, a few at two, and I found one at three levels in a Vivisimo search for "psychology" on February 24, 2004.

Unlike HITS, which relies on hyperlinks and is therefore limited largely to web sites, clustering can be used to sort digitalized documents of any sort. Anyone can use Vivisimo, for example, to get clustered search results of the PubMed database of medical literature, the New York Times, eBay, and several other databases. It also markets its clustering engine as a way to help corporations and other organizations bring order to their large volumes of poorly organized memos, files and reports.

As techniques to sort documents at several hierarchical levels of similarity, clustering and HITS seem to bear all the markings of taxonomic classification. However, a crucial difference divides them from human-devised, taxonomic schemes. And it is precisely that: the latter are *schemes.* That is to say, they are structured categories that have been laid down in advance and into which documents, web sites, things of any description are classified. Clustering and HITS do not use predefined categories. They begin with a concrete array of documents and sort them into *ad hoc* categories according to their degrees of similarity to each other.[10]

This may not seem to be fundamentally different from human classification. After all, taxonomic schemes have to begin *somewhere,* and their beginnings are not unlike clustering. One or more human beings observe an array of objects, perceive degrees of similarity among them, and sort them into categories accordingly. But the difference is this: the human scheme becomes a taxonomic framework into which new items are subsequently slotted, and if its categories seem to make good sense and it can successfully accommodate quite a lot of new items, then it becomes institutionalized and considerable pressure is exerted to fit more and more items into it. HITS and clustering never form lasting taxonomic schemes. Their categories are created afresh with each inspection of an array.

Anyone can observe this with their own eyes by watching the metasearch engines listed above at work. On February 24, 2004, I ran two searches for "American civil war generals" on Dogpile, within minutes of each other. Although the queries were identical, the first search returned fifty-five web sites and the second sixty-four. The explanation is that different numbers of findings were reported from some of the search engines consulted, because the allotted time ran out or for some other reason. Most interesting, not only do the numbers of results change, but so do the categories into which they are clustered. Both of the searches returned clusters titled "Soldiers," "Pictures," and several others. But "Confederate," which with seven sites was the largest cluster to emerge from the first trial, did not appear at all in the second trial. Nor did "Reenactment, Gettysburg," a cluster with three sites in the first trial. In revenge, the three-site cluster "Army, Major" appeared in the second trial but not the first. The same sort of thing happened when I performed the identical search with Vivisimo.

The reason for variations such as this is that clustering and HITS do not attempt to mimic the kind of classification born of human intelligence that subsumes specifics under generalities.

Instead, they utilize artificial intelligence's strong point, most obvious in keyword searching, of matching specifics. In HITS the specifics are hyperlinks to and from web sites; in clustering they are the contents of documents. The same principles operate in both, but it will be simpler to confine this brief analysis of how they work to clustering. Keyword searching is a kind of indexing. Its aim is to locate items in a database that match the terms of a search query. The query may be as short as a single number or word, it may be a set of terms conditioned by Boolean operators and proximity controls, or it may be indefinitely long: a sentence, an entire page of text, or whatever. Clustering, we have seen, works by comparing all documents in a data set, two by two. This too is a form of matching, or indexing. It is as if each document in the data set were considered to be a keyword query used to search all the others. But the critical difference is that in an ordinary keyword search, only documents having exact matches with the query will be returned. With clustering, because an entire document is used as a search query, no other document in the data set will be an exact match for it. But some documents will be closer to it than others. Thus in clustering the search is for approximate rather than exact matches. The more closely another document in the data set approaches the document used as the query (as determined by the familiar ranking criteria such as number of shared words and the importance and location of those words), the greater the similarity between them. Clustering uses the degrees of similarity among documents in the set to group them into categories, often with hierarchical levels. And yet, because there are no predetermined categories, it is not really classifying. At bottom both keyword searching and clustering depend on weighted matching, and therefore they should be understood as variations on the single technique of indexing.

What I wish to stress about HITS and clustering is that, unlike the other applications we have examined, their use of artificial intelligence does not attempt to mimic human intelligence.

