Discovering Semantic Patterns in Bibliographically Coupled Documents

JIAN QIN

ABSTRACT

ISSUES IN DISCOVERING KNOWLEDGE IN BIBLIOGRAPHIC databases are addressed. An example of semantic pattern analysis is used to demonstrate the methodological aspects of knowledge discovery in bibliographic databases. The semantic pattern analysis is based on the keywords selected from the documents grouped by bibliographical coupling. The frequency distribution patterns suggest the existence of a common intellectual base with a wide range of specialties and marginal areas in the antibiotic resistance literature. The resulting values for keyword density per rank show a difference of ten times between the specialty and marginal keyword densities. The possibilities and further studies of incorporating knowledge discovery results into information retrieval are discussed.

INTRODUCTION

Knowledge discovery in databases (KDD) is considered a process of nontrivial extraction of implicit, previously unknown, and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases (Chen, Han, & Yu, 1996, p. 866). Most research on KDD has focused on applications in business operations and well-structured data. Knowledge discovery in textual databases has been underemphasized (Trybula, 1997). Among the limited publications on KD in textual databases, the full-text document data are the primary source of analysis. Lent, Agrawal, and Srikant (1997) developed a patent mining system at IBM for identifying trends in large textual databases over a period of time. They

Jian Qin, School of Information Studies, Syracuse University, 4-206 Center for Science and Technology, Syracuse, NY 13244
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 109-132

used sequential pattern mining to identify recurring phrases and generate histories of phrases, after which they then extracted phrases that satisfied a specific trend. Discovering associations among the keywords in texts is another area of research in KD in textual databases. Using background knowledge about the relationships of keywords, Feldman and Hirsh (1996) studied associations among the keywords or concepts representing the documents. The knowledge base they built supplies unary or binary relations among the keywords representing the documents. Feldman, Dagan, and Hirsh (1998) developed a system for Knowledge Discovery in Text (KDT) that extracts keywords to represent document contents and allows users to browse a list of keywords that co-occur with another keyword(s) for knowledge discovery purposes.

Mining in full-text documents attempts to extract useful associations and patterns for representing the document content, including clustering, categorization, summarization, and feature extraction. While many studies using data from bibliographic databases were not conducted in terms of KDD or data mining, they nevertheless bear the marks of KDD's techniques and analysis. Such examples can be found in citation and cocitation analysis (Kassler, 1965; Small, 1973; Small & Sweeney, 1985; Braam, Moed, & van Raan, 1991), keyword classifications (Sparck Jones & Jackson, 1970), investigation of indexing similarities between keywords and controlled vocabularies (Shaw, 1990; Qin, in press), and author mapping (Logan & Shaw, 1987). Discovering knowledge through mining textual data in bibliographic databases presents more problems than mining numerical data. One problem is that most fields in a bibliographic database have long character strings—e.g., author name, title, affiliation, journal title, and indexing terms (from both keywords and controlled vocabularies). Such long strings are usually difficult for statistical packages or data mining software to perform computational tasks. Unlike the full-text document source, bibliographic data are semi-structured. Although it may be an advantage over completely unstructured full-text documents, it also creates a challenge for mining tools that the data in the structured fields should not be mixed up when extracting data sets and performing analysis. Linguistic problems (such as singulars and plurals, stems and suffixes) and inconsistencies in abbreviating journal titles and institution names can also be challenging issues in mining bibliographic data. To obtain valid and reliable data for discovering trends and patterns in subject fields and research, data preprocessing and cleansing can become very time-consuming and both labor and intellectually intensive. However, the most challenging issue remains whether there is a chance for information retrieval systems to "be extended to become knowledge discovery systems," or whether "the kinds of record existing in bibliographical and textual databases offer any possibility of analysis in ways similar to those in more structured factual databases" (Vickery, 1997, pp. 119-20).

This study selected a set of bibliographic records as the data source for discovering semantic patterns among the keywords in these records. The purpose of this keyword analysis was to discover if any semantic patterns existed in the keywords extracted from bibliographically coupled documents regarding antibiotic resistance in pneumonia.

Also, if such patterns did exist, how the discovered knowledge about a subject field can be used to improve the effectiveness of knowledge representation and information retrieval. A preliminary test of antibiotic resistance in pneumonia literature found that documents citing the same publication not only co-cited other publications but also contained semantically similar or same keywords in the titles of cited publications. The frequency distributions of these keywords characterized three distinctive strata: a very small number of keywords falling into the highest frequency region, a relatively larger group with moderate occurrences, and a majority of them appearing only once or twice. If the terms occurring most frequently represent the intellectual base in this subject area (Small, 1973; Small & Sweeney, 1985) and the ones with medium occurrences represent the specialties, then the terms occurring least frequently represent the marginal terms. These marginal terms may be the links between the mainstream of the antibiotic resistance research to the less overt but promising research. The citation-semantic analysis is aimed at discovering semantic patterns of the antibiotic resistance literature so that the analysis process and semantic patterns can be programmed into tools that can assist information searchers in building search queries and customizing their post-search analysis. Specifically, this project studied whether the distribution follows the three strata described earlier, how such distribution can be measured, and to what extent the keywords in these strata reflect the research front in antibiotic resistance. The methods used to preprocess and analyze the data are discussed in detail in the following sections.

Research Design

The first and most important step in KDD is to clarify what kinds of knowledge are to be discovered, because this decides what types of data or database one needs to work on and what techniques to use for discovering the knowledge anticipated. In general, mining data in any type of database includes association rule generalization, multilevel data characterization, data classification, data clustering, pattern-based similarity search, and mining path traversal patterns (Chen, Han, & Yu, 1996). This project was to identify semantic patterns in antibiotic resistance literature, which would be based on the frequency analysis of keyword occurrences. To achieve this goal, one can obtain a set of working data either by selecting keywords directly from individual records or by obtaining a more coherent pool(s) of source documents by applying a citation restriction such as bibliographical coupling. When the bibliographical coupling method is

used to select source documents, at least one similar publication is cited in all the source documents of a bibliographical coupling pool. By this criterion, the documents can be considered coherent in content. Because of this, the keyword data were collected from pools of source documents through bibliographical coupling.

Data Collection

The *Science Citation Index* (*SCI*) database was used to collect data. The following search query was formulated to achieve relative precision and recall:

SELECT (ANTIBIOTIC? (W) RESISTAN?) AND PNEUMONI?

