Available online at www.sciencedirect.com
SCIENCE DIRECT®
Energy 30 (2005) 709–730
ENERGY
www.elsevier.com/locate/energy

Power source roadmaps using bibliometrics and database tomography

R.N. Kostoff[a,*], R. Tshiteya[b], K.M. Pfeil[a], J.A. Humenik[c], G. Karypis[d]

[a]*Office of Naval Research, Code 35, 800 North Quincy Street, Arlington, VA 22217, USA*
[b]*DDL OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600, Mclean, VA 22102, USA*
[c]*Noesis, Inc., 10440 Balls Ford Road, Suite 250, Manassas, VA 20109, USA*
[d]*Computer Science and Engineering Department, 4-192 EE/CS Building, 200 Union Street, S.E., University of Minnesota, Minneapolis, MN 55455, USA*

Received 15 August 2002

Abstract

Database Tomography (DT) is a textual database analysis system consisting of two major components: (1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment (2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a Power Sources database derived from the Science Citation Index. Phrase frequency analysis by the technical domain experts provided the pervasive technical themes of the Power Sources database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the Power Sources literature supplemented the DT results with author/journal/institution/country publication and citation data.
Published by Elsevier Ltd.

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 commensurate 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 direct human intelligence gathering cannot be substituted, many techniques have become available that can support and complement it. In particular, techniques that identify, select,

* Corresponding author. Tel.: +1-703-696-4198; fax: +1-703-696-4274.
E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).

0360-5442/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/j.energy.2004.04.058

gather, cull, and interpret large amounts of technological information semi-automatically can expand greatly the capabilities of human beings in performing technical intelligence.

One such technique being developed by different researchers for these, and many other, applications is text mining (the extraction of useful information from large volumes of text). Its component capabilities of computational linguistics and bibliometrics were the main analytical techniques used for the present study, and these capabilities can be summarized as follows.

Science and technology (S&T) computational linguistics [1–4] is a process for extracting useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from frequently occurring technical phrases. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Computational linguistics can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [5–7];
- Potential discovery and innovation based on merging common linkages among very disparate literatures [8–11];
- Uncovering unexpected asymmetries from the technical literature [12,13];
- Estimating global levels of effort in S&T sub-disciplines [14–16];
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their impact factors [15,17];
- Tracking myriad research impacts across time and applications areas [18,19].

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the database using computational linguistics and bibliometrics, and integrates the processed information.

Evaluative bibliometrics [20–22] uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that (1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, (2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and (3) the citations from papers to papers, from patents to patents and from patents to papers provide indicators of intellectual linkages between the organizations which are producing the patents and papers, and knowledge linkage between their subject areas [23]. Evaluative bibliometrics can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain;
- Identify experts for innovation-enhancing technical workshops and review panels;
- Develop site visitation strategies for assessment of prolific organizations globally;
- Identify impacts (literature citations) of individuals, research units, organizations, and countries.

One computational linguistics approach developed by the first author's group is Database Tomography (DT) [24], a system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and phrase proximities from the textual databases, 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 multi-word technical phrases) provides the pervasive technical themes of a database, and the phrase

proximity (physical closeness of the multi-word 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 describes use of the DT process, supplemented by literature bibliometric analyses, to derive technical intelligence from the published literature of Power Sources S&T.

Power Sources, as defined by the authors for this study, consists of systems and processes for generating and converting power, and storing energy. It is defined operationally by a query with two components: (1) a *phrase-based query*, obtained by the iterative technique referenced in the next paragraph; and (2) a *journal-title-based query*, obtained by identifying non-technology-specific power source journals from the Science Citation Index (SCI) journal listing under Energy and Fuels whose articles were deemed highly relevant to the Power Sources topic.

To execute the study reported in this paper, a database of relevant Power Sources articles is generated using the iterative search approach of Simulated Nucleation [25]. Then, the database is analyzed to produce the following characteristics and key features of the Power Sources field: recent prolific Power Sources authors; journals that contain numerous Power Sources papers; institutions that produce numerous Power Sources papers; keywords most frequently specified by the Power Sources authors; authors, papers and journals cited most frequently; pervasive technical themes of Power Sources; and relationships among the pervasive themes and sub-themes.

2. Background

2.1. Overview

The information sciences background for the approach used in this paper is presented in Ref. [26]. This reference shows the unique features of the computer and co-word-based DT process relative to other roadmap techniques. It 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 evolution of DT from its co-word roots to its present form.

The DT method in its entirety requires generically three distinct steps. The first step is identification of the main themes of the text being analyzed. The second 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. Time evolution of themes has not yet been studied.

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles [27], the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases consist of journal article abstracts and associated bibliometric information (authors, journals, addresses, etc.), the final results have also included relationships among the technical themes and authors, journals, institutions, etc. [26,28–32].