Scorpion and Northern Light try to recapitulate the human way of classifying by subsuming particulars into pre-established general categories represented by their control documents or controlled vocabularies. But, as I have argued already, that objective is doomed to failure in principle because the control documents are not really general but are just other particulars. Thus they compare particulars with each other, which is indexing and not human-style classifying. For their parts, HITS and clustering also utilize the unique indexing capacity of artificial intelligence to match particulars. But, because they do not take any documents to be definitive of pre-established categories, the groupings they generate—and the relations among those groupings—are freshly created for each search. This is a distinctly new way of organizing information, fundamentally different from human techniques. Indeed, it would be nearly as difficult for human intelligence to organize information by HITS and clustering as it is for artificial intelligence to mimic the human way of classifying.

The difference between classification and indexing is captured by Gilles Deleuze and Félix Guattari's distinction between arborescent (tree-like) and rhizomatic structures. "Unlike trees or their roots, the rhizome connects any point to any other point, and its traits are not necessarily linked to traits of the same nature….In contrast to centered (even polycentric) systems with hierarchical modes of communication and preestablished paths, the rhizome is an acentered, nonhierarchical, nonsignifying system without a General and without an organizing memory or central automaton, defined solely by a circulation of states" (Deleuze and Guattari 1987:21). Classification is arborescent, while indexing (especially automated searching) is rhizomatic.

Classificatory and indexical worldviews

Most people are not like anthropologists and other social scientists. They do not say that the worldview they hold is merely one of many constructions that human cultures in different times and places have imposed on reality. Instead, they believe that their worldview is the self-evidently accurate representation of things as they are; they think the world really is the way the information they have about it says it is. Information presented in terms of classification is structured and received differently than when it is presented in terms of indexing, and this distinguishes two types of worldview. We will call them classificatory and indexical.

To assemble information in terms of classification is to seek the place of one's interest in a pre-existing scheme. From this perspective the truth is out there, things unequivocally are what they are, and the relations between them are clear, permanent, and, in Deleuze and Guattari's terms, arborescent. The person who would learn something, or make a new contribution to knowledge, must relate it to the structure of established knowledge. Established knowledge is taken to be certain, which is why proposed paradigm shifts provoke stiff resistance and why those that are ultimately successful are considered to be momentous developments. The certainty built into this view of things also means that when people encounter ways of thinking and behaving different from their own, their typical reaction is to assume that the alien ways are at best misguided, and at worst heretical and evil. Divergent notions about the structure of reality mean that many different worldviews are included within the classificatory type, and they often find themselves at odds with each other. Examples are fundamentalist versions of Christianity, Judaism and Islam, or laissez-faire capitalism and communism. Their tendency to absolutism is simultaneously what defines them as a family of worldviews and what makes them prone to squabbles.

The indexical type of worldview is rhizomatic; it contrasts with the classificatory type at every point. To assemble information by indexing is a matter of sifting existing knowledge—be it the contents of a book, the holdings of a library, the sites on the World Wide Web—to precipitate what is relevant to one's interest. Searchers do not articulate what they want to know in terms of what is out there; they organize what is out there in terms of what they want to know. Readers of a book, for example, may not be interested in the specific way the author presents a body of information but still anticipate that something in the work is relevant to their interests. In that event they decline to read the full text, going instead only to those pages that the index indicates may be useful to them. This has important implications for the meaning of the text. Such a reader may remain oblivious to the meaning the author intended to convey by writing it. On the other hand, the text (better, selected parts of it) may hold meaning for the reader quite apart from what the author intended. It is of course nothing new for readers to miss the author's point, or to detect meanings in a text that the author did not put there. But that is more likely to occur when the reader accesses the text by means of an index than when following the author's argument from start to finish.

