2005

What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results

Tina Gross
St. Cloud State University, tmgross@stcloudstate.edu

Arlene G. Taylor
University of Pittsburgh - Main Campus, ataylor@sis.pitt.edu

Recommended Citation
Tina Gross and Arlene G. Taylor, "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results," College & Research Libraries 66, 3 (2005): 212-30.

What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results

Tina Gross and Arlene G. Taylor

Using controlled vocabulary in the creation and searching of library catalogs has evoked a great deal of debate because it is expensive to provide. Leading to this study were suggestions that because most users seem to search by keyword, subject headings could be removed from catalog records to save space and cost. This study asked, what proportion of records retrieved by a keyword search has a keyword only in a subject heading field and thus would not be retrieved if there were no subject headings? It was found that more than one-third of records retrieved by successful keyword searches would be lost if subject headings were not present, and many individual cases exist in which 80, 90, and even 100 percent of the retrieved records would not be retrieved in the absence of subject headings.

Once upon a time in library land, most searching of catalogs was done to find authors and titles. In fact, Ruth French Strout told us that in the 1830s in Great Britain statements were made that "classified catalogs and indexes were not needed because living librarians were better than subject catalogs… [and] any intelligent man who was sufficiently interested in a subject to want to consult material on it could just as well use author entries as subject, for he would, of course, know the names of all the authors who had written in his field."[1] This attitude prevailed through most of the twentieth century, even though Charles Cutter had persuaded American librarians to use subject headings in dictionary catalogs by the beginning of that century. Many catalog use studies have shown that most searches were for known items or at least for a known author. Although a few studies have shown that the majority of searches were subject searches, especially in public libraries, these studies have tended to be ignored.[2]

In the early 1990s, soon after online catalogs became relatively common, many librarians were quite surprised to

Tina Gross is a Hispanic/Latin American Languages Cataloger in the University Library System at the University of Pittsburgh; e-mail: tinag@pitt.edu. Arlene G. Taylor is a Professor Emerita in the School of Information Sciences at the University of Pittsburgh; e-mail: ataylor@mail.sis.pitt.edu. *The authors would like to thank Kevin Furniss for his assistance in obtaining keyword searches from the transaction log of the catalog of the library at Winthrop University, Rock Hill, S.C. We also wish to acknowledge the assistance of Jean Brumfield, Donna Capezzuto, Ximena Miranda, and Jo Tavener for their assistance in searching PittCat and counting occurrences of keywords in records. Finally, we want to thank Juan Pablo Zuluaga for his assistance with statistics.*

learn from various transaction log studies that a high proportion of searches in catalogs was for subject matter. At that time, subject searching still consisted of left-anchored searches for exact subject heading strings. Users could not yet browse lists of headings to find the exact string; therefore, many searches retrieved no hits. In his longitudinal study, Ray R. Larson found a gradual decline in subject searching but said it was obvious that the decline in subject index use percentages was being offset by the use of the title keyword index. That is, users were still trying to do subject searching, but because they knew so little about the controlled vocabulary, they did not know how to search it. (At that time, one could search titles by keyword, but almost no catalogs allowed keyword searching of subject headings, or indeed, any record fields other than title fields.) Larson concluded, "The subject index, even after the decline discussed above, is still one of the most commonly used search access points in the online catalog."[3]

In 2005, most online catalogs can search every field in a record, although moving from catalog to catalog can be quite confusing, with the definition of "keyword search" being quite different as to which fields are included in that search. However, students in schools of library and information science tell us that librarians often have recommended to them not to attempt subject searching but, instead, to use keyword searching when they wish to find information on a subject. This attitude has led to the suggestion (in at least one academic library) that subject headings should be stripped from the bibliographic records in the catalog. The argument was that thousands of subject headings needlessly take up gigabytes of space because users hardly ever search for subject headings. (And an unspoken cost saving, of course, would be that catalogers would not need to provide subject headings for new records.) The suggestion to remove subject headings was troubling to some experienced librarians who have observed that some keyword searches retrieve records in which one or more sought-after word(s) is found only in a subject string in a subject-heading field. That is, at least one keyword of a search is only in a subject field, not in any other field in the record, and thus if the subject headings were to be stripped out of current records and not added to new records, these records would not be found in response to that keyword search. But no one knew how often this happens.

Review of the Literature

In 1994, Jennifer Rowley reviewed the literature on the century-old debate about the use of controlled vocabulary versus the use of natural language for subject searching.[4] She divided the history of the debate into four eras:

1. Introduction of controlled vocabulary
2. Comparisons of indexing languages to determine which was best
3. Case studies of limited generalizability along with a general realization that perhaps the best subject searching was done by using both natural language (keyword) searching and controlled vocabulary searching in parallel
4. Development of systems for end users (including OPACs and indexing databases) and attempts to develop expert system techniques to support the representation of meaning.

Rowley mentioned work that was proceeding with artificial intelligence techniques that might someday integrate controlled indexing languages into the knowledge base of an expert system. However, she acknowledged that information retrieval, in practice, was still based on a mixture of natural and controlled indexing languages and that searchers were required to decide how much use of each would be an optimal combination in a search strategy.

Only a few articles have discussed the debate in the years following Rowley's thorough review. In 1995, Joy Tillotson investigated whether keyword searching

produced useful results, whether people who used keyword searches for subject searching were satisfied with the results, and whether OPAC interfaces available at that time offered and explained both keyword searching and controlled vocabulary searching.[5] She took failed subject heading searches (as found in transaction logs in a small library catalog and a large library catalog) and redid them as keyword searches. She then judged relevancy of the retrievals and found between 63 and 73 percent average likely relevancy. Her next step was to ask users about satisfaction with keyword searching. Her study produced too few responses from which to draw significant conclusions, but she stated, "Part of what happened is that people resorted to keyword searches when an exact search failed and then found nothing they liked with the keyword search either."[6] She concluded that both kinds of searches should be available. Tillotson's final step was to check available OPAC interfaces to determine how much help was given to users. She found that OPACs mostly provided both kinds of search but did not offer explanations for them or help with unsuccessful searches.

In the same year that Tillotson's article appeared, Monica Cahill McJunkin reported her study of title keyword searches.[7] She noted that the scope of the study did not involve comparing title keyword searching with subject searching, but, interestingly, she used the subject headings that were on the retrieved records to judge the relevancy of the responses. She observed that "Many exact subject heading matches were missed by title keyword searches."[8]

Also in 1995, Arlene G. Taylor reviewed the state of the art of subject access in library catalogs at the time.[9] Included was a section on controlled vocabulary versus keywords, in which the advantages and disadvantages of controlled vocabulary searching and keyword searching were reviewed. Concern was expressed about the metadata schemes then being developed with elements for subject terminology, but with little or no concern for controlled vocabulary. A particular problem involves the creation of metadata for images and objects using whatever words come to mind at the moment, rather than relying on controlled vocabulary. In such cases, of course, there is often no text provided by an author and titles may not be provided either.