These most recent DT/bibliometrics studies were conducted in the technical fields of: (1) Near-earth space (NES) [28]; (2) Hypersonic and supersonic flow over aerodynamic bodies (HSF) [26];

Table 1
Dt studies of topical fields

Topical area	Number of sci articles	Years covered
Near-earth space (NES)	5480	1993–mid 1996
Hypersonics (HSF)	1284	1993–mid 1996
Chemistry (JACS)	2150	1994
Fullerenes (FUL)	10,515	1991–mid 1998
Aircraft (AIR)	4346	1991–mid 1998
Hydrodynamics (HYD)	4608	1991–mid 1998
Electrochem Power (ECHEM)	6985	1991–mid 2001
Research Assessment (RIA)	2300	1991–beg 1995
Electric Power Sources (EPS)	20,835	1991–late 2000

(3) Chemistry (JACS) [29] as represented by the Journal of the American Chemical Society; (4) Fullerenes (FUL) [30]; (5) Aircraft (AIR) [31]; (6) Hydrodynamic flow over surfaces (HYD); (7) Electrochemical Power Sources (ECHEM) [32]; and (8) the non-technical field of research impact assessment (RIA) [29]. Overall parameters of these studies from the SCI database results and the current EPS study are shown in Table 1.

2.2. Unique study features

The study reported in the present paper is in the latter (journal article abstract) category. It differs from the previous published papers in this category [26,28–32] in four respects. First, the topical domain (power sources) is completely different. Second, a more rigorous technical theme clustering approach is used. Third, the phrase-based query approach has been supplemented by the journal-title-based query approach. Fourth, since estimation of relative global levels of emphasis in power sources was desired, a generic power sources query was used in both the phrase-based and journal-title-based queries (e.g. ELECTRICITY PRODUCTION), rather than using power source-specific terms (e.g. FUEL CELL). A companion study [32] examines the more specific sub-area of ELECTROCHEMICAL POWER SOURCES using specific terms rather than the generic terms.

3. Database generation

The key step in the power source literature analysis is the generation of the database. There are three key elements to database generation: the overall objectives, the approach selected, and the database used. Each of these elements is described.

3.1. Overall study objectives

The main objective was to identify global S&T that had both direct and indirect relations to Power Sources. One sub-objective was to estimate the overall level of global effort in Power Sources S&T, as reflected by the emphases in the published literature. Another sub-objective was to determine whether any radically new power sources were under development.

It was believed that if known specific technical terms were used for the query, there would be three negative impacts relative to the objectives above. First, the query would be biased toward the specific technologies reflected in the query, and the records retrieved would reflect this bias. The relative global efforts devoted toward each technology would have little credibility. Second, the use of specific technical terms in the query would identify advances made in existing technologies, but might not access radically new technologies. Third, the query size would have been unmanageable, and unusable in present search engines. An unpublished study of controlled fusion energy resulted in a query of hundreds of terms after only the first iteration. The companion study to the present study, on the topic of electrochemical power sources, generated a query with hundreds of terms. Summing this experience over all the source, converter, and storage technologies contained within the umbrella of power sources S&T would have generated thousands of query terms.

Thus, it was decided to use generic energy or power-related terms for the query, relatively independent of any specific power supply, conversion, or storage system (e.g. ELECTRICITY PRODUCTION vs LIGHT-WATER REACTOR). This approach would retrieve documents that described technologies specifically related to power production, conversion, and storage. The journal-based approach was added to retrieve documents related to power production, but where the author may not have used specific terminology relating the technology to power production in the write-up. The concept was to identify power source journals that were generic, not source specific, and add their articles to the phrase-based query database.

However, even with the use of both approaches, one class of articles will not be retrieved. These are power source-related articles that do not contain the generic terms relating them to power sources, nor are published in a journal with a dedicated power source emphasis. Thus, an article on a new scientific phenomenon potentially related to power sources that was published in, for example, Science or Nature would not appear in this retrieval. To retrieve such articles, a detailed technology-specific query is required, such as the type developed in the companion study on Electrochemical Power Sources [32].

3.2. Databases and approach

The SCI [33] was the database used for the present study. The approach used for query development was the DT-based iterative relevance feedback concept [25].

The database consists of selected journal records (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for power source articles. At the time the present paper was written, the Web version of the SCI accessed about 5600 journals (mainly in physical, engineering, and life sciences basic research).

The SCI database selected represents a fraction of the available Power Source (mainly research) literature, that in turn represents a fraction of the Power Source S&T actually performed globally [34]. It does not include the large body of classified literature, or company proprietary technology literature. It does not include technical reports or books or patents on Power Sources. It covers a finite slice of time (1991 to late 2000). The database used represents the bulk of the peer-reviewed high quality Power Source science and technology documented.