The query was executed in May 1996 and resulted in a total of 360 postings. After ranking by CR (Cited Reference) field, the number of records was reduced to 340 due to the fact that some records did not include references. In Figure 1, these articles are represented by a_1, a_2, a_3, . . . , a_n. A total of 8,753 publications (c_1, c_2, c_3, . . ., c_k in Figure 1) were cited in 340 papers. The highest frequency that a paper was cited was seventy-two times, which means the largest pool of source documents identified via bibliographical coupling contained seventy-two articles (see Table 1). The pools with the same number of source documents were treated as the same rank. All thirty-three ranks in this data set were grouped into three categories: 1 through 10 were large pools, those from 11 to 20 the medium, and the rest the small. The first five pools of source documents were selected from each category for extracting keyword data because of the time constraints for the project. Separate keyword files (i.e., w_1, w_2, w_3, . . ., w_j, in Figure 1) were downloaded for each pool of documents.

Table 1.
Top 10 Most Frequently Cited Documents in Antibiotic Resistance in Pneumonia Literature

Rank	*Frequency of Being Cited*	*Author Name and Source*
1	72	KLUGMAN KP, 1990, V3, P171, CLIN MICROBIOL REV
2	45	MARTON A, 1991, V163, P542, J INFECT DIS
3	41	JACOBS MR, 1978, V299, P735, NEW ENGL J MED
4	38	FENOLL A, 1991, V13, P56, REV INFECT DIS
5	34	HANSMAN D, 1967, V2, P264, LANCET
6	33	APPELBAUM PC, 1992, V15, P77, CLIN INFECT DIS
7	32	SPIKA JS, 1991, V163, P1273, J INFECT DIS
8	28	PALLARES R, 1987, V317, P18, NEW ENGL J MED
9	26	APPELBAUM PC, 1987, V6, P367, EUR J CLIN MICRO
9	26	WARD J, 1981, V3, P254, REV INFECT DIS
10	25	JORGENSEN JH, 1990, V34, P2075, ANTIMICROB AGE
10	25	MUNOZ R, 1991, V164, P302, J INFECT DIS
10	25	PHILIPPON A, 1989, V33, P1131, ANTIMICROB AGEN

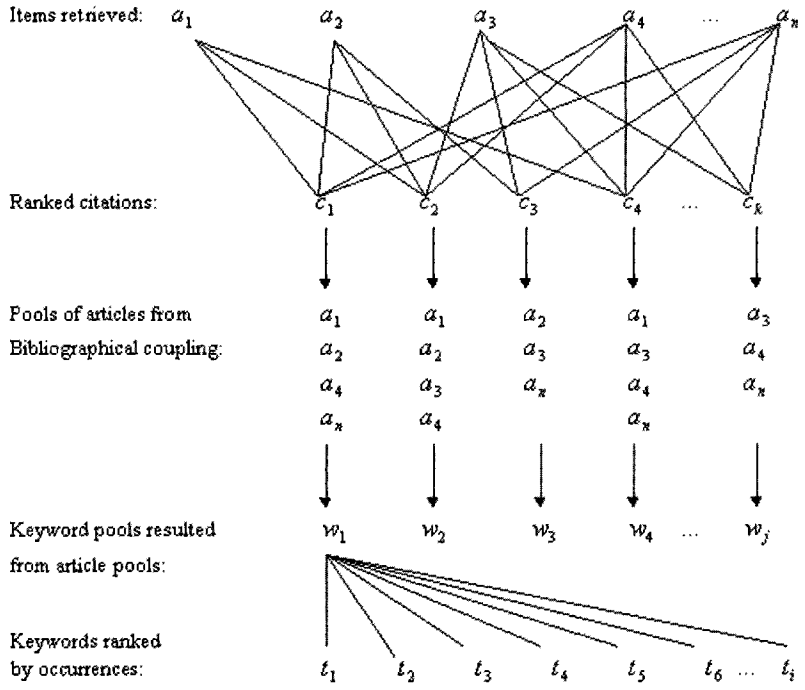

Figure 1. Flowchart of Keyword Extraction.

Data Preprocessing

The first step in preprocessing is cleaning the downloaded keyword files and converting them into tables. This can be done easily with a word processor's FIND and REPLACE functions. Macros or programs can also be written to read the text files into database tables. Data preprocessed through either way would need to be checked for errors, missing values, and the irregular labels missed by the REPLACE command. The next step is then to assign to keywords the text codes that can be computed by analytic tools (see Appendix). As mentioned earlier, textual data mining faces difficulty in handling long character strings and normalizing terms linguistically. Long strings would not be suitable for calculating frequencies or performing other statistical analysis. The text codes designed for the keywords in this subject field are mnemonic and, in most cases, comprehensible without the help of the original forms. A dictionary or a knowledge base for linking text codes to their keywords can be built for automatic coding. In coding keywords for this data set, a general rule was made to maintain as much of the original form and semantics of the keywords as possible. Other coding rules were set as follows:

- The same codes were assigned to both singular and plural forms of the same keywords, e.g., invit = invitro activity/invitro activities, child = child/children.
- The same codes were assigned to those having the same stem but different suffixes, e.g., pn-r = penicillin-resistance/penicillin-resistant, ther = therapeutic and therapy.
- Key phrases were coded by the noun with its modifying adjective or noun as a modifier, e.g., pnu-k = klebsiella pneumoniae, pnu-r = resistant pnuemoniae, pnu-s = streptococcus pnuemoniae.
- A few keywords that were semantically the same but morphologically different were given the same code for the purpose of joining those with the same meanings. Only two keywords fell into this category in this data set: child was used for coding child, children, infants, and pediatric patients; and 3rdw for third-world and developing-countries.

The text coding process was done semi-manually since building the initial code dictionary often needs human intelligence to analyze and translate a keyword or phrase into an appropriate code. The coding consistency (i.e., the same keyword is given the same code or vice versa throughout the data set) was double checked by sorting the data in the order of keyword and text code and then the order of text code and keyword.

Data Analysis

Data analysis in KDD processes is associated with data generalization and summarization which "presents the general characteristics or a summarized high-level view over a set of user-specified data in a database" (Chen, Han, & Yu, 1996, p. 866). The semantic patterns of keywords can be generalized from different perspectives—the simple frequency of occurrences and co-occurrences, or the number of unique keywords per rank (frequency), each of which uses a different measure to analyze the data. The simple frequency of occurrences counts how many times a keyword appears in a bibliographic coupling pool. It draws a high-level view of the semantic patterns from keyword frequency distribution. How often a keyword occurred is often decided by the size of the keyword pool. Obviously, the larger a keyword pool is, the more likely it is for a keyword to occur more frequently. When comparing the simple keyword frequency of a large pool of source documents with that of a smaller one, the result can be misleading because of the uneven bases for comparison. A more meaningful and reliable measure would be relative occurrences—i.e., percentage of times that a keyword appears in the total occurrences.