This applies even more to digitalized information because, as we have seen, indexing in the form of keyword searching is more powerful than any print index. Readers can search any digitalized text (or database containing millions of texts) for keywords, and they can select the index terms (keywords) themselves rather than being restricted to those chosen by an indexer. In effect, with every keyword search the user creates a new index of the text or database. What sites, books or articles emerge depends entirely on the topic selected by the searcher. It is not even sensible to speak of a "topic" as having any enduring presence, because what passes as

such, when sorted electronically, expands, contracts, and is reconfigured according to any number of diverse criteria used to design the specific search strategy.

Material accessed by indexing is often just a list, with no internal organization at all. Or, if a technique such as HITS or clustering discerns some pattern in it, that pattern has no prior standing of its own. It is entirely *ad hoc*, emerging out of that particular body of information and customized to it. In either event, users have more opportunity and responsibility to interpret information assembled by indexing than when classification feeds it to them in prepackaged categories. Therefore, despite the anxiety of some, the growth of artificial intelligence does not threaten to restrict or replace human intelligence. On the contrary, it frees human intelligence from the constraints of received categories to think more imaginatively. As John Henry Merryman prophetically wrote in 1977, "One of the most attractive features of the LEXIS system[11]…is that it liberates the researcher from [pre-established] indexes and opens up an enormous range of possible avenues of access to the literature" (1977:426, see also Bowker and Star 1999:292). When asked what appealed to her about computerized legal research, Roberta Shaffer replied: "Being liberated. Having the choice between looking at something using someone else's taxonomy…versus letting your own mind create the taxonomies. With the books, you don't have the freedom to think of it the way *you* think of it. You're constrained by how somebody else chose to present it" (quoted in Halvorson 2000:114-15). [12]

Indexical worldviews feature more open-mindedness than classificatory ones. This is a by-product of the fluid, variable, *ad hoc* quality of the findings of index searches, and also of the greater responsibility, in the absence of preset categories, for users to devise their own interpretations of those findings. When information is assembled by means of indexing, people become accustomed to arrays of information in novel combinations, and they develop something

of the flexibility of mind required to make sense of them. This loosens the commitment to a particular, received set of ideas and values characteristic of classificatory worldviews. People become more aware of alternative perspectives and interpretations and more open to their possible merits.

The cultural impact of indexing

Evidence of the flexible, contingent, open-minded quality of indexical worldview is visible in many sectors of contemporary society. We will briefly review a few of them. I make no claim that automation is the sole cause of this; indeed, some of the changes to be discussed began before widespread computing. Nevertheless, in every case it is possible to identify automation as an important contributing factor

The law.—About a quarter of a century ago American law began a major transformation with the introduction of the computerized legal research services LEXIS and WESTLAW. Previously legal research had been conducted with a set of print resources such as legal encyclopedias, Restatements, treatises, and the "key number system." All of these provided access to legal information via classificatory schemes that divided the law into a number of categories. The key number system in particular classified the points of law addressed in appellate level judicial opinions according to a scheme of over 400 categories, each with its subcategories. This enabled attorneys to locate cases of interest to them from any time period and jurisdiction.[13] LEXIS and WESTLAW automated legal research by placing case law, legal journals and other resources online, where they were subject to powerful electronic keyword searching strategies.

The impact of automation on legal research has been immense. Manual research using "the books" was made obsolete as it became possible to do in minutes what had previously

required hours of tedious work. Hyperlinks allow attorneys searching for favorable precedent instantly to move from one opinion to another as they review cases similar to the one they are working on. Hyperlinked footnotes in law review articles enable readers to go directly to other relevant works as they build a knowledge base for their own work. Certainly the ease of following hyperlinks in both of these situations results in lawyers actually consulting more cited cases and publications than they would have done when it required finding the relevant volumes in the library. However, the impact of computerized legal research is more fundamental than just doing the same kind of things as before, only more faster and more thoroughly. The transformation in how information is located has important consequences for what that information means.

The law looks different depending on the means used to research it (Berring 1986:29, 33). Non-automated techniques such as encyclopedias, treatises and the key number system are classified indexes. Much as other encyclopedias and library cataloging systems, they organize the law in a hierarchical system of categories that also serve as devices for finding legal information. For those imbued with such research techniques, the classificatory scheme underlying them reveals what the structure of the law really is. A good example is legal positivism: the view that the law exists in its own right and is out there, waiting to be discovered.