In 1996, Brendan J. Wyly reported his investigation of a transaction log of a system that required users to give another command to the system in order to obtain location and circulation information for particular items after they had done a search for bibliographic records.[10] He hypothesized that a searcher's decision to obtain location information indicated that the searcher believed the record represented something worth pursuing. This was interesting because, as Wyly pointed out, other transaction analyses rated success as being whether a search retrieved anything and considered zero-hit searches to be "failures." He observed that such "failure," taken together with actions that follow it, might actually lead to success, as in the example of a user getting zero hits with a subject search for "Canoeing" and then using the word as a title keyword and discovering the subject heading "Canoes and canoeing" on a retrieved record. The searcher then may return to a subject search using "Canoes and canoeing" and be successful. Wyly stated, "Communication involves 'failure' because it necessarily involves feedback and learning. Online catalogs are communication devices."[11] He measured "success" as being a searcher's decision to obtain location information. He was able to link the decision to follow up with location information to the access point that had been used to find the bibliographic record in the first place. Of all such "successful" searches, about 30 percent were subject heading searches and about 25 percent were title keyword searches.

In 1997, Charles R. Hildreth reported the results of a study of keyword and

Boolean searching by users of an online catalog.[12] He found that "users of this online catalog search more often by keyword than any other type of search, their keyword searches fail more often than not, and a majority of these users do not understand how the system processes their keyword searches."[13] Although he did not discuss the presence or use of subject headings, his finding about the failure of keyword searches is relevant to this research.

A study reported in 1998 by Henk J. Voorbij comes closest to dealing with the question addressed by this study.[14] Voorbij indicated that because controlled vocabulary requires subject indexing, which is often conducted by highly paid employees, he wanted to learn whether the presence of controlled terms led to better results than searching by uncontrolled terms (title keywords, for the most part). He conducted two studies. In the first study, descriptors (i.e., controlled vocabulary) and title keywords were compared, and in the second study, subject searches on the same topics were performed using title keywords and subject descriptors. In comparing descriptors and title keywords, subject librarians were asked to judge whether the descriptor was the same (or almost the same) as a title word; whether the descriptor was a synonym; whether the descriptor was broader, narrower, or related; or whether the concept expressed by the descriptor appeared in the title at all. He then asked the participants to judge whether addition of the descriptors to the records resulted in enhancements that were "slight" or "considerable." The overall results showed that 37 percent of the records were considerably enhanced by a subject descriptor and another 12 percent were slightly enhanced.

The second study reported by Voorbij in the same article compared subject descriptor searches with title keyword searches for the same topics.[15] Each searcher conducted both a broad subject search and a narrow subject search, first using title keywords and then descriptors. He found that recall for searches conducted by using descriptors was 86.9 percent and recall for keyword searches was 48.2 percent. Voorbij offered two explanations for this large difference: (1) titles, although hardly ever completely meaningless, do not always offer sufficient clues for keyword searching; and (2) subject descriptors control the vocabulary, thus compensating for the wide diversity of ways to express a topic.

Research Question

The research question guiding this study was, What proportion of records retrieved by a keyword search has a keyword only in a subject heading field and thus would not be retrieved if there were no subject headings? The purpose of the study was to take an initial step toward finding the answer to this research question. Using captured searches from a transaction log, a series of keyword searches was performed to determine what proportion of the records retrieved by each user's search had a keyword only in a subject heading field and thus would not be retrieved if the subject headings were not there.

Methodology

The search terms used were obtained from a transaction log of 3,397 keyword searches from the catalog of the library at Winthrop University, Rock Hill, S.C., captured March 18–24, 2000. Some searches consisted of a single term each; others consisted of phrases or a string of two or more words. There were many repetitions of identical searches among the 3,397; 2,270 of the searches were unique. A sample of 227 of these searches was selected by using a common statistical formula for determining sample size.[16]

Keyword searches on each set of terms in the sample were performed in Pitt*Cat*, the University of Pittsburgh's OPAC, which contains more than three million titles from all of the university's libraries, including those on four regional campuses. To minimize the impact of duplicate holdings while including a broad range

of materials, the searches were limited to the holdings of the University Library System (which at the time the searches were performed consisted of fourteen libraries located on the main Pittsburgh campus and a remote storage facility) and the Law and Health Sciences libraries. The words "a," "an," "and," "by," "for," "from," "in," "of," "on," "or," "the," "to," and "with" were treated as stop words and omitted.

It was necessary to limit the searches to English because the vast majority of bibliographic records for foreign-language materials with English-language subject headings could only contain many of the English-language search terms from the sample in their subject headings. A very high proportion, in some cases 100 percent, of records for non-English-language materials could not be retrieved with English-language keyword searching in the absence of subject headings. This crucial factor makes subject headings even more essential for many bilingual users, but it was necessary to exclude foreign-language materials from this study because their inclusion could be viewed as "stacking" the results. For example, a keyword search for *literature brazil,* limited to English, would lose 33.2 percent of the hits it currently retrieves if the subject fields were not there.[17] The same search including materials in all languages would lose 56.7 percent of its hits. If the searches had not been limited to English, the results would have had less broad applicability and would have been representative only of libraries with a relatively high proportion of foreign-language materials.

In addition to completed records, Pitt*Cat* contains provisional acquisitions records with minimal bibliographic information. Because they contain no subject headings, their presence may have resulted a slightly smaller proportion of records retrieved with keywords only in a subject heading field, but there was no practical way to exclude them.

For each term or set of terms, the following kinds of data were collected:

- Number of hits with all keyword(s) anywhere
- Number of hits with all keyword(s) and at least one in subject, but not all in title
- Number of records (or of the first

FIGURE 1
Record with Keywords in Subject Headings and Also in Title

Gold, silver, and bronze : metal sculpture of the Roman baroque / Jennifer...