The frequency of co-occurrences is useful for measuring the importance of a keyword in the subject area, but it needs to be used with care. This data set was divided into large, medium, and small groups of source document pools. A complete coordination of all possible co-occurrences

would involve those between groups 1 and 2, 1 and 3, 2 and 3, and among all three. Even though a keyword may appear in two or three groups at the same time, its frequency of occurrences may vary greatly in different groups. There were also large variations in the numbers of total ranks or frequencies of keyword occurrences: thirty-three in the large group, twenty-four in the medium, and eleven in the small. These can lead to an invalid comparison for the same keyword with the same rank number but in different groups. For instance, a keyword ranked at eleven in the large group, which was considered high in its group, would have meant the lowest rank in the small group.

To normalize the frequency of occurrences, a measure of keyword density per rank was used. The keyword density per rank can be interpreted as the ratio of the number of unique keywords to the number of ranks at which the unique keywords occurred. It can be expressed in the following formula:

$$D(t) = \frac{1}{r_i} \sum_{i=1}^{n} t_i \qquad [1]$$

Where $D(t)$ = Average number of keywords $t_1, t_2, \ldots, t_i$ per rank,

r_i = Number of ranks,

$\sum_{i=1}^{n} t_i$ = Total number of unique keywords included from ranks 1 through n.

Figure 2 shows how the keyword density was calculated.

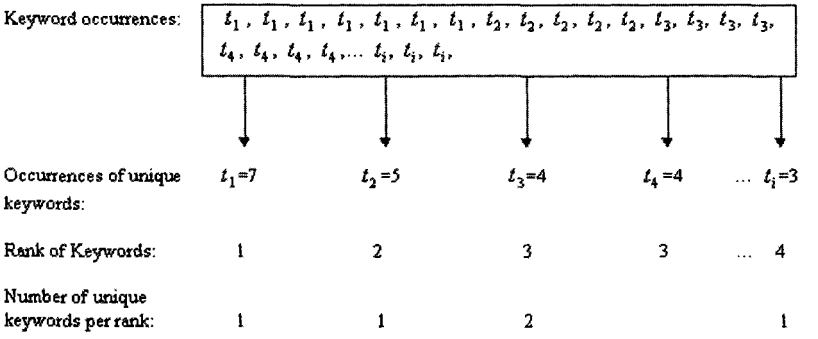

Figure 2. Computation of Keyword Density per Rank.

This measure eliminates the defects of simple frequency and co-occurrences and focuses on how many unique keywords scatter in a region.

This region is denoted by the frequency rank, and its size can be set according to the distribution shape. In Equation [1], the least possible *D(t)* is 1, that is, both the number of unique keywords and the number of ranks are the same. For example, three unique keywords were found to have appeared in three different frequencies (or frequency ranks), then 3/3 = 1. The largest possible *D(t)* can be an infinite in theory, which means that all unique keywords appeared at the same *one* rank. It is clear that the keyword density will increase as a rank contains more unique keywords.

FINDINGS

Frequency Distribution

There were a total of 2,994 keywords in the fifteen pools of source documents. The number of keywords in the large group (source document pools 1-5) consists of 54.5 percent of the total. The medium group had slightly over 40 percent keywords, and the small group only about 10 percent (see Table 2). The decrease in the number of keywords was mainly due to the decrease in the size of document pools; the average number of keywords (7) per record remained approximately the same for each pool. Nonetheless, the frequency distribution of keywords in all three groups was very similar: a majority of the keywords appeared less than five times in each of the groups; as the occurrences increased, the percentage of keywords decreased (see Figure 3).

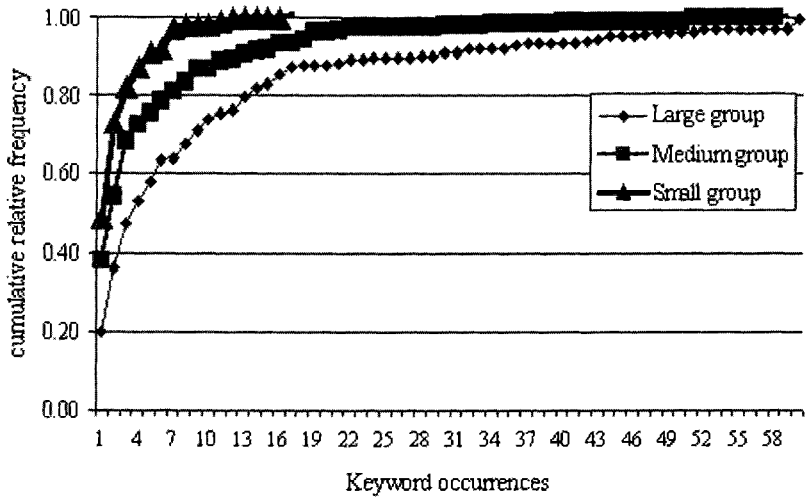

Figure 3. Cumulative Relative Frequency Distribution of Keywords in Three Groups.

Table 2.
NUMBER OF KEYWORDS IN INDEXING RECORDS FOR THE SOURCE DOCUMENTS IDENTIFIED THROUGH BIBLIOGRAPHICAL COUPLING

Group Size by Number of Documents	*Document Pool*	*Number of Documents*	*Number of Keywords*	*Percentage*	*Cumulative Percentage*
Large (Pools of source documents identified through bibliographic coupling)	1	72	512	17.1	17.1
	2	45	316	10.6	27.7
	3	41	291	9.7	37.4
	4	38	273	9.1	46.5
	5	34	241	8.0	54.5
Medium	6	25	200	6.7	61.2
	7	25	208	7.0	68.2
	8	23	207	7.0	75.2
	9	23	171	5.7	80.9
	10	23	202	6.7	87.6
Small	11	11	78	2.6	90.2
	12	11	77	2.6	92.8
	13	10	73	2.4	95.2
	14	10	67	2.2	97.4
	15	10	78	2.6	100.0
Total		401	2,994	100.0	

Although the percentage of keywords declined dramatically as the group size decreased, all three groups shared the same top three keywords—antibiotic resistance, antimicrobial resistance, and streptococcus-pneumoniae. This suggests that a common "intellectual base" existed among all three groups (see Table 3). The percentage of these three keywords dropped in medium and small groups compared to the large group. A close examination of data revealed that the lower occurrences were caused mainly by fluctuations in individual groups (see Figure 4). Figure 4 suggests that such fluctuations became wider as the group size shrunk to the next level.