In contrast, lawyers who regularly use LEXIS and WESTLAW can design highly customized searches that pinpoint and juxtapose information in ways that would be impossible with the key number system or any other classified index. An attorney wanting to learn about cases involving a particular kind of factual situation would be able to search for that using LEXIS or WESTLAW more easily than in the print reference sources, which are organized

according to legal principles rather than factual circumstances. Or the attorney might be interested what happens in cases where two or more points of law are simultaneously in play. Automated searching would allow them to be found directly, while using the traditional tools would involve a more tedious process of separately noting down cases and articles that involve each of the relevant points and then manually comparing the resulting lists for overlaps. Legal research of any sort, be it in case law, regulatory law, or the academic literature, is being weaned away from the hierarchical categories embedded in the traditional research tools. As a result, lawyers are coming to think of the law as a collection of facts and principles that can be assembled, disassembled and reassembled in a variety of ways for different purposes (Bintliff 1996:345-46). This could call into question the notion that the law actually *has* an intrinsic, hierarchical organization, and that would signal a basic change in the perception of legal knowledge and of the law itself (Katsh 1989:221-2). In essence, the advent of automated research techniques is a pivotal factor in a shift in legal worldview from classificatory to indexical.[14]

The academy.—Something similar has been happening in scholarly research and education. Distinct branches of knowledge have been recognized ever since the quadrivium (arithmetic, music, geometry, and astronomy) and trivium (grammar, rhetoric, and logic) of classical antiquity. The expansion of recorded knowledge with the passage of time and inventions such as printing made it evident that the Renaissance Man could not endure long after the Renaissance, and scholars became disciplinary specialists. Prior to the twentieth century science justified itself mainly as adding to knowledge, filling in our understanding of the world. This was conducted in the context of the various disciplines, each of which had carved out a part of the world as its domain for investigation. This was entirely characteristic of the classificatory

worldview. The disciplines themselves represent the classification of things in the world together with the methods for studying them, and there was a tendency toward exclusiveness, each discipline developing more or less in isolation from the others.

Twentieth-century research turned more toward practical applications in war and industry (Dolby 1979:187-88) and to focused questions in basic research. This shifted science from discipline-orientation toward problem-orientation. The ascendance of problem-oriented research is much in evidence today. For example, the National Science Foundation influences the overall course of scientific research in the United States by allocating a portion of its funding resources to certain areas it designates as priorities. Typically these call for interdisciplinary research. One of the newest is the Human and Social Dynamics priority area. The program description reads, in part, "Revolutionary technologies and ideas that are the product of human minds have created a more closely linked world, within which there is almost instantaneous transmission of information that feeds a global economy. But it is also a world of change, uncertainty, and disruption that leaves many uncertain how to respond….Scientific understanding of the dynamics of mental processes, individual behavior, and social activity increasingly requires partnerships that span the different science, engineering, and education communities."[15] Private enterprises such as pharmaceutical and aeronautic companies and other federal agencies that support scientific research—the Department of Defense, the Department of Energy, the National Institutes of Health—are even more directive in defining the particular applications of scientific research that they are willing to fund, many of which also involve interdisciplinary teamwork.

Automated techniques such as keyword searching are highly compatible with problem-oriented, interdisciplinary research. They instantly bring together information on any topic from a variety of specialized fields. This juxtaposition of information makes the hitherto

unrecognized relevance of research on a topic in one field apparent for work being done in another. Possibilities for new insights derived from sharing findings and methods and for future collaboration leap into view. The work of each researcher may still be specialized, in some cases more than ever. But automated information retrieval enables them to become aware of what others are doing. They perceive common ground, upon which the differences between the contributions of scholars from different fields become recognized as complementary rather than compartmentalized. Disciplinary separation gives way to interdisciplinary cooperation.

