SCIENCE AND TECHNOLOGY TEXT MINING: ANALYTICAL CHEMISTRY

By

Dr. Ronald N. Kostoff
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217
Phone: 703-696-4198
Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

Dr. Ronald A. DeMarco
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217
Internet: demarcr@onr.navy.mil

[***The views in this article are solely those of the authors and do not necessarily represent the views of the Department of the Navy or any of its components***]

ABSTRACT

Text mining is the extraction of useful information from large volumes of literature. This report addresses text mining in the context of the science and technology literature. It describes the major text mining components, and shows its myriad applications in support of science and technology. To show some of the text mining products, illustrative examples from diverse literatures, but (mainly) from analytical chemistry, will be presented.

KEYWORDS: text mining; information retrieval; bibliometrics; computational linguistics; information processing; information integration; science and technology; clustering; relevance feedback; technical infrastructure; literature-based discovery; analytical chemistry; taxonomy; co-occurrence; phrase frequency.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

Public reporting burder for this collection of information is estibated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burder to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (FROM - TO)
15-07-2003	Technical	xx-xx-2000 to xx-xx-2001

4. TITLE AND SUBTITLE
SCIENCE AND TECHNOLOGY TEXT MINING
ANALYTICAL CHEMISTRY
Unclassified

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kostoff, Ronald N ;
DeMarco, Ronald A ;

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Text mining is the extraction of useful information from large volumes of literature. This report addresses text mining in the context of the science and technology literature. It describes the major text mining components, and shows its myriad applications in support of science and technology. To show some of the text mining products, illustrative examples from diverse literatures, but (mainly) from analytical chemistry, will be presented.

15. SUBJECT TERMS

text mining; information retrieval; bibliometrics; computational linguistics; information processing; information integration; science and technology; clustering; relevance feedback; technical infrastructure; literature-based discovery; analytical chemistry; taxonomy; co-occurrence; phrase frequency.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	Same as Report (SAR)	25	Kostoff, Ronald kostofr@onr.navy.mil
Unclassified	Unclassified	Unclassified			19b. TELEPHONE NUMBER International Area Code Area Code Telephone Number 703696-4198 DSN -

BACKGROUND

The technical literature is the storage medium for science and technology (S&T) knowledge. Rapid advancement of S&T depends on the efficiency of knowledge extraction from this literature, including both infrastructure (authors, journals, institutions) and thematic (technical thrusts, relationships) information. Relative to global S&T, questions of interest centeraround:

- what S&T is being performed,
- who is performing the S&T,
- where is it being performed, and
- what messages and heretofore undiscovered information can be extracted from the global literature.

The expert analysts can then judge what is not being done, and recommend what should be done differently.

In the past, the technical community used the thorough but inefficient approach of visually scanning printed and electronic technical literature to identify relevant documents, then reading the relevant documents (with no decision aids) to extract the information. Now, techniques have been developed to perform the pre-selection of relevant literature semi-automatically, and to order the intrinsic technical concepts and their relationships to provide a framework for an integrated analysis. These techniques are encompassed under the umbrella of S&T text mining.

This article defines text mining, describes its major components, and shows its myriad applications to support all types of S&T functions. Text mining can benefit S&T performers, managers, sponsors, administrators, evaluators, and oversight organizations. It can serve as a catalyst to enhance peer review, metrics, road-mapping, and other decision aids. It could allow comprehensive roadmaps for strategic planning to be constructed, and thereby serve as a foundation for international policy assessment. Text mining can support workshops and S&T reviews by identifying the key performers in disciplines related to those being evaluated. It can identify productive sites to be visited in global S&T evaluations. It can identify new information groupings, to provide novel technical insights that could lead to discovery and innovation. In parallel, this could lead to promising new S&T opportunities, and new research directions. To illustrate some of the text

mining products, illustrative examples from diverse literatures, but (mainly) analytical chemistry, will be presented.

DEFINITIONS

We define S&T text mining as the extraction of information from technical literature. There are three major components under our definition: 1) Information Retrieval; 2) Information Processing; 3) Information Integration.

Information retrieval is the extraction of records from the source technical literatures. High quality information retrieval produces both comprehensive and highly relevant records. It is the foundational step in text mining. The most sophisticated information processing cannot compensate for insufficient core records retrieved.

Information processing is the extraction of patterns from the retrieved records. Our definition includes three components: 1) Bibliometrics; 2) Computational Linguistics; 3) Clustering. For multi-field structured records, with some free-text fields (such as paper Abstracts), *bibliometrics* is the extraction of the technical discipline infrastructure (authors, journals, organizations) as represented by the core records. *Computational linguistics* is the computer-based extraction of technical themes and their relationships. Computational linguistics is complex for technical literature analysis, because the technical phraseology appears as a foreign language to the computer. *Clustering* is the grouping of common technical themes, and could be executed as phrase pattern groupings or actual document groupings.

Information integration is the synergistic combination of the information processing computer output with the reading of the retrieved relevant records. The information processing output serves as a framework for the analysis, and the insights from reading the records enhance the skeleton structure to provide a logical integrated product.

More detailed descriptions of text mining can be found in (1) and (2).

APPLICATIONS

A few of the myriad existing and potential S&T text mining applications will be summarized.

1) RETRIEVE DOCUMENTS

Text mining can substantially improve the comprehensiveness and relevance of records retrieved from databases. There are many approaches to information retrieval. Annual conferences focus on comparing various techniques for their comprehensiveness and S/N of records retrieved (3, 4). Most high quality methods include some type of relevance feedback . This is an iterative method where a test query is generated, records are retrieved, and then patterns from the relevant and non-relevant records are used to modify the query for increased comprehensiveness and precision. These patterns are typically linguistic phrase and phrase combination patterns, but could also include infrastructure patterns such as author/ journal/ organization, etc (5).

2) IDENTIFY INFRASTRUCTURE

The infrastructure of a technical discipline consists of the authors, journals, organizations and other groups or facilities that contribute to the advancement and maintenance of the discipline. To obtain this infrastructure, scientometric studies without text mining typically assemble this literature-based information for a given discipline (e.g., 6), sometimes including temporal trends. However, text mining can identify these infrastructure elements, and in addition provide their specific relationships to the total technical discipline or to sub-discipline areas. This information is valuable for inviting the right people and discipline combinations to workshops and S&T reviews. It is also very valuable for planning a site visitation strategy for global discipline evaluations.