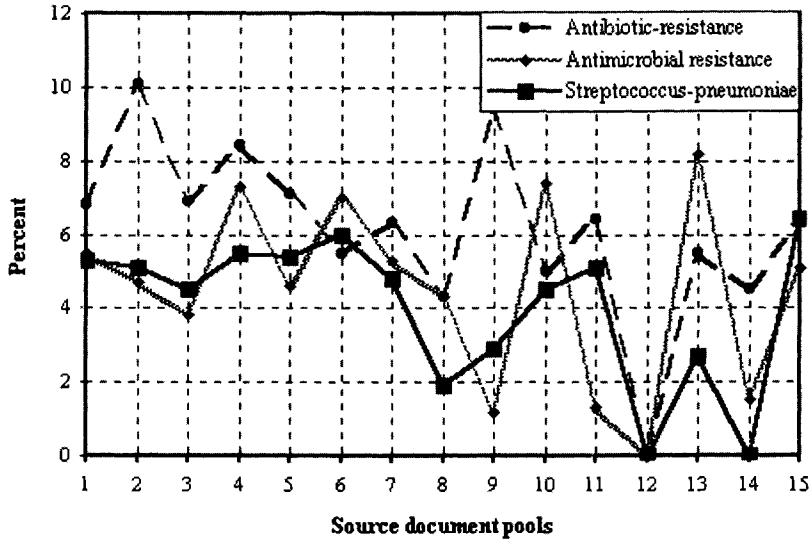

Figure 4. Frequency Distribution of the Intellectual-Base Keywords in 15 Document Pools (1-5 = Large Group, 6-10 = Medium Group, and 11-15 = Small Group).

Co-Occurrence of Keywords

In addition to the base keywords, other keywords co-occurred in either all three groups or two of the three. The largest number of keywords (eighty-five) co-occurred in all three groups. Only seven keywords co-occurred in both the large and small groups besides the eighty-five (see Table 4). The number of unique keywords that occurred in only one group was surprisingly similar: 25, 33, and 33 in the large, medium, and small groups respectively. Among the eighty-five keywords in "all groups" in Table 4, there existed large variations that the same keyword appeared with varied frequencies in different groups. The highest occurrences concentrated in the large group, then declined as the rank of the document pool went down (see Table 5). For example, "children" had sixty-nine occurrences in the large group but decreased to twenty-six and nine respectively in the medium and small groups. While the numbers of unique keywords in the three groups did not differ significantly ($p < 0.05$), the relative occurrences (4.2, 2.6, 2.5 respectively) show that more records in the large group had this keyword. This similar phenomenon happened throughout most other co-occurring keywords in either all three groups or any of the two groups together. Very few keywords that occurred in the small group outnumbered the occurrences in the medium or large group—i.e., although keywords co-occurred in different groups, they did not appear at the same frequency. Keywords co-occurring in only two groups were mostly those with lower frequencies. Figure 4 depicts the number of

Table 3.
RELATIVE FREQUENCIES OF KEYWORDS IN THE FIRST 25TH PERCENTILES IN THREE GROUPS

Keywords	*Large Group*		*Medium Group*		*Small Group*	
	Rank	*Rel. freq.*	*Rank*	*Rel. freq.*	*Rank*	*Rel. freq.*
Antibiotic-resistance	1	7.8	1	6.0	1	4.6
Antimicrobial resistance	2	5.2	2	5.2	2	3.2
Streptococcus-pneumoniae	3	5.1	3	4.0	3	2.9
Children/Infants/Pediatric patients	4	4.2	6	2.6	4	2.2
Susceptibility	5	3.7			6	1.9
Infection/infections					6	1.9
Day-care center/centers			7	2.2		
United States			5	3.0	6	1.9
Haemophilus-influenza			4	3.4		
3rd-generation cephalosporins					5	2.1
Escherichia-co					6	1.9
Mechanically ventilated patients					6	1.9
Penicillin-binding protein			7	2.2		

Table 4.
NUMBER OF UNIQUE KEYWORDS THAT OCCURRED OR CO-OCCURRED IN DIFFERENT GROUPS

	Large Group	*Medium Group*	*Small Group*	*All Groups*
Large Group	25	46	7	
Medium Group	46	33	28	
Small Group	7	28	33	
All Groups	85	85	85	85
Total	163	192	153	

TABLE 5.
PORTION OF THE FREQUENCY AND PERCENTAGE OF THE KEYWORDS THAT CO-OCCURRED

Keywords Occurring in All Three Groups	*Large Group* Freq.	%	*Medium Group* Freq.	%	*Small Group* Freq.	%
Children/Infants/ Pediatric patients	69	4.2	26	2.6	9	2.5
Susceptibility	60	3.7	3	0.3	7	1.9
Pneumococci	47	2.9	21	2.1	4	1.1
Infection/infections	44	2.7	13	1.3	7	1.9
Day-care	42	2.6	22	2.2	5	1.3
United States	37	2.3	6	0.6	7	1.9
Haemophilus-influenza	32	2.0	34	3.4	3	0.8
Therapy/therapeutic	30	1.8	1	0.1	7	1.9
Penicillin resistance	28	1.7	19	1.9	4	1.1
Penicillin-binding protein	24	1.5	22	2.2	5	1.3
Penicillin	22	1.3	6	0.6	1	0.3
Strains	21	1.3	2	0.2	1	0.3
Disease	18	1.1	6	0.6	3	0.8
Epidemiology	17	1.0	11	1.1	2	0.5
Failure	17	1.0	9	0.9	4	1.1
Otitis-media	16	1.0	9	0.9	1	0.3
Vaccine/conjugate vaccine	16	1.0	2	0.2	2	0.5

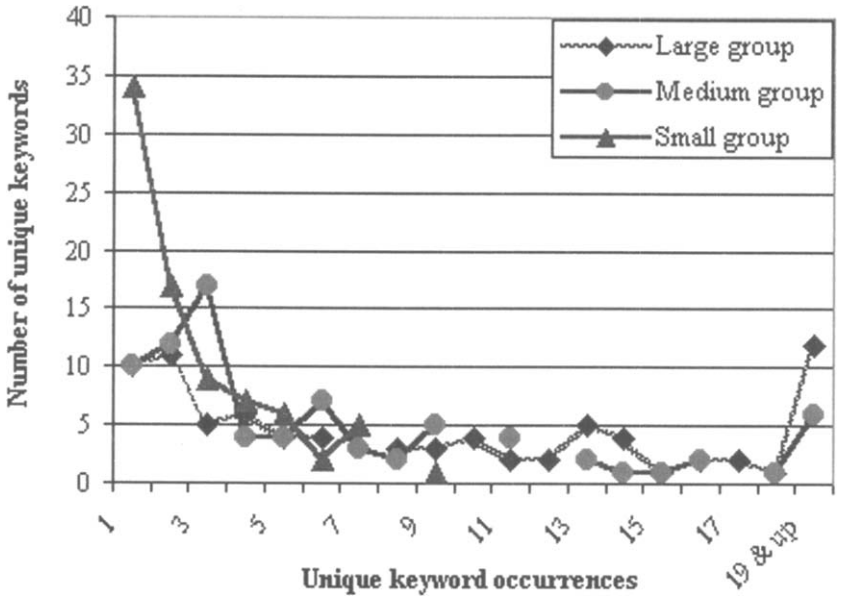

Figure 5. Frequency Distribution of Unique Keywords Occurring in All Three Groups.

unique keywords having occurrences one through more than nineteen. Most unique keywords in the large group occurred more frequently but much less frequently (only one, two, or three times) in the small group.

