NORTH-HOLLAND

Technological Forecasting & Social Change
68 (2001) 223–253

Technological Forecasting and Social Change

Text mining using database tomography and bibliometrics: A review[☆]

Ronald N. Kostoff[a,*], Darrell Ray Toothman[b], Henry J. Eberhart[c], James A. Humenik[d]

[a]*Office of Naval Research, 800 N. Quincy Street, Arlington, VA 22217, USA*
[b]*RSIS, Inc., McLean, VA 22102, USA*
[c]*Ridgecrest, CA 93555, USA*
[d]*NOESIS, Inc., Manassas, VA 20109, USA*

Received 13 November 2000; received in revised form 23 January 2001; accepted 25 January 2001

Abstract

Database tomography (DT) is a textual database analysis system consisting of two major components: (1) algorithms for extracting multiword phrase frequencies and phrase proximities (physical closeness of the multiword technical phrases) from any type of large textual database, to augment (2) interpretative capabilities of the expert human analyst. DT has been used to derive technical intelligence from a variety of textual database sources, most recently the published technical literature as exemplified by the Science Citation Index (SCI) and the Engineering Compendex (EC). Phrase frequency analysis (the occurrence frequency of multiword technical phrases) provides the pervasive technical themes of the topical databases of interest, and phrase proximity analysis provides the relationships among the pervasive technical themes. In the structured published literature databases, bibliometric analysis of the database records supplements the DT results by identifying: the recent most prolific topical area authors; the journals that contain numerous topical area papers; the institutions that produce numerous topical area papers; the keywords specified most frequently by the topical area authors; the authors whose works are cited most frequently in the topical area papers; and the particular papers and journals cited most frequently in the topical area papers. This review paper summarizes: (1) the theory and background development of DT; (2) past published and unpublished literature study results; (3) present application activities; (4) potential expansion to new DT

[☆] The views in this paper are solely those of the authors and do not represent the views of the Department of the Navy, any of its components, NOESIS or RSIS.

[*] Corresponding author. Tel.: +1-703-696-4198; fax: +1-703-696-4274.

E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).

PII: S0040-1625(01)00133-0

applications. In addition, application of DT to technology forecasting is addressed.

Keywords: Database tomography; Text mining; Bibliometrics; Innovation; Information retrieval; Information extraction; Cluster; Taxonomies

1. Introduction

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a concomitant increase in the need for scientific and technical intelligence to insure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, there have become available many techniques that can support and complement direct human intelligence gathering. In particular, techniques that identify, select, gather, cull, and interpret large amounts of technological information semiautonomously can expand greatly the capabilities of human beings for performing technical intelligence.

One such technique is database tomography (DT) [1–5], a system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multiword phrase frequencies from the textual databases and performing phrase proximity analyses, coupled with the topical expert human analyst to interpret the results and convert large volumes of disorganized data to ordered information. Phrase frequency analysis (occurrence frequency of multiword technical phrases) provides the pervasive technical themes of a database, and phrase proximity (physical closeness of the multiword technical phrases) analysis provides the relationships among pervasive technical themes, as well as among technical themes and authors/journals/institutions/countries, etc. The present paper reviews the evolution and applications of the DT process, including recent augmentation of DT capabilities by literature bibliometric analyses and more intensive use of topical domain experts, to derive technical intelligence from textual databases.

What is the importance of applying DT and bibliometrics to a topical field of interest? The roadmap, or guide, of this field produced by DT and bibliometrics provides the demographics and a macroscopic view of the total field in the global context of allied fields. This allows specific starting points to be chosen rationally for more detailed investigations into a specific topic of interest. DT and bibliometrics do not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but allow these detailed efforts to be executed more efficiently.

In addition, combination of DT-based literature analysis for discovery and innovation [6] with innovation workshops can help identify promising science and technology directions for research managers and promising science and technology opportunities for research performers.

2. Background

2.1. Overview

DT can be perceived as a computer-based process for developing roadmaps to the textual representation of topical areas. A more detailed and comprehensive survey of science and technology roadmaps can be found in Ref. [7], and the unique features of the co-word-based DT process relative to other roadmap techniques are summarized in Ref. [8]. This latter reference describes the two main roadmap categories (expert-based and computer-based), summarizes the different approaches to computer-based roadmaps (citation and co-occurrence techniques), presents the key features of classical co-word analysis, and shows the detailed evolution of DT from its co-word roots to its present form. The development of DT will now be summarized.

2.2. Development of DT

In 1990–1991, experiments performed at the Office of Naval Research [9] showed that the frequency with which phrases appeared in full-text narrative technical documents was related to the main themes of the text. The phrases with the highest frequencies of appearance represented the main, 'pervasive' themes of the text. In addition, the experiments showed that the physical proximity of the phrases was related to the thematic proximity. These experiments formed the basis of DT.

The DT method in its entirety requires generically four distinct steps. The first, and most time-consuming, step is the extraction of the text to be analyzed from the source databases [10]. The second step is identification of the main themes of the text being analyzed. The third step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first three steps will be summarized now. Time evolutions of themes have not yet been performed.

First, for the more recent journal paper-based studies, an initial test query applied to the source technical databases retrieves a sample of papers. Evaluations of the sample by technical experts result in two groups of papers. One group is judged by the domain topical expert(s) to be relevant to the subject matter; the other is judged to be nonrelevant. Gradations of relevancy or nonrelevancy are not considered (although they could be, and weighting, or correction, factors then applied). An initial database of titles, keywords, and Abstracts is created for each of the two groups of papers. Phrase frequency and proximity analyses are performed on this textual database for each group. The high-frequency single-, double-, and triple-word phrases characteristic of the relevant group, and their Boolean combinations, are then added to the query to expand the number of papers retrieved. Similar phrases characteristic of the nonrelevant group are added to the query as negation terms (NOT Boolean) to contract the number of papers retrieved. The process is repeated on the new database of titles, keywords, and Abstracts obtained from the search. A few more iterations are performed until the number of records retrieved stabilizes (convergence).

Second, once the relevant records are retrieved, the frequencies of appearance in the total text of all single-word phrases (e.g., matrix), adjacent double-word phrases (e.g., metal matrix), and adjacent triple-word phrases (e.g., metal matrix composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database of relevant records.

Third, for each theme phrase, the frequencies of phrases within $\pm M$ (nominally 50) words of the theme phrase for every occurrence in the full text are computed, and a phrase frequency dictionary is constructed. This dictionary contains the phrases physically (and thematically) close to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert(s) for each dictionary (hereafter called cluster) yielding, among many results, those subthemes closely related to and supportive of the main cluster theme.

Fourth, threshold values are assigned to the numerical indices, and these indices are used to filter out those phrases most closely related to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and subthemes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full-text analysis allows an understanding of the theme interrelationships not heretofore possible with previous text Abstraction techniques (using index words, key words, etc.).

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles [2,3], the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting subthrust areas (both high- and low-frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article Abstracts and associated infrastructure/bibliometric information (authors, journals, addresses, etc.), the final results have also included relationships among the technical themes and authors, journals, institutions, etc. (e.g., Refs. [8,11]). The present paper summarizes the methods and outputs of these two generic types of DT applications.

2.3. Evolution of DT into text mining

Recent evaluations of real-world text mining applications (unpublished) across a number of organizations showed a strong decoupling of the text mining research performer from the text mining user. The performer tended to focus on the development of exotic automated techniques, to the relative exclusion of the components of judgement necessary for user credibility and acceptance. Consequently, the text mining techniques actually employed by most of the potential users examined involved reading copious numbers of articles obtained by the simplest of queries. The more recent journal Abstract-based DT processes reported in latter sections of this paper represent the framework of a text mining approach that will couple the text mining research and associated computer technology processes much more closely with the text mining user. Strategic database maps will be developed on the front end of the process using bibliometrics and DT, with heavy involvement from topical domain experts (either users or their proxies) in the DT component of strategic map generation. The strategic maps themselves will then be used as guidelines for detailed expert analysis of segments of the total database. The authors believe that this is the proper use of automated techniques for text mining: to augment and amplify the capabilities of the expert by providing insights to the database structure and contents, not to replace the experts by a combination of machines and nonexperts.

3. DT applications

3.1. Initial applications to technical reports

The initial DT applications utilized the phrase frequency algorithms to identify pervasive themes from text, then, at a later stage, incorporated the phrase proximity algorithms to identify theme relationships. Two of these applications will be summarized; more detailed descriptions can be found in Ref. [2].

3.1.1. Theme identification and interrelationships: promising research opportunities database

The multiword phrase frequency generator was applied to a known and modest-sized ($\sim$600KB) compendium of reports of promising research opportunities for the Navy developed by National Academy of Sciences panels and Navy internal experts. When the resultant single-, double-, and triple-word phrases were ordered by frequency, a clear picture of the pervasive themes (themes that in many cases cut across different disciplines) of the total text emerged. This computerized scanning of the database provided a starting point for the development of technical guidance that was eventually sent to members of the Navy research management community.

Phrase proximity analyses were performed on an upgraded version of the promising research opportunities database. The objective was to identify the relationships among these thrusts in order to see what multidisciplinary thrusts are emerging.

Initially, a single-, double-, and triple-word phrase frequency analysis was performed. The 20 highest frequency technical single-word phrases, and 30 highest frequency double-word phrases, were defined as themes and extracted, and phrase frequency dictionaries (clusters) were generated for each one of these themes. Triple-word phrases were not chosen because of the relatively small size of the text database and the consequent relatively small number of high-frequency phrases in the dictionary.

The contents of each cluster were structured into different categories. In terms of cluster categorization, as a compromise between detail and conciseness, each cluster could be subdivided into from two to four categories. For those themes that were fairly specific, such as Integer Programming, subcategorization was straightforward. For those themes that were fairly general, and perhaps ambiguous in meaning, such as a homonym like Current, subcategorization was much more difficult, and an integrated set of categories was in some cases impossible. Usually, though not always, the single-word phrase themes were harder to categorize because of the broader implications of the themes. The conclusion to be drawn is that cluster subcategorization is useful for integrating the disparate members into related topical groups when a focused theme exists, but subcategorization serves less of a purpose when the theme is diffuse.