To extract the relevant articles from the SCI, the phrase-based query and the journal-title-based query were used, and two disjoint databases were generated. For application of the phrase-based query, the Title, Keyword, and Abstract fields were searched using phrases relevant to power sources. The resultant

Abstracts were culled to those relevant to power sources. The search was performed with the aid of two powerful DT tools (multi-word phrase frequency analysis and phrase proximity analysis) using the process of Simulated Nucleation [25].

An initial query of generic power source-related terms produced two groups of papers: one group was judged by domain experts to be relevant to the subject matter, the other was judged to be non-relevant. Gradations of relevancy or non-relevancy were not considered. An initial database of Titles, Keywords, and Abstracts was created for each of the two groups of papers. Phrase frequency and proximity analyses were 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, were then added to the query to expand the papers retrieved. Similar phrases characteristic of the non-relevant group were effectively subtracted from the query to contract the papers retrieved. The process was repeated on the new database of Titles, Keywords, and Abstracts obtained from the search. A few more iterations were performed until the number of records retrieved stabilized (convergence). The final approximately 400 term phrase-based query used for the Power Source study is shown in Ref. [35].

The query consists of two components. The first component consists of phrases and phrase combinations designed to access mainly relevant records (e.g. bio-mass energy, power conversion, energy storage). The second component consists of phrases and phrase combinations designed to remove non-relevant records (e.g. leptin, lunch, spawning, muscle, women). Thus, the first component increases the comprehensiveness of the retrieval (recall), while the second component increases the signal-to-noise ratio (precision) by removing the noise.

For application of the journal-title-based query to the SCI database, articles contained in the 68 journals classified by the SCI under the category Energy and Fuels were sampled. Those journals that were not power-source specific, and that contained a very high fraction of articles deemed relevant to the Power Source topic, were identified, and all their articles were included in the retrieved database. The final journal title-based query used for the Power Source study contains 11 journals, and is shown in Ref. [35].

The authors believe that queries of these magnitudes and complexities are required when a tailored database of relevant records that encompasses the broader aspects of target disciplines is needed. 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 [2,8–10], then even more complex queries using Simulated Nucleation may be required.

4. Results

The results from the publications bibliometric analyses are presented in Section 4.1, followed by the results from the citations bibliometrics analysis in Section 4.2. Results from the DT analyses are shown in Section 4.3. The SCI 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. Due to space limitations, not all results could be presented in this paper. To access all the results, as well as the technical details of all analytical processes, the reader is referred to Ref. [35].

4.1. Publication statistics on authors, journals, organizations, countries

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, since these papers are published in the (typically) high caliber journals accessed by the SCI.

4.1.1. Author frequency results

Of the 10 most prolific authors listed in Table 2, four are from India, three are from the UK, and one each from the USA, Japan, and Saudi Arabia. All are from universities. This prolific author country distribution differs radically from any in previous studies [26,28–32], with the high concentration from India. These prolific author countries in previous text mining studies tended to be dominated by Northern America countries (United States and Canada), the most developed Western European nations (UK, Germany, France, Italy), and the major oriental Asian countries (Japan, China, South Korea). In these previous text mining studies, the prolific author country distributions tended to align with the prolific country distributions. In the present paper, the prolific country distributions follow the conventional pattern above (shown later), contrary to the prolific author country distributions. The electrochemical power sources study [32] showed 65% of the prolific authors from the Far East, mainly Japan and China.

Because of the nature of the query used in the present study, many traditional energy production and conversion technologies were included (solar cooking, solar drying, solar distillation, biomass, coal combustion, etc.). A review of thousands of Abstracts confirmed that much of the Power Sources S&T focused on relatively low technology traditional approaches, especially research from the developing countries. The most prolific Indian authors addressed the solar and biomass topics. Interestingly, the most prolific British authors all concentrated on coal, including combustion, properties, and gasification.

4.1.2. Journals containing most power sources papers

There were 1422 different journals represented. This is twice the number of journals from any of the previous studies [26,28–32], and again reflects the multi-disciplined nature of EPS. There was an average of 14.64 papers per journal. This number is somewhat inflated compared to the journal averages

Table 2
Most prolific authors (present institution listed)

Author name	Institution	Country	No. of papers
Wu C.	US Naval Academy	USA	71
Kandiyoti R.	University of London	UK	69
Tiwari GN.	Indian Institute of Technology	India	62
Dincer I.	King Fahd University	Saudi Arabia	61
Garg HP.	Indian Institute of Technology	India	49
Kandpal TC.	Indian Institute of Technology	India	48
Snape CE.	University of Nottingham	UK	43
Williams A.	University of Leeds	UK	42
Ishikawa M.	Yamaguchi University	Japan	41
Kumar S.	Indian Institute of Technology	India	39