The Keyword Density

To compute the keyword density per rank, the frequency distribution of keyword occurrences was plotted for each of the three groups after the intellectual base keywords had been excluded. Figure 6 reveals a sharp turn at four, which was then used as a dividing point between the marginal and specialty keywords in the sample. In other words, keywords occurring three or fewer times in the sample were assumed to be marginal in the subject under study, and those with four or more times to be the specialties. Applying Formula [1] in the Data Analysis section, the keyword density was calculated according to the data in Table 6. When calculating the keyword density, ranks that had no keyword occurrences were treated as missing cases and ignored because only the actual number of frequency ranks reflected the keyword density. Thus the number of ranks for specialty keywords in the large group would be 42 minus 3 (intellectual base ranks) minus 3 (marginal ranks) minus 5 (missing cases) equal 31, and so forth for the other two groups. Results in Table 7 show that the density for marginal keywords is approximately ten times greater

than those of specialty keywords in all three groups. Further studies are needed to explore whether this is only a coincidence for this particular data set or a phenomenon existing across disciplines.

Table 6.
FREQUENCY DISTRIBUTION OF KEYWORD OCCURRENCES IN THREE GROUPS EXCLUDING THE THREE INTELLECTUAL BASE KEYWORDS

		Number of Unique Keywords (t_i)					*Number of Unique Keywords (t_i)*		
No.	*Keyword Occurrences*	*Large Group*	*Medium Group*	*Small Group*	*No.*	*Keyword Occurrences*	*Large Group*	*Medium Group*	*Small Group*
1	1	33	72	83	22	24	1		
2	2	26	30	24	23	28	1		
3	3	18	26	15	24	30	1	1	
4	4	10	9	9	25	32	2		
5	5	8	5	7	26	34		1	
6	6	9	8	2	27	36	1		
7	7	1	4	8	28	37	1		
8	8	6	5	1	29	40		1	
9	9	5	6	1	30	42		1	
10	10	5			31	43	1		
11	11	2	4	1	32	44	1		
12	12	2	2	1	33	47	1		
13	13	5	2		34	48	1		
14	14	4	2		35	51	1		
15	15	1	1		36	52	1		
16	16	4	3		37	59	1		
17	17	3		1	38	60	1		
18	18	1	2		39	69	1		
19	19		3		40	84	1		
20	21	1	1		41	85	1		
21	22	1	2		42	129	1		
						Total	163	192	153

Table 7.
KEYWORD DENSITY IN GROUPS

Keyword Density (D)	*Intellectual Base Keywords (i)*	*Specialty Keywords (s)*	*Marginal Keywords (m)*
Large Group (l)	D(li)=3/3=1	D(ls)=83/31=2.68	D(lm)=77/3=25.67
Medium Group (m)	D(mi)=3/3=1	D(ms)=61/19=3.21	D(mm)=128/3=42.67
Small Group (s)	D(si)=3/3=1	D(ss)=28/6=4.67	D(sm)=122/3=40.67

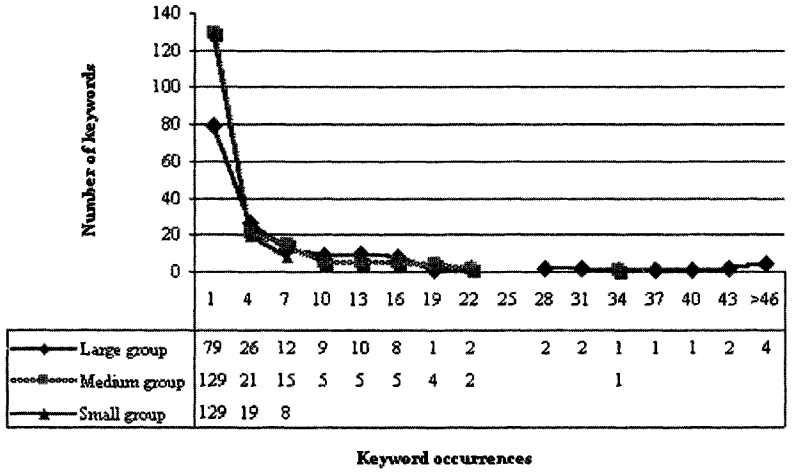

Figure 6. Frequency Distribution of Keyword Occurrences in Three Groups.

A further examination was made for keywords in the specialty and marginal groups. Several patterns emerged in the specialty keywords (see Tables 8, 9, and 10):

- Keywords co-occurring in two or three groups tended to be more generic or disciplinarily generic than non-co-occurring ones. Examples included children, day-care, failure, infections, prevalence, United States, genes.
- There were more microbial names and related infections in the keywords co-occurring than in the ones not co-occurring. Examples included pneumococci, enterococcus/enterococci, Escherichia-coli, Klebsiella-pneumonia, Neisseria-gonorrhoea, haemophilus-influenza, streptococcus-pneumonicoccal meningitis, Branhamella-catarrhalis.
- There was a clear tendency in the keywords both co-occurring and non-co-occurring (the latter happened in the first two groups only) that antibiotic resistance in pneumonia was investigated from perspectives of genetics (binding protein gene, penicillin-binding proteins, multiresistant clone), microbiology (invitro activities), and immunology (pneumococcal polysaccharide). However, this tendency in co-occurring keywords seemed to be more toward pharmaceutical aspects in relation to the microbes and infections they caused (spectrum beta-lactam, chloramphenicol therapy, third-generation cephalosporins), and more pathologically oriented in non-co-occurring keywords (serotype distribution, strains, antimicrobial susceptibility, plasmids).
- The keyword density in double digits were generally more specific than those in single digits, though there did exist a few general ones (see Table 10).

