

A Simple Scheme for Book Classification Using Wikipedia

Andromeda Yelton

Editor's note: This article is the winner of the LITA/Ex Libris Student Writing Award, 2010.

Because the rate at which documents are being generated outstrips librarians' ability to catalog them, an accurate, automated scheme of subject classification is desirable. However, simplistic word-counting schemes miss many important concepts; librarians must enrich algorithms with background knowledge to escape basic problems such as polysemy and synonymy. I have developed a script that uses Wikipedia as context for analyzing the subjects of nonfiction books. Though a simple method built quickly from freely available parts, it is partially successful, suggesting the promise of such an approach for future research.

As the amount of information in the world increases at an ever-more-astonishing rate, it becomes both more important to be able to sort out desirable information and more egregiously daunting to manually catalog every document. It is impossible even to keep up with all the documents in a bounded scope, such as academic journals; there were more than twenty-thousand peer-reviewed academic journals in publication in 2003.¹ Therefore a scheme of reliable, automated subject classification would be of great benefit.

However, there are many barriers to such a scheme. Naive word-counting schemes isolate common words, but not necessarily important ones. Worse, the words for the most important concepts of a text may never occur in the text.

How can this problem be addressed? First, the most characteristic (not necessarily the most common) words in a text need to be identified—words that particularly distinguish it from other texts. Some corpus that connects words to ideas is required—in essence, a way to automatically look up ideas likely to be associated with some particular set of words. Fortunately, there is such a corpus: Wikipedia.

What, after all, is a Wikipedia article, but an idea (its title) followed by a set of words (the article text) that characterize that title? Furthermore, the other elements of my scheme were readily available. For many books, Amazon lists Statistically Improbable Phrases (SIPs)—that is, phrases that are found “a large number of times in a particular book relative to all Search Inside! books.”² And Google provides a way to find pages highly relevant to a given phrase. If I used Google to query Wikipedia for a book's SIPs (using the query form “site:en.wikipedia.org SIP”), would Wikipedia's page titles tell me something useful about the subject(s) of the book?

Background

Hanne Albrechtsen outlines three types of strategies for subject analysis: simplistic, content-oriented, and requirements-oriented.³ In the simplistic approach, “subjects [are] absolute objective entities that can be derived as direct linguistic abstractions of documents.” The content-oriented model includes an interpretive step, identifying subjects not explicitly stated in the document. Requirements-oriented approaches look at documents as instruments of communication; thus they anticipate users' potential information needs and consider the meanings that documents may derive from their context. (See, for instance, the work of Hjørland and Mai.⁴) Albrechtsen posits that only the simplistic model, which has obvious weaknesses, is amenable to automated analysis.

The difficulty in moving beyond a simplistic approach, then, lies in the ability to capture things not stated, or at least not stated in proportion to their importance. Synonymy and polysemy complicate the task. Background knowledge is needed to draw inferences from text to larger meaning. These would be insuperable barriers if computers limited to simple word counts. However, thesauri, ontologies, and related tools can help computers as well as humans in addressing these problems; indeed, a great deal of research has been done in this area. For instance, enriching metadata with Princeton University's WordNet and the National Library of Medicine's Medical Subject Headings (MeSH) is a common tactic,⁵ and the Yahoo! category structure has been used as an ontology for automated document classification.⁶ Several projects have used Library of Congress Classification (LCC), Dewey Decimal Classification (DDC), and similar library tools for automated text classification, but their results have not been thoroughly reported.⁷

All of these tools have had problems, though, with issues such as coverage, currency, and cost. This has motivated research into the use of Wikipedia in their stead. Since Wikipedia's founding in 2001, it has grown prodigiously, encompassing more than 3 million articles in its English edition alone as of this writing; this gives it unparalleled coverage.

Wikipedia also has many thesaurus-like features. Redirects function as “see” references by linking synonyms to preferred terms. Disambiguation pages deal with homonyms. The polyhierarchical category structure provides broader and narrower term relationships; the vast majority of pages belong to at least one category. Links between pages function as related-term indicators.

Andromeda Yelton (andromeda.yelton@gmail.com) graduated from the Graduate School of Library and Information Science, Simmons College, Boston, in May 2010.

Because of this thesaurus structure, all of which can be harvested and used automatically, many researchers have used Wikipedia for metadata enrichment, text clustering and classification, and the like.