Table 3
Journals from query-derived component of database containing most papers

Journal names	No. of papers
J Engng Gas Turbines Power, Trans ASME	200
Int J Hydrog Energy	186
J Propul Power	140
Biomass Bioenerg	134
Combust Sci Technol	121
Brennst-Warme-Kraft	119
IEEE Trans Magn	108
Combust Flame	103
Energy Policy	102
Solar Energy	98
Appl Energy	90
Combust Explos	88
J Appl Phys	82
Solid State Ion	75
Fusion Technol	71
J Electrochem Soc	67
IEEE Trans Energy Convers	62
JSME Int J Ser B: Fluids Therm Engng	58
Appl Therm Engng	57
IEEE Trans Power Syst	55

from these other text mining studies. In the journal-derived component of the present study, all the papers in 11 journals were used. Nevertheless, even for those journals identified by the query-derived component of the database, the journals containing the most Power Source papers had in some cases an order of magnitude of more papers than the average (Table 3).

The journals cover a wide range of energy themes. These include *Combustion/Propulsion* (Journal of Propulsion and Power, Combustion Science and Technology, Combustion and Flame, Combustion and Explosion), *Converters* (Journal of Engineering for Gas Turbines and Power-Transactions of the ASME, Brennstoff-Warme-Kraft, IEEE Transactions of Energy Conversion, IEEE Transactions of Power Systems), *Thermal Engineering* (Applied Thermal Engineering, JSME International Journal Series B, Fluids Thermal Engineering), *Renewables* (International Journal of Hydrogen Energy, Biomass and Bioenergy, Solar Energy), *Electrochemistry* (Solid State Ionics, Journal of the Electrochemical Society), *Physics/Magnetics* (IEEE Transactions on Magnetics, Journal of Applied Physics, Fusion Technology), and *General/Policy* (Energy Policy, Applied Energy). They do not cover the most fundamental science journals (e.g. Science, Nature, Physics of Fluids, Journal of Chemical Physics), since the query had a power/energy sources focus.

4.1.3. Institutions producing most power sources papers

Of the 10 most prolific institutions listed in Table 4, four are from the Far East, two are from Western Europe, two from the USA, one from Eastern Europe, and one from the Middle East. Five are universities, and the remaining five institutions are research institutes. Compared to previous studies [26,28–32], the ratios of research institutes to universities is relatively high in this study. Typically, the ratio of research institutes to universities has been in the vicinity of 10–20%. The higher ratio in

Table 4
Prolific Institutions

Institution names	Country	No. of papers
Indian Institute of Technology	India	415
CSIC	Spain	186
Pennsylvania State University	USA	172
Russian Academy of Science	Russia	164
Tohoku University	Japan	163
Argonne National Laboratory	USA	142
CSIRO	Australia	137
King Fahd University Petroleum and Minerals	Saudi Arabia	137
University of Leeds	UK	127
University of Tokyo	Japan	122

the present study is indicative of the applied focus of the query and retrievals, where it would be expected that more of the effort would be conducted in research institutes or industry.

4.1.4. Countries producing most power sources papers

There are 78 different countries listed in the results. The country bibliometric results are summarized in Table 5. The dominance of a handful of countries is clearly evident.

Table 5
Prolific countries

Country	No. of papers	Population (millions)	Gross domestic product ($Billions)	No. of papers/ population	No. of papers/ gross domestic product
USA	5285	278	9963	19.01079	0.530463
Japan	2269	127	3150	17.86614	0.720317
England	1358	60	1360	22.63333	0.998529
India	1196	1030	2200	1.161165	0.543636
Germany	1141	83	1936	13.74699	0.58936
Canada	997	31	775	32.16129	1.286452
France	813	59	1448	13.77966	0.561464
Australia	603	19	445	31.73684	1.355056
Peoples Republic of China	586	1284	4500	0.456386	0.130222
Italy	559	58	1273	9.637931	0.43912
Spain	498	40	720	12.45	0.691667
Turkey	474	66	444	7.181818	1.067568
Russia	464	145	1120	3.2	0.414286
Sweden	382	9	197	42.44444	1.939086
Netherlands	353	16	388	22.0625	0.909794
South Korea	316	48	765	6.583333	0.413072
Egypt	294	68	247	4.323529	1.190283
Poland	256	39	328	6.564103	0.780488
Saudi Arabia	248	23	232	10.78261	1.068966
Greece	225	11	182	20.45455	1.236264

There appear to be three dominant groups in the 20 most prolific countries. The US and Japan constitute the most dominant group. England, India, Germany, Canada, and France constitute the next group, and the remaining countries constitute the third group. This is the prolific country distribution pattern typical of past text mining studies [26,28–32].