Table 8.
SPECIALTY KEYWORDS THAT CO-OCCURRED IN TWO OR THREE GROUPS

		Large Group		*Medium Group*		*Small Group*	
No.	*Keywords*	*Freq.*	*%*	*Freq.*	*%*	*Freq.*	*%*
1	Children/Infants/ Pediatric patients	69	4.2	26	2.6	9	2.4
2	Pneumococci	47	2.9	21	2.1	4	1.1
3	Infection/infections	44	2.7	13	1.3	7	1.9
4	Day-care/Day-care centers	42	2.6	22	2.2	5	1.3
5	United States	37	2.3	6	0.6	7	1.9
6	Penicillin resistance	28	1.7	19	1.9	4	1.1
7	Penicillin-binding proteins	24	1.5	22	2.2	5	1.3
8	Failure	17	1.0	9	0.9	4	1.1
9	Gene/genes	14	0.9	15	1.5	4	1.1
10	Pneumococcal polysaccharide	14	0.9	7	0.7	7	1.9
11	Prevalence	13	0.8	11	1.1	5	1.3
12	Enterococcus/ enterococci	6	0.4	16	1.6	4	1.1
13	Escherichia-coli	5	0.3	16	1.6	7	1.9
14	Spectrum beta-lactam	5	0.3	18	1.8	5	1.3
15	Klebsiella-pneumoni	4	0.2	9	0.9	6	1.6
16	Neisseria-gonorrhoea	4	0.2	4	0.4	6	1.6
17	Meningitis	52	3.2	16	1.6		
18	Chloramphenicol therapy	36	2.5	8	0.8		
19	Haemophilus-influenza	32	2.0	34	3.4		
20	Penicillin	22	1.3	6	0.6		
21	Disease	18	1.1	6	0.6		
22	Epidemiology	17	1.0	11	1.1		
23	Binding protein gene	16	1.0	14	1.4		
24	Otitis-media	16	1.0	9	0.9		
25	Streptococcus-pneumonicoccal meningitis	15	0.9	14	1.4		
26	Bacterial-meningitis	14	0.9	11	1.1		
27	Bacteria	13	0.8	6	0.6		
28	Beta-lactam antibiotics	13	0.8	8	0.8		
29	Branhamella-catarrhalis	12	0.7	7	0.7		
30	Multiresistant clone	12	0.7	8	0.8		
31	Antibiotics	11	0.7	9	0.9		
32	Influenzae type-b	11	0.7	11	1.1		
33	Invitro activities	10	0.6	5	0.5		
34	Carriage	9	0.6	4	0.4		
35	Diagnose	9	0.6	4	0.4		

36	Resistance	9	0.6	6	0.6		
37	Erythromycin	8	0.5	4	0.4		
38	Staphylococcus-aureu	8	0.5	11	1.1		
39	Influenzae	4	0.2	5	0.5		
40	Tuberculosis	4	0.2	6	0.6		
41	Susceptibility	60	3.7			7	1.9
42	Therapy/ Therapeutic	30	1.8			7	1.9
43	Emergence	8	0.5			4	1.1
44	Protective efficacy	5	0.3			4	1.1
45	3rd-generation cephalosporins			6	0.6	8	2.2
46	Enterobacter/ Enterobacteriaceae			7	0.7	5	1.3
47	Respiratory-tract infection			4	0.4	4	1.1

Table 9.
SPECIALTY KEYWORDS OCCURRING IN A SINGLE GROUP

Keywords Unique in the Large Group				*Keywords Unique in the Medium*			
No.	*Group*	*Freq.*	*%*	*No.*	*Group*	*Freq.*	*%*
1	Serotype distribution	48	2.9	37	UK	30	3
2	Spain	32	2.0	38	Sulbactam	19	1.9
3	Strains	21	1.3	39	Resistant staphyl-ococcuci	18	1.8
4	New-Guinea/ Papua-New-Guinea	17	1.0	40	Nucleotide-sequences	13	1.3
5	Vaccine/Conjugate vaccine	16	1.0	41	Sri-Lanka	12	1.2
6	Pneumococcal meningitis	14	0.9	42	Cephalosporins	9	0.9
7	Systemic infections	13	0.8	43	Pneumococ-cal serotype	9	0.9
8	Penicillin-resistant pneumoniae	13	0.8	44	Tetracycline	8	0.8
9	Resistant pneumo-coccus pneumoniae	10	0.6	45	Transpeptidase	8	0.8
10	Upper respiratory -tract	10	0.6	46	Mechanism	6	0.6
11	Community-acquired pnuemoniae	10	0.6	47	Catarrhalis beta-lactamase	5	0.5
12	Immune-deficiency syndrome	9	0.6	48	Pneumococcal vaccine	5	0.5
13	Antibody	9	0.6	49	Postsplenectomy sepsis	5	0.5
14	Vancomycin	8	0.5	50	Ampicillin	4	0.4
15	Horizontal transfer	8	0.5	51	Gram-negative bacilli	4	0.4
16	Antimicrobial susceptibility	8	0.5	52	Plasmid/plasmids	4	0.4

continued on page 126

table 9 continued

No.	*Keywords Unique in the Medium Group*	*Freq.*	%	No.	*Keywords Unique in the Small Group*	*Freq.*	%
17	Hungary	7	0.4	53	Mechanically ventilated patients	7	1.9
18	Iceland	6	0.4				
19	Invasive disease	6	0.4	54	Patients/critically ill patients	7	1.9
20	Patterns	6	0.4				
21	Pneumococcal infections	6	0.4	55	Intensive-care unit (AIDS)	5	1.3
22	Acquired immuno defficiency syndrome	6	0.4	56	Nosocomial infection	5	1.3
				57	Transferable resistance	4	1.1
23	Bacterial pneumonia	6	0.4				
24	Cerebrospinal-fluid	6	0.4				
25	Co-trimoxazole	6	0.4				
26	Requiring hospitalization	5	0.3				
27	Sensitivity	5	0.3				
28	Septic arthritis	5	0.3				
29	Human-Immuno deficiency Virus (HIV)	5	0.3				
30	Capsular poly-saccharide	5	0.3				
31	Molecular epidemiology	4	0.2				
32	Invasive pneum-ococcal infections	4	0.2				
33	Anemia	4	0.2				
34	Binding proteins	4	0.2				
35	Capsular types	4	0.2				
36	Ciprofloxacin	4	0.2				

Table 10.
MARGINAL KEYWORDS THAT CO-OCCURRED IN EITHER ALL THREE OR TWO OF THE THREE GROUPS

Keywords	*Large Group*		*Medium Group*		*Small Group*	
	Freq.	%	*Freq.*	%	*Freq.*	%
Pseudomonas-aeruginosa	3	0.2	3	0.3	2	0.5
Management	3	0.2	3	0.3	1	0.3
Etiology	3	0.2	1	0.1	1	0.3
Isolate/clinical isolate	2	0.1	2	0.2	1	0.3
Coagulase-negatives	2	0.1	2	0.2	1	0.3
Aminoglycoside resistance	2	0.1	1	0.1	2	0.6
Legionnaires-disease	2	0.1	1	0.1	1	0.3
Microdilution system	2	0.1	1	0.1	1	0.3
Blood cultures	2	0.1	2	0.2	1	0.3
Trimethoprim-sulfame	1	0.1	2	0.2	1	0.3
Methicillin-resistant	1	0.1	2	0.2	1	0.3