3) IDENTIFY TECHNICAL THEMES/ RELATIONSHIPS

Phrase pattern analyses through computational linguistics allow technical themes, their inter-relationships, their relationships with the infrastructure, and technical taxonomies to be identified. These are important for understanding the structure of a discipline, the linkages among people/ organizations/ sub-disciplines, and being able to estimate adequacies and deficiencies of S&T in sub-technology areas. Taxonomies can be generated manually from visual text analysis, or automatically through advanced text clustering techniques.

4) DISCOVERY FROM LITERATURE

Generically, literature-based discovery consists of examining relationships between linked, overlapping literatures, and discovering relationships or

promising opportunities not obtainable from reading each literature separately. The general theory behind this approach, applied to two separate literatures, is based upon the following considerations (7).

Assume that two literatures can be generated, the first literature AB having a central theme "a" and sub-themes "b," and the second literature family BC having a central theme(s) "b" and sub-themes "c." From these combinations, linkages can be generated through the "b" themes which connect both literatures (e.g., AB-->BC). Those linkages that connect the disjoint components of the two literatures (e.g., the components of AB and BC whose intersection is zero) are candidates for discovery, since the disjoint themes "c" identified in literature BC could not have been obtained from reading literature AB alone.

Successful performance of this generic approach can lead to new treatments for illnesses, new materials for different applications, extrapolation of ideas from one discipline to a disparately related discipline, and identification of promising new S&T opportunities and research directions. Some studies and concept papers have been published (2, 7, 8, 9, 10, 11, 12, 13).

TECHNIQUES AND ILLUSTRATIVE EXAMPLES

This section provides illustrative examples of S&T text mining techniques. It starts with an example of a query developed for a recent Aircraft S&T study, and shows some of the lessons learned from the query development. The section then proceeds to show some bibliometrics results. Most of these are from a database of papers published recently in *Analytical Chemistry*, and the journal bibliometrics are from a Mass Spectrometry query. Computational linguistics examples are taken from a variety of sources, related to analytical chemistry where possible.

1) RECORD RETRIEVAL QUERY, AIRCRAFT TECHNOLOGY

In the typical S&T text mining analyses performed by the first author, the starting point is the generation of a record retrieval query. A query development example is provided from a recent text mining study of the Aircraft S&T literature (14) in order to illustrate an important point about query complexity.

The study's focus was the S&T of the aircraft platform. The query philosophy was to start with the term AIRCRAFT, then add terms that would expand the number of Aircraft S&T papers retrieved and would eliminate papers not relevant to Aircraft S&T. Two databases were examined, the Science Citation Index (SCI-basic research, 5300 journals accessed) and the Engineering Compendex (EC-technology development, 2600 journals accessed). The SCI record retrieval query required 207 terms (separate phrases and phrase combinations) and 3 iterations to develop, while the EC query required 13 terms and one iteration. The SCI query retrieved 4,346 relevant records, while the EC query retrieved 15,673 relevant records.

Because of the technology focus of the EC, most of the papers retrieved using an AIRCRAFT or HELICOPTER type query term focused on the S&T of the platform itself, and were aligned with the study goals. Because of the research focus of the SCI, many of the papers retrieved focused on the science that could be performed from the aircraft platform, rather than the S&T of the platform, and were not aligned with the study goals. Therefore, no adjustments were required to the EC query, whereas, with the SCI, many NOT Boolean terms were required to eliminate aircraft papers not aligned with the main study objectives. It is analogous to the selection of a mathematical coordinate system for solving a physical problem. If the grid lines are well aligned with the physical problem to be solved, the equations will be relatively simple. If the grid lines are not well aligned, the equations will contain a large number of terms required to translate between the geometry of the physical problem and the geometry of the coordinate system.

The most important message to be extracted from the aircraft and parallel studies is that ***the information retrieval query size depends on the objectives of the study, and the contents of the database relative to the study objectives.*** The query size should not be pre-determined, but should result from the attainment of the comprehensiveness and precision objectives.

Another important message is that substantial manual labor is required to examine the thousands of detailed technical phrases that result from the computational linguistics analyses of the free text, and to make judgements about the applicability of these phrases to inclusion in the final query. Because these queries are applied to multi-discipline source databases such

as the Science Citation Index, an understanding of the use of these phrases in other technical disciplines is required for successful query development. Thus, the person or team developing a query for a specific technical sub-discipline requires broader technical knowledge than in the target discipline alone.

2) BIBLIOMETRICS

-MOST PROLIFIC AUTHORS, *ANALYTICAL CHEMISTRY*

As a simple example of a bibliometrics output, records of the 2000 most recent articles (as defined by the SCI) published in the journal *Analytical Chemistry* (June 1998-August 2000) were extracted from the SCI. There were 5072 authors listed. The most prolific authors, and the number of papers on which they were listed, include: Ramsey JM (19), Smith RD (18), Wang J (17), Jacobson SC (14), Yeung ES (12), Anderson GA (11), Umezawa Y (11), Carr PW (11), Guillame YC (10), Peyrin E (10), Sweedler JV (10). These are rather impressive numbers for a two-year publication period in a prestigious journal.

The author distribution function is shown on Figure 1. Most of the authors have only one or two publications. Previous technical discipline studies