For example, in the cluster from this study whose theme is ACOUSTIC, the first four phrases in the cluster ranked by one of the statistical indices were the single-word phrases PROPAGATION, SCATTERING, OCEAN, and BOTTOM. While they contain far less information than, say, 'WAVEGUIDE INVERSE SCATTERING,' 'OCEANOGRAPHIC SAMPLING NETWORK,' or 'COASTAL TRANSITION ZONE,' they do provide some broad structuring and categorization for the cluster, as well as the potential for broader overlap with other clusters. In fact, the set of single-word phrases included in the ACOUSTIC cluster (PROPAGATION, SCATTERING, OCEAN, BOTTOM, SENSORS, ARCTIC, WATER, WAVE, MODELING, ENERGY, DATA) provided a reasonable taxonomy for categorizing the double and triple-word phrases contained in the ACOUSTIC cluster.

Interestingly enough, these first four single-word phrases in the Acoustic cluster appeared to cover the main two subthemes within the cluster, namely, acoustic propagation within the ocean environment, and acoustic wave interactions with the boundaries (mainly ocean bottom).

The subcategorizations were performed for each of the 50 clusters to identify subtheme interrelationships. It was also desired to categorize the total database by a few relatively independent themes. Toward this end, megaclusters, or groupings of similar clusters were generated, to provide an orthogonalized taxonomy of the total database. Each megacluster consisted of clusters that had a threshold number of phrases in common with at least one other cluster in the megacluster. Based on the detailed clustering background provided in Zamir's thesis [12] on suffix tree clustering, the single-link megacluster generation process described above appears to have been the first application of multiword technical phrases to the clustering process.

Three high-level megaclusters were evident from the analyses. The first could be broadly categorized as Ocean Sciences, the second as Information Sciences, and the third as Materials. At the time of this study, the Office of Naval Research (ONR) identified three specific areas of emphasis (core competencies) in its investment strategy, namely, Ocean Sciences, Materials, and Information Sciences. Thus, the three broad megacluster areas identified by an analysis of experts' recommendations to ONR of promising research opportunities coincided with ONR's stated areas of emphasis.

3.1.2. Theme identification and interrelationships: former Soviet Union database

For the decade of the 1980s, assessments were made of selected areas of foreign applied science by the Foreign Applied Sciences Assessment Center (FASAC), a multiagency supported project. Panels were contracted to assess the foreign literature (mainly Soviet) in the chosen area, and then write a report. The 35 reports on Soviet applied science were combined into one text database, and subjected to DT analysis [3]. While the focus of the FASAC study was identifying technical thrusts and their interrelationships, the raw data obtained by the extraction algorithms allowed the user to relate technical thrusts to institutions, journals, people, geographical locations, and other categories.

The phrase frequency generator (the technical phrases output from the phrase and proximity generators are shown as CAPS in the remainder of this paper) was applied to the FASAC database. High technical content phrases were arbitrarily categorized in bins of similar science thrusts, and an applied research taxonomy was generated. It consisted of *Information* (IMAGE PROCESSING, PATTERN RECOGNITION, SIGNAL PROCESSING, ARTIFICIAL INTELLIGENCE, etc.), *Physics* (SHOCK WAVES, RADIO WAVES, CHARGED PARTICLE ACCELERATORS, OPTICAL PHASE CONJUGATION, etc.), *Environment* (INTERNAL WAVES, OCEANIC PHYSICS, SEA SURFACE, IONOSPHERIC MODIFICATION, etc.), and *Materials* (THIN FILM, COMPOSITE MATERIALS, FRACTURE MECHANICS, SOLID FUEL CHEMISTRY, etc.). Compared to other databases examined, there appeared to be a relatively higher occurrence of Physics-related terms and of terms related to Combustion/Explosion and its consequences.

The phrase proximity generator was applied to the FASAC database subsequent to the phrase frequency generator, and a taxonomy of the full FASAC database was obtained by the megacluster analysis described for the promising opportunities database. The cluster overlaps were determined, and those clusters that had three or more overlaps (three or more common members) were combined to form strings of related clusters, or megaclusters. Normalization, or adjustment of the overlap threshold criteria for different cluster sizes, was not performed.

The results of the multiword phrase frequency analysis performed on the total FASAC database allowed a high-level science taxonomy of four broad categories to be generated: Information, Physics, Environment, and Materials. A phrase proximity analysis on the 60 highest frequency pervasive themes identified by the multiword phrase frequency analysis, and a subsequent (effective) renormalization of the pervasive themes due to linkages among subthemes allowed nine 'umbrella' themes (megaclusters) to be generated: Ionospheric Heating/Modification; Image/Optical Processing; Air–Sea Interface; Low Observable; Explosive Combustion; Particle Beams; Automatic/Remote Control; Frequency Standards; Radar Cross Section. Based on the results and interpretation of the multiword phrase frequency and proximity analyses, it could be concluded that the FASAC database used in this study is a compendium of those aspects of FSU science of interest to the US for strategic and military purposes. The microlevel analysis of selected theme clusters, showing how the cluster members related to each theme, reinforced this conclusion and provided more detail about those aspects of each theme on which FASAC concentrated.

For example, many classes of materials were researched and developed in the FSU. Yet the materials subcategory in the FASAC analysis focuses on FSU capabilities in energetic materials and coatings to reduce radar cross sections, both important classes from a military viewpoint. The main environmental focus is air–sea interface, with little mention of the terrestrial environment. Coupled with the information category focus on image and optical processing, and the secondary information category focus on remote control, it could be concluded that the FASAC concern was FSU capability in sensing the ocean for ship and submarine activity, and remotely processing and interpreting this information. The secondary environmental focus of FASAC was on the ionosphere, specifically on FSU capabilities for modifying the ionosphere through high power radio wave heating and exploiting its use as a communication medium. One focus of the physics category is particle beams, that could have dual applications of high energy directed weapons and heaters for magnetically confined plasmas and inertial fusion targets.

3.2. Recent applications to journal literature databases

In contrast to the initial studies reported above, the recent efforts have used databases of journal paper summaries, such as the Science Citation Index (SCI) and the Engineering Compendex (EC). The DT algorithms for phrase frequency and proximity analysis are supplemented by bibliographic and scientometric analyses to provide deeper insight into the structure and thematic relationships of the topical area of interest. The results of seven studies in the topical areas of Research Impact Assessment-RIA [13], Chemistry-JACS [13], Near-Earth Space-NES [8], Hypersonic/Supersonic Flow-HSF [11], Fullerenes-FUL [14], Aircraft-AIR [15], and Surface Hydrodynamics-HYD [16] will be summarized and integrated where possible.

3.2.1. Database generation

The key step in these literature analyses is the generation of the database to which the information processing algorithms will be applied. For most (not all) of these studies, the database consisted of selected journal and conference proceeding records (including authors, titles, journals, author addresses, author keywords, Abstract narratives, and references cited for each paper) obtained by searching the SCI and the EC for topical articles. The CD-ROM version of the SCI (used for the earlier studies) accesses about 3200 journals (mainly in physical, engineering, and life sciences basic research), while the Web version of the SCI (used in the more recent studies) accesses about 5300 journals. The EC accesses about 2600 journals and conference proceedings (mainly in applied research and technology).

The databases that have been selected for these studies typically represent a fraction of the available topical literature. They do not include the large body of classified literature, or company proprietary technology literature. They do not include the large body of technical reports on the topical area. They typically have covered a finite slice of time (early 1990s to present). It is the authors' perception that the databases used have represented the bulk of the peer-reviewed high-quality topical science and technology literature, and have served as a representative sample of all the relevant topical science and technology in recent times.

To extract the relevant articles from the SCI and EC, the Title, Keyword, and Abstract fields have been searched using phrases relevant to the topical area, although different procedures were used to search the Title and Abstract fields [10]. The resultant Abstracts were culled to those relevant to the topic. The searches have been performed with the aid of two powerful DT tools (multiword phrase frequency analysis and phrase proximity analysis) using the iterative process of Simulated Nucleation [10].

In most studies, the final query contained over 200 terms. The authors believe that queries of these magnitudes and complexities are required to provide a tailored database of relevant records that encompass the broader aspects of target disciplines. In particular, if it is desired to enhance the transfer of ideas across disparate disciplines, and thereby stimulate the potential for innovation and discovery from complementary literatures [10], then even more complex queries using Simulated Nucleation may be required.

3.2.2. Results

The results from the publications bibliometric analyses are followed by the results from the citations bibliometrics analysis in Section 3.2.2.1. Results from the DT analyses are shown in Section 3.2.2.2. The SCI and EC bibliometric fields incorporated into the database included, for each paper, the Author, Journal, Institution, and Keywords. In addition, the SCI included references for each paper.

The bibliometrics sections have two components. Important numerical indicators that illuminate some aspect of the topical research literature (e.g., average authors per paper, number of journals, papers per institution) are tabulated, and distribution functions of publication and citation parameters (e.g., numbers of authors $f(n)$ who publish n papers) are compared with those of other technical discipline studies that used a similar approach.

In the full published papers on DT-bibliometrics, the DT sections contain four components. First, the high-frequency Keywords are grouped into 'natural' categories, and the picture they provide of the topical literature (S&T, open literature, unclassified, nonproprietary) is described. Second, the high-frequency phrases from the Abstracts are grouped into 'natural' categories, and the picture they provide of the topical literature is presented. Third, the high numerical indicator phrases from the proximity analyses of the Abstracts and other portions of the database (author names, article titles, journal names, author addresses) are grouped into 'natural' categories, and the picture they provide of the topical literature is shown. Fourth, the technical expert's analyses and interpretation of all the Abstracts, enhanced by the computer-driven results from the three previous components, are summarized. In the present review paper, only the first three components are presented, due to space limitations.