Refractory periodont	1	0.1	1	0.1	1	0.3
Outer-membrane permeability	1	0.1	3	0.3	1	0.3
Outbreak	1	0.1	3	0.3	1	0.3
Nosocomial outbreak	1	0.1	3	0.3	2	0.5
Colonization	1	0.1	3	0.3	3	0.8
2x	2	0.1	2	0.2		
Anti-inflammatory agent	3	0.2	1	0.1		
Antibiotic-therapy	1	0.1	1	0.1		
Aspiration	1	0.1	1	0.1		
Broth	1	0.1	1	0.1		
Cefamandole	3	0.2	1	0.1		
Ceftriaxone	3	0.2	1	0.1		
Clarithromycin	1	0.1	2	0.2		
Clindamycin	1	0.1	1	0.1		
Clones	2	0.1	2	0.2		
Common organization	2	0.1	1	0.1		
D-alanine ligase	2	0.1	3	0.3		
Directions	1	0.1	1	0.1		
Group-a	1	0.1	1	0.1		
High-level resistance	2	0.1	3	0.3		
Invasive pneumococcal infections	1	0.1	1	0.1		
Nasopharyngeal carriage	3	0.2	2	0.2		
Neisseria-meningitis	1	0.1	2	0.2		
Pathogen	1	0.1	1	0.1		
Populations	3	0.2	2	0.2		
Quinolones	1	0.1	1	0.1		
South-Africa	2	0.1	1	0.1		
Spread	3	0.2	1	0.1		
Streptococcus-pneumoniae strains	2	0.1	1	0.1		
Structural-changes	1	0.1	1	0.1		
Ampicillin	3	0.2			1	0.3
Antimicrobial agents	3	0.2			1	0.3
Bacterium legionella	2	0.1			1	0.3
Catarrhalis beta-lactamase	3	0.2			1	0.3
Cephalosporins	2	0.1			3	0.8
Dirithromycin	1	0.1			1	0.3
Mechanism	2	0.1			2	0.5
Norfloxacin	1	0.1			1	0.3
Nucleotide-sequences	2	0.1			2	0.5
Plasmid/plasmids	2	0.1			2	0.5
Pneumococcal vaccine	1	0.1			1	0.3
Spectrum	1	0.1			1	0.3
Affairs-medical-center			2	0.2	1	0.3
Anemia			1	0.1	1	0.3
Antibody			3	0.3	1	0.3
Aztreonam			1	0.1	1	0.3
Broad-spectrum cepha			1	0.1	1	0.3
Calcoaceticus var anitratus			2	0.2	1	0.3
Capsular polysaccharide			3	0.3	1	0.3

continued on page 128

table 10 continued

Keywords	*Large Group* *Freq.*	%	*Medium Group* *Freq.*	%	*Small Group* *Freq.*	%
Ceftazidime resistance			3	0.3	1	0.3
Cerebrospinal-fluid			3	0.3	2	0.5
Ciprofloxacin			3	0.3	1	0.3
Classification			2	0.2	1	0.3
Community-acquired pnuemoniae			2	0.2	2	0.5
Digestive-tract			1	0.1	3	0.8
DNA			3	0.3	3	0.8
Enzymatic resistance			3	0.3	2	0.5
Horizontal transfer			1	0.1	2	0.5
Identification			1	0.1	1	0.3
Imipenem-cilastatin			3	0.3	3	0.8
India			1	0.1	1	0.3
Nursing-home patient			3	0.3	1	0.3
Patterns			1	0.1	1	0.3
Penicillin-resistant pneumoniae			3	0.3	1	0.3
Pneumococcal meningitis			3	0.3	3	0.8
Resistant pneumococcus pneumoniae			3	0.3	1	0.3
Salmonella-typhi			1	0.1	1	0.3
Selective decontamination			2	0.2	3	0.8
Staphylococcus-aureu			3	0.3	2	0.5
Steady-state treatment			2	0.2	1	0.3
Strains			2	0.2	1	0.3
Substitution			1	0.1	1	0.3
Systemic infections			2	0.2	1	0.3
Third-world countries			1	0.1	2	0.5
Transposition			1	0.1	1	0.3
Upper respiratory-tract			2	0.2	3	0.8
Vancomycin			1	0.1	1	0.3

CONCLUSION

Knowledge discovery in bibliographic databases is distinctive compared to KD in full-text document and numerical databases. One challenge is transforming semi-structured textual data into the types and structures suitable for calculations and modeling. In the case of subject keywords, all the idiosyncrasies existing in natural language, including suffixes, different spellings for the same word, and synonyms, need to be normalized before analysis.

Similar work on this type of term normalization has been done in automatic indexing, such as stem stripping (Paice, 1990; Porter, 1980). Harman and Candela (1990) argue that term normalization such as stem stripping is not worth the effort for large full-text databases because this operation has little impact on other methods (e.g., frequency counts) of

indexing. While this may be true in indexing full-text documents, preprocessing of this kind is a necessity in discovering knowledge from bibliographic databases. The reason is obvious: semantic analysis of subject keywords needs to have accurate data to draw reliable and valid conclusions. The unnormalized keyword data set can present false patterns or trends in keywords. Although term normalization is a time-consuming operation, it can be improved by making use of the prior research and database technology mentioned earlier. In this study, as the initial text code base took shape, coding became easier and quicker as more and more text codes were established. It was found that semantic coding, while resembling vocabulary control, is different from vocabulary control in indexing. Semantic coding groups semantically same and/or similar keywords together by using simple codes that can be easily constructed. Using these codes, original terms can be preserved with semantic value-added processing, thus there is no need for using totally different "controlled" terms to substitute the keywords used in the publications. It quantifies the information analysis process and turns the jobs requiring expert knowledge into relatively simple tasks. This method is particularly suitable for specialized and interdisciplinary subject fields.

Presentation of the knowledge discovered is an important part of the KDD process. Visualizing the patterns, trends, and associations in a subject field can be very challenging because of the size of the screen and the number of text values that one data field can contain. This study of semantic patterns in keywords was by no means a large one in scale, but the total number of keywords made it difficult to draw any legible charts for the whole data set. An inclusion of even one group of keywords would clutter the chart badly and cause the keywords on the chart to be unrecognizable. Substituting long keywords with shorter and mnemonic text codes normalized the inconsistencies in keywords as well as leaving more room for visual presentation of the knowledge discovered.

The semantic patterns discovered in this data set suggest that different keyword density regions may be used as a controlling mechanism for better targeted searching. Traditionally, query expansion is one of the main techniques used to improve retrieval performance (Sparck Jones & Jackson, 1970; Salton, Fox, & Voorhees, 1983; Salton & Buckley, 1990; Harman, 1992). Query expansion allows searchers to browse the indexing term list or give relevance feedback to searchers through frequency ranking or term weighting. While performance was reported to have improved to a high percentage in small testing collections, it is still unproven that these techniques would achieve the same performance in large collections in the real world (Korfhage, 1997). Keyword density may provide a new solution to this uncertainty because it is computed on the basis of a collection of keywords extracted from bibliographically

coupled source documents. By implementing keyword density analysis into an algorithm, it is possible that a simple search query to the database(s) would generate a group of keywords stratified by their density regions. Information searchers can then select keywords from different density regions according to their own definition of relevance.