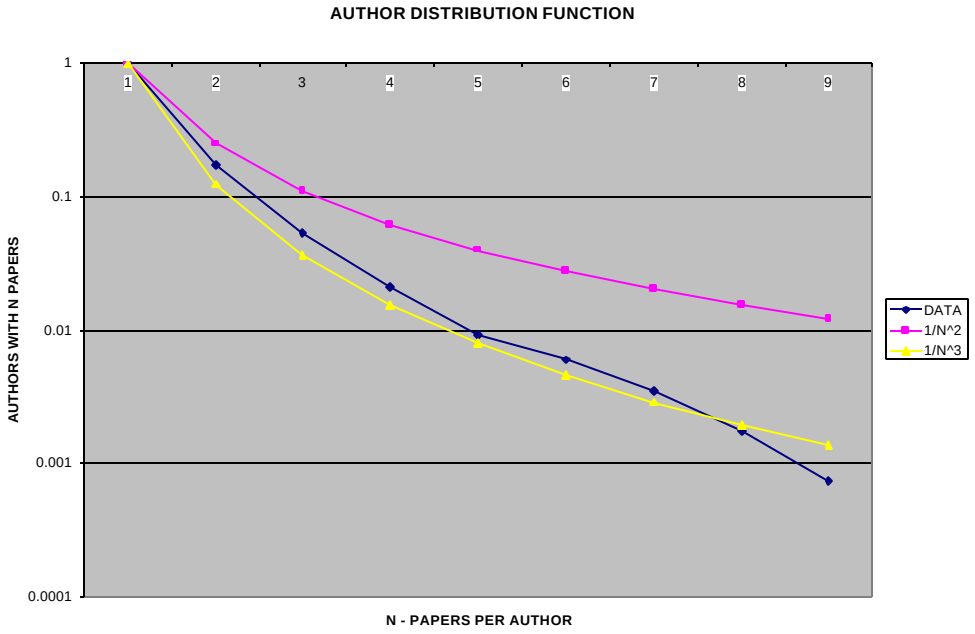

FIGURE 1

(14, 15, 16) show author distribution functions that range from 1/N^2 to 1/N^3. The present author distribution function is within that range, closer to 1/N^3.

-MOST CITED AUTHORS, *ANALYTICAL CHEMISTRY*
There were 22200 different authors cited from the same *Analytical Chemistry* database. The most cited authors include Jacobson SC (164), Giddings JC (123), Wang J (115), Bakker E (106), Grate JW (93), Bard AJ (87). There is reasonable correlation between the top 20 or so prolific authors and the top 20 cited authors, showing that many of the pioneers of present-day analytical chemistry thrust areas are still quite active. It should be re-emphasized that these integrated author citation numbers reflect only references contained in the 2000 most recent *Analytical Chemistry* articles, and an author's total citations from all sources could be substantially greater. An independent check of Bard AJ in the SCI, for example, showed tens of thousands of citations for all papers, as opposed to the 87 listed for this study.

-MOST PROLIFIC JOURNALS, MASS SPECTROMETRY
In this example, records of the 2000 most recent papers referenced in the SCI, and containing the term mass spectrometry (the highest frequency technique phrase from the 2000 records extracted from the journal *Analytical Chemistry* above) in the title were extracted. There were 377 journals listed. The journals containing the most mass spectrometry papers include *Rapid Communications In Mass Spectrometry* (224), *Journal Of Chromatography: A* (157), *Analytical Chemistry* (138), *Journal Of Mass Spectrometry* (93), *Journal Of Chromatography*: B (93), *Journal Of Analytic And Atomic Spectrometry* (75), and *Journal Of The American Society Of Mass Spectrometry* (65). The journal frequency decreases rapidly after this group. The first three journals appear to form the top core group, and the next four form the second core group.

We re-emphasize that in our standard text mining studies of a discipline, the iteratively-developed query used for the records from which the bibliometrics are derived would typically involve substantial time and effort, and contain hundreds of terms, not just one (mass spectrometry) as in this illustrative example.

-MOST CITED JOURNALS, *ANALYTICAL CHEMISTRY*
There were 6177 different journals/ sources cited by the 2000 *Analytical Chemistry* papers. The most cited journals include *Analytical Chemistry* (9107), *Journal of Chromatography: A* (1525), *Journal of Chromatography* (1427), *Journal of the American Chemical Society* (1334), *Analytic Chim Acta* (1177), *Rapid Communications in Mass Spectrometry* (901), *Journal of Electroanalytical Chemistry* (889), and *Science* (806). These rankings reflect two characteristic phenomena seen in previous studies. The journal in which the citing papers are published tends to be cited frequently, and the more fundamental journals tend to be cited with higher frequency than the applied journals.

-MOST PROLIFIC INSTITUTIONS, *ANALYTICAL CHEMISTRY*
The most prolific organizations were identified from the 2000 *Analytical Chemistry* papers database. The organization names, and the number of articles on which they were listed, include: Univ Calif (all campuses, and including LASL and LANL) (83), Oak Ridge Natl Lab (45), Univ Michigan (36), Univ Texas (32), Univ Tokyo (31), Univ Washington (27), Iowa State Univ (27), Univ Alberta (26), Univ N Carolina (25), Indiana Univ (25), Univ Florida (23), Univ Illinois (22), Texas A&M Univ (20), Univ Lund (18), Texas Tech Univ (17), Sandia Natl Labs (17), Univ Tennessee (16), Cornell Univ (15).

This example illustrates some of the limitations of metrics in general, and bibliometrics in particular. The institutions listed tend to be large, and one would expect large numbers of outputs. There is no indication of efficiency; i.e., output per unit of resources. There is no indication of output quality, other than the papers exceeded the obviously high threshold required for publication in *Analytical Chemistry*. Because of space limitations, organizational sub-units could not be listed. Thus, the high achievements of a sub-unit may not be reflective of the institution overall.

-MOST PROLIFIC COUNTRIES, *ANALYTICAL CHEMISTRY*
The most prolific countries were identified from the 2000 *Analytical Chemistry* papers database. The country names, and the number of articles on which they were listed, include: USA (1098), Japan (156), Germany (129), Canada (118), England (96), Switzerland (62), Sweden (59), France (53), Spain (53), Netherlands (44). When all countries are included, the USA has as many listings as all other countries combined. This dominance by the USA is characteristic of total discipline study bibliometrics obtained

previously, although the dominance is slightly exaggerated in *Analytical Chemistry*.

-MOST CITED PAPERS, *ANALYTICAL CHEMISTRY*
There were 35243 different papers cited by the 2000 *Analytical Chemistry* papers. The most cited papers include Jacobson SC, *Analytical Chemistry*, 1994; Fenn JB, *Science*, 1989; Harrison DJ, *Science*, 1993; Hjerten S, *Journal of Chromatography*, 1985; and Karas M, *Analytical Chemistry*, 1988. Of the ten most highly cited papers, half were in the 1980s and half were in the 1990s. This reflects a relatively dynamic field.