The meaning of the term 'natural' in the previous paragraph is that these categories were not prescribed beforehand. From observation of the hundreds of different phrases and their frequencies by topical domain experts, categories that appeared to be useful for interpreting and describing the main literature findings emerged. These categories were not necessarily the same for each component.

The analytical approaches taken for the first three components are based on their fundamental data structures. The Keyword and Abstract phrase frequencies are essentially quantity measures. They lend themselves to 'binning,' and addressing adequacies and deficiencies in levels of S&T activity in the different technical subcategories. They do not contain relational information, and therefore offer little insight into S&T linkages. The phrase proximity results are essentially relational measures, although some of the proximity results imply levels of effort that support specific S&T areas. The phrase proximity results mainly offer insight into S&T linkages, and have the potential to help identify innovative concepts from disparate disciplines [6]. The phrase proximity results also offer insight into linkages between S&T categories and supporting infrastructures (performers, institutions, journals, etc.). Thus, the Keyword and Abstract phrase frequency analyses will be addressed to adequacy of effort, and the phrase proximity analyses will be addressed to intra-S&T/inter-S&T−infrastructure relationships primarily and supporting levels of effort secondarily.

Also, one might expect that each of the four components that derived from the same base of relevant records would produce the same overall conclusions, with perhaps the level of detail and some relational information differing among the components. This was not always the case; sometimes there were substantially different conclusions drawn from the components. This has implications for how the literature should be accessed, and how the literature should be interpreted when accessing only summary perspectives.

3.2.2.1. Results from the bibliometrics analyses. First, the results of the bibliometrics analyses will be presented, then followed by the results of the DT analyses. The SCI and EC bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, and keywords. In addition, the SCI included references for each paper.

The bibliometrics results are compared for the seven DT studies that have been performed. These results are presented for the SCI only. See Table 1 for the number of articles and their time period in these studies (RIA, NES, JACS, HSF, AIR, HYD, and FUL).

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred due to their publication in the (typically) high-caliber journals accessed by the SCI.

Prolific authors. In each study, the Author field was separated from the database, and a frequency count of author appearances was made. The most prolific authors were listed in order of decreasing publications. However, because of the database limitations described above, there may have been excellent researchers writing in the various topical fields who were omitted from the list.

Table 2 compares the SCI author bibliometric statistics for the different studies. These studies are listed, proceeding from left to right, in approximate order of the (subjectively estimated) science/technology ratio of the underlying field. Thus, the leftmost field listed,

Table 1
DT studies of topical fields

Topical Area	Number of SCI articles	Years covered
CHEMISTRY (JACS)	2,150	1994
NEAR-EARTH SPACE (NES)	5,480	1993–mid 1996
HYPERSONICS (HSF)	1,284	1993–mid 1996
RESEARCH ASSESSMENT (RIA)	2,300	1991–EARLY 1995
FULLERENES (FUL)	10,515	1991–mid 1998
AIRCRAFT (AIR)	4,346	1991–mid 1998
HYDRODYNAMICS (HYD)	4,608	1991–mid 1998

Table 2
Author bibliometrics — SCI

	Metric/Study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of authors	12,837	6535	12,453	7,869	2483	6619	2975
Number of author listings	41,167	8151	18,474	10,558	3372	9085	3868
Average number of listings per author	3.2	1.2	1.5	1.3	1.38	1.4	1.3
Number of papers retrieved	10,515	2150	5,481	4,608	1284	4346	2300
Average number of author listings per paper	3.92	3.79	3.37	2.29	2.63	2.09	1.68

FUL, is estimated to be the most fundamental (based on the specific query used and the themes of the papers retrieved), and the rightmost technical field, AIR, is estimated as the most applied. RIA, the rightmost column, is not a technical field, and is listed for completeness only.

In Table 2, five variables/figures of merit are presented for each study. The number of authors represents the total number of different names contained in the author blocks, while the number of author listings is the sum over all authors of the number of times each author's name was listed in an author block. The average number of (author) listings per author is the ratio of the above two quantities. The number of papers retrieved is the total number of relevant papers that comprised the database and was used for the analyses, while the average number of author listings per paper is the number of author listings divided by the number of papers retrieved.

Fig. 1 shows the distribution function of SCI author listing frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abcissa is the number of author listings *n*, and the ordinate is the number of authors who have author listing *n*. In each case, the distribution function has been normalized to the number of authors who have one listing in the respective databases. The graph is plotted on a semilog scale to stretch the lower ordinate region.

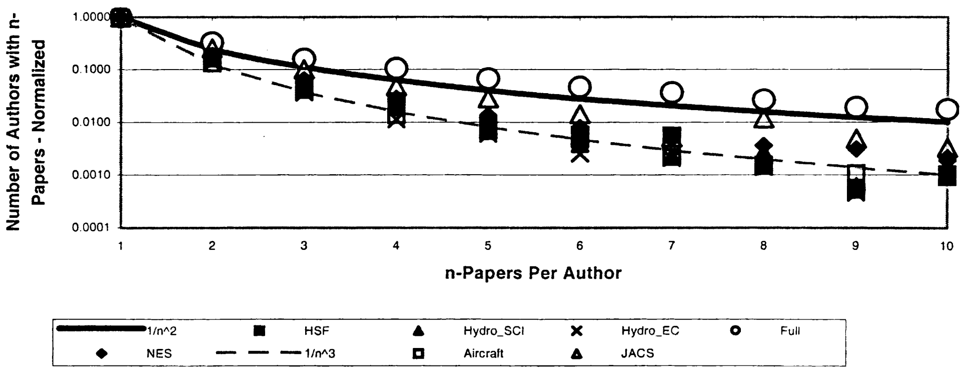

Fig. 1. Author distribution.

Table 3
Journal bibliometrics — SCI

	Metric/study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of papers retrieved	10515	2150	5481	4608	1284	4346	2300
Number of journals	680	1	628	675	277	713	645
Average number of papers per journal	15.46	2150	8.73	6.83	4.6	6.10	3.57
Bradford's law — ratio between groups	2.2		2	1.5	3	3.1	

The solid line on Fig. 1 is the nominal ($1/n^2$) Lotka's law [17] distribution. With the exception of the FUL data, all of the experimental data decline much steeper than the ($1/n^2$) law predicts, centering about a ($1/n^3$) distribution. One interpretation of this observation is that Lotka concentrated on only the very core journals in the disciplines studied. These journals tend to accept relatively more contributions from the prolific and recognized researchers than the non-core journals.

Journals containing most topical papers. A similar process was used to develop a frequency count of journal appearances. Table 3 compares the SCI journal bibliometric statistics for the different studies.

Four variables/figures of merit are presented for each study. The number of journals represents the total number of different journal names contained in the source blocks. The average number of papers per journal is the ratio of total papers retrieved to total number of journals. The Bradford's law [18] metric derives from the following definition/restatement of the law: If the journals for a bibliography are grouped in order of decreasing publications, such that each group of journals contains the same number of papers, then the ratio of number of journals in each successive group will be a constant greater than unity. The Bradford's law metric in Table 3 is this ratio between journal groups.

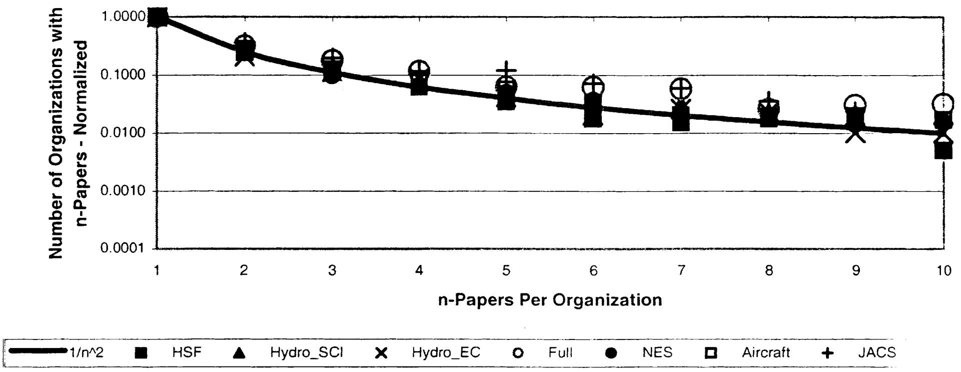

Fig. 2. Journal distribution.

One unexpected finding is the closeness of the magnitudes of number of journals for the different studies. Of the seven different topics studied, using different experts and different queries and different versions of the SCI and having different science/technology ratios, the total number of journals for five of those topics is within about 10% of 650. In fact, for four of those five journals, the total number of journals is within about 5% of 650.

Fig. 2 shows the distribution function of SCI journal frequency for the FUL, AIR, HYD, HSF, NES, and RIA databases. The JACS database was derived from one journal only, *The Journal of the American Chemical Society*, and therefore was not applicable to this chart. The abcissa is the number of papers n from the relevant database published in a given journal, and the ordinate is the number of journals that contain *n* papers. In each case, the distribution function has been normalized to the number of journals that contain one relevant paper. Again, because of the strong initial gradients, the graph is plotted on a semilog scale.