Problem-oriented, interdisciplinary research is representative of the indexical worldview. It is *ad hoc* and open-ended. It consists of sifting a wide range for data for material that seems relevant, recognizing that what is or is not relevant may change as the investigation proceeds, and building a conclusion on the basis of facts and concepts that may never have been combined in that way before. What Ethan Katsh said with reference to the law is equally apt for scholars in other disciplines: "Speed and convenience may be the attraction for new computer users and the justification for purchasing hardware and software, but most users at some point find themselves using information differently, possessing information that they would not have had previously, asking questions they might not have asked previously and working with people they might not have had contact with before" (Katsh 1993:443).

Similar developments have occurred in education. Scholars, thinking that what they teach ought to be more in line with what they do as researchers, began to design curricula and degree programs that spill over the borders between different disciplines as traditionally defined. Students began to seek training that would feed their personal interests. The result of such rethinking on the parts of both teachers and students is the institutionalization of interdisciplinary studies. Programs in human development combine biology with psychology and other social

sciences. Natural, biological and social sciences all figure in programs in ecology, and some of them also incorporate history and literature. Programs in cultural studies, women's studies, gay and lesbian studies, peace and conflict studies draw upon history, literature, philosophy, and several of the social sciences in various combinations. In 2004 Princeton University inaugurated a new freshman/sophomore science curriculum designed to teach chemistry, physics, biology, and computation in an integrated fashion over four semesters. The sequence will feature a "just-in-time" approach that, in common with manufacturing procedures that provide materials only when they are needed, will introduce concepts and methods at the moment they will be used to address specific questions rather than presenting them at the beginning with the assurance that they will come in handy later.

In common with other instances of indexical worldview, interdisciplinary learning and research amplify human intelligence because they throw up unanticipated combinations of information upon which the human mind is called to exercise its peculiar powers of interpretation. Knowledge-seekers and knowledge itself are liberated from classificatory assumptions that assign a place for everything in advance. The flexibility of interpretation is joined by a flexibility of evaluation, as novel and alternative ways of thinking and behaving are considered on their own merits instead of being measured against pre-existing standards.

Business and manufacturing.—Finally, interdisciplinary research teams find their counterparts in *ad hoc* groups that form specifically to accomplish certain tasks in many business and manufacturing enterprises. Often these are "virtual teams" with members located in many different places, who interact via email and other kinds of electronic communication (Coleman 1997, Maznevski and Chudoba 2000). So pervasive is this trend that The Gartner Group forecasts that by 2005, 80% of all global knowledge work will be delivered by virtual project

teams “that work together but are physically apart. Their activities are often time-bound—they come together to accomplish a specific task and when their objective is met, they disband, with members joining other newly forming project teams” (Kaplan 2002:3).

A number of other on-the-fly responses have been designed to maximize efficiency and respond to consumer wants with maximum speed. Kawasaki plants use “a ‘just in time’ supply method which eliminates expensive warehousing and over-ordering of parts....For instance, certain parts and pieces are made on special presses located right on the assembly line. This means no shortages or excess inventory on these items for more efficiency and less cost. In many cases, it also means the worker makes the part he assembles, and thus enjoys a full sense of accomplishment.” [16] In some places, “just in time” techniques also govern allocation of machinery and human resources. “In the industrial districts of Italy and Germany, researchers discovered congeries of firms making apparel or ceramics that cooperated so intensely that they seemed to blur the line between market and organization. Using flexible-production methods to tailor products to rapidly fluctuating demand, these companies worked together on a routine basis, sharing workers, outsourcing to one another during times of high demand, even loaning machinery as the situation required” (DiMaggio 2001:19, citing Sabel and Zeitlin 1997).[17]

The organizational structure of business firms is also changing. “Hierarchy is an approach to Organization that is beginning to lose its once unquestionable authority where it exists in its more extreme form; in a multilevel hierarchy, which gives to rise to multilevel bureaucracy, and absolute hierarchy, where all work is determined by downward assignment and where peers play no part in distributing work among themselves” (Belbin 1996:vi). Vertically-oriented hierarchy is being replaced by more horizontal kinds of organization. This breaks down compartmentalization by encouraging direct communication and collaboration between different

units or departments. Much of this communication takes place at the level of middle management, giving people in such positions greater decision-making powers while requiring a wider range of competencies and more initiative from them (Farquhar 1998, see also Ostroff 1999). All these examples indicate that in the business world too, the qualities of flexibility, indeterminacy, contingency, and liberation of human intelligence that set the indexical worldview apart from the classificatory type are on the rise.