Again, the numbers of citations from the limited citing population do an injustice to total paper citations. The 1989 paper by Fenn JB, for example, was listed with 37 citations, but had total citations from all sources of almost 1350. Additionally, the 1980 paper by Bard AJ was listed with 25 citations, but had total citations from all sources of over 4000. Our more comprehensive discipline studies generate numbers more consonant with total citations from all sources.

3) COMPUTATIONAL LINGUISTICS

-LITERATURE-BASED DISCOVERY, RAYNAUDS SYNDROME
Some initial applications of literature-based discovery have been published in the medical literature (7, 8, 9, 10, 11, 12), and conceptual papers have been published as well (2, 13). An early study (7) was focused on identifying treatments for Raynauds Syndrome, a circulatory disease. Assume that two literatures can be generated, the first literature AB containing potential treatments and having a central theme "a" and sub-themes "b," and the second literature family BC containing the Raynauds Syndrome papers and having a central theme(s) "b" and sub-themes "c." One interesting discovery was that dietary eicosapentaenoic acid (theme "a" from literature AB) can decrease blood viscosity (theme "b" from both literatures AB and literatures BC) and alleviate symptoms of Raynaud's disease (theme "c" from literature BC). There was no mention of eicosapentaenoic acid in the Raynaud's Syndrome literature, but the acid was linked to the disease through the blood viscosity themes in both literatures.

Subsequent medical experiments confirmed the validity of this literature-based discovery (17). A web site (8) overviews the process used to generate this discovery, and contains software that allows the user to experiment with

the technique. A 1998 article in The Scientist outlines perceptions of different knowledgeable individuals on Swanson and Smalheiser's general technique (18).

-IDENTIFYING TECHNICAL THEMES/ TAXONOMIES, AIRCRAFT
The analysis of phrase patterns allows both technical themes to be identified, and technical taxonomies to be generated. In the recent Aircraft S&T study (14), use of phrase pattern analysis and taxonomy generation showed that levels of emphasis of aircraft sub-technologies could be estimated on a global basis, and judgements of technology 'adequacy' and 'deficiency' could be made. The following procedure was used.

Single word, adjacent double word, and adjacent triple word phrases were extracted from the Abstracts of the retrieved papers, and their frequencies of occurrence in the text were computed. A taxonomy whose categories covered the technical scope of the phrases was manually generated, and the phrases and their associated occurrence frequencies were placed in the appropriate categories. The frequencies in each category were summed, thereby providing an estimate of levels of category technical emphasis on a global basis. These summed category frequencies were compared with their counterparts extracted from 'requirements' and 'needs' documents, and estimates of aircraft technology adequacy and deficiency were made.

One cautionary note. The technology perspective of aircraft S&T was a function of the database record field (Abstract, Keywords, Title) examined, because the different record fields were used by their authors for different purposes. Multiple record fields must be examined to provide a balanced perspective of the overall technology.

Specifically, a high frequency focal area from the aircraft study Keyword field analysis was concentrated on the mature technology issues of longevity and maintenance; this view of the aircraft literature was not evident from the high frequency Abstract phrases. The lower frequency Abstract phrases had to be accessed to identify any thrusts in this mature technology/ longevity/maintenance area. Also, the Abstract phrases from the aircraft study contained heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the Keywords. The indexers may view much of the testing as a means to the end, and their Keywords reflect the ultimate objectives or applications rather than the detailed approaches for reaching these

objectives. However, there was also emphasis on high aircraft performance characteristics in the Abstract phrases, a category conspicuously absent from the Keywords. In fact, the presence of mature technology and longevity descriptors in the Keywords, coupled with the absence of high performance descriptors, provides a very different picture of aircraft literature research from the presence of high performance descriptors in the Abstract phrases, coupled with the absence of mature technology and longevity/maintenance descriptors.

Keywords are author/ indexer summary judgements of the main focus of the paper's contents, and represent a higher level description of the contents than the actual words in the paper/ Abstract. Thus, an Abstract that describes the details of 'corrosion' research may be viewed by the Keyword indexer as focused on 'maintenance' as its broader objective. Another explanation is that maintenance and longevity issues are politically popular now, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

The above findings have strong implications for Web-based information retrieval. The Web is an undefined conglomeration of Keyword fields, Abstract fields, many unstructured fields, and many different databases with widely varying contents. Yet the above lessons demonstrate that the record retrieval queries and technical perspectives obtained are strong functions of each of these elements. The credibility of the record retrieval and analysis quality from any Web text mining must be questioned, and its improvement will require substantial amounts of intensive detailed labor.

The aircraft (and other) study showed conclusively that, for a high quality text mining study, close involvement of the technical domain expert(s) is required in all stages. This is especially true for the computational linguistics and taxonomy generation portions. To insure that multiple perspectives are incorporated into the study, such that maximum data anomalies will be detected, multiple domain experts with diverse backgrounds and text mining experts who have analyzed many different disciplines are required.

-IDENTIFYING THEMES/ THEME RELATIONSHIPS/ TAXONOMIES,
ANALYTICAL CHEMISTRY
--THEMES

All single, double, and triple word phrases in the 2000 most recent *Analytical Chemistry* paper Abstracts were examined. A cohesive picture of analytical chemistry could be drawn. In the following summarization, phrases that appeared in the Abstracts are capitalized.

The main emphasis was on ANALYSIS SEPARATION METHODS for DETECTING COMPOUND MASS and CONCENTRATIONS. In order of emphasis (phrase frequency), the main TECHNIQUES in this MEASUREMENT-based discipline include MASS SPECTROMETRY, CAPILLARY ELECTROPHORESIS, and CHROMATOGRAPHY (mainly LIQUID, but some GAS), and hybrid combinations of these techniques. Also in order of emphasis, the main MASS SPECTROMETRY techniques include ELECTROSPRAY IONIZATION, MATRIX-ASSISTED LASER DESORPTION, TANDEM, TIME-OF-FLIGHT, ION CYCLOTRON RESONANCE, and QUADRUPOLE ION TRAP. Other high emphasis ANALYSIS techniques include CYCLIC VOLTAMMETRY, CAPILLARY ELECTROCHROMATOGRAPHY, ATOMIC FORCE MICROSCOPY, MICELLAR ELECTROKINETIC CHROMATOGRAPHY, LASER-INDUCED FLUORESCENCE DETECTION, ATOMIC ABSORPTION SPECTROMETRY, SUPERCRITICAL FLUID CHROMATOGRAPHY, X-RAY PHOTOELECTRON SPECTROSCOPY, and SCANNING ELECTROCHEMICAL MICROSCOPY. *There is negligible mention of theoretical approaches and analyses.*

The main types of COMPOUNDS (ANALYTES) examined include, in order of emphasis, AMINO ACIDS, ORGANIC COMPOUNDS, AROMATIC HYDROCARBONS, GLUCOSE OXIDASE, HYDROGEN PEROXIDE, ASCORBIC ACID, and HORSERADISH PEROXIDASE. These ANALYTES reflect the strong biomedical thrust of analytical chemistry today.