For example, Han and Zhao wanted to automatically disambiguate names found online but faced many problems familiar to librarians: “The traditional methods measure the similarity using the bag of words (*BOW*) model. The *BOW*, however, ignores all the semantic relations such as social relatedness between named entities, associative relatedness between concepts, polysemy and synonymy between key terms. So the *BOW* cannot reflect the actual similarity.” To counter this, they constructed a semantic model from information on Wikipedia about the associative relationships of various ideas. They then used this model to find relationships between information found in the context of the target name in different pages. This enabled them to accurately group pages pertaining to particular individuals.⁸

Carmel, Roitman, and Zwerdling used page categories and titles to enhance labeling of document clusters. Although many algorithms exist for sorting large sets of documents into smaller, interrelated clusters, there is less work on labeling those clusters usefully. By extracting cluster keywords, using them to query Wikipedia, and algorithmically analyzing the results, they created a system whose top five recommendations contained the human-generated cluster label more than 85 percent of the time.⁹

Schönhofen looked at the same problem I examine—identifying document topics with Wikipedia data—but he used a different approach. He calculated the relatedness between categories and words from titles of pages belonging to those categories. He then used that relatedness to determine how strongly words from a target document predicted various Wikipedia categories. He found that although his results were skewed by how well-represented topics were on Wikipedia, “for 86 percent of articles, the top 20 ranked categories contain at least one of the original ones, with the top ranked category correct for 48 percent of articles.”¹⁰

Wikipedia has also been used as an ontology to improve clustering of documents in a corpus,¹¹ to automatically generate domain-specific thesauri,¹² and to improve Wikipedia itself by suggesting appropriate categories for articles.¹³

In short, Wikipedia has many uses for metadata enrichment. While text classification is one of these potential uses, and one with promise, it is under-explored at present. Additionally, this exploration takes place almost entirely in the proceedings of computer science conferences, often without reference to library science concepts or in a place where librarians would be likely to benefit from it. This paper aims to bridge that gap.

An Initial Test Case

To explore whether my method was feasible, I needed to try it on a test case. I chose Stephen Hawking’s *A Brief History of Time*, a relatively accessible meditation on the origin and fate of the universe, classified under “cosmology” by the Library of Congress. I began by looking up its SIPs on Amazon.com. Noticing that Amazon also lists Capitalized Phrases (CAPs)—“people, places, events, or important topics mentioned frequently in a book”—I included those as well (see table 1).¹⁴

I then queried Wikipedia via Google for each of these phrases, using queries such as “site:en.wikipedia.org ‘grand unification theory.’” I selected the top three Wikipedia article hits for each phrase. This yielded a list of sixty-one distinct items with several interesting properties:

- Four items appeared twice (Arrow of time, Entropy [arrow of time], Inflation [cosmology], Richard Feynman). However, nothing appeared more than twice; that is, nothing definitively stood out.
- Many items on the list were clearly relevant to *Brief History*, although often at too small a level of granularity to be good subject headings (e.g., Black hole, Second law of thermodynamics, Time in physics).
- Some items, while not unrelated, were wrong as subject classifications (e.g., List of Solar System objects by size, Nobel Prize in Physics).
- Some items were at best amusingly, and at worst bafflingly, unrelated (e.g., Alpha Centauri [Doctor Who], Electoral district [Canada], James K. Polk, United States men’s national soccer team).
- In addition, I had to discard some of the top Google hits because they were not articles but Wikipedia special pages, such as “talk” pages devoted to discussion of an article.

This test showed that I needed an approach that would give me candidate subject headers at a higher level of granularity. I also needed to be able to draw a brighter line between candidates and noncandidates. The presence of noncandidates was not in itself distressing—any automated approach will consider avenues a human would not—but not having a clear basis for discarding low-probability descriptors was a problem.

As it happens, Wikipedia itself offers candidate subject headers at a higher level of granularity via its categories system. Most articles belong to one or more categories, which are groups of pages belonging to the same list or topic.¹⁵ I hoped that by harvesting categories from the sixty-one pages I had discovered, I could improve my method.