Table 4
Institution bibliometrics — SCI

	Metric/Study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of papers retrieved	10,515	2150	5,481	4608	1284	4346	2300
Number of institutions	2,168	750	10,435	1905	661	1484	1125
Average number of papers per institution	4.85	2.9	0.53	2.42	1.94	2.93	2
Average number of authors per institution	5.92	8.7	1.19	4.13	3.76	4.46	2.64
A. Country bibliometrics — SCI							
Upper							
Number of countries	63	44	105	78	53	56	56
Ratio of US papers to five nearest producers	0.73	2.5	1.94	1.32	1.6	1.74	2.47
Lower							
Rank							
1	US-5861	US-2040	US-5266	US-2708	US-1677	US-2771	US-1595
2	JP-2840	JP-276	UK-660	UK-560	RU-230	UK-507	UK-279
3	GR-1500	CN-168	FR-614	RU-420	JP-224	GR-250	CN-138
4	CH-1363	GR-148	JP-549	JP-388	FR-161	FR-218	NL-80
5	RU-1177	FR-116	CN-476	FR-377	GR-143	JP-218	GR-79
6	FR-1117	UK-109	GR-471	GR-314	UK-143	RU-163	FR-71
7	UK-1001	IT-97	RU-370	CN-252	IT-66	CN-133	AU-69
8	IT-586	SP-58	IT-274	IT-164	TW-57	ID-112	SP-58
9	ID-474	ST-53	AU-207	ID-149	CH-52	AU-86	HU-46
10	ST-418g	IS-48	ID-203	TW-132	ID-49/ CN-49	IS-84	BE-45

Code: US, United States; UK, United Kingdom; CN, Canada; NL, Netherlands; GR, Germany; FR, France; JP, Japan; RU, Russia; CH, China; IT, Italy; ID, India; AU, Australia; SP, Spain; HU, Hungary; BE, Belgium; ST, Switzerland; IS, Israel; TW, Taiwan.

The solid line in Fig. 2 is a ($1/n^2$) distribution, and represents a lower bound of all the experimental data.

Institutions producing most topical papers. A similar process was used to develop a frequency count of institutional appearances. Table 4 compares the SCI institutional bibliometric statistics for the different studies.

Four variables/figures of merit are presented for each study. The number of institutions represents the total number of different institution names contained in the address blocks. The average number of papers per institution is the ratio of total papers retrieved to total number of institutions. The average number of authors per institution is the ratio of total number of authors to total number of institutions.

Fig. 3 shows the distribution function of SCI institution frequency for the HSF, NES, JACS, AIR, HYD, and FUL databases. The abcissa is the number of papers n in the database produced by a given institution, and the ordinate is the number of institutions that produced n relevant papers. In each case, the distribution function has been normalized to the number of institutions that produced one relevant paper.

The data center around a ($1/n^2$) distribution remarkably well. For a ($1/n^2$) distribution, the number of organizations that generate three papers is about 11% of the organizations that generate one paper only. Also, integrating this distribution function shows that more than 67% of the papers result from organizations that produce three or less papers.

Countries producing most topical papers. The countries producing the most topical papers were listed in each study. Table 4A presents the bibliometric statistics for all the studies performed. The upper component of Table 4A has two metrics. Number of countries refers to the total number of countries listed in all the address blocks. Ratio of US papers to five nearest producers is the number of US listings in the address blocks of all papers divided by the number of listings in rank order of the next five countries. The lower component of Table 4A contains the top 10 country listings for each study in rank order.

The dominance of a handful of countries is clearly evident in all the studies, especially the dominance of the United States. In many cases, the US is almost an order of magnitude more prolific than its nearest competitor in terms of absolute numbers of papers produced, and in

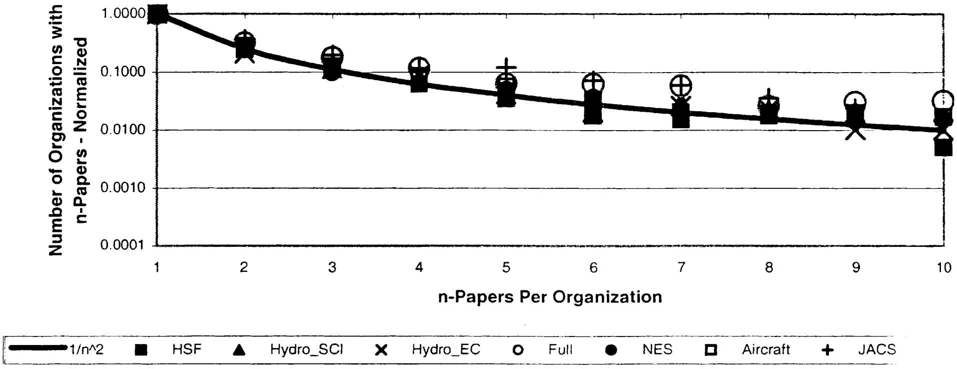

Fig. 3. Organization distribution.

some cases is as prolific as its nearest major competitors combined. In fact, the total number of country listings summed across all seven studies is 21,307 for the US, and 15,515 for all other countries combined.

Most cited authors, papers, years, and journals. The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics, much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers [19–21].

The citations in all the SCI papers were aggregated, the authors, specific papers, years, journals, and countries cited most frequently were identified, and were presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations.

Most cited authors. Table 5 compares the bibliometric statistics for the different studies. Seven variables/figures of merit are presented for each study.

The number of citations represents the total numbers of references in all papers retrieved. The average number of citations per paper is the ratio of total number of citations to total number of papers retrieved. The number of authors cited is the total number of different first authors cited. The average number of citations per author cited is the ratio of total number of citations to total number of authors cited. The average number of citations per author is the ratio of references to authors.

Fig. 4 shows the distribution function of author citation frequency for the FUL, NES, HSF, JACS, AIR, and HYD databases. The abscissa is the total number of citations n received by a given author, and the ordinate is the number of authors that received n total citations. In each case, the distribution function has been normalized to the number of authors that received one citation.

The data cluster very closely around a $(1/n^2)$ distribution, making this distribution far more universal than the somewhat discipline-dependent author publishing distribution.

Table 5
Cited author bibliometrics — SCI

	Metric/Study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of papers retrieved	10,515	2,150	5,481	4,608	1,284	4,346	2,300
Number of citations	263,844	5,000 +	140,662	82,395	26,768	45,744	7,000 +
Average number of citations per paper	25.1	39.5	25.7	17.9	20.9	10.5	16.1
Number of authors cited	33,579	32,450	42,094	26,322	11,138	21,868	18,140
Average number of citations per author cited	7.86	2.62	3.34	3.13	2.4	2.09	2
Number of authors	12,837	6,535	12,453	7,869	2,483	6,619	2,975
Average number of citations per author	20.6	13	11.3	10.5	10.8	6.9	12.4

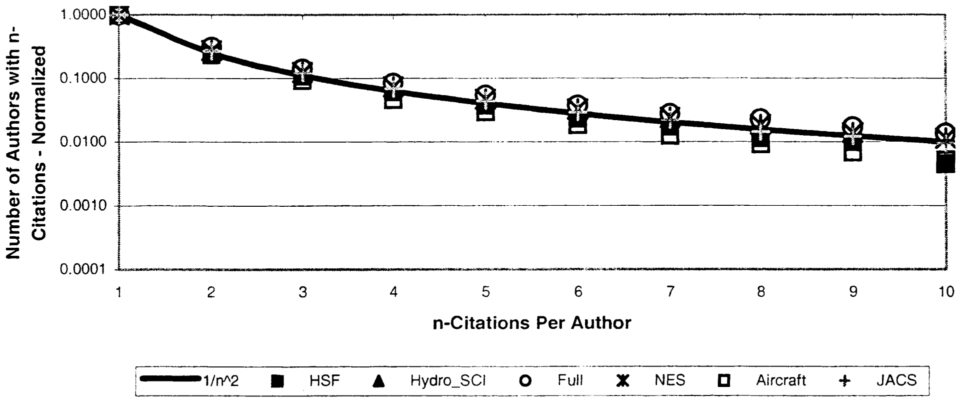

Fig. 4. Cited author distribution.

Most cited papers. Table 6 compares the bibliometric statistics for the different studies. Four variables/figures of merit are presented for each study.

The number of different papers cited is the total number of different papers referenced by the papers in the database. The average number of citations per cited paper is the ratio of number of citations to number of different papers cited. The average number of papers cited per author cited is the ratio of total papers cited to total authors cited.

Fig. 5 shows the distribution function of paper citation frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abscissa is the total number of citations n received by a given paper, and the ordinate is the number of papers that received n total citations. In each case, the distribution function has been normalized to the number of papers that received one citation.

For five of the six topical fields presented, the data follow a $(1/n^3)$ distribution very closely, as contrasted with the $(1/n^2)$ distribution for author citations. Examination of the five topical studies that produced the five sets of data showed that each of the highly cited authors had a wide range of citations for his/her different papers. For any given highly cited author, most papers will receive few citations. It is the infusion of numbers of lowly cited papers from the highly cited authors that expands the pool of lowly cited papers in Fig. 5, and results in the

Table 6
Cited paper bibliometrics — SCI

	Metric/Study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of citations	263,844	5,000 +	40,662	82,395	26,768	45,744	37,000 +
Number of different papers cited	75,890	64,800	93,194	57,618	20,950	38,792	30,400
Average number of citations per cited paper	3.48	1.31	1.51	1.43	1.27	1.18	1.22
Average number of papers cited per author cited	2.26	2	2.21	2.19	1.88	1.77	1.68

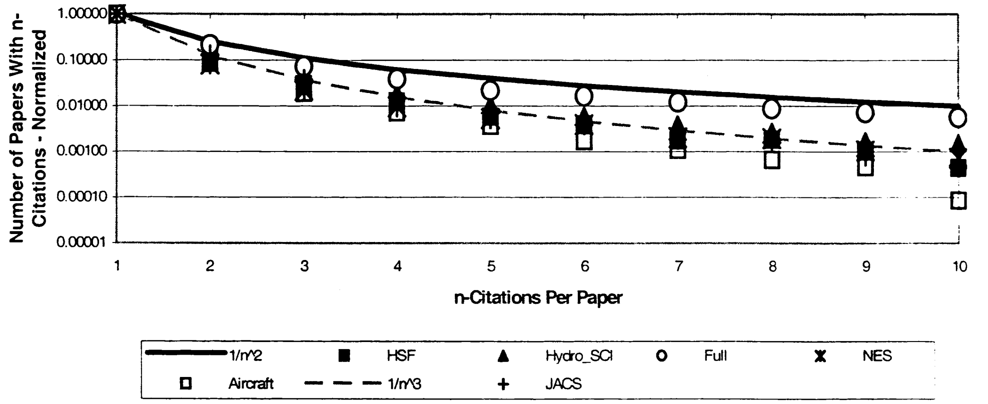

Fig. 5. Cited paper distribution.

conversion of the $(1/n^2)$ distribution of Fig. 4 to the $(1/n^3)$ distribution of Fig. 5. This effect appears to transcend the five different science and technology topical fields, and to be almost universal based on the limited data presented for the six topical science and technology fields. This relation, the Kostoff–Eberhart–Toothman (KET) law [11], can be restated as follows: for a topical science and technology field, the ratio of the normalized number of authors with n citations per author to the normalized number of papers with n citations per paper is n, for low to moderate values of n.