--THEME RELATIONSHIPS/ TAXONOMIES

While the phrase frequency analyses of technical text identify both major and obscure technical themes, they do not identify the relationships among the themes, or among the themes and the infrastructure. To achieve these goals, some type of co-occurrence or clustering analysis is required (2, 19, 20). In the aircraft study described in the previous section, or in the fullerene study (16), the clustering/ taxonomy generation, was performed manually. The experts assigned phrases to categories.

A more quantitative approach to defining categories, or clusters, is based on co-occurrence analysis. In co-occurrence analysis, a domain is selected (e.g., sentence, paragraph, Abstract, etc). The number of times multiple elements co-occur in that domain, the co-occurrence frequency, is the key indicator of the relatedness of those elements. Thus, if the retrieved record database consists of 100 records, and ANALYTICAL and CHEMISTRY co-occur in 30 of those records, then the co-occurrence frequency of ANALYTICAL and CHEMISTRY is 30. A symmetrical matrix can be constructed with phrase headings along the rows and columns, and the matrix element ANALYTICAL, CHEMISTRY would have a value of 30. For analysis purposes, it is very useful to non-dimensionalize the matrix elements, so that the co-occurrence frequency can be related to total occurrence frequency of each phrase in the total text.

Co-occurrence matrices do not have to be symmetrical. Author-phrase, author-organization, organization-phrase matrices could also be constructed. Analysis of these matrix patterns is very useful for understanding linkages that may not be immediately obvious from reading the literature. Figure 2 shows a dimensionalized co-occurrence matrix for the highest frequency phrases from the *Analytical Chemistry* retrieved records.

FIGURE 2

	Detection	separation	sensitivity	compounds	ions	electrode	spectra	measurements	water	mass spectrometry	proteins	detection limits	experiments	chromatography	spectrometry
detection	397	62	71	44	24	30	21	28	30	24	23	39	20	23	18
separation	62	215	19	31	12	7	11	8	12	15	22	19	9	21	13
sensitivity	71	19	175	20	9	21	14	12	8	9	17	16	6	7	10
compounds	44	31	20	155	4	3	14	6	10	9	1	15	4	12	14
ions	24	12	9	4	136	10	16	11	7	20	9	7	10	3	14
electrode	30	7	21	3	10	131	1	12	3	2	1	10	7		1
spectra	21	11	14	14	16	1	129	8	6	8	6	6	13	7	13
measurements	28	8	12	6	11	12	8	127	6	4	7	7	9	3	4
water	30	12	8	10	7	3	6	6	118	6	1	7	2	9	5
mass spectrometry	24	15	9	9	20	2	8	4	6	117	14	5	7	5	13
proteins	23	22	17	1	9	1	6	7	1	14	113	5	11	4	12
detection limits	39	19	16	15	7	10	6	7	7	5	5	107	3	8	7
experiments	20	9	6	4	10	7	13	9	2	7	11	3	104	6	7
chromatography	23	21	7	12	3		7	3	9	5	4	8	6	98	6
spectrometry	18	13	10	14	14	1	13	4	5	13	12	7	7	6	97

Clustering of phrases or authors or organizations can be useful for understanding and displaying linkages. There are many statistical packages available for clustering; finding an appropriate algorithm is relatively straight-forward. Defining the appropriate association metrics, and the cluster hierarchies, is much more complex. Phrase clustering has a fractal characteristic, wherein clusters can be appropriately formed in different phrase frequency regimes. The cluster structures are similar, but the high frequency phrase clusters contain relatively less detail than the low frequency phrase clusters. High frequency phrases tend to be single words with modest detail, while low frequency phrases tend to be multi-words, with correspondingly greater detail.

A clustering analysis was performed of the above co-occurrence matrix. A sub-set of the results is shown in Figure 3, as a four level hierarchical structure. The choice of the four levels, consisting of two, four, eight, and sixteen clusters, respectively, was made after detailed inspection of the dendogram. This is a tree-type schematic of the clustering process that accompanies the computer output, and it shows how each phrase is linked to other phrases in the clustering process. In the following discussion, phrases that appear in the clusters are capitalized. Only the top two levels will be discussed in any detail, with references made to specific clusters in the bottom two levels as appropriate.

FIGURE 3

# Records	CL2	CL4	CL8	CL16	
175	A-F	A-C	AB	A	Sensitivity
107	A-F	A-C	AB	A	detection limits
94	A-F	A-C	AB	A	resolution
90	A-F	A-C	AB	A	capillary electrophoresis
50	A-F	A-C	AB	A	acids
215	A-F	A-C	AB	B	separation
155	A-F	A-C	AB	B	compounds
93	A-F	A-C	AB	B	mixture
70	A-F	A-C	AB	B	temperature
50	A-F	A-C	AB	B	mobile phase
42	A-F	A-C	AB	B	liquid chromatography
41	A-F	A-C	AB	B	stationary phase
44	A-F	A-C	C	C	particles
136	A-F	D-F	DE	D	ions
117	A-F	D-F	DE	D	mass spectrometry
113	A-F	D-F	DE	D	proteins
104	A-F	D-F	DE	D	experiments