From modernity to postmodernity

The contrast between the classificatory and indexical worldviews, with their emphases on established structures vs flexibility and indeterminacy, is another way of expressing the difference between modern and postmodern epistemologies. My argument has been that increasing reliance on automation entails the expansion of the indexical worldview at the expense of the classificatory worldview. If this is correct, it follows that postmodern epistemology is on the rise in contemporary culture.

The intriguing aspect of this is that it is not because the intellectual arguments in favor of postmodernism have been that persuasive. In fact, the case for postmodernism has fared poorly within the academy, and worse with the general public. The reason is that theoretical presentations of postmodernism usually corrode all systems of belief. Their claims about flexibility and indeterminacy often take the form of a frontal assault against the notion that anything can be accepted as true. This cuts all anchors and sets everything adrift. In just one of many passages expressing his belief that contemporary culture has come to a state of utter exhaustion, for example, Jean Baudrillard wrote: "we have nothing else now but objects in which not to believe" (1998:3). That smacks of a nihilism from which most people recoil.

The indexical worldview that flows from the automation of information also features flexibility and indeterminacy, but in a matter that is less threatening and more effective than postmodernism. The main reason is that the consequences of automation inhabit the realm of habitual behavior rather than that of philosophy, where postmodernism dwells. According to Karl Marx, V. Gordon Childe, Leslie White, Marvin Harris and most others who have studied the matter, significant social change seldom originates at the level of ideas. Instead, the seeds of change are typically found in the material conditions of life. That is where, often as a result of some technological innovation, people begin to do things differently. Transformations of habitual behavior then foment change in other areas of life: social and political organization, religion and ideology. Think, for example, of the far-reaching ramifications that came in the wake of technological innovations such as the domestication of plants and animals, the smelting of iron, gunpowder, the printing press, the steam engine, automobiles and television, to name just a few.

In the case before us, the technological innovation of computers led to changes in habitual behavior as their use became general, and those behavioral changes are producing changes in how people think. In the simplest possible terms, thinking about a problem or issue consists of assembling information relevant to it, and subjecting that information to interpretation or analysis. Prior to the introduction of automated technology, both of these tasks were carried out by human intelligence. Human intelligence organizes large bodies of information by classifying them into a manageable number of categories. The classificatory schemes are not invented by each individual from whole cloth; they are acquired from, shared with, and passed on to other individuals. That is, they are cultural in nature. The information needed to address a particular issue or problem is assembled from the cultural categories most relevant to it.

Therefore, by determining what information will be presented to human intelligence for interpretation and analysis, cultural categories play an important role in the thought process.

With the invention and widespread dissemination of automated information technology, the part of the thinking process that consists of assembling information relevant to a problem or issue is being assumed by artificial intelligence. Habitual behavior changes as people become accustomed to using computers to assemble information: learning how to formulate automated search queries and how to assess the results. Artificial intelligence assembles information by means of indexing, the prime example we have considered being the retrieval of textual information by searching for keywords and phrases. A keyword query is customized to the precise issue or problem under investigation, and matches for the query are sought in one or more databases without regard to any classification of the material in those databases. As before, once assembled, the information is presented to human intelligence for interpretation and analysis. But information assembled by artificial intelligence is less conditioned and constrained by culture's classificatory categories than information retrieved with preautomated techniques. Therefore it is more likely to include unanticipated contents and juxtapositions. These stimulate human intelligence to develop interpretations that are not prefigured by the received categories of culture. The result is greater flexibility and creativity of thought, similar in many respects to what happens when artists free themselves from standardized ways of looking and thinking. With the retreat of fixed certainties and the rise of flexibility and indeterminacy, culture is inexorably moving from a condition of modernity to postmodernity. Not, however, in the sense that people are becoming philosophical nihilists, with no incentive to believe anything. Instead, it is a more positive version of postmodernity, grounded in habitual behavior that features greater flexibility and creativity in many sectors of daily life, including interdisciplinary approaches to

education and research, developments in law, and new procedures and forms of organization in business and manufacturing.