Most cited journals. Table 7 compares the bibliometric statistics for the different studies. Seven variables/figures of merit are presented for each study.

The number of different journals/sources cited is the total number of different journals and other sources referenced by the papers in the database. The average number of citations per cited journal is the ratio of number of citations to number of different journals and other sources cited. The average number of journals cited per author is the ratio of total journals and

Table 7
Cited journal bibliometrics — SCI

	Metric/Study						
	FUL	JACS	NES	HYD	HSF	AIR	RIA
Number of citations	263,844	85,000 +	140,662	82,395	26,768	45,744	37,000 +
Number of different journals/sources cited	13,294	6,725	28,740	21,523	9,498	21,518	
Average number of citations per cited journal	19.85	12.6	4.89	3.83	2.82	2.13	
Number of authors	12,837	6,535	12,453	7,869	2,483	6,619	2,975
Average number of journals cited per author	1.04	1.03	2.31	2.74	3.83	3.25	0.00
Number of authors cited	33,579	32,450	42,094	26,322	11,138	21,868	18,140
Average number of authors cited per journal cited	2.53	4.83	1.46	1.22	1.17	1.02	

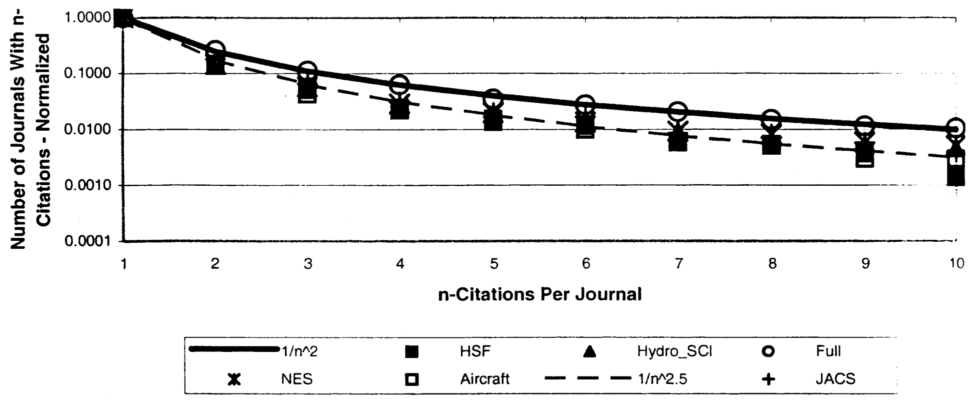

Fig. 6. Cited journal distribution.

other sources cited to total authors. The average number of authors cited per journal cited is the total number of authors cited to total number of journals and other sources cited.

Fig. 6 shows the distribution function of journal citation frequency for the NES, JACS, HSF, AIR, HYD, and FUL databases. The abscissa is the total number of citations *n* received by a given journal, and the ordinate is the number of journals that received *n* total citations. In each case, the distribution function has been normalized to the number of journals that received one citation. The data follow approximately a $(1/n^{2.5})$ distribution.

There are some important implications to be drawn from these journal distribution functions and tabulated metrics with regard to text mining, and these conclusions will be addressed briefly. During the development of the Bradford's law metric of Table 3, the number of journals in successive isopaper groups was computed. In addition, the number of journals in successive isocitation groups was computed for NES, HSF, and AIR, to ascertain whether a Bradford's law for citations was operable. The ratio between isocitation groups was less regular than the ratio between isopaper groups, and seemed to vary between 1.5 and 2 for the three studies.

However, a very important message can be extracted from this data, namely, that a potential substantial capability increase (for an organization involved in S&T) from a successful text mining program is possible. Consider the aircraft results as an example (while actual numbers may differ among disciplines, the conclusions drawn are probably applicable to any technical discipline). There are over 700 different journals that contain aircraft-related papers. The core (first) journal group (for the Bradford's law computation) contains three journals. There are about six journal groups that contain the total number of over 700 journals, the first five groups being isopaper, and the last somewhat less (essentially, the remainder). Thus, the core journal group contains about 18–20% of the total number of papers. For a technical manager or performer to be considered a true expert in all aspects of aircraft S&T, this individual would have to be familiar with the results from the aircraft papers in most of the more than 700 journals. One would suspect that bench-level aircraft experts, such as field managers, don't read more than the first two core groups on a regular basis, and this is probably a very generous estimate. Thus, these experts may be familiar with 30–40% of the relevant literature within the focused field; they would be far less familiar with complementary disparate-discipline literatures from which novel concepts could be extrapolated to benefit aircraft S&T.

In addition, one would suspect that program managers, at the federal level or in the field, who have broad responsibilities for aircraft S&T development (or of any technical discipline/multidiscipline development), do not have time to read much more than the main core group, if that much. Thus, they are probably familiar with 10% of the relevant literature, or less, and probably far less familiar with the disparate discipline literature.

One might argue that most of the good papers are contained in the first or second core journal groups, and all that is required for effective coverage is to read the journal papers in the first one or two groups. However, if citations are used as one measure of quality, the results show that citations are at least as widely spread out among the journals as actual publications. In fact, because the most highly cited journals are not necessarily those with the most publications, the spreading among journals may be broader than the results above suggest.

One might further argue that the previous paragraph aggregates citations over papers to draw journal citation conclusions; that the most highly cited papers are contained in the first or second core groups, and all that is required for effective coverage is to read the first one or two groups. Again, the data do not support this assertion.

The 10 most highly cited papers in the aircraft study were examined. It was found that none of these 10 were contained in the first core group journals, and only one of these 10 was contained in the second core group. One could argue that aircraft is a very broad field, and citations would more likely be aimed at papers in focused specialty journals in the lower groups than at the broader coverage journals in the higher groups. The 10 most highly cited papers in the hypersonics study were then examined; hypersonics constituted a more focused technical area. It was found that 2 of these 10 were contained in the first core group, and 4 of these 10 were contained in the first and second core groups. If one assumes that literature coverage should encompass the more fundamental highly cited papers/journals as well as the more applied perhaps less cited papers/journals, then it is important that all these types of journals be included in maintaining cognizance of the technical field of interest.

Obviously, citations are not the only measure of quality, and journal research papers accessed by the SCI are not the only source of useful literature information. Technical reports accessed by NTIS, technology papers/conference proceedings accessed by EC, program narratives accessed by RADIUS, and patents accessed by the patent database are other sources of useful information. The presence of these other quality measures besides citations, and the presence of other data sources, further expands the number of articles/documents to be read to maintain currency in the quality S&T, and results in even a smaller fraction of the literature accessed by any individual.

Thus, based on the results from these three different SCI bibliometric approaches (publications, aggregate citations, highly cited papers), one can conclude that (at least for the fields examined) confining one's reading to the first one or two core journal groups will exclude many high-quality documents. Text mining can make the user aware of these omitted papers in the target field, and, equally important, can make the user aware of papers in disparate disciplines that could impact the target field.

The argument could then be made that the literature is only one source of information. All the other useful sources are in fact accessed through proposals, workshops, site visits, and contacts. However, all these other sources are limiting as well. Consider workshops, for example. They contain a small fraction of the technical community; they tend to attract many repeat performers; they may or may not be representative of the community, depending on how they were selected and the size of the workshop. In most workshops, the focus is on a limited target discipline. Representatives from disparate disciplines who could impact the target discipline with innovative concepts are usually not present. The attendees tend to use the workshop, or expert panel, as a forum to sell their own approaches. Their willingness to share real cutting-edge approaches in an open forum (or any forum) is questionable. Workshops tend to be dominated by forceful personalities, adding further skewing to their results.

However, text mining could potentially support and add value to workshops and expert panels as well, and complement their strengths to provide a more comprehensive and balanced product. In conclusion, this brief discussion shows by example that text mining allows informed access to a wide body of literature not accessed presently. It demonstrates further that this nonaccessed literature has high-quality components and is important; therefore, its availability through text mining offers a potential new or enhanced capability to support program management.

3.2.2.2. Database tomography results. Pervasive themes — most frequently used Abstract phrases. High-frequency single-, double-, and triple-word phrases from the text of the SCI/EC databases whose technical content were deemed by topical experts to be significant were identified as the pervasive themes. Nontechnical content phrases, trivial phrases (automatically), etc., were eliminated from the analysis. In this particular exercise, each database was split into two parts, Titles and Abstracts, and the analysis was done on each part. Since the highest frequency phrases from the Title and Abstract databases tended to be very similar, only raw data outputs from the Abstract database were presented. This section of the results also attempted to construct a global picture of the total database from these high-frequency phrases.

Abstract phrase frequency perspective of NES. From a global perspective, the SCI database portrays the major applications of NES to be REMOTE SENSING of the SEA and EARTH SURFACE from SATELLITES using SAR and HIGH RESOLUTION RADIOMETRY to obtain TEMPERATURES and ICE information, RADIATION BUDGETS, and VEGETATION and CROP information, as well as NAVIGATION using GPS.

The EC database confirms the SCI thrusts listed, but in addition shows a major technological emphasis on SATELLITE COMMUNICATIONS and associated hardware. This supports the findings from a similar analysis of Keywords, and suggests that the issues in communication satellites tend to focus around technology rather than research.