77	A-F	D-F	DE	D	laser
70	A-F	D-F	DE	D	peptides
36	A-F	D-F	DE	D	dissociation
63	A-F	D-F	DE	E	instrument
31	A-F	D-F	DE	E	plasma
52	A-F	D-F	F	F	complexes
37	A-F	D-F	F	F	distribution
35	A-F	D-F	F	F	assay
131	G-P	G-N	G-I	G	electrode
127	G-P	G-N	G-I	G	measurements
77	G-P	G-N	G-I	G	membrane
69	G-P	G-N	G-I	G	sensor
62	G-P	G-N	G-I	G	cell
52	G-P	G-N	G-I	G	film
45	G-P	G-N	G-I	G	enzyme
40	G-P	G-N	G-I	G	glucose
93	G-P	G-N	G-I	H	pH
61	G-P	G-N	G-I	H	oxidation
58	G-P	G-N	G-I	H	reduction
40	G-P	G-N	G-I	H	aqueous solution
35	G-P	G-N	G-I	I	theory
34	G-P	G-N	G-I	I	equation
33	G-P	G-N	G-I	I	diffusion
118	G-P	G-N	JK	J	water
37	G-P	G-N	JK	J	methanol
36	G-P	G-N	JK	J	sodium
89	G-P	G-N	JK	K	reaction
39	G-P	G-N	JK	K	carbon
36	G-P	G-N	JK	K	oxygen
96	G-P	G-N	L-N	L	molecules
82	G-P	G-N	L-N	L	fluorescence
32	G-P	G-N	L-N	L	dye
87	G-P	G-N	L-N	M	rate
44	G-P	G-N	L-N	M	surfaces
41	G-P	G-N	L-N	M	adsorption
37	G-P	G-N	L-N	N	transfer
33	G-P	G-N	L-N	N	Kinetics
52	G-P	OP	OP	O	polymer
47	G-P	OP	OP	P	volume
42	G-P	OP	OP	P	stability
36	G-P	OP	OP	P	flow
32	G-P	OP	OP	P	coating

The top hierarchical level, CL2, was the computer output for a two cluster assignment. The first cluster's members have the attribute A-F in column CL2. The cluster has a central theme of physical SEPARATION of COMPOUNDS in MIXTURES for the purpose of obtaining composition DISTRIBUTIONS, and is focused around physical SEPARATION techniques such as ELECTROPHORESIS, CHROMATOGRAPHY, and SPECTROMETRY. The second cluster's members have the attribute G-P in

column CL2. It has a central theme of direct REACTION RATE MEASUREMENTS of compounds in mixtures for the purpose of obtaining composition or specific compound distributions, and is focused around MEMBRANE or FILM-COATED SENSOR-ELECTRODE techniques that utilize SURFACE ADSORPTION, electron-TRANSFER KINETICS, and REACTION-RATE dependent processes.

The first cluster contains no specific mention of any theory-related terms, and concentrates on experimental techniques (CAPILLARY ELECTROPHORESIS, LIQUID CHROMATOGRAPHY, MASS SPECTROMETRY, ASSAY), experimental tools (LASER, SPECTROMETER, INSTRUMENT, DETECTOR), experimental processes and issues (DETECTION LIMITS, SENSITIVITY, RESOLUTION, SEPARATION, EXPERIMENTS), and experimental variables (TEMPERATURE, PRESSURE, IONIZATION, DISSOCIATION). These experiment-specific phrases tend to have high frequencies.

The second cluster contains mainly experimental-based phrases (ELECTRODES, SENSOR, CELL, FILMS, SUBSTRATE, SOLUTIONS, DYE, POLYMERS), and they also occur with high frequencies. Theory-related terms appear concentrated in two sub-clusters (I, N), all having relatively low frequencies in the mid-30s. Thus, the discipline of analytical chemistry, as portrayed by the literature in *Analytical Chemistry*, appears to be heavy on experiments and light on theory. Of all the disciplines text mined by the first author, analytical chemistry has the strongest imbalance between experiment and theory, as perceived through the literature.

The MASS SPECTROMETRY sub-cluster (D-F) focuses on PROTEINS and PEPTIDES, the MEMBRANE-coated ELECTRODE sub-cluster (G) focuses on ENZYME and GLUCOSE, and the element REACTION sub-cluster (JK) focuses on WATER, METHANOL, SODIUM, CARBON, OXYGEN. These focal areas span both clusters, and reflect the strong emphasis of analytical chemistry on biological, biomedical, and biochemical processes.

The next hierarchical level, CL4, was the computer output for a four cluster assignment. The first cluster, attribute A-C, focuses on the solution-based physical SEPARATION processes of ELECTROPHORESIS and CHROMATOGRAPHY. The second cluster, attribute D-F, focuses on the ION mass-based SEPARATION process of MASS SPECTROMETRY.

While these two clusters can be differentiated on the basis of fundamental physical operational characteristics, their linkage under the first cluster in the top level hierarchy reflects the potential for tandem operation. LIQUID CHROMATOGRAPHY and CAPILLARY ELECTROPHORESIS have been combined operationally as capillary electrochromatography, exploiting the best features of both processes, and capillary electrochromatography has been combined with MASS SPECTROMETRY to further exploit the best features of both processes.

The third cluster in the second hierarchical level, attribute G-N, focuses on the analysis of SURFACE REACTION/ ADSORPTION/ TRANSFER processes at MEMBRANE or FILM-coated ELECTRODES to identify SPECIES concentrations in mixtures. The fourth cluster in the second hierarchical level, attribute OP, focuses on the STABILITY of POLYMER-COATED sensors and cells under different FLOW and cell-VOLUME conditions.

The contents of the four level taxonomy described above are based on the 97 highest frequency technical phrases. As such, they are valuable for displaying the overall thematic structure of analytical chemistry, but are limited in describing the categories in any detail. To obtain the more detailed phrases associated with any of the clusters above, the following procedure is used. The detailed phrases that co-occur strongly with the highest frequency phrases (these low frequency detailed phrases only occur in close physical proximity to specific high frequency phrases) are identified through the use of numerical indicators. These low frequency detailed phrases are then placed in the same clusters as the high frequency phrases. This procedure allows both the 'needles-in-haystacks' (represented by the low frequency phrases) and the more generic descriptors (represented by the high frequency phrases) to co-exist, and places the low frequency phrases in the broader high frequency context.

To obtain a more detailed technical understanding of the clusters and their contents, the lower frequency phrases in each cluster need to be identified. A different matrix element non-dimensional quantity is required, one that remains relatively invariant to distance from the matrix origin. In addition, a different approach for clustering the low frequency phrases in the sparse matrix regions is required, one that relates the very detailed low frequency phrases to the more general high frequency phrases that define the cluster

structure. In this way, the low frequency phrases can be placed in their appropriate cluster taxonomy categories.