Of these top 20 countries, two are from North America, five are from the Far East, nine are from Western Europe, two are from Eastern Europe, and two are from the Middle East. South America and Africa are not represented.

Weighting these regions by number of papers, the ranking is North America (6282), Western Europe (5803), Far East (4970), Eastern Europe (720), and Middle East (542). When total population and GDP are taken into account, some dramatic changes occur. For papers per unit of population in the top 20, the top five are mainly Western European and English-speaking nations (Sweden, Canada, Australia, UK, Netherlands), and the bottom five are dominated by Asia and Eastern Europe (China, India, Russia, Egypt, Poland). For papers per unit of GDP in the top 20, the top five are mainly developed nations (Sweden, Australia, Canada, Greece, Egypt), and the bottom five are a more amorphous mix (China, South Korea, Russia, Italy, USA). Interestingly, for all three productivity measures, Canada, Australia, and Sweden rank high.

4.2. Citation statistics on authors, papers, 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 [36], 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 [37,38].

The citations in all the retrieved SCI papers were aggregated, the authors, specific papers, years, journals, and countries cited most frequently were identified, and were presented in order to decreasing frequency. A small percentage of any of these categories received large numbers of citations. From the citation year results, the most recent papers tended to be the most highly cited. This reflected rapidly evolving fields of research.

4.2.1. Most cited authors

Of the 20 most cited authors listed in Table 6, eight are from the USA, four are from Japan, five are from Western Europe, one from Israel, one from Bulgaria, and one from China. This is a far different distribution from the most prolific authors, where half were from Asia, and 10% from the USA. There are a number of potential reasons for this difference, including difference in quality and late entry into the research discipline. In another three or four years, when the papers from present-day authors have accumulated sufficient citations, firmer conclusions about quality can be drawn.

Ten of the most cited authors worked on fossil fuels (mainly coal, mainly combustion), five worked in thermodynamics, three worked on batteries (mainly lithium), one worked on solar, and one worked on polymers.

The lists of most prolific authors and most highly cited authors only had one name in common (Wu, C). This phenomenon of minimal intersection has been observed in all other text mining studies performed by the first author. The time frame of interest for most prolific authors is present time, whereas the time frame of interest for most cited authors can span many decades. Researchers who may very well have been prolific when their most citable work was done may no longer be

Table 6
Most cited authors (cited by other papers in this database only)

Author	Topic	Institution	Country	No. of cites
Solomon PR.	Coal pyrolysis	Adv Fuel Res, Inc.	USA	510
Pavlov D.	Lead–acid batteries	Bulgarian Acad Sci	Bulgaria	420
Bejan A.	Thermodynamics	Duke Univ	USA	405
Aurbach D.	Lithium batteries	Bar Ilan Univ	Israel	367
Larsen JW.	Coal pyrolysis	Lehigh Univ	USA	355
Mochida I.	Carbon applications	Kyushu Univ	Japan	292
Ohzuku T.	Lithium batteries	Osaka City Univ	Japan	274
Suuberg EM.	Coal pyrolysis	Brown Univ	USA	245
Nishioka M.	Combustion	Nagoya Univ	Japan	233
Wu C.	Thermodynamics	US Naval Academy	USA	230
Duffie JA.	Solar heating	Univ Wisconsin	USA	221
Vankrevelen DW.	Polymers	Akzo Res and Engrng	Holland	206
Devos A.	Thermodynamics	State Univ Ghent	Belgium	198
Suzuki T.	Coal pyrolysis	Kyoto Univ	JAPAN	196
Painter PC.	Coal properties	Penn State Univ	USA	194
Li CZ.	Coal pyrolysis	Univ London Imper Coll	UK	193
Sabbah R.	Combustion thermodynamics	CNRS	France	190
Herod AA.	Coal combustion	Univ London Imper Coll	UK	190
Chen JC.	Thermodynamics	Xiamen Univ	China	185
Huffman GP.	Fossil combustion	Univ Kentucky	USA	184

prolific. They may have left the discipline, may have assumed non-research duties, or may have slowed down. As the gap between their most citable work and the present widens, the validity of this statement increases.

Sixteen of the authors' institutions are universities, two are government-sponsored research laboratories, and two are private companies. The appearance of the companies on this list is another differentiator from the list of most prolific authors.

The citation data for authors and journals represents citations generated only by the specific records extracted from the SCI database for this study. It does not represent all the citations received by the references in those records; these references in the database records could have been cited additionally by papers in other technical disciplines.

4.2.2. Most cited papers

The most highly cited papers are listed in Table 7.