This yielded a list of more than three hundred categories. Unsurprisingly, this list mostly comprised irrelevant

Table 1. SIPs and CAPs for *A Brief History of Time*

SIPs	grand unification energy, complete unified theory, thermodynamic arrow, psychological arrow, primordial black holes, boundary proposal, hot big bang model, big bang singularity, more quarks, contracting phase, sum over histories
CAPs	Alpha Centauri, Solar System, Nobel Prize, North Pole, United States, Edwin Hubble, Royal Society, Richard Feynman, Milky Way, Roger Penrose, First World War, Weak Anthropic Principle

candidates (“wars involving the states and peoples of Asia,” “video games with expansion packs,” “organizations based in Sweden,” among many others). Many categories played a clear role in the Wikipedia ecology of knowledge but were not suitable as general-purpose subject headers (“living people,” “1849 deaths”). Strikingly, though, the vast majority of candidates occurred only once. Only forty-two occurred twice, fifteen occurred three times, and one occurred twelve times: “physical cosmology.”

Twelve occurrences, four times as many as the next candidate, looked like a bright line. And “physical cosmology” is an excellent description of *Brief History*—arguably better than LCSH’s “cosmology.” The approach looked promising.

Automating Further Test Cases

The next step was to test an extensive variety of books to see if the method was more broadly applicable. However, running searches and collating queries for even one book is tedious; investigating a large number by hand was prohibitive. Therefore I wrote a categorization script (see appendix) that performs the following steps:¹⁶

- reads in a file of statistically improbable phrases¹⁷
- runs Google queries against Wikipedia for all of them¹⁸
- selects the top hits after filtering out some common Wikipedia nonarticles, such as “category” and “user” pages
- harvests these articles’ categories
- sorts these categories by their frequency of occurrence

This algorithm did not filter out Wikipedia administrative categories, as creating a list of them would have been prohibitively time-consuming. However, it would be

computationally trivial to do so, given such a list. (The list need not be exhaustive as long as it exhaustively described category types; for instance, the same regular expression could filter out both “articles with unsourced statements from October 2009” and “articles with unsourced statements from May 2008.”) At this stage of research, however, I simply ignored these categories in analyzing my results.

To find a variety of books to test, I used older *New York Times* nonfiction bestseller lists because brand-new books are less likely to have SIPs available on Amazon.¹⁹ These lists were heavily slanted toward autobiography, but also included history, politics, and social science topics.

Results

Of the thirty books I examined (the top fifteen each from paperback and hardback nonfiction lists), twenty-one had SIPs and CAPs available on Amazon. I ran my script against each of these phrase sets and calculated three measures for each resulting category list:

- Precision (P): of the top categories, how many were synonyms or near-synonyms of the book’s LCSHs?
- Recall (R): of the book’s LCSHs, how many had synonyms or near-synonyms among the top categories?
- Right-but-wrongs (RbW): of the top categories, how many are reminiscent of the LCSHs without actually being synonymous? These included narrower terms (e.g., the category “African_American_actors” when the LCSHs included “Actors—United States—Biography”), broader terms (e.g., “American_folk_singers” vs. “Dylan, Bob, 1941–”), related terms (e.g., “The_Chronicles_of_Narnia_books” vs. “Lion, the Witch and the Wardrobe (Motion picture)”), and examples (“Killian_documents_controversy” vs. “United States—Politics and government—2001–2009”).

I considered the “top categories” for each book to be the five that most commonly occurred (excluding Wikipedia administrative categories), with the following exceptions:

- Because I had no basis to distinguish between them, I included all equally popular categories, even if that would bring the total to more than five. Thus, for example, for the book *Collapse*, the most common category occurred seven times, followed by two categories with five appearances and six categories with four. Rather than arbitrarily selecting two of the six four-occurrence categories to bring the total to five, I examined all nine top categories.
- If there were more than five LCSHs, I expanded the number of categories accordingly, so as not to

misleadingly increase recall statistics.

- I did not consider any categories with fewer than four occurrences, even if that left me with fewer than five top categories to consider. The lists of three-, two-, and one-occurrence categories were very long and almost entirely composed of unrelated items.

I also considered, subjectively, the degree of overlap between the LCSHs and the top Wikipedia categories. I chose four degrees of overlap:

- *strong*: the top categories were largely relevant and included synonyms or near-synonyms for the LCSH
- *near miss*: some categories suggested the LCSH but missed its key points, such as

“Continental_Army_generals” vs. “United States—History—Revolution, 1775–1783.”

- *weak*: some categories treated the same subject as the LCSH but not at all in the same way
- *wrong*: the categories were actively misleading

The results are displayed in table 2.