The following high frequency phrases were selected as generic themes for the low frequency phrases: ELECTROPHORESIS, CAPILLARY ELECTROPHORESIS, CHROMATOGRAPHY, LIQUID CHROMATOGRAPHY, SPECTROMETRY, MASS SPECTROMETRY, FILM, MEMBRANE, THEORY, FLUORESCENCE, KINETICS, POLYMER.

Four types of results were obtained with the lower frequency phrases.

1) Many of the lower frequency phrases were closely associated with one higher frequency phrase only (e.g., [High Frequency Phrase, Lower Frequency Phrases] CHROMATOGRAPHY: ATOMIC EMISSION DETECTION, PULSED AMPEROMETRIC DETECTION, BLOCK COPOLYMERS, IONIC LIQUIDS, VEGETABLE OIL TRIGLYCERIDES, GAS-LIQUID PARTITION COEFFICIENTS, IMMOBILIZED METAL AFFINITY; ELECTROPHORESIS: HIGH MASS RESOLUTION, MAPPING OF PROTEINS, ULTRATHIN SLAB GEL, POLYACRYLAMIDE GEL, PROTEIN KINASE, METHADONE, POLYPROPOLYNE HOLLOW FIBER, STEADY-STATE TRANSMEMBRANE CURRENTS);

The phrases in this category, on average, tend to be longer and more detailed/ specific than the phrases in any of the other categories. They also tend to be the lowest frequency phrases, and their length and detail characteristics are consonant with the very lowest frequency phrases.

2) Some lower frequency phrases were unique to a second tier cluster (e.g., [High Frequency Phrases in Second Tier Cluster: Low Frequency Phrases] CHROMATOGRAPHY, ELECTROPHORESIS: SEPARATION TECHNIQUES; FLUORESCENCE, KINETICS: ANISOTROPY AND INTENSITY, C-18 MODIFIED SILICA, DETERMINING THE EXCITED-STATE; KINETICS, MEMBRANE: CHARGE TRANSFER; FILM, MEMBRANE: INTERNAL SENSING SOLUTION; FLUORESCENCE, MEMBRANE: NEUTRAL SODIUM IONOPHORE).

3) A few lower frequency phrases were unique to a first tier cluster (e.g., [Low Frequency Phrases: High Frequency Phrases in First Tier Cluster] CAPTURE NEGATIVE ION, GINSENG, HYDROXYLATIONS AND N-OXIDATIONS, METHYL VINYL KETONE, SEPARATION OF PROTEINS: CHROMATOGRAPHY, MASS SPECTROMETRY; DIGESTION OF PROTEINS: ELECTROPHORESIS, MASS SPECTROMETRY; FRACTIONAL FREE VOLUME, GLUTAMATE MICROSENSORS, NON-STEADY-STATE MASS TRANSFER, STM IMAGES: FILM, POLYMER; SILICONE RUBBER: MEMBRANE, POLYMER.

4) Only a few lower frequency phrases were shared by all first tier clusters (e.g., [Low Frequency Phrases: High Frequency Phrases] ALLELES, APTEMER, COLON ADENOCARCINOMA CELL, EMITTED FLUORESCENT LIGHT, LASER BEAM SCANNING: ELECTROPHORESIS, FLUORESCENCE; CHARACTERISTIC FUNCTION METHOD, STOCHASTIC MODEL: CHROMATOGRAPHY, KINETICS; CHLORINATED AROMATIC COMPOUNDS, HOLLOW WAVEGUIDE SAMPLER: SPECTROMETRY, FILM; DERIVATIZED WITH METHYLAMINE, POLY METHACRYLIC ACID, SIZE-EXCLUSION CHROMATOGRAPHY, UNIVERSAL CALIBRATION: CHROMATOGRAPHY, POLYMER; MICRODIALYSIS JUNCTION: MASS SPECTROMETRY, MEMBRANE; PULSED RF PLASMA: MASS SPECTROMETRY, POLYMER; REDOX CYCLING: ELECTROPHORESIS, FILM).

As a general rule, the low frequency phrases in this category tend to be relatively generic, at least compared to phrases in the other three categories.

BARRIERS TO S&T TEXT MINING IMPLEMENTATION

Despite the myriad potential applications of text mining to the advancement of S&T, the surface of this powerful technique has barely been scratched. There exist many barriers to its widespread implementation, and these will be outlined. These barriers include: 1) lack of incentives; 2) lack of awareness of available text mining capabilities; 3) database limitations; 4) lack of coordination in technical community; 5) text mining not integrated with business operations.

1) Lack of Incentives

A substantial effort is required to obtain high quality information retrieval and text mining. The computer can produce thousands of phrases and phrase patterns from the core text. Human expertise is required to sift out the nuggets from the large background clutter. Unfortunately, there are presently few, if any, rewards for expending the effort on high quality text mining, and there are essentially no penalties for doing low quality text mining. In addition, the 'not-invented-here' syndrome is a strong dis-incentive for expending substantial effort to determine S&T performed elsewhere.

2) Lack of Awareness of Available Text Mining Capabilities
S&T personnel are unaware of required or available processes and tools for, and subsequent potential benefits from, high quality information retrieval and text mining. How many readers of *Analytical Chemistry* had any familiarity with text mining before reading this article?

3) Database Limitations
The base data available restricts what can be obtained from text mining. There is over $500 Billion of S&T being performed globally on an annual basis. Only a very modest fraction of this S&T is documented (21). Of the S&T documented, only a modest fraction is accessed by the major S&T databases (Science Citation Index, Engineering Compendex, NTIS Technical Reports, etc). Of this accessed documented S&T, only a modest fraction is available to the user because of cost, restricted access, inclusion of data fields not uniform across databases, lack of awareness, and user unfriendliness of the software. A major factor driving this step and the previous step is that the contents of the databases are determined by the database developers, not the S&T sponsors or the users. Of the available accessed documented S&T, only a modest fraction is available to the information processing software due to poor information retrieval techniques, and poor text-to-phrase conversion techniques.