Abstract phrase frequency perspective of HSF. The summary from the HSF study showed that, from a global perspective, the SCI database portrays the major focus of HSF research to be SUPERSONIC/HYPERSONIC MACH NUMBER flows over simple shapes (FLAT PLATE, LEADING EDGE) at ANGLES OF ATTACK containing BOW SHOCKS and OBLIQUE SHOCKS. The experimental focus is WIND TUNNEL TESTS with measurement emphasis on SURFACE HEAT TRANSFER and SURFACE PRESSURE DISTRIBUTIONS within the VISCOUS SHOCK LAYER, SHEAR LAYER, SUPERSONIC MIXING LAYER, and TURBULENT BOUNDARY LAYER; the analytical focus is NUMERICAL SIMULATION by COMPUTATIONAL FLUID DYNAMICS with FINITE ELEMENT and MONTE CARLO SIMULATION, using the COMPRESSIBLE NAVIER–STOKES EQUATIONS to model the VISCOUS SHOCK LAYER and near-body region, and using the EULER EQUATIONS to model the outer inviscid region. Conspicuous by their absence are exotic gas mixtures (helium, hydrogen, etc.) that would simulate other planetary atmospheres, exotic body shapes that would simulate novel vehicle designs, and real gas effects (dissociation, ionization, radiation, etc.) that would accompany very high Mach numbers characteristic of planetary entry speeds.

Phrase proximity analysis — relationships among themes. *Background.* To obtain the theme and subtheme relationships, a phrase proximity analysis is performed about each theme phrase. Typically, 40 to 60 multiword phrase themes are selected from a multiword phrase analysis of the type shown above. For each theme phrase, the frequencies of phrases within ±50 words of the theme phrase for every occurrence in the full text are computed. A phrase frequency dictionary is constructed that shows the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those subthemes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the most closely related phrases to the cluster theme (e.g., see Table 8 for part of a typical filtered cluster from the study).

For purposes of analysis, the cluster members in a given theme are segregated by their values of inclusion indices I_i and I_j. I_i is the ratio of C_{ij} to C_i, and is the inclusion index based

Table 8
HSF inclusion theme phrase “boundary layer” — Abstract database — sort by I_i

C_{ij}	C_i	I_i (C_{ij}/C_i)	I_j (C_{ij}/C_j)	E_{ij} (I_iI_j)	Cluster member
14	18	0.778	0.022	0.0174	expansion corner
7	9	0.778	0.011	0.0087	fully turbulent
6	8	0.750	0.010	0.0072	separation point
12	16	0.750	0.019	0.0144	linear stability theory
5	7	0.714	0.008	0.0057	bleed configurations

Code: C_{ij} is co-occurrence frequency, or number of times cluster member appears within ±50 words of cluster theme in total text; C_i is absolute occurrence frequency of cluster member; C_j is absolute occurrence frequency of cluster theme; I_i, the cluster member inclusion index, is ratio of C_{ij} to C_j; I_j, the cluster theme inclusion index, is ratio of C_{ij} to C_j, and E_{ij}, the equivalence index, is product of inclusion index based on cluster member I_i (C_{ij}/C_i) and inclusion index based on cluster theme I_j (C_{ij}/C_j). E_{ij} bears some similarity to the mutual information method from computational linguistics, that compares the probability of two words occurring together with the probability of the words occurring separately.

on the cluster member. I_j is the ratio of C_{ij} to C_j, and is the Inclusion Index based on the theme phrase. I_i and I_j are categorized as either high or low. The dividing points between high and low I_i and I_j are the middle of the "knee" of the distribution functions of numbers of cluster members versus values of I_i and I_j. All cluster members with I_i greater than or equal to approximately 0.5 were defined as having high I_i. All cluster members with I_j greater than or equal to 0.1 were defined as having high I_j.

Analysis. The full-text database was split into two databases. One was the Abstract narrative database (referred to as ABSTRACT in the phrase proximity analysis below), and phrase proximity analysis of this database yielded mainly topical theme relationships. The other database (referred to as BLOCK below) consisted of records (one for each published paper) containing four fields: author(s), title, journal name, author(s) institutional address(es). Phrase proximity analysis of this database yielded not only topical theme relationships from the proximal title words, but also relationships among technical themes and authors, journals, and institutions.

Because of space limitations in the reported studies, only one theme was chosen to illustrate the phrase proximity analysis for each study. The theme selected tended to be high frequency in both the Abstracts and Titles, and was always a central theme of the study. The theme was analyzed for the BLOCK and ABSTRACT database components. Further, for each of these database components, the cluster theme was analyzed from the two perspectives of high I_i low I_j and low I_i high I_j. The phrase proximity analysis process consisted of providing the experts with two lists of cluster members, one sorted by I_i and the other by I_j. By visual examination of these lists, the experts constructed categories of related items, and these relationships were reported in their respective studies.

The types and numbers of categories possible are limited by the perceptual capabilities of the experts, and could vary substantially among experts. This issue of category definition is a good example of the advantages and challenges of the full-text procedure reported in this paper relative to the key word or index word approaches used in most co-word-based analyses. Full text provides many more degrees of freedom relative to index words, and therefore many more possibilities of different relational categories. However, analysts with the ability to perceive large numbers of relationships, especially the highest value relationships, are required to obtain maximal benefit from the increased degrees of freedom.

For example, the expert used in the HSF study had experience in very high speed hypersonic flow (typically Mach Number > 20) from the space program. Consequently, the analytical perspective and especially the perceived literature gaps (no exotic gas mixtures characteristic of planetary atmospheres, no high-temperature dissociative and radiative phenomena) were reflective of high-speed phenomena, and might not have been easily identified by an expert with the lower speed lower temperature military hypersonic flow experience (typically Mach Number $\sim 6-8$, in terrestrial atmospheres). Conversely, a military hypersonics expert might have readily perceived gaps not immediately identifiable by the space hypersonics expert. Thus, a fully credible analysis requires not only domain knowledge by the analyst(s), but probably domain knowledge representing a diversity of backgrounds in the target literature. More generally, because the iterative information retrieval process allows documents from disparate disciplines to be accessed and analyzed, experts knowledgeable in those disparate disciplines are required for a fully credible analysis as well.

Taxonomies. The use of full text by DT, compared to the use of index or key words by classical co-word analysis, allows many different types of taxonomies, or classifications into component categories, to be generated. Such categorizations, analogous to the independent axes of a mathematical coordinate system, allow the underlying structure of a field to be portrayed more clearly, leading to more focused analytical and management analyses. Two separate taxonomies will be discussed here.

The first taxonomy derives from the phrase frequencies. The authors examined the phrase frequency outputs, then arbitrarily grouped the high-frequency phrases into different, relatively independent, categories for which all remaining terms would be accounted. In the NES study, one taxonomy was developed for the SCI phrase frequencies, and another taxonomy for the EC phrase frequencies. Table 9 presents the results for the NES study.

The second taxonomy derives from the phrase frequency and proximity analyses, and will be presented here. From the phrase frequency analysis, about 60 high-frequency technical phrases were identified by the domain expert as pervasive themes. A proximity analysis was done for each of these high-frequency phrases. A phrase frequency dictionary, or cluster, was generated for each phrase. This cluster contained those phrases that were in close physical proximity to the pervasive theme throughout the text.

The degree of overlap among clusters was computed by counting the number of shared phrases, in the following manner. The cluster sizes were normalized, and each normalized cluster pair that shared more than a threshold number of common phrases was viewed as overlapping. These overlapping clusters were viewed as links in a chain, with the different

Table 9
Space taxonomy — SCI— phrase frequency-based

* Space platform (e.g., satellite, spacecraft)
* Satellite function (e.g., mapping, navigation)
* Satellite type (e.g., Geosat, Landsat)
* Measuring instrument (e.g., radiometer, microwave imager)
* Region examined (e.g., sea, boundary layer)
* Location examined (e.g., North Atlantic, Southern Hemisphere)
* Variable measured (e.g., temperature, soil moisture)
* Variable derived (e.g., radiation budget, general circulation)
* Analytical tool (e.g., data processing, mathematical models)
* Products (e.g., time series, sea ice maps)
* Space environment (e.g., solar wind, magnetic field)

Space taxonomy — e.g., phrase frequency based

Same as 1a, but add:
* Satellite configuration (geostationary satellites, tethered satellite system)
* Satellite state (attitude determination, high elevation angle)
* Satellite subsystems (solar cells, attitude control system)

Table 10
Formation of megathemes

Theme name/no. of overlaps	60	50	40	35	30	25	20	15
SHOCK TUNNEL	A60	A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HIGH TEMPERATURE	A60	A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HYPERSONIC FLIGHT		A50	A40	A35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
BOUNDARY LAYERS			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
OBLIQUE SHOCK			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
REYNOLDS NUMBER			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
TRAILING EDGE			B40	B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC AIRCRAFT				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
PRESSURE GRADIENT				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC SPEEDS				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
TURBULENT BOUNDARY LAYER				B35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
VISCOUS SHOCK LAYER				C35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
PERFECT GAS				C35	ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SURFACE PRESSURE					ABC30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC NOZZLE EXIT				D35	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
NUMERICAL SIMULATION				D36	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
WIND TUNNELS				D37	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
FLOW CONDITIONS				D38	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
GAS FLOW				D39	D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
HYPERSONIC VEHICLES					D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC COMBUSTION					D30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
GROWTH RATE				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SHEAR LAYERS				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC JETS				E35	E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
CONVECTIVE MACH					E30	ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
LARGE SCALE						ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
SUPERSONIC FLOWS						ABCDE25	ABCDEFGH20	ABCDEFGHIJK15
DIRECT SIMULATION		F50	F40	F35	F30	F25	ABCDEFGH20	ABCDEFGHIJK15
MONTE CARLO		F50	F40	F35	F30	F25	ABCDEFGH20	ABCDEFGHIJK15
FLOW VISUALIZATION						G25	ABCDEFGH20	ABCDEFGHIJK15
PRESSURE MEASUREMENTS						G25	ABCDEFGH20	ABCDEFGHIJK15
ANGLE OF ATTACK						H25	ABCDEFGH20	ABCDEFGHIJK15
BLUNT BODY						H25	ABCDEFGH20	ABCDEFGHIJK15
BOW SHOCK							ABCDEFGH20	ABCDEFGHIJK15
MIXING LAYER			I40	I35	I30	I25	I20	ABCDEFGHIJK15
SUPERSONIC MIXING			I40	I35	I30	I25	I20	ABCDEFGHIJK15
CHEMICAL REACTION			J40	J35	J30	J25	JK20	ABCDEFGHIJK15
HEAT RELEASE			J40	J35	J30	J25	JK20	ABCDEFGHIJK15
MASS FLOW				K35	K30	K25	JK20	ABCDEFGHIJK15
NUMERICAL SOLUTIONS				K35	K30	K25	JK20	ABCDEFGHIJK15
COMPUTATIONAL FLUID DYNAMICS								ABCDEFGHIJK15
FREESTREAM MACH								ABCDEFGHIJK15