Discussion

The results of this test were decidedly more mixed than those of my initial test case. On some books the Wikipedia method performed remarkably well; on

Table 2. Results (sorted by percentage of relevant categories).

Book	P	R	RbW	Subjective Quality
<i>Chronicles</i> , Bob Dylan	0.2	0.5	0.8	strong
<i>The Chronicles of Narnia: The Lion, the Witch and the Wardrobe Official Illustrated Movie Companion</i> , Perry Moore	0.25	1	0.625	strong
<i>1776</i> , David McCullough	0	0	0.8	near miss
<i>100 People Who Are Screwing Up America</i> , Bernard Goldberg	0	0	0.625	weak
<i>The Bob Dylan Scrapbook, 1956–1966</i> , with text by Robert Santelli	0.2	0.5	0.4	strong
<i>Three Weeks With My Brother</i> , Nicholas Sparks	0	0	0.57	weak
<i>Mother Angelica</i> , Raymond Arroyo	0.07	0.33	0.43	near miss
<i>Confessions of a Video Vixen</i> , Karrine Steffans	0.25	0.33	0.25	weak
<i>The Fairtax Book</i> , Neal Boortz and John Linder	0.17	0.33	0.33	strong
<i>Never Have Your Dog Stuffed</i> , Alan Alda	0	0	0.43	weak
<i>The World is Flat</i> , Thomas L. Friedman	0.4	0.5	0	near miss
<i>The Tender Bar</i> , J. R. Moehringer	0	0	0.2	wrong
<i>The Tipping Point</i> , Malcolm Gladwell	0	0	0.2	wrong
<i>Collapse</i> , Jared Diamond	0	0	0.11	weak
<i>Blink</i> , Malcolm Gladwell	0	0	0	wrong
<i>Freakonomics</i> , Steven D. Levitt and Stephen J. Dubner	0	0	0	wrong
<i>Guns, Germs, and Steel</i> , Jared Diamond	0	0	0	weak
<i>Magical Thinking</i> , Augusten Burroughs	0	0	0	wrong
<i>A Million Little Pieces</i> , James Frey	0	0	0	wrong
<i>Worth More Dead</i> , Ann Rule	0	0	0	wrong
<i>Tuesdays With Morrie</i> , Mitch Albom	No category with more than 4 occurrences			

others, it performed very poorly. However, there are several patterns here:

Many of these books were autobiographies, and the method was ineffective on nearly all of these.²⁰ A key feature of autobiographies, of course, is that they are typically written in the first person, and thus lack any term for the major subject—the author’s name. Biography, by contrast, is rife with this term. This suggests that including titles and authors along with SIPs and CAPs may be wise. Additionally, it might require making better use of Wikipedia as an ontology to look for related concepts (rather in the manner that Han and Zhao used it for name disambiguation).²¹

Books that treat a single, well-defined subject are easier to analyze than those with more sprawling coverage. In particular, books that treat a concept via a sequence of illustrative essays (e.g., *Tipping Point*, *Freakonomics*) do not work well at all. SIPs may apply only to particular chapters rather than to the book as a whole, and the algorithm tends to pick out topics of particular chapters (e.g., for *Freakonomics*, the fascinating chapter on Sudhir Venkatesh’s work on “Gangs_in_Chicago, _Illinois”²²) rather than the connecting threads of the entire book (e.g. “Economics—Sociological aspects”). The tactics suggested for autobiography might help here as well.

My subjective impressions were usually, but not always, borne out by the statistics. This is because some of the RbWs were strongly related to one another and suggested to a human observer a coherent narrative, whereas others picked out minor or dissimilar aspects of the book.

There was one more interesting, and promising, pattern: my subjective impressions of the quality of the categories were strongly predicted by the frequency of the most common category. Remember that in the *Brief History* example, the most common category, “physical cosmology,” occurred twelve times, conspicuously more than any of its other categories. Therefore I looked at how many times the top category for each book occurred in my results. I averaged this number for each subjective quality group; the results are in table 3.

In other words, the easier it was to draw a bright line between common and uncommon categories, the more likely the results were to be good descriptions of the work. This suggests that a system such as this could be used with very little modification to streamline categorization. For example, it could automatically categorize works when it met a high confidence threshold (when, for instance, the most common category has double-digit occurrence), suggest categories for a human to accept or reject at moderate confidence, and decline to help at low confidence.