The semantic patterns found in non-co-occurring and co-occurring keywords suggest that it is necessary and possible to design new search tools that will deliver "analyzed" search results to users. In retrieving information from terabyte databases, the most challenging task in information retrieval is probably how to find most relevant information, in a manageable amount, in the easiest way. Conventional information systems have applied various sophisticated methods to accomplish this task but were limited by their design which requires equally sophisticated search techniques to find the information and leaves the information filtration to users themselves. The development of science and the growth of scientific literature has made filtering relevant information more difficult than ever in highly interdisciplinary scientific research areas. The semantic pattern analysis of the keywords from bibliographical coupling shows a possibility that simple semantic processing to natural language (keywords extracted from citations in this case) may be programmed and serve as a tool for providing "analyzed" search results to users.

The results in this study are only preliminary. It is unknown whether the semantic patterns identified in this data set are a coincidence or a common phenomenon across subject fields. Further studies are needed to discover whether the subject category of keywords is related to the density region and whether the stratified keyword distribution and density can contribute to customizing the selection of a targeted group of documents and post-search analysis.

APPENDIX

EXAMPLES OF KEYWORDS AND THEIR SEMANTIC CODES IN THE SAMPLE

Keyword	*Code*	*Keyword*	*Code*
2x	2x	HUMAN-IMMUNO-DEFICIENCY VIRUS	hiv
ANTI-INFLAMM-ATORY AGENT	agnt	HUNGARY	hungary
AID	Saids	IMMUNE-DEFICIENCY SYNDROM	Eids
ANTIMICROBIAL SUSCEPTIBILITY	sus-am	INVASIVE PNEUMO-COCCAL INFECTION	Sinf-pnu
ASPIRATION	aspirat	PNEUMOCOCCAL INFECTIONS	inf-pnu
BARCELONA	barc	INVASIVE DISEASE	invasiv
BINDING PROTEINS	bp	MENINGITIS/ MENINGEAL	meningi
BINDING PROTEIN GENE	bpg	MOLECULAR EPIDEMIOLOGY	epi-mol
BROTH	broth	NEW-GUINEA/ PAPUA-NEW	ng
CAPSULAR TYPES	capt	COMMON ORGAN-IZATION	org
CARRIAGE	carrig	PATHOGEN	pathoge
NASOPHARYNGEAL CARRIGE	carr-n	BACTERIAL PNEUMONIA	pnu-b
CEFAMANDOLE	cefam	QUINOLONES	quinolo
CEFTRIAXONE	ceftri	HIGH-LEVEL RESISTANCE	res-hi
CHLORAMPHEN -ICOL THERAPY	ther-chl	SOUTH-AFRICA	sa
CLARITHROMYCIN	clarith	SENSITIVITY	sensiti
CLINDAMYCIN	clindam	POSTSPLE- SEPSIS NECTOMY	sepsi-p
CLONES	clone	SPREAD	spread
MULTIRESISTANT CLONE	clone-m	STREPTOCOCCUS -PNEUMONIAE STRAINS	stra-s
DIRECTIONS	direct	SOUTH-AFRICAN STRAIN	stra-sa
D-ALANINE LIGASE	ligase	STRUCTURAL-CHANGES	struct
ERYTHROMYCIN	erythr	TETRACYCLINE	tetracy
GROUP-A	grp-a	ANTIBIOTIC-THERAPY	ther-a

References

Braam, R. R.; Moed, H. F.; & van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis. Part I: Structural aspects. *Journal of the American Society for Information Science, 42*(4), 233-251.

Braam, R. R.; Moed, H. F.; & van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis. Part II: Dynamical aspects. *Journal of the American Society for Information Science, 42*(4), 252-266.

Chen, M. Y.; Han, J.; & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8,* 866-883.

Feldman, R., & Hirsh, H. (1996). Mining associations in text in the presence of background knowledge. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *KDD '96* (Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, August 1996) (pp. 343-346). Menlo Park, CA: AAAI Press.

Feldman, R.; Dagan, I.; & Hirsh, H. (1998). Mining text using keyword distributions. *Journal of Intelligent Information Systems, 10*(3), 281-300.

Harman, D. (1992). User-friendly systems instead of user-friendly front-ends: Four end-user systems employing probabilistic ranking: PRISE, CITE, MUSCAT and News Retrieval Tool. *Journal of the American Society for Information Science, 43*(2), 164-174.

Harman, D. K., & Candela, G. (1990). Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *Journal of the American Society for Information Science, 41*(8), 581-589.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation, 16*(3), 223-233.

Korfhage, R. R. (1997). *Information storage and retrieval.* New York: Wiley.

Lent, B.; Agrawal, R.; & Srikant, R. (1997). Discovering trends in text databases. In D. Heckerman, H. Mannila, & D. Pregibon (Eds.), *KDD '97* (Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, Newport Beach, California, August 14-17, 1997) (pp. 227-230). Menlo Park, CA: AAAI Press.

Logan, E. L., & Shaw, W. M. (1987). An investigation of the coauthor graph. *Journal of the American Society for Information Science, 38*(4), 262-268.

Paice, C. (1990). Another stemmer: Natural language processing. *SIGIR Forum, 24*(3), 56-61.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.

Qin, J. (In press). Indexing similarities in a keyword database and a controlled vocabulary database: Antibiotic resistance in pneumonia. *Journal of the American Society for Information Science.*

Salton, G.; Fox, E. A.; & Voorhees, E. M. (1985). Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science, 36*(2), 200-210.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science, 41*(4), 288-297.

Shaw, W. M. (1990). Subject indexing and citation indexing: Clustering structure in the cystic fibrosis document. *Information Processing and Management, 26*(6), 693-718.

Small, H. (1973). Co-citation in scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations: I. A comparison of methods. *Scientometrics, 7*(3/6), 391-409.

Sparck Jones, K., & Jackson, D. M. (1970). The use of automatically-obtained keyword classifications for information retrieval. *Information Storage and Retrieval, 5*(4), 175-201.

Travis, J. (1994). Reviving the antibiotic miracle. *Science, 264*(5157), 360-362.

Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science and Technology, 32,* 197-229.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.

Additional Reference

Small, H.; Sweeney, E.; & Greenlee, E. (1985). Clustering the *Science Citation Index* using co-citations: II. Mapping science. *Scientometrics, 8*(5/6), 321-340.