(*continued on next page*)

Table 10 (*continued*)

Theme name/no. of overlaps	60	50	40	35	30	25	20	15
PRESSURE DISTRIBUTION								ABCDEFGHIJK15
STATIC PRESSURE								ABCDEFGHIJK15
TOTAL PRESSURE								ABCDEFGHIJK15
EXPERIMENTAL STUDY							L20	L15
HEAT FLUX							L20	L15
EULER EQUATIONS							M20	M15
FINITE VOLUME							M20	M15
BOUNDARY CONDITIONS								
FINITE ELEMENT								
TURBULENCE MODEL								
TURBULENT FLOW								

chains being relatively independent. Each chain was then defined as a megacluster, or category of the larger taxonomy.

The threshold level for cluster overlap was determined by varying the number of common phrases parametrically, and observing the patterns of aggregation of the clusters. The parametric variation of these patterns of aggregation is shown in Table 10. The leftmost column is the cluster, or theme, name. The numeric column headings (e.g., 60, 50) represent the number of overlaps, or common phrases, among normalized clusters. The alphanumeric matrix entries are the members of the different chains. The alphabetical characters of the matrix entry identify the chain, and the numerical characters of the matrix entry identify the minimum number of overlaps.

For example, in the first column (60), there is one chain (A). It has two links/themes/clusters (SHOCK TUNNEL, HIGH TEMPERATURE), and the themes/clusters (normalized to 100 cluster component phrase members) have at least 60 phrase members in common. As another example, in the column with heading 30, there are seven chains (ABC, D, E, F, I, J, K), and every link/theme/cluster in each chain has at least 30 phrase members in common with at least one other link in the chain. The largest chain (ABC) is an amalgamation of three component chains (A, B, C) that were formed previously. One value of following the chain formations parametrically is that the strong link associations evidenced by the component chains A, B, C can be readily identified.

Obviously, many taxonomies are possible with this approach, depending on the final threshold value selected. If the threshold value is set too high (e.g., > 60), there will be a large number of independent categories, and the taxonomy will be unwieldy for any practical use. If the threshold value is set too low (e.g., < 15), all the categories tend to merge, and the taxonomy does not provide much information. The results from Table 10, modified by the judgement and experience of the authors, are used to form the useful taxonomy shown in Table 11. The phrases preceeded by an asterisk (*) are the megacluster themes, and the phrases preceded by a hyphen (-) are their component cluster themes.

This taxonomy reflects very accurately the thrust areas of hypersonic and supersonic flow over aerodynamic bodies. The HYPERSONIC EXPERIMENTS measurements emphasize high pressure and temperature conditions, focusing on heat flux data and flow visualization techniques, and making increased use of shock tunnels relative to free-flight

Table 11
HSF taxonomy — megaclusters

* HYPERSONIC EXPERIMENTS	* COMPUTATIONAL FLUID DYNAMICS
–SHOCK TUNNEL	COMPUTATIONAL FLUID DYNAMICS
–HYPERSONIC FLIGHT	–DYNAMIC PRESSURE
–HIGH TEMPERATURE	–DIRECT SIMULATION
–FLOW VISUALIZATION	–MONTE CARLO
–PRESSURE MEASUREMENTS	–EULER EQUATIONS
–STATIC PRESSURE	–FINITE VOLUME
–EXPERIMENTAL STUDY	–BOUNDARY CONDITIONS
–HEAT FLUX	–FINITE ELEMENT
* BOUNDARY LAYER	* SHOCK LAYER
–BOUNDARY LAYERS	–SHOCK LAYER
–SURFACE PRESSURE	–VISCOUS SHOCK LAYER
–PRESSURE GRADIENT	–PERFECT GAS
–OBLIQUE SHOCK	–SUBSONIC AND SUPERSONIC
–REYNOLDS NUMBER	–SUPERSONIC AIRCRAFT
–TRAILING EDGE	–SUPERSONIC SPEEDS
–SUPERSONIC FLOWS	
* SHEAR AND MIXING LAYER	* NOZZLE FLOW
–SHEAR LAYERS	–NOZZLE EXIT
–TOTAL PRESSURE	–SUPERSONIC NOZZLE
–SUPERSONIC JETS	–FLOW CONDITIONS
–GROWTH RATE	–GAS FLOW
–CONVECTIVE MACH	–PRESSURE DISTRIBUTION
–FREESTREAM MACH	–NUMERICAL SIMULATION
–MIXING LAYER	–HYPERSONIC VEHICLES
–SUPERSONIC MIXING	–SUPERSONIC COMBUSTION
	–WIND TUNNELS
* TURBULENT FLOW	* ASYMMETRICAL FLOW
–TURBULENT FLOW	–ANGLE OF ATTACK
–TURBULENT BOUNDARY LAYER	–BLUNT BODY
–TURBULENCE MODEL	–BOW SHOCK
–LARGE SCALE	
* INTERNAL ENERGY PRODUCTION	
–HEAT RELEASE	
–CHEMICAL REACTION	
–MASS FLOW	
–NUMERICAL SOLUTIONS	

experiments. The COMPUTATIONAL FLUID DYNAMICS approaches, that are assuming a greater portion of HSF research, encompass finite volume and finite element techniques and Monte Carlo simulations as well. The three postshock regions of SHOCK LAYER, SHEAR AND MIXING LAYER, and BOUNDARY LAYER, each constitute emphasis areas with unique subthrust areas. As the threshold conditions for overlapping phrases were further reduced, these three areas shortly merged into one, paralleling their intrinsic physical connectivity. TURBULENT FLOW with its high mixing and heat flux rates is of primary interest, while ASYMMETRICAL FLOW with its potentially higher lift coefficients assumes increasing importance for improving hypersonic vehicle performance. NOZZLE FLOW has a dual importance: the study and control of high-speed flow from actual aircraft and missile nozzle exits to maximize thrust and minimize fuel consumption, and similar studies of laboratory nozzle flows to understand the flowfield fluid dynamics and improve the nozzle as a high-speed flow source. Finally, INTERNAL ENERGY PRODUCTION is important for studying high-speed combustion, as well as the reaction and dissociation chemistry of high-speed gases.

The most recent DT studies add a co-occurrence matrix-based clustering approach, and the results of both clustering approaches are considered when structuring the final bottom-up taxonomy.

3.2.2.3. Applications of DT to technology forecasting. Two of the credible major approaches to technology forecasting are 1) the use of expert workshops for group dynamic approaches and 2) the use of experts for literature-based innovation and discovery. For the latter approach, some of the most revolutionary discoveries from TM/information retrieval have occurred in the medical field, resulting from linking disparate literatures to the primary target literature [22–27].

However, each of these two major approaches has deficiencies when conducted in isolation. As stated previously, workshops typically access a very small fraction of the relevant technical community, can be skewed by group dynamics, and contain little incentive for participants to share innovative concepts. The literature-based approaches include documented material only, and the documentation may reflect work performed a year or more previously.

A 1999 paper by the first author recommended combining these two approaches to eliminate their individual weaknesses and exploit their synergies [6]. In this tandem approach, literature-based innovation and discovery using DT would be performed initially. Based on the results of this initial step, a workshop would then be assembled using the linked disciplines from the literature-based study for the structure, and the experts identified from the literature-based study as the participants. The 1999 paper provides an example of the tandem process for Autonomous Flying Systems, although the literature-based component did not operationalize all the concepts listed in the theoretical section of the paper. Appendix 1A of the unabridged Web version of the 1999 paper [6] contains a preliminary proposal that resulted from the workshop.

4. Conclusions

In all the DT/bibliometrics studies described in the present document (NES, JACS, HSF, FUL, AIR, HYD, RIA), there was a concentration of output in the top authors, journals, institutions, and countries. The top authors had an order of magnitude larger number of listings than the average, as did the top journals and top institutions. While there is a wide range among disciplines in the number of papers retrieved, the average number of author listings per paper decreases steadily proceeding from the most basic fields to the most applied. The three most fundamental fields examined (FUL, JACS, NES) tend to be experiment-dominated, with much less effort devoted to computational modeling. In many cases, these experiments require expensive equipment and large teams of researchers because of their complexity, and this is reflected in the large numbers of authors on the papers produced. Conversely, the three most applied fields examined (AIR, HYDRO, HSF) focus on substituting computational modeling (e.g., Computational Fluid Dynamics) for experiments because of the prohibitive costs of wind/water tunnel tests and flight/sea tests. These computer-based studies can be performed by one or two individuals at their desks, and the resulting papers tend to be authored by these one or two persons.