Title: Gold, silver, and bronze : metal sculpture of the Roman baroque / Jennifer Montagu.
Gold, silver & bronze
Author: Montagu, Jennifer.
Published: Princeton, N.J. : Princeton University Press, c1996.
Series: Bollingen series ; XXXV, 39. The A.W. Mellon lectures in the fine arts ; 1990
Bollingen series ; 35, 39.
A.W. Mellon lectures in the fine arts ; 1990.
ISBN: 0691027366 (CL : acid-free paper)
Physical Description: xvii, 262 p. : ill. (some col.) ; 29 cm.
LC Subject Heading(s) (limits do not apply): Metal sculpture, Baroque--Italy--Rome.
Metal sculpture, Italian--Italy--Rome.
Notes: Includes bibliographical references (p. [249]-255) and index.
Contents: 1. Introduction
2. Roman Bronzes around 1600
3. Adoration of the Host. Tabernacles of the Baroque
4. Medal. Some Drawings for the Hamerani
5. Silver Tribute from a Prince to a Grand Duke. The 'Piatti di San Giovanni'
6. Giardini and Giardoni. Silversmiths and Founders of the Eighteenth Century
7. From Rome to Lisbon. The Patronage of John V

App. A The Tabernacles of Jacopo del Duca

App. I The Ciborium of Sta Maria Maggiore

App. II The Ciborium of Sta Maria in Vallicella

App. III The Tabernacle of the Cappella Aldobrandini in Bologna

App. IV The Virgin and Child for Lisbon
App. V The Virgin of the Immaculate Conception for Lisbon.

fifty records) with at least one keyword in subject only

For example, the search "metal sculpture" had nineteen hits with the keywords anywhere. For a result as small as this one, it would have been possible to examine each hit manually to determine where the keywords appeared and which ones had one of the two words only in a subject field. In figure 1, for example, one can see that both keywords are in the title as well as the subject headings. This record would still have been retrieved if the subject headings had not been present.

Many of the sets retrieved were very large, and so to improve accuracy and reduce the number of records that would have to be viewed, a second search was performed to eliminate as many hits as possible that contained all of the keywords somewhere other than in subject fields. The second search performed on each set of keywords was for the number of hits containing all of the keywords, with at least one keyword in the subject fields, but not all of them in the title. (See figure 2.) This step removed the hits containing all of the keywords in a title field, a large subset of the hits that would still be retrieved if the records did not have subject headings. The second search (as shown in figure 2) was designed to eliminate records such as the one shown in figure 1 from the set of hits that needed to be examined manually.

Because keywords can appear in many fields (subject, title, author, series, notes, publication, physical description , etc.), it was still necessary for us to view the remaining hits. It could be the case that a keyword appeared in a subject field and not in the title, but also appeared in a contents note, a corporate author's name, or a publisher's name. In that case, the record would still be retrieved if the subject headings were not there.

For "metal sculpture," the result of the second search was ten hits. Manual examination of these ten hits found that three of them would still have been retrieved if the subject fields were not present because not all of the keywords appeared in the title, but all appeared in the record somewhere other than the subject fields. For example, in figure 3, "metal" appears in the title field and "sculpture" is in the author field.

The other seven hits had at least one of the keywords in a subject field only, such as in figure 4, where both "metal" and "sculpture" appear only in subject fields. Therefore, seven out of the total nineteen hits, or 36.8 percent, would not have been retrieved in the absence of subject headings. That is, they would be lost to a keyword search for "metal sculpture."

FIGURE 2
Second Search Performed to Reduce Hits Needing to Be Viewed Manually

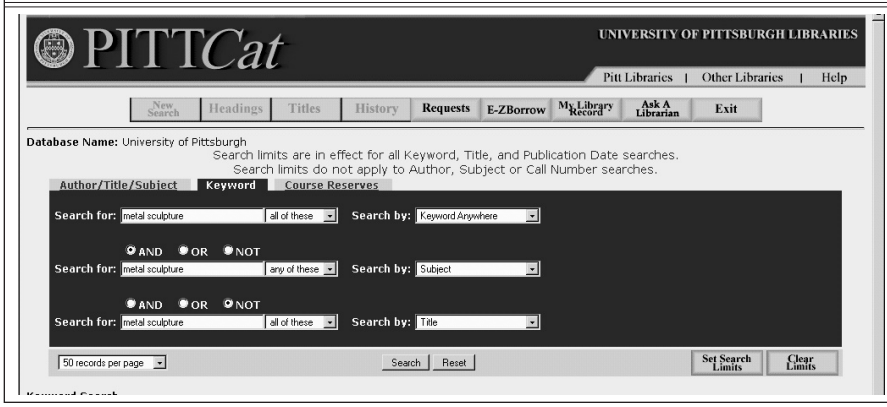


FIGURE 3
Record with All Keywords in Fields Other Than Subject Headings ("metal" in title, "sculpture" in both author and subject heading fields)

Enduring memory, in stone, in metal, in beauty.

Title: Enduring memory, in stone, in metal, in beauty.
Author: National Sculpture Society (U.S.)
Published: [New York, 1946?]
Physical Description: [52] p. illus., plates. 35 cm.
LC Subject Heading(s) (limits do not apply): Sepulchral monuments.
Sculpture.

When the retrieved set for a search was larger than fifty, only the first fifty records were viewed and the percentage of hits that would be lost from them was used to determine the percentage for the entire set. The first fifty were used rather than sampling because Pitt*Cat* displays results of keyword searches in reverse chronological order and thus the most recent, and presumably the most useful, hits appear first.

For example, the search "crime policy" had 388 hits with all of the keywords anywhere and 218 with all keywords in the record and at least one keyword in a subject heading, but not all of the keywords in a title field. Of these 218, forty-two of the first fifty had at least one keyword in a subject field only. These forty-two represent 84 percent of fifty. By applying this proportion to the entire set of 218, it was projected that the total number of hits with at least one keyword only in a subject field would be 183.1.

The final step for retrieved sets greater than fifty was to determine the percentage of hits that would be missed out of the total number of hits. For "crime policy," there were 388 hits with the keywords anywhere and a projected 183.1 hits with at least one keyword in a subject field only. Therefore, for the search "crime policy," an estimated 47.2 percent of the hits would not have been retrieved without the presence of subject headings.

Of the 227 searches selected for the sample, forty-one did not yield valid results and were rejected for the analysis (approx. 18% of the sample). Nine of these were searches that retrieved more than 10,000 hits, the maximum that Pitt*Cat* will display. (See table 1.) Given that the total number of hits for these searches was unknown, the proportion of hits lost could not be determined. Thirty-two of the searches retrieved no hits at all. (See table 2.) Many of these appeared to be typos or spelling errors. Others looked perfectly legitimate but retrieved no

FIGURE 4
Record with Keywords Only in Subject Headings

Paul Wayland Bartlett and the art of patination / Carol P. Adil, Henry A....