4) Lack of Coordination in Technical Community
Database development, data input quality and structure, and data dissemination require horizontal co-operation among global entities, and vertical co-operation among the full spectrum of S&T sponsors, database developers, journal publishers and editors, and research performers and managers. There is no coordinated agreement and support for the full data development and dissemination cycle. The paradox exists that co-operation among competitors is required for the common good.

5) Text Mining not Integrated with Business Operations

Organizationally, text mining and other decision aids are not treated as an integral part of the S&T strategic management process (22). Rather, it is treated as an ad hoc add-on, in isolation from other management decision aids. The downside of such an approach is that the study objectives are driven by the data available from ordinary business operations, rather than the study objectives driving the data necessary to quantify the business performance metrics.

CONCLUSIONS

Text mining comprises a system of algorithms and procedures that, when coupled with expert human analysts, can extract highly useful information from technical text. The typical iteratively-generated queries used in our studies contain a few hundred phrases/ phrase combinations. These queries are more than an order of magnitude larger than those used by the average researcher for literature searches. Queries of this length are required for comprehensive and highly relevant retrievals of the target literature, related literatures, and disparate literatures with some common thread. The quality of the retrieved literature limits the potential quality of any subsequent information processing, whether it is bibliometrics, computational linguistics, or literature-based discovery and innovation. Development of these high-quality queries requires time and some cost, and participation of both technical domain and information technology experts.

The bibliometrics analyses in our studies are useful for identifying credible experts for workshops and review panels, and for planning itineraries of productive individuals and organizations to be visited. The wide spectrum discipline database generated by the enhanced query allows more innovation-oriented workshops to be conducted (13). through identifying more related technical disciplines, and the leading experts in these disciplines.

Computational linguistics allow the generation of taxonomies for the specific technical disciplines being examined, allow global levels of emphasis in the target discipline to be estimated, and allow technology-infrastructure relationships to be determined. The taxonomies are useful for providing deeper analytical insight into the technologies of interest, for insuring that program management categories reflect the separation of

technologies, and for improving the management of multi-discipline science and technology areas. The level of emphasis estimations are useful for determining adequacies and deficiencies in technology sub-areas on a global basis, and inputting to investment strategy determinations. The technology-infrastructure relationships are valuable for intelligence applications, as well as for the reasons provided under the bibliometrics section, but applied to sub-areas of the technology.

The final benefit addressed is one that has occurred in every one of the text mining studies that have been performed, and its value cannot be stressed too strongly. From an organization's long-range strategic viewpoint, the main output from these text mining studies is the technical expert(s) who has had his/ her horizons and perspectives broadened substantially as a result of participating in the full text mining process, and who can use this expanded knowledge to better support the conduct and the management of the S&T. While the text mining tools/ processes/ protocols/ tangible products are important, they are of lesser importance to the organization's long-term strategic health relative to the expert with advanced capabilities.

Text mining has enormous potential to support the rapid advancement of S&T. High quality S&T text mining requires substantial time and effort. There exist a number of barriers to its wide-scale implementation. They all originate from the absence of serious global agreements to develop the databases, train skilled personnel in S&T text mining, develop affordable high quality text mining techniques for a variety of applications, and implement prototype demonstrations of these techniques.

REFERENCES

(1) Losiewicz, P.; Oard, D.; Kostoff, R. N. *Journal of Intelligent Information Systems*. 2000. 15: 2. 99-119.

(2) Hearst, M. A. *Proceedings of ACL '99, the 37th Annual Meeting of the Association of Computational Linguistics, 1999.* University of Maryland, June 20-26.

(3) PROCEEDINGS SIGIR 99, *Proceedings of the Conference on the 22nd International Conference on Research and Development in Information Retrieval*, eds., Hearst, M., Gey, F., and Tong, R., ACM Press, New York, N.Y., 1999.

(4) NIST Special Publication 500 240; The Sixth Text Retrieval Conference (TREC-6), NTIS, eds**.,** *E. M. Voorhees and D. K. Harman*, order number PB98-148166, Department of Commerce, National Institute of Standards and Technology, 1998.
(5) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. *Journal of Information Science* 1997. 23:4. 301-311.
(6) Braun, T.; Bujdoso, E.; Schubert, A. *Literature of Analytical Chemistry: A Scientometric Evaluation*. CRC, CRC Press, 1987.
(7) Swanson, D.R. *Perspect Biol Med*. 1986. 30: 1. 7-18.
(8) Swanson, D.R.: Smalheiser, N.R. *Artif Intell. 1997.* 91:2. 183-203.
(9) Swanson, D.R. *Abstr Pap Am Chem S. 1999.* 217. U551-U551.
(10) Smalheiser, N.R.; Swanson, D.R. *Neurosci Res Commun*. 1994. 15: (1). 1-9.

(11) Smalheiser, N.R.; Swanson, D.R. *Comput Meth Prog Bio*. 1998. 57: (3). 149-153.
(12) Smalheiser, N.R.; Swanson, D.R. *Arch Gen Psychiat.* 1998. 55 :8. 752-753.
(13) Kostoff, R. N. *Technovation.* 1999. 19: 10. 593-604.
(14) Kostoff, R. N.; Green, K. A.; Toothman, D. R.; Humenik, J. *Journal of Aircraft* 2000. 37:4. 727-730.
(15) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. *Information Processing and Management* 1998. 34:1. 69-85.
(16) Kostoff, R. N.; Braun, T.; Schubert, A.; Toothman, D. R.; Humenik, J. *Journal of Chemical Information and Computer Science*. 2000. 40. 19-39.
(17) Gordon, M. D.; Lindsay, R. K. *JASIS.* 1996. **47**:2. 116-128.
(18) Finn, R. *The Scientist.* 1998. 12:10. 12-13.
(19) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. *Journal of the American Society for Information Science*. 1999. 50:5. 427-447.
(20) Zamir, O. *PhD Thesis* . University of Washington, 1999.
(21) Kostoff, R. N. *The Scientist*, 2000. 14:9. 6-6.
(22) Kostoff, R. N.; Geisler, E. *Technology Analysis and Strategic Management*. 1999. 11:4. 493-525.
(23) Kostoff, R. N.; DeMarco, R. A. *Analytical Chemistry.* 2001. 73:13. 370-378A.