It was also interesting to me that—unlike my initial test case—none of the bestsellers were scientific or technical works. It is possible that the jargon-intensive nature of science makes it easier to categorize accurately, hence

Table 3. Category Frequency and Subjective Quality

Subjective Quality of Categories	Frequencies of Most Common Category	Average Frequency of Most Common Category
strong	6, 12, 16, 19	13.25
near miss	5, 5, 7, 10	6.75
weak	4, 5, 6, 7, 8	6
wrong	3, 4, 4, 5, 5, 5, 7, 7	5

my method’s success with *A Brief History of Time*. I tested another technical, jargon-intensive work (N. Gregory Mankiw’s *Macroeconomics* textbook), and found that the method also worked very well, giving categories such as “macroeconomics” and “economics terminology” with high frequency. Therefore a system of this nature, even if not usable for a broad-based collection, might be very useful for scientific or other jargon-intensive content such as a database of journal articles.

Future Research

The method outlined in this paper is intended to be a proof of concept using readily available tools. The following work might move it closer to a real-world application:

- A configurable system for providing statistically improbable phrases; there are many options.²³ This would provide the user with more control over, and understanding of, SIP generation (instead of the Amazon black box), as well as providing output that could integrate directly with the script.
- A richer understanding of the Wikipedia category system. Some categories (e.g., “all articles with unsourced statements”) are clearly useful only for Wikipedia administrative purposes, not as document descriptors; others (e.g., “physical cosmology”) are excellent subject candidates; others have unclear value as subjects or require some modification (e.g., “environmental non-fiction books,” “macroeconomics stubs”). Many of these could be filtered out or reformatted automatically.
- Greater use of Wikipedia as an ontology. For example, a map of the category hierarchies might help locate headers at a useful level of granularity, or to find the overarching meaning suggested by several headers by finding their common broader terms. A more thorough understanding of Wikipedia’s relational structure might help disambiguate terms.²⁴

- A special-case system for handling books and authors that have their own article pages on Wikipedia.

In addition, a large-scale project might want to work from downloaded snapshots of Wikipedia (via <http://download.wikimedia.org/>), which could be run on local hardware rather than burdening their servers. This would require using something other than Google for relevance ranking (there are many options), with a corresponding revision of the categorization script.

Conclusions

Even a simple system, quickly assembled from freely available parts, can have modest success in identifying book categories. Although my system is not ready for real-world applications, it demonstrates that an approach of this type has potential, especially for collections limited to certain genres. Given the staggering volume of documents now being generated, automated classification is an important avenue to explore.

I close with a philosophical point. Although I have characterized this work throughout as automated classification, and it certainly feels automated to me when I use the script, it does in fact still rely on human judgment. Wikipedia's category structure and its articles linking text to title concepts are wholly human-created. Even Google's PageRank system for determining relevancy rests on human input, using web links to pages as votes for them (like a vast citation index) and the texts of these links as indicators of page content.²⁵ My algorithm therefore does not operate in lieu of human judgment. Rather, it lets me *leverage* human judgment in a dramatically more efficient, if also more problematic, fashion than traditional subject cataloging. With the volume of content spiraling ever further beyond our ability to individually catalog documents—even in bounded contexts like academic databases, which strongly benefit from such cataloging—we must use human judgment in high-leverage ways if we are to have a hope of applying subject cataloging everywhere it is expected.