Title: Paul Wayland Bartlett and the art of patination / Carol P. Adil, Henry A. DePhillips, Jr.
Author: Adil, Carol P.
Contributors: DePhillips, Henry A.
Paul Wayland Bartlett Society.
Published: Wethersfield, Conn. : Paul Wayland Bartlett Society, c1991.
Physical Description: xiii, 80 p., [1] p. of plates : ill., port. ; 23 cm.
LC Subject Heading(s) (limits do not apply): Bartlett, Paul Wayland, 1865-1925.
Patina of metals.
Metal castings.
Sculpture, American.
Notes: "Books of reference [used by Bartlett]": p. 80.
Includes bibliography (p. ix-x).

TABLE 1
Searches Yielding More Than 10,000 Hits

diseases	civilization	welfare
space	electronic	1925
assessment	us trade	military

results. The analysis was performed on the remaining 186 valid searches. A list of these, along with data from the searches, may be found in the appendix.

Findings

The mean proportion of hits that would be lost in the absence of subject headings was 35.9 percent, and the median was 30.2 percent. The total percentage of all hits that would be lost if subject headings were not present, combining all of the searches, was 35.4 percent (36,319 out of 102,580 hits).

Because the average proportion of hits that would be lost increases as the number of keywords increases up to three, it was appropriate to consider whether the number of keywords included in a search might have an impact on the proportion of hits that would be lost if there were no subject headings. Searches with three keywords would lose an average of 44.9 percent of retrieved hits if the subject fields were not present, considerably higher than the overall average. (See table 3.) However, the median proportions of hits that would be lost by number of keywords does not display the same pattern, and regression analysis did not suggest any significant difference depending on the number of keywords.

There were many searches where the percentage of the hits that would not be found in the absence of subject headings was much higher than the averages. (See table 4.)

For about 31.7 percent of the searches, the percentage of hits with a keyword only in a subject field was 50 percent or greater. This means that for about three out of every ten successful keyword searches, half or more of the hits now retrieved would not be retrieved if there were no subject headings. For about four

TABLE 2
Searches Yielding Zero Hits

overcrowding classrooms	vintriloquism
cunningham imogene	artist michealangelou
elementry school foriegn language	south carolina government publication teachers
health care ukraine	simple science projects kids
hollecaust	pet doctors
mississippi river flora fauna	reviews screwtape letters
medgar wiley evars	baum l frank lymon frank 1856 1919
helathcare	games elementary student
music math link	excersize
neoclassic theaters	geometric patterns
nicaragua 1789 1914	female health care administrators
morning after pill	appeal situation comedies
racial identification terminology	teleproductions technology
pearstein philip	theater historyt
turbo charger	thomas eddison
winthrop college baseball	capital punishment china

TABLE 3
Results by Number of Keywords in Search

	All Searches	1 Keyword	2 Keywords	3 Keywords	4 or More Keywords
# of searches	186	44	98	30	14
Median # of hits	66	390	57.5	39.5	9
Average % lost	35.9%	26.0%	37.3%	44.9%	38.0%
Median % lost	30.2%	19.7%	36.6%	34.7%	26.5%

of every ten successful searches, more than 40 percent of hits would be lost; and for half of all successful searches, more than a third of hits would be lost.

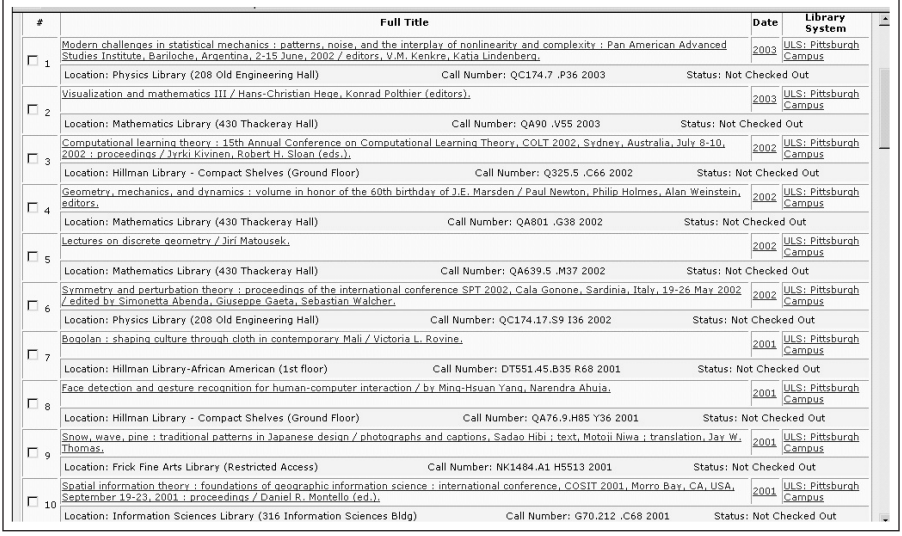
Since the time this study was conducted, the University of Pittsburgh library system has begun adding tables of contents (TOC) to many English-language monograph records with Blackwell's Table of Contents Enrichment Service. As more libraries use such TOC record enhancement services, this relatively new advance may rapidly become widespread. The positive aspect of such enhancement is that bibliographic records can be augmented substantially by providing chapter-level access, thus making it easier for users to assess the relevance of materials to their particular needs. These records also may include highly specific search terms not typically present in a traditional MARC record.

It seemed prudent to consider how this change might affect the results of this study, so several searches from the original sample were searched again in the TOC-enhanced catalog. It appears that even if all catalog records included a complete contents note, subject headings would still be essential. Although the inclusion of TOCs increases the number of hits and decreases the chances that a search will produce no hits at all, it also reduces precision; that is, it increases the number of irrelevant hits.[18] To return to one of the earlier examples, the search "metal sculpture" now yields considerably more hits, but among the first twenty-five results displayed are many items retrieved solely because of the

TABLE 4
Individual Searches with High Percentages of Hits Lost without Subject Headings

Keyword(s)	Number of Hits	Number of Hits With a Keyword in Subject Headings Only	% of Hits Retrieved That Would Be Missed Without Subject Headings
airplanes military parts	23	23	100%
businesswomen	173	171	98.8%
divorced people	55	51	92.7%
baptists united states	916	848.8	92.7%
horror films	402	332.8	82.8%
mass media politics	372	292.5	78.6%
history slang	22	17	77.3%
storytelling books	65	46.4	71.4%
hispanic americans	762	543.7	71.4%