The Bradford's law results mean that in the fundamental fields there are more core discipline-oriented journals in which researchers would be motivated to publish relative to those in the applied fields. This conclusion is substantiated further by a more detailed examination of the numbers presented in the FUL and HSF examples, where it is shown that there is more depth in the FUL core than in the HSF core. The journals in which researchers are motivated to publish penetrates much deeper into the total FUL journal body relative to the total HSF body. In other words, there are more good fundamental research journals available for publication in FUL than there are in HSF. The dominance of a handful of countries is clearly evident in all the studies, especially the dominance of the US. In many cases, the US is almost an order of magnitude more prolific than its nearest competitor in terms of absolute numbers of papers produced, and in some cases is as prolific as its nearest major competitors combined. In fact, the total number of country listings summed across all seven studies is 21,307 for the US, and 15,515 for all other countries combined.

Generically, the western democracies (UK, Germany, France, Canada) tend to be the most prolific on the basis of absolute numbers of country listings produced; normalizations to population or GNP were not performed. In addition, Japan is in the first JACS, FUL, HSF, NES, and HYD tiers, and second AIR tier; Russia is in the first HYD and HSF tiers, and second FUL, NES, and AIR tiers; the People's Republic of China is in the second FUL, JACS, NES, and RIA tiers; and India is in the third FUL, NES, HYD, and AIR tiers. Most of the cited authors were cited once, and perhaps an order of magnitude less were cited twice. A relatively few percent received large numbers of citations. These observations held for cited journals and papers as well.

The most cited authors, while prolific, were usually not the most prolific authors, and vice versa. This relation between most prolific and most cited authors was the norm for most studies. One notable exception occurred in the FUL study, where Kroto was the most highly cited author, and the second most prolific author. This may be an anomaly of a young dynamic research discipline, where the discipline founders and pioneers are still very active.

However, in the nominal case, part of the difference between most prolific and cited authors may be due to the time lag between the highly cited authors' productivity at the time their highly cited papers were written and their productivity today, as well as the phase in their career of the prolific authors. Another partial explanation may be the intrinsic nature of the papers; the large numbers of papers produced may reflect more applied papers, which lend themselves more to shorter-term production-line type output. Stated differently, the time and effort required to produce a fundamental seminal highly cited paper probably do not allow overly high volumes of papers to be produced.

There is a definite trend in average number of citations per cited journal, decreasing sharply from the fundamental fields to the applied fields. One needs to make a distinction here between the journals in which authors publish and the journals that they cite. As the Bradford's law results showed, there were more credible journals in which the researchers could publish in the fundamental fields compared to the applied fields. However, in the case of citations, there is a wider variety of journals that the researchers in the applied fields will access (both basic and applied journals) than the researchers in the fundamental fields will access (basic). Therefore, it would be expected that the researchers in fundamental fields (who cite more frequently as shown above, and who cite a narrower group of journals than their applied counterparts) would have a substantially higher value of this 'citations per cited journal' metric than their applied counterparts.

This difference in breadth of journals cited between the researchers in basic and applied fields, discussed in the previous paragraph, is substantiated and displayed most dramatically by the average number of journals cited per author metric. The metric increases sharply from the fundamental fields to the applied fields.

The metric 'average number of authors cited per journal cited' trends downward as the fields become more applied. The researchers in the more applied fields tend to cite from a wider variety of journals than their counterparts in the more fundamental fields, and the denominator of this metric therefore increases as the fields become more applied.

Publication and citation frequency distribution results were presented for authors, papers, journals, and organizations for all the fields studies. With the exception of the author distributions, most of the distributions transcended the different topical fields, appearing to be topic-independent, and differed modestly ($\sim 1/n^2$ vs. $\sim 1/n^3$) for the type of distribution function (author, journal, etc.).

Two types of computational linguistics tools were used for DT results, phrase frequency analyses and phrase proximity analyses. These tools were applied to the paper Abstracts, Keywords, and infrastructure (authors/titles/institutions/countries) databases. The Keyword and Abstract phrase frequencies are essentially quantity measures. They lend themselves to 'binning,' and addressing adequacies and deficiencies in levels of S&T activity in the different technical subcategories. They do not contain relational information, and therefore offer little insight into S&T linkages. The phrase proximity results are essentially relational measures, although some of the proximity results imply levels of effort that support specific S&T areas. The phrase proximity results mainly offer insight into S&T linkages, and have the potential to help identify innovative concepts from disparate disciplines [6]. The phrase proximity results also offer insight into linkages between S&T categories and supporting infrastructures (performers, institutions, journals, etc.). Thus, the Keyword and Abstract phrase frequency analyses were addressed to adequacy of effort, and the phrase proximity analyses were addressed to intra-S&T/inter-

S&T–infrastructure relationships primarily and supporting levels of effort secondarily. This paper has presented a number of advantages of using DT and bibliometrics for deriving technical intelligence from the published literature. Large amounts of data can be accessed and analyzed, well beyond what a finite group of expert panels could analyze in a reasonable time period. Preconceived biases tend to be minimized in generating roadmaps. Compared to standard co-word analysis, DT uses full text, not index words, and can make maximum use of the rich semantic relationships among the words. It also has the potential of identifying low occurrence frequency but highly theme related phrases, which are 'needles-in-a-haystack' a capability unavailable to any of the other co-occurrence methods. Combined with bibliometric analyses, DT identifies not only the technical themes and their relationships, but relationships among technical themes and authors, journals, institutions, and countries. Unlike other roadmap development processes, DT generates the roadmap in a 'bottom-up' approach. Unlike other taxonomy development processes, DT can generate many different types of taxonomies (because it uses full text, not key words) in a 'bottom-up' process, not the typical arbitrary 'top-down' taxonomy specification process. Compared to co-citation analysis, DT can use any type of text, not only published literature, and it is a more direct approach to identifying themes and their relationships.

The maximum potential of the DT and bibliometrics combination can be achieved when these two approaches are combined with expert analysis of selected portions of the database. If a manager, for example, wants to identify high-quality research thrusts as well as science and technology gaps in specific technical areas, then an initial DT and bibliometrics analysis will provide a contextual view of work in the larger technical area; i.e., a strategic roadmap. With this strategic map in hand, the manager can then commission detailed analysis of selected Abstracts to assess the quality of work done as well as identify work that needs to be one (promising opportunities). Adding a workshop to the DT-bibliometrics combination provides a unique approach to innovation and discovery, and potentially to technology forecasting.

References

[1] R.N. Kostoff, Database tomography: multidisciplinary research thrusts from co-word analysis, Proceedings: Portland International Conference on Management of Engineering and Technology, 1991.

[2] R.N. Kostoff, Research impact assessment, Proceedings: Third International Conference on Management of Technology, Miami, FL, 1992, (Larger text available from author).

[3] R.N. Kostoff, Database tomography for technical intelligence, Compet. Intell. Rev. 4 (1) (1993).

[4] R.N. Kostoff, Database tomography: Origins and applications, Compet. Intell. Rev. 5 (1) (1994) (Special issue on technology).

[5] Kostoff, R.N. et al., System and Method for Database Tomography, US Patent Number 5440481, 1995.

[6] R.N. Kostoff, Science and Technology Innovation, Technovation 19 (1999) (October; Also, R.N. Kostoff, Science and Technology Innovation. www.scicentral.com).

[7] R.N. Kostoff, Science and Technology Roadmaps, http://www.dtic.mil/dtic/kostoff/index.html. Also, R.N. Kostoff, R.R. Scholler, Science and Technology Roadmaps, IEEE Trans. Eng. Manage. 48 (2) (2001) 132–143 (May).

[8] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature, Inf. Process. Manage. 34 (1) (1998).

[9] R.N. Kostoff, Word Frequency Analysis of Text Databases. ONR Memorandum 5000 Ser 10P4/1443. April 12, 1991.

[10] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Database tomography for information retrieval, J. Inf. Sci. 23 (4) (1997).

[11] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography, JASIS, 1999 (15 April).

[12] O. Zamir, Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. PhD Thesis, University of Washington, 1999.

[13] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, R. Pellenbarg, Database tomography for technical intelligence: Comparative roadmaps of the research impact assessment literature and the Journal of the American Chemical Society, Scientometrics 40 (1) (1997).

[14] R.N. Kostoff, T. Braun, A. Schubert, D.R. Toothman, J. Humenik, Fullerene roadmaps using bibliometrics and database tomography, J. Chem. Inf. Comput. Sci. (2000) (Jan–Feb).

[15] R.N. Kostoff, K.A. Green, D.R. Toothman, J. Humenik, Database tomography applied to an aircraft science and technology investment strategy, J. Aircr. 37 (4) (2000) (July–August).

[16] R.N. Kostoff, D. Coder, S. Wells, D.R. Toothman, Surface hydrodynamics roadmaps using bibliometrics and database tomography, J. Ship Res. (unpublished).

[17] A.J. Lotka, The frequency distribution of scientific productivity, J. Wash. Acad. Sci. 16 (1926).

[18] S.C. Bradford, Sources of information on specific subjects, Engineering 137 (1934).

[19] R.N. Kostoff, Citation analysis cross-field normalization: A new paradigm, Scientometrics 39 (3) (1997).

[20] R.N. Kostoff, Use and misuse of citation analysis in research evaluation, Scientometrics 43 (1) (1998).

[21] M. MacRoberts, B. MacRoberts, Problems of citation analysis, Scientometrics 36 (3) (1996) (July–August).

[22] N.R. Smalheiser, D.R. Swanson, Assessing a gap in the biomedical literature — magnesium-deficiency and neurologic disease, Neurosci. Res. Commun. 15 (1) (1994).

[23] N.R. Smalheiser, D.R. Swanson, Calcium-independent phospholipase A (2) and schizophrenia, Arch. Gen. Psychiatry 55 (8) (1998).

[24] N.R. Smalheiser, D.R. Swanson, Using ARROWSMITH: A computer assisted approach to formulating and assessing scientific hypotheses, Comput. Methods Programs Biomed. 57 (3) (1998).

[25] D.R. Swanson, Fish oil, Raynauds syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1) (1986).

[26] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: A stimulus to scientific discovery, Artif. Intell. 91 (2) (1997).

[27] D.R. Swanson, Computer-assisted search for novel implicit connections in text databases, Abstr. Pap.-Am. Chem. Soc. 217 (1999).