References and Notes

1. Carol Tenopir. "Online Databases—Online Scholarly Journals: How Many?" *Library Journal* (Feb. 1, 2004), <http://www.libraryjournal.com/article/CA374956.html> (accessed Mar. 13, 2010).
2. "Amazon.com Statistically Improbable Phrases," Amazon.com, http://www.amazon.com/gp/search-inside/sipshelp.html/ref=sib_sip_help (accessed Mar. 13, 2010).
3. Hanne Albrechtsen. "Subject Analysis and Indexing: From Automated Indexing to Domain Analysis," *The Indexer*, 18, no. 4 (1993): 219.
4. Birger Hjørland, "The Concept of Subject in Information Science," *Journal of Documentation* 48, no. 2 (1992): 172; Jens-Erik Mai, "Classification in Context: Relativity, Reality, and Representation," *Knowledge Organization* 31, no. 1 (2004): 39; Jens-Erik Mai, "Actors, Domains, and Constraints in the Design and Construction of Controlled Vocabularies," *Knowledge Organization* 35, no. 1 (2008): 16.
5. Xiaohua Hu et al., "Exploiting Wikipedia as External Knowledge for Document Clustering," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009* (New York: ACM, 2009): 389.
6. Yannis Labrou and Tim Finin, "Yahoo! as an Ontology—Using Yahoo! Categories to Describe Documents," in *Proceedings of the Eighth International Conference on Information and Knowledge Management, Kansas City, MO, USA 1999* (New York: ACM, 1999): 180.
7. Kwan Yi, "Automated Text Classification using Library Classification Schemes: Trends, Issues, and Challenges," *International Cataloging & Bibliographic Control* 36, no. 4 (2007): 78.
8. Xianpei Han and Jun Zhao, "Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge," in *Proceeding of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009* (New York: ACM, 2009): 215.
9. David Carmel, Haggai Roitman, and Naama Zwerdling, "Enhancing Cluster Labeling using Wikipedia," in *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, USA* (New York: ACM, 2009): 139.
10. Peter Schönhofen, "Identifying Document Topics using the Wikipedia Category Network," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, China, 18–22 December 2006* (Los Alamitos, Calif.: IEEE Computer Society, 2007).
11. Hu et al., "Exploiting Wikipedia."
12. David Milne, Olena Medelyan, and Ian H. Witten, "Mining Domain-Specific Thesauri from Wikipedia: A Case Study," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, 22–26 December 2006* (Washington, D.C.: IEEE Computer Society, 2006): 442.
13. Zeno Gantner and Lars Schmidt-Thieme, "Automatic Content-Based Categorization of Wikipedia Articles," in *Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009, 7 August 2009, Suntec, Singapore* (Morristown, N.J.: Association for Computational Linguistics, 2009): 32.
14. "Amazon.com Capitalized Phrases," Amazon.com, http://www.amazon.com/gp/search-inside/capshelp.html/ref=sib_caps_help (accessed Mar. 13, 2010).
15. For more on the epistemological and technical roles of categories in Wikipedia, see <http://en.wikipedia.org/wiki/Wikipedia:Category>.
16. Two sources greatly helped the script-writing process: William Steinmetz, *Wicked Cool PHP: Real-World Scripts that Solve Difficult Problems* (San Francisco: No Starch, 2008); and the documentation at <http://php.net>.
17. Not all books on Amazon.com have SIPs, and books that do may only have them for one edition, although many editions may be found separately on the site. There is not a readily apparent pattern determining which edition features SIPs. Therefore

this step cannot be automated.

18. Be aware that running automated queries without permission is an explicit violation of Google's Terms of Service. See Google Webmaster Central, "Automated Queries," <http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=66357> (accessed Mar. 13, 2010). Before using this script, obtain an API key, which confers this permission. AJAX web search API keys can be instantly and freely obtained via <http://code.google.com/apis/ajaxsearch/web.html>.

19. "Hardcover Nonfiction," *New York Times*, Oct. 9, 2005, http://www.nytimes.com/2005/10/09/books/bestseller/1009besthardnonfiction.html?_r=1 (accessed Mar. 13, 2010); "Paperback nonfiction," *New York Times*, Oct. 9, 2005, http://www.nytimes.com/2005/10/09/books/bestseller/1009bestpaperonfiction.html?_r=1 (accessed Mar. 13, 2010).

20. For the purposes of this discussion I consider the

problematic *Million Little Pieces* to be autobiography, as it has that writing style, and as its LCSH treats it thus.

21. Han and Zhao, "Named Entity Disambiguation."

22. Sudhir Venkatesh, *Off the Books: The Underground Economy of the Urban Poor* (Cambridge: Harvard Univ. Pr., 2006).

23. See Karen Coyle, "Machine Indexing," *The Journal of Academic Librarianship* 34, no. 6 (2008): 530. She gives as examples PhraseRate (<http://ivia.ucr.edu/projects/PhraseRate/>), KEA (<http://www.nzdl.org/Kea/>), and Extractor (<http://extractor.com/>).

24. Per Han and Zhao, "Named Entity Disambiguation."

25. Lawrence Page et al., "The PageRank Citation Ranking: Bringing Order to the Web," Stanford InfoLab (1999), <http://ilpubs.stanford.edu:8090/422/> (accessed Mar. 13, 2010). This paper precedes the launch of Google; as the title indicates, the citation index is one of Google's foundational ideas.