FIGURE 5
First Ten Hits for "Geometric Patterns" Post-TOC Enrichment



#	Full Title	Date	Library System
1	Modern challenges in statistical mechanics : patterns, noise, and the interplay of nonlinearity and complexity : Pan American Advanced Studies Institute, Bariloche, Argentina, 2-15 June, 2002 / editors, V.M. Kenkre, Katia Lindenberg.	2003	ULS: Pittsburgh Campus
	Location: Physics Library (208 Old Engineering Hall) Call Number: QC174.7 .P36 2003 Status: Not Checked Out		
2	Visualization and mathematics III / Hans-Christian Hege, Konrad Polthier (editors).	2003	ULS: Pittsburgh Campus
	Location: Mathematics Library (430 Thackeray Hall) Call Number: QA90 .V55 2003 Status: Not Checked Out		
3	Computational learning theory : 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002 : proceedings / Jyrki Kivinen, Robert H. Sloan (eds.).	2002	ULS: Pittsburgh Campus
	Location: Hillman Library - Compact Shelves (Ground Floor) Call Number: Q325.5 .C66 2002 Status: Not Checked Out		
4	Geometry, mechanics, and dynamics : volume in honor of the 60th birthday of J.E. Marsden / Paul Newton, Philip Holmes, Alan Weinstein, editors.	2002	ULS: Pittsburgh Campus
	Location: Mathematics Library (430 Thackeray Hall) Call Number: QA801 .G38 2002 Status: Not Checked Out		
5	Lectures on discrete geometry / Jiri Matousek.	2002	ULS: Pittsburgh Campus
	Location: Mathematics Library (430 Thackeray Hall) Call Number: QA639.5 .M37 2002 Status: Not Checked Out		
6	Symmetry and perturbation theory : proceedings of the international conference SPT 2002, Cala Gonone, Sardinia, Italy, 19-26 May 2002 / edited by Simonetta Abenda, Giuseppe Gaeta, Sebastian Walcher.	2002	ULS: Pittsburgh Campus
	Location: Physics Library (208 Old Engineering Hall) Call Number: QC174.17.S9 I36 2002 Status: Not Checked Out		
7	Bogolan : shaping culture through cloth in contemporary Mali / Victoria L. Rovine.	2001	ULS: Pittsburgh Campus
	Location: Hillman Library-African American (1st floor) Call Number: DT551.45.B35 R68 2001 Status: Not Checked Out		
8	Face detection and gesture recognition for human-computer interaction / by Ming-Hsuan Yang, Narendra Ahuja.	2001	ULS: Pittsburgh Campus
	Location: Hillman Library - Compact Shelves (Ground Floor) Call Number: QA76.9.H85 Y36 2001 Status: Not Checked Out		
9	Snow, wave, pine : traditional patterns in Japanese design / photographs and captions, Sadao Hibi ; text, Motoji Niwa ; translation, Jay W. Thomas.	2001	ULS: Pittsburgh Campus
	Location: Frick Fine Arts Library (Restricted Access) Call Number: NK1484.A1 H5513 2001 Status: Not Checked Out		
10	Spatial information theory : foundations of geographic information science : international conference, COSIT 2001, Morro Bay, CA, USA, September 19-23, 2001 : proceedings / Daniel R. Montello (ed.).	2001	ULS: Pittsburgh Campus
	Location: Information Sciences Library (316 Information Sciences Bldg) Call Number: G70.212 .C68 2001 Status: Not Checked Out		

TOCs and summaries, items probably irrelevant to the user doing a search on "metal sculpture," such as:

- "Jazz modernism: from Ellington and Armstrong to Matisse and Joyce"
- "Collected poems "
- "Rapid prototyping casebook"
- "Animaculture" [book of poems]
- "The wound-dresser's dream"
- "Answered prayers: miracles and milagros along the border"

Many of these nonrelevant hits could be excluded from the results if the user performed a phrase search, selecting "as a phrase" from the drop-down menu instead of using the default "all of these." However, this also would eliminate potentially relevant hits where the words do not appear as a phrase, such as those in figures 3 and 4. The "as a phrase" search for "metal sculpture" now has thirteen hits, five of which have the phrase only in subject headings, and one ("Jazz Modernism") that, in the added summary, uses the concept of scrap-metal sculpture as a metaphor for the rebuilding of Tin Pan Alley (in jazz).

One search that yielded no hits at the time of the study, "morning after pill," now retrieves two hits of questionable relevance: "Paper trail: common sense in uncommon times" (which includes essays titled "Good Morning Spamerica," "After 20 Years of Cultivation…, "A Pill for What Haunts You") and "Fear of dreaming: the selected poems of Jim Carroll" (which includes poems titled "Morning," "After St. John of the Cross," and "Blue Pill"). It still retrieves zero hits "as a phrase." Another search that retrieved no hits, "geometric patterns," provides a good illustration of both the benefits and the drawbacks of including TOCs. Such a specific topic is not well represented by subject headings. Although it had zero hits earlier, the search now retrieves more than fifty records. However, many of them do not appear relevant, including eight out of the first ten results displayed. (See figures 5 and 6.)

Although a sophisticated searcher would likely make "geometric patterns" a phrase search and find more satisfying results, this example reflects what might be the experience of the average user, who tends to use the default settings without fully understanding them.[19] Moreover, there are many keyword searches for

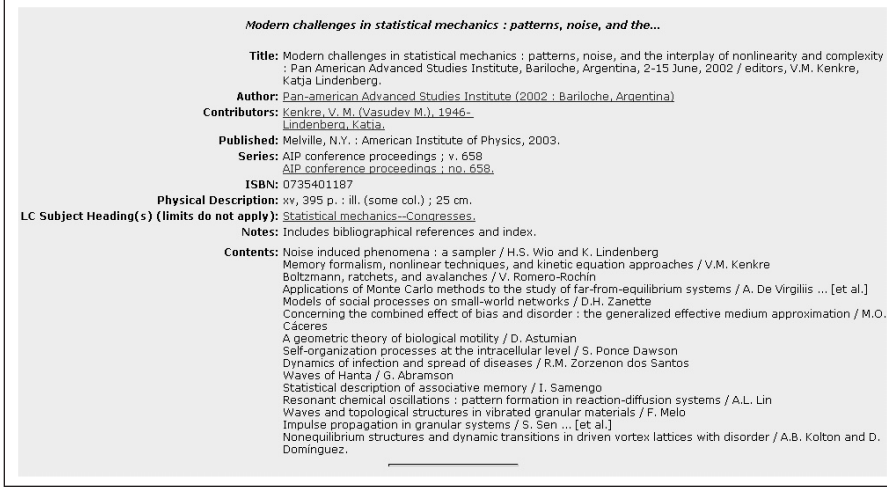
which performing a phrase search would be of no use, most obviously one-word searches, which comprised 23.7 percent of the searches in this study. For example, the first ten hits for the search "athletes" now include:

- "Confronting the body: the politics of physicality in colonial and postcolonial India" (includes chapter " Schools, athletes and confrontation: the student body in colonial India")
- "Diagnosis and management of hypertrophic cardiomyopathy" (one of its 31 chapters is "Cardiovascular causes of sudden death, preparticipation screening, and criteria for disqualification in young athletes")
- "The dietitian's guide to vegetarian diets: issues and applications" (includes the word "athletes" in the summary)
- "Diversity issues in American colleges and universities: case studies for higher education and student affairs professionals" (includes "Advising African American Student Athletes")
- "Legal medicine" (one of its 75 articles is "Competitive Athletes: Cardiovascular Preparticipation Screening")
- "Multiple literacies for the 21st century" (one of its 23 chapters is "Conceptual Diversity across Multiple Contexts: Student Athletes on the Court and in the Classroom")
- "Nutritional concerns of women" (one of its 21 chapters is "Nutritional Concerns of Female Recreational Athletes")

Seven out of the first ten do not appear relevant for someone doing a general search on athletes. The first hit returned, however, is: "The bases were loaded (and so was I): up close and personal with the greatest names in sports." If users performing this search opened this record (assuming that, as in Pitt*Cat*, the subject headings are displayed in the initial "brief" view), they would see the subject heading "Athletes—Biography." If users clicked on it, they would retrieve a list of subject headings from which they could select or scroll forward or backward for more, a far more user-friendly result than the list of records retrieved by the keyword search only. (See figure 7.) Unfortunately, if the records did not have subject headings, this possibility would not exist.

Future research needs to be conducted to determine the full effect of the addition of TOC data and summaries to catalog records. Especially important will be an at-

FIGURE 6
First Record Displayed for the Search "Geometric Patterns"



Modern challenges in statistical mechanics : patterns, noise, and the...

Title: Modern challenges in statistical mechanics : patterns, noise, and the interplay of nonlinearity and complexity : Pan American Advanced Studies Institute, Bariloche, Argentina, 2-15 June, 2002 / editors, V.M. Kenkre, Katja Lindenberg.
Author: Pan-american Advanced Studies Institute (2002 : Bariloche, Argentina)
Contributors: Kenkre, V. M. (Vasudev M.), 1946-
Lindenberg, Katja.
Published: Melville, N.Y. : American Institute of Physics, 2003.
Series: AIP conference proceedings ; v. 658
AIP conference proceedings ; no. 658.
ISBN: 0735401187
Physical Description: xv, 395 p. : ill. (some col.) ; 25 cm.
LC Subject Heading(s) (limits do not apply): Statistical mechanics--Congresses.
Notes: Includes bibliographical references and index.
Contents: Noise induced phenomena : a sampler / H.S. Wio and K. Lindenberg
Memory formalism, nonlinear techniques, and kinetic equation approaches / V.M. Kenkre
Boltzmann, ratchets, and avalanches / V. Romero-Rochín
Applications of Monte Carlo methods to the study of far-from-equilibrium systems / A. De Virgiliis ... [et al.]
Models of social processes on small-world networks / D.H. Zanette
Concerning the combined effect of bias and disorder : the generalized effective medium approximation / M.O. Cáceres
A geometric theory of biological motility / D. Astumian
Self-organization processes at the intracellular level / S. Ponce Dawson
Dynamics of infection and spread of diseases / R.M. Zorzenon dos Santos
Waves of Hanta / G. Abramson
Statistical description of associative memory / I. Samengo
Resonant chemical oscillations : pattern formation in reaction-diffusion systems / A.L. Lin
Waves and topological structures in vibrated granular materials / F. Melo
Impulse propagation in granular systems / S. Sen ... [et al.]
Nonequilibrium structures and dynamic transitions in driven vortex lattices with disorder / A.B. Kolton and D. Domínguez.

FIGURE 7
List of Subject Headings Retrieved through a Subject Heading Link in a Bibliographic Record

#	Titles	Headings	Headings Type
1	17	Athletes Biography	LC subject headings
2	1	Athletes Biography Dictionaries.	LC subject headings
3	1	Athletes Biography Juvenile literature.	LC subject headings
4	1	Athletes, Black	LC subject headings
5	1	Athletes, Black Biography.	LC subject headings
6	1	Athletes, Black Biography Dictionaries.	LC subject headings
7	1	Athletes, Black Canada Public opinion History 20th century.	LC subject headings
8	1	Athletes, Black Pennsylvania Pittsburgh Interviews.	LC subject headings
9	1	Athletes, Black Portraits.	LC subject headings
10	2	Athletes Canada Biography.	LC subject headings
11	4	Athletes China.	LC subject headings
12	1	Athletes Competitions Fiction.	LC subject headings
13	1	Athletes Conduct of life.	LC subject headings
14	1	Athletes Congresses.	LC subject headings
15	5	Athletes Counseling of.	LC subject headings
16	1	Athletes, Cuban.	LC subject headings
17	1	Athletes Czechoslovakia.	LC subject headings
18	8	Athletes Diseases.	LC subject headings
19	1	Athletes Drug testing Bibliography.	LC subject headings
see also 20	5	Athletes Drug use	LC subject headings
21	1	Athletes Drug use.	LC subject headings for children
22	1	Athletes Drug use Bibliography.	LC subject headings
23	5	Athletes Drug use United States.	LC subject headings
24	1	Athletes Education, Higher.	LC subject headings
25	3	Athletes Education United States.	LC subject headings
26	1	Athlètes féminins Santé et hygiène.	Repertoire des vedettes-matiere
27	4	Athletes Fiction.	LC subject headings for children
28	1	Athletes Finance, Personal United States Congresses.	LC subject headings
29	3	Athletes Greece.	LC subject headings
30	2	Athletes Greece Biography Dictionaries.	LC subject headings
31	1	Athletes Greece History.	LC subject headings

tempt to determine the effect on precision of the dramatic increase in recall that is occurring with this addition.

Conclusion

This study found that if subject headings were to be removed from or no longer included in catalog records, users performing keyword searches would miss more than one third of the hits they currently retrieve. On average, 35.9 percent of hits would not be found. (Although establishing precision was not the aim of this study, it is likely that this missing 35.9 percent would include a high proportion of relevant hits.) These findings are consistent with that of Voorbij, whose study concluded that 37 percent of the records used in his study were "considerably enhanced" by a subject descriptor and another 12 percent were slightly enhanced.