The theme of each paper is shown in italics on the line after the paper listing. The order of paper listings is inverse number of citations by other papers in the extracted database analyzed. The total number of citations from the SCI paper listing, a more accurate measure of total impact, is shown in the last column on the right. Papers more closely linked to energy applications, such as those on coal, capture many of the total citations (about half) within the present database. The more fundamental science-oriented papers tend to be referenced by myriad disciplines, and the papers within the present database capture a much smaller fraction of the total citations (in some cases, near 10% of the total).

Table 7
Most cited papers (total citations listed in SCI)

Author	Year	Journal	Volume	Sci cites	Total
Curzon FL. *Carnot engine efficiency at maximum power output*	1975	Am J Phys	V43	154	366
Miller JA. *Modeling nitrogen chemistry in combustion*	1989	Prog energy combust	V15	90	825
Solum MS. *Solid state NMR of Argonne premium coals*	1989	Energy Fuel	V3	83	170
Vorres KS. *Argonne premium coal*	1990	Energy Fuel	V4	82	153
Fong R. *Lithium intercalation into carbon*	1990	J Electrochem Soc	V137	68	346
Larsen JW. *Structure of bituminous coals*	1985	J Org Chem	V50	59	125
Solomon PR. *Argonne premium coal analysis*	1990	Energ Fuel	V4	59	143
Iino M. *Coal extraction*	1988	Fuel	V67	56	112
Ohzuku T. *Manganese dioxide in lithium non-aqueous cell*	1990	J Electrochem Soc	V137	54	336
Nishioka M. *Aromatic structures in coals*	1990	Energ Fuel	V4	51	80

Energy and Fuels contains the most papers, four out of the 10 listed. Most of the journals are fundamental science journals, and most of the topics have a fundamental science theme. Most of the papers are from the 1989–1990 time frame. This reflects a dynamic research field, with seminal works being performed in the recent past.

Six papers focus on coal issues, one on combustion, one on thermodynamics, and two on secondary lithium battery issues. Thus, the intellectual heritage focus is on conversion to electricity with a thermal step, as opposed to direct conversion to electricity. Even though the text analysis will show later a significant effort on renewables, this level of effort is not reflected in the intellectual heritage.

4.2.3. Most cited journals

As shown in Table 8, the journal Fuel received almost as many citations as the next three journals combined. Most of the highly cited journals are fossil fuel/combustion oriented or electrochemical power source oriented. These are followed by some fundamental Chemistry and Physics journals. The only renewables journal interspersed is Solar Energy. These results are fully in line with those of the most cited authors and papers, and suggest that consensus seminal works have yet to be established for many of the renewables areas.

Table 8
Most cited journals (cited by other papers in this database only)

Journal	Times cited
Fuel	15,013
J Electrochem Soc	6600
Energy Fuel	6317
J Power Sources	4238
Solar Energy	2957
Combust Flame	2611
Solid State Ionics	1922
J Chem Phys	1752
Carbon	1686
J Appl Phys	1654
J Phys Chem, US	1652
Fuel Process Technol	1573
Electrochim Acta	1558
Combust Sci Technol	1523
J Am Chem Soc	1511
Energy	1466
Ind Eng Chem Res	1426
Anal Chem	1412
J Catal	1371
Nature	1358

The authors end this bibliometrics section by recommending that the reader interested in researching the topical field of interest would be well-advised to, first, obtain the highly-cited papers listed and, second, peruse those sources that are highly cited and/or contain large numbers of recently published papers.

4.3. Database tomography results

There are two major analytic methods used in this section to generate taxonomies of the SCI databases: non-statistical clustering, based on manual assignment of phrases to categories, and statistical clustering, based on algorithmic assignment of phrases or documents to categories. Non-statistical clustering is performed on the Keywords and Abstracts fields. Due to space limitations, the Keywords and Abstracts results are contained in Ref. [35]. Statistical clustering is performed on the Abstracts field only, and only the document clustering results are presented here. The phrase clustering results are contained in Ref. [35].

4.3.1. Statistical clustering

Two generic types of statistical clustering were performed, concept clustering and document clustering. In concept clustering, words/phrases are combined into groups based on their co-occurrence in documents. In document clustering, documents are combined into groups based on their text similarity. Document clustering yields number of documents in each cluster directly, a proxy metric for level of emphasis in each taxonomy category.

4.3.1.1. Abstract Journal and Query-based Taxonomies. As previously mentioned, the EPS database was constructed with two queries:

1. A Journal Title query where all SCI articles (1991–2000 inclusive) from 11 identified relevant energy journals were retrieved (JOURNAL QUERY);
2. A Phrase query, where SCI articles were retrieved by searching Title/Keywords/Abstract fields with a query of phrases and phrase combinations (PHRASE QUERY).