Appendix. PHP Script for Automated Classification

```
<?php
/* This script takes two arguments: the first, a file with
comma-separated key phrases for a text; the second, the
number of Wikipedia hits to be included in classifying the
text. (The second argument is optional and will default
to 3 if not specified.)
*/
/* First, let's test the command-line arguments we were given,
if any.
Let's make sure we have a file to operate on.
*/
if (!array_key_exists(1, $argv)) {
echo "Please specify a .txt file which contains comma-separated key phrases.";
die;
}
if (!is_file($argv[1])) {
echo "I'm sorry; the first argument doesn't appear to be a file.";
die;
}
/* If the user has specified how far down we should plumb our
Google search, we'll go with it. Otherwise we need to set a
value for this parameter.
*/
if (!array_key_exists(2, $argv)) {
$argv[2] =3;
}
/* The Google default only returns 4 hits per query, so don't
let the user demand more.
*/
if ($argv[2] >4) {
echo "I'm sorry; the number specified cannot be more than 4.";
die;
}
// Next, turn our comma-separated list into an array.
```

Appendix. PHP Script for Automated Classification (continued)

```
$sip_temp = fopen($argv[1], 'r');
$sip_list = "";
while (! feof($sip_temp)) {
    $sip_list .= fgets($sip_temp, 5000);
}
fclose($sip_temp);
$sip_array = explode(' ', $sip_list);
/* Here we access Google search results for our SIPs and CAPs.
It is a violation of the Google Terms of Service to run
automated queries without permission. Obtain an AJAX API key
via http://code.google.com.
*/
$apikey = 'your_key_goes_here';
foreach($sip_array as $query) {
    /* In multiword terms, change spaces to + so as not to
break the google search.
*/
    $query = str_replace(" ", "+", $query);
    $googresult="http://ajax.googleapis.com/ajax/services/search/web?v=1.0&q=site%3Aen.wikipedia.org+$query&key=$apikey";
    $googdata = file_get_contents($googresult);
    // pick out the URLs we want and put them into the array $links
    preg_match_all('url:" [^"]*"li', $googdata, $links);
    /* Strip out some crud from the JSON syntax to get just
URLs
*/
    $links[0] = str_replace("\ url\".\\" " ", $links[0]);
    $links[0] = str_replace("\\" " ", $links[0]);
    /* Here we step through the links in the page Google
returned to us and find the top Wikipedia articles among
the results
*/
    $i=0;
    foreach($links[0] as $testlink) {
        /* These variables test to see if we have hit a
Wikipedia special page instead of an article.
There are many more flavors of special page, but
these are the most likely to show up in the first
few hits.
*/
        $filetest = strpos($testlink, 'wiki/File:');
        $cattest = strpos($testlink, 'wiki/Category:');
        $usertest = strpos($testlink, 'wiki/User');
        $talktest = strpos($testlink, 'wiki/Talk:');
        $disambtest = strpos($testlink, '(disambiguation)');
        $templatetest = strpos($testlink, 'wiki/Template_');
        if (!$filetest && !$cattest && !$usertest && !$talktest && !$disambtest && !$templatetest) {
            $wikipages[] = $testlink;
            $i++;
        }
    }
    /* Once we've accumulated as many article pages as the
user asked for, stop adding links to the $wikipages
array.
*/
}
```

Appendix. PHP Script for Automated Classification (continued)

```
if ($i == $argv[2]) {
break;
}
//This closes the foreach loop which steps through $links
}
// This closes the foreach loop which steps through $sip_array
}
/* For each page that we identified in the above step, let's
find the categories it belongs to.
*/
$mastercatarray = array();
foreach ($wikipages as $targetpage) {
// Scrape category information from the article page.
$wikiscrape = file_get_contents($targetpage);
preg_match_all("/wiki/Category.[^\"]+|",,$wikiscrape, $categories);
foreach ($categories[0] as $catstring) {
/* Strip out the "wiki/Category:" at the beginning of
each string
*/
$catstring = substr($catstring, 15);
/* Keep count of how many times we've seen this
category.
*/
if (array_key_exists($catstring, $mastercatarray)) {
$mastercatarray[$catstring]++;
} else {
$mastercatarray[$catstring] =1;
}
}
}
// Sort by value: most popular categories first.
arsort($mastercatarray);
echo "The top categories are:\n";
print_r($mastercatarray);
?>
```