Of course, the loss of hits would be in addition to the loss of other functions and advantages provided by subject headings and controlled vocabulary in general, summarized by Voorbij as:

1) enhancing of the bibliographic record of a publication;

2) grouping synonyms, other ways to express a topic, and terms in foreign languages under the same heading;

3) suggesting other entries by cross-references;

4) reducing irrelevant hits.[20]

Without subject headings, a user whose keyword search produced an overwhelming number of hits with a high proportion of "false" ones would have few options in trying to find a smaller, more relevant set of hits. Subject headings allow users to perform additional searches using headings found in records they deem relevant, providing a simple means to limit retrieval to materials more likely to be relevant. This is especially true now that performing such a subject search can be done in most catalogs just by clicking on the heading. And, finally, as has been found by this research, subject headings allow the retrieval of relevant records that could not be retrieved with some keyword searches because one or more

of the words being sought do not appear anywhere else in the record except in a subject heading.

What might we lose if subject headings were not added to bibliographic records? We would lose more than one-third of the retrievals that users now see in response to their keyword searches and, in addition, we would lose a powerful tool for narrowing retrievals to the most relevant hits. And, arguably, a much larger proportion of the lost one-third would be relevant to the users than is found in the remaining two-thirds that would be retrieved.

Notes

1. Ruth French Strout, "The Development of the Catalog and Cataloging Codes," *Library Quarterly* 26 (Oct. 1956): 267–68.

2. Karen Markey, *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs* (Dublin, Ohio: OCLC Online Computer Library Center, 1984).

3. Ray R. Larson, "The Decline of Subject Searching: Long-term Trends and Patterns of Index Use in an Online Catalog," *Journal of the American Society for Information Science* 41 (Apr. 1991): 207.

4. Jennifer Rowley, "The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research," *Journal of Information Science* 20, no. 2 (1994): 108–19.

5. Joy Tillotson, "Is Keyword Searching the Answer?" *College & Research Libraries* 56 (May 1995): 199–206.

6. Ibid., 203.

7. Monica Cahill McJunkin, "Precision and Recall in Title Keyword Searches," *Information Technology and Libraries* 14 (1995): 161–71.

8. Ibid., 170.

9. Arlene G. Taylor, "On the Subject of Subjects," *Journal of Academic Librarianship* 21, no. 6 (Nov. 1995): 484–90.

10. Brendan J. Wyly, "From Access Points to Materials: A Transaction Log Analysis of Access Point Value for Online Catalog Users," *Library Resources &Technical Services* 40, no. 3 (July 1996): 211–36.

11. Ibid., 214.

12. Charles R. Hildreth, "The Use and Understanding of Keyword Searching in a University Online Catalog," *Information Technology and Libraries* 16 (June 1997): 52–62.

13. Ibid., 61.

14. Henk J. Voorbij, "Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences," *Journal of Documentation* 54, no. 4 (Sept. 1998): 466–76.

15. Ibid., 470–75.

16. David S. Moore and George P. McCabe, *Introduction to the Practice of Statistics,* 2nd ed. (New York: Freeman, 1993), 438. The formula used was $N = (z^*\sigma/m)^2$, with the values $z^* = 1.96$ for 95% confidence; $\sigma = .3$ for the standard deviation estimated from preliminary searching, and $m = .04$ for a 4% margin of error. The resulting equation was $((1.96)(.3)/.04)^2 = 216.09$. Because the total number of unique searches (2,270) divided by the desired sample size (216.09) came out to 10.5, we decided for simplicity's sake to select every tenth unique search for the sample, although this made the sample size slightly larger than it needed to be to achieve 95% confidence and a 4% margin of error.

17. Search performed in 2004 after inception of TOC enhancements to PittCat.

18. Rowley, "The Controlled versus Natural Indexing Language Debate Revisited," 114; Tillotson, "Is Keyword Searching the Answer," 199; McJunkin, "Precision and Recalling title Keyword Searches," 163; Hildreth, "The Use and Understanding of Keyword Searching in a University Online Catalog," 61.

19. Steve Jones et al., "A Transaction Log Analysis of a Digital Library," *International Journal on Digital Libraries* 3, no. 2 (2000): 155–56.

20. Voorbij, "Title Keywords and Subject Descriptors," 475–76.

APPENDIX
The Sample

Keyword(s)	No. of Total Hits with Keyword Anywhere	No. of Hits with Keyword Anywhere and 1 or More in Subjects But Not All in Title	No. of 1st 50 Records in Column 3 Not Retrieved if at Least 1 Word Not in Subject	No. of Total Records with a Keyword in Subject Only	Proportion of Col. 2 Records with a Keyword in Subject Only
photography printing processes	10	10	10	10	1
stuttering therapy methods	5	5	5	5	1
juvenile folk tales	71	71	50	71	1
dwellings remodeling	25	25	25	25	1
airplanes military parts	23	23	23	23	1
illumination books manuscripts celtic	20	20	20	20	1
labor productivity private service united states	3	3	3	3	1
jamaicas history	2	2	2	2	1
businesswomen	173	171	50	171	0.988439
television serials	43	41	40	40	0.930233
divorced people	55	51	50	51	0.927273
baptists united states	916	903	47	848.82	0.926659
automobile travel	126	117	49	114.66	0.91
indian pottery	243	225	49	220.5	0.907407
afro american actors	10	10	9	9	0.9
act philosophy	165	148	48	142.08	0.861091
lesson planning	111	95	50	95	0.855856
interprofessional relations	102	95	45	85.5	0.838235
attitude psychology	574	542	44	476.96	0.830941
roman civilization	331	280	49	274.4	0.829003
horror films	402	354	47	332.76	0.827761
manic depressive illness	217	191	47	179.54	0.827373
educational games	188	164	47	154.16	0.82
mass media politics	372	325	45	292.5	0.78629
history slang	22	17	17	17	0.772727
schenkerian analysis	16	13	12	12	0.75
humus	18	13	13	13	0.722222