Subsequently, taxonomies were developed for each database (JOURNAL QUERY and PHRASE QUERY). In this section, the two component taxonomy results are presented to elucidate the differences between the JOURNAL QUERY and PHRASE QUERY databases approaches.

In each case, the taxonomies were developed through document clustering of the database Abstracts.

4.3.2. Document clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [39–48]. The approach presented in this section is based on a partitional clustering algorithm [49,50] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements. Thirty-two individual clusters were chosen for the query-based database and the journal-based database. The 32 clusters for each type of database are presented in Ref. [35].

CLUTO also agglommorates the 32 clusters in a hierarchical tree (taxonomy) structure. The taxonomies for each of the two databases are presented here.

4.3.3. Query-based database taxonomy

Table 9 shows a four-level hierarchical taxonomy for the query-based database. The left-most column is the highest taxonomy level, and each column to the right is the next lowest level. The number of records in each category is shown in parenthesis.

The first level taxonomy can be sub-divided into two approximately equal categories: Power Generation/Energy Storage, and Energy Conversion. Power Generation/Energy Storage (4843) focuses on the systems aspects of energy generation and storage, while Energy Conversion (4527) focuses on the direct and indirect conversion of energy to electricity.

For the second level taxonomy, each first level category is divided into two sub-categories. Power Generation/Energy Storage is divided into Fossil Remediation and Replacement Systems (1443 records, focusing on remediation of CO_2 emissions from fossil plants, as well as renewable source systems to replace the CO_2-emitting fossil plants), and Power Plant Heating and Storage Systems (3400 records, focusing on heating and energy storage systems, and nuclear power generation systems). Energy Conversion is divided almost equally into Direct Conversion (2117 records, focusing on the direct conversion of energy sources to electrical power), and Thermal Step Conversion/Combustion (2410 records, focusing on conversion with a thermal step (such as combustion)).

Table 9
Four level taxonomy: query database

Level 1	Level 2	Level 3	Level 4
Power generation/energy storage (4843)	Fossil remediation and replacement systems (1443)	Biomass and renewable generation (1052)	Wind and solar generation (297)
			Biomass generation (755)
		CO_2 emissions from fossil generation (391)	CO_2 emissions from fossil generation (391)
	Power plant heating and storage systems (3400)	Nuclear power generation (976)	Nuclear and fusion (479)
			Steam turbine plant (497)
		Heating and energy storage (2424)	Heat engine storage (996)
			Power system control and battery storage (1428)
Energy conversion (4527)	Direct conversion (2117)	Magnetic field conversion (625)	Material magnetic properties (184)
			Magnetic field structures (441)
		Electrochemical and photochemical conversion (1492)	Material electrical properties (691)
			Fuel cells and photovoltaics (801)
	Thermal step conversion (2410)	Catalytic combustion (1251)	Catalytic reactions (690)
			Coal particle bed combustion (561)
		Engine droplet combustion (1159)	Droplet combustion (680)
			Diesel engine combustion (479)

All second level categories are sub-divided into form 8 third level categories, and the third level categories are sub-divided into form 16 fourth level categories. The category headings for the third and fourth levels are sufficiently detailed that no further description is required.

4.3.4. Journal-based database taxonomy

Table 10 shows a fourth-level hierarchical taxonomy for the journal-based database. The first level taxonomy can be sub-divided into two categories, Fossil Remediation and Replacement Systems, Turbine Conversion (6294 records, focusing partially on remediation of CO_2 emissions from fossil plants, mainly on renewable source systems to replace the CO_2-emitting fossil plants, emphasizing turbine conversion), and Fossil Generation and Storage (5860 records, focusing on fossil-based power plants and mainly battery storage systems).

For the second level taxonomy, each first level category is divided into two sub-categories. Fossil Remediation and Replacement Systems is divided into Solar Thermal (2623 records, focusing on solar

Table 10
Four level taxonomy: journal database

Level 1	Level 2	Level 3	Level 4
Fossil remediation and replacement systems, turbine conversion (6294)	Solar thermal (2623)	Heating and cooling modeling (1633)	Heat transfer modeling (1009)
			Heat pump systems (624)
		Solar collectors (990)	Solar collector systems (673)
			Solar radiation data (317)
	CO_2 remediation and other low emission replacement systems, turbine conversion (3671)	Power plant production, turbine conversion, wind, photovoltaics, geothermal (2444)	Energy consumption and production (1036)
			Wind, turbine conversion, photovoltaics, biomass, and geothermal power (1408)
		Fuel cells and CO_2 emissions (1227)	CO_2 emissions from vehicles (669)
			Vehicle fuel cells (558)
Fossil generation and storage (5860)	Batteries (1890)	Lithium and nickel (1419)	Nickel batteries (745)
			Lithium batteries (674)
		Lead–acid batteries (471)	Lead-acid batteries (471)
	Fossil generation (3970)	Coal (3048)	Coal extraction, liquefaction, gasification, pyrolysis (2325)
			Fluidized bed catalysis (723)
		Oil (922)	Multiple oil sources (489)
			Asphaltene structure and properties (433)