---

[1] As it happens, the full text of *Following the Equator* is now available at http://www.gutenberg.net/etext/2895. I did search it for "round man" on September 15, 2004, and found the quote very quickly.
[2] Keith Shafer, "A Brief Introduction to Scorpion" (http://orc.rsch.oclc.org:6109/bintro.html, visited 6/2/02).
[3] www.public.iastate.edu/~CYBERSTACKS/Aristotle.htm, visited 3/27/02.
[4] Northern Light began to narrow its services in the spring of 2002 when it ceased searching the Internet and limited itself to documents in its own "general collection," which it delivered to users on a pay-as-you-go basis. (For the most part the cost was modest—from $1 to $4 per document. The highest fee I have seen is $1600.) The search engine virtually disappeared when its owning company went bankrupt in 2003, but later that year it began restructuring under new ownership for use by business customers (Hollmer 2003). In late 2004 the fee-based Northern Light Business Research Engine was launched, catering to market researchers, competitive intelligence professionals, sales people, product developers, strategic planners, and information center professionals (http://www.northernlight.com/library.html, visited 12-5-04).

[5] Another product that also automatically classifies documents according to a predefined taxonomic scheme is "Quiver" (Wiggins 2002:61-63, 65). It is more customized than Northern Light, being designed to categorize corporate Intranet documents according to schemes devised by each customer. Although it is on the market it can hardly be said to be in widespread use because its price is $125,000 and up.
[6] See Dreyfus (1992:xvi-xxx) for an argument that the Cyc Project is misguided in principle.
[7] I am using the same example as my source, "Hits and Misses" (1998). It states that the search for "jaguar" will also turn up documents about a football team. But no sites pertaining to the Jacksonville Jaguars were included in the first ten pages of my search on Google on February 26, 2004. They do appear in abundance in a search for "jaguars."
[8] A single document might end up in more than one cluster if it is similar to other documents in one cluster in one way, and to documents in another cluster in another way.
[9] http://vivisimo.com/faq/Technology.html, consulted June 25, 2004.
[10] Another application of potential relevance here is DolphinSearch. This is a product that uses neural net technology to ascertain the meaning of words and documents through pattern recognition, which enables it to weed out irrelevant documents in a web search, and also automatically to classify documents (Wiggins 2002:63-65, Roitblat 2000). Its parallel with clustering and HITS is that Roitblat stresses that DolphinSearch classifies not according to some pre-established taxonomic scheme, but according to the particular needs of the user at the moment (2000:3-4, 9). As with Quiver, DolphinSearch is hardly in general use, being limited to those large law firms and corporations that can afford its six-figure cost.
[11] LEXIS is an automated legal research service that features powerful keyword searching capacities.
[12] Another way of expressing this is with the concept "informate," a term attributed to Professor Shosana Zuboff of the Harvard Business School. Automation encourages the development of new insights by "informating" data, i.e. establishing an environment in which data may be applied in a variety of ways other than the original intention (DiMarco 1997).
[13] See Doyle (1992:231-2), Grossman (1994:76-81, 83), Cohen (1985:34-47, 60-70) and Hanson (2002) for more detailed descriptions of these resources.
[14] More thorough discussions of these transformations in the law may be found in Bast and Pyle (2001) and Hanson (2002).
[15] http://www.nsf.gov/home/crssprgm/hsd/start.htm, visited 3/23/04.
[16] (http://www.kawasaki.com/about/kmm_lincoln.html, visited April 29, 2004).
[17] DiMaggio gives a publication date of 1996 for Sabel and Zeitlin's book, but all references to the work that I could find indicate that it was published in 1997.