kinetic sculpture	7	5	5	5	0.714286
storytelling books	65	58	40	46.4	0.713846
hispanic americans	762	697	39	543.66	0.713465
cubans	99	67	50	67	0.676768
punic wars	12	8	8	8	0.666667
psychological measurement instruments	9	7	6	6	0.666667
plastics craft	3	3	2	2	0.666667
computers	7302	4808	50	4808	0.65845
preventive health services	427	351	40	280.8	0.657611
violence motion pictures	52	41	34	34	0.653846
self directed work teams	17	11	11	11	0.647059
motion pictures behavior	31	21	20	20	0.645161
catholic church	5158	4599	36	3311.28	0.64197
mongolia history	56	41	35	35	0.625
desert reclamation	21	14	13	13	0.619048
infibulation	18	12	11	11	0.611111
religion brazil	45	33	27	27	0.6
women administration	627	435	43	374.1	0.596651
history can	971	734	39	572.52	0.589619
organizational sociology	198	166	34	112.88	0.570101
greenhouse	520	296	50	296	0.569231
us government publications	1936	1616	34	1098.88	0.567603
athletes	466	285	46	262.2	0.562661
television broadcasting	2226	1343	46	1235.56	0.555058
robots	422	260	45	234	0.554502
international socialist congress	39	21	21	21	0.538462
schools prayer	72	42	37	37	0.513889
surrealism	269	136	50	136	0.505576
nerves	460	231	50	231	0.502174
prayer schools	70	40	35	35	0.5
united states divorce rates	10	8	5	5	0.5
musical competitions	2	2	1	1	0.5

solzhenitsyn aleksandr isaevich 1918	99	56	44	49.28	0.497778
music influences	96	66	36	47.52	0.495
secret societies	84	41	41	41	0.488095
crime policy	388	218	42	183.12	0.471959
slavery america	396	221	42	185.64	0.468788
art sculpture	1567	1105	33	729.3	0.465412
ethics business	611	353	40	282.4	0.462193
ball games	35	25	16	16	0.457143
animal farm	58	30	25	25	0.431034
u s trade policy	3278	2235	31	1385.7	0.422727
gravitation	261	109	50	109	0.417625
education bilingual	1312	728	37	538.72	0.41061
fabric history	66	37	27	27	0.409091
political conventions	112	73	31	45.26	0.404107
voter characteristics	5	2	2	2	0.4
college students	3174	1892	33	1248.72	0.393422
fitness	731	283	50	283	0.387141
video games	57	27	22	22	0.385965
lightning war	13	6	5	5	0.384615
yugoslavia	1885	737	48	707.52	0.375342
farm engines	8	3	3	3	0.375
oil pollution	453	290	29	168.2	0.371302
metal sculpture	19	10	7	7	0.368421
general relativity physics	191	174	20	69.6	0.364398
film criticism	930	846	20	338.4	0.363871
africa north	788	354	40	283.2	0.359391
furniture	664	224	50	224	0.337349
censorship television	24	16	8	8	0.333333
deception advertising	12	5	4	4	0.333333
teaching foreign language	1286	1180	18	424.8	0.330327
women movies	66	32	21	21	0.318182
bosnia	545	173	50	173	0.317431
advertising	2450	841	46	773.72	0.315804
medieval	7550	2436	47	2289.84	0.30329
child sexual abuse	815	583	21	244.86	0.300442
language development problems	40	20	12	12	0.3

tales	6504	1911	50	1911	0.293819
abortion	1092	344	46	316.48	0.289817
paper manufacture	52	17	15	15	0.288462
china history opium war 1840 1842	25	25	7	7	0.28
agnosticism	29	11	8	8	0.275862
black power movement	29	1	8	8	0.275862
louise nevelson	15	5	4	4	0.266667
corporal punishment	34	10	9	9	0.264706
public school	4929	1561	41	1280.02	0.259692
law enforcement	3774	1479	32	946.56	0.250811
installation art	48	25	12	12	0.25
annual reviews physical chemistry	8	3	2	2	0.25
popular music college students	4	3	1	1	0.25
causes crimean war 1853 1856	4	4	1	1	0.25
music culture	752	375	24	180	0.239362
united states trading japan	18	13	4	4	0.222222
dance	3085	752	44	661.76	0.214509
judy chicago	47	11	10	10	0.212766
drug addiction	396	183	23	84.18	0.212576
bronze	607	127	50	127	0.209226
english official language	108	33	22	22	0.203704
real estate financing	63	25	12	12	0.190476
teamwork	163	30	30	30	0.184049
frank rizzo	8	2	1	1	0.125
machining	105	14	13	13	0.12381
body art	261	124	13	32.24	0.123525
communications	8634	1069	47	1004.86	0.116384
steinbeck john	190	54	20	21.6	0.113684
opera	2139	236	50	236	0.110332
womens glass ceiling	10	1	1	1	0.1
mozart	510	64	39	49.92	0.097882
marines	167	17	16	16	0.095808
historical romance	106	27	10	10	0.09434
rothko	33	4	3	3	0.09090
art degas	50	32	9	9	0.18

sexual violence	301	114	23	52.44	0.174219
living hard	35	6	6	6	0.171429
campaign 2000	108	41	18	18	0.166667
girls women sports	47	28	7	7	0.148936
alternative treatments	28	8	4	4	0.142857
religious denomina-tions	44	7	6	6	0.136364
eating disorders	466	174	18	62.64	0.134421
islam china	23	3	3	3	0.130435
responsibility	2995	409	47	384.46	0.128367
black white photog-raphy	29	5	1	1	0.034483
noguchi	91	3	2	2	0.021978
degas	94	3	2	2	0.021277
seuss	48	1	1	1	0.020833
charlie brown	49	2	1	1	0.020408
lee smith	384	18	7	7	0.018229
ramsey	526	9	8	8	0.015209
nader	209	4	2	2	0.009569
eighties	358	2	2	2	0.005587
encyclopedia	3596	7	4	4	0.001112
carson david	81	4	0	0	0
arnold lobel	34	0	0	0	0
gormley	33	0	0	0	0
bilingual education act	31	21	0	0	0
collins phil	14	0	0	0	0
food webs	13	1	0	0	0
american zoologist	5	0	0	0	0
speech impediments	5	0	0	0	0
programs about college students	5	2	0	0	0
screwtape letters	5	0	0	0	0
nutrient cycle	4	0	0	0	0
jargons	4	0	0	0	0
william r hearst	3	1	0	0	0
habitual offenders	3	0	0	0	0
effects music lyrics	3	1	0	0	0
sexism music	2	1	0	0	0
4mat	2	0	0	0	0

counseling native americans	2	1	0	0	0
constituion	2	0	0	0	0
women health care managers	2	1	0	0	0
torture devices	1	0	0	0	0
bereavement instruments	1	0	0	0	0
society view college students	1	0	0	0	0
pierre bonnard	1	1	0	0	0
how make resume	1	0	0	0	0
sports quotes	1	0	0	0	0
children television	1	0	0	0	0