collectors for heating and cooling applications), and CO_2 Remediation and other Low Emission Replacement Systems, Turbine Conversion (3671 records, focused on CO_2 emission reduction and other mainly renewable low emission power generating systems, emphasizing turbine conversion). Fossil Generation and Storage is divided into Fossil Generation (3970 records, focusing on fossil fuel sources and conversion technologies), and Batteries (1890 records, focusing on battery development).

All second level categories are sub-divided into form 8 third level categories, and the third level categories are sub-divided into form 16 fourth level categories. The category headings for the third and fourth levels are sufficiently detailed that no further description is required.

4.3.5. Comparison of query and journal-based database taxonomies

With the exception of the Journal of Power Sources, the journal query approach accessed generic energy related journals that, for the most part, focused on applied energy research. These journals

reported on the numerous processes that utilize energy, and the potential that developed/developing energy sources/conversion methods could provide. Many of the contributors were from the developing countries, where those types of technologies could be readily produced and implemented.

This is substantially different from the articles retrieved using the specific phrase query, where the focus was well distributed among existing and developing primary sources of energy and the fundamental technology issues with converting these sources in various energy-requiring applications. The contributors reflected, on average, the more developed countries, that have the resources to both develop and implement these technologies.

The query taxonomy is more integrated structurally, and the major theme components tend to be complementary. The journal taxonomy is more disjoint, and thematic groupings are sometimes heterogeneous. The linkage between the documents in the query taxonomy is based on the query phrases, whereas the linkage between the documents in the journal taxonomy is their publication in discrete journals. Since the document clustering process is based on text similarity, and the query document linkage is query text similarity, the document clustering is more compatible with the query-based database. In addition, the query database taxonomy has much more of a high technology focus than the journal database taxonomy. The major technology differences that support this conclusion are presented here.

4.3.5.1. Nuclear. Nuclear power has modest representation in the query database compared to renewables and fossil, and no representation in the journal database. The reasons for low frequencies related to Nuclear are as follows.

There are three major journal types in the SCI that serve as sources of papers. First, there are the fundamental multi-discipline journals, such as Science and Nature. These journals would contain papers focused on the fundamental energy conversion phenomena. Because of the high tech nature of these journals, they would have a higher fraction of nuclear-related articles than are reflected in the Keyword analysis of the present study. These papers would have a higher probability of being accessed through phenomena-related terms, rather than the specific energy production and conversion terms in the query used to generate part of the overall database in this study.

The second journal type is generic power-oriented. These journals constituted the journal-derived component of the total database used in this study, and are listed in Section 1. The journals in this category contain basic and applied research papers, but on average, as will be shown later, tend to emphasize fossil, electrochemical, and traditional renewables, with very modest representation of fusion, fission, MHD, and more exotic renewables.

The third journal type is specific power-oriented, and 30 journals in this category are listed in Table 11. These journals were not added to the total database in full, as were the generic power-oriented, for the reasons provided in the database generation section. Their representation in the total database derived from their papers that were accessed by the query. Half of these journals were devoted to nuclear energy and power. It appears that the nuclear S&T community publishes mainly in the first and third types of journals, especially in their dedicated literatures for the more applied S&T.

Thus, the observation that nuclear documents are a small fraction of the fossil and renewables documents should not be interpreted that nuclear source S&T is not being performed or is not important. The proper interpretation is that when power source-related nuclear S&T is examined within the overall power source-related S&T, the high and low tech non-nuclear S&T performed

Table 11
Specific power-oriented journals from SCI

Journals
Journal of the American Oil Chemists Society
Oil Shale
Energy Exploration and Exploitation
Petroleum Science and Technology
Chemistry and Petroleum Engineering
Sekiyu Gakkaishi
Petroleum Chemistry
Pipeline Gas Journal
Biomass and Bioenergy
Solar Energy
Solar Energy Materials and Solar Cells
Journal of Solar Energy Engineering
Progress in Photovoltaics
Journal of Wind Engineering and Industrial Aerodynamics
Journal of Nuclear Materials
Nuclear Energy, Journal of the British Nuclear Energy Society
Annals of Nuclear Energy
Nuclear Engineering international
Progress in Nuclear Energy
Nuclear Science and Engineering
Fusion Technology
Fusion Engineering and Design
Nuclear Fusion
Plasma Physics and Controlled Fusion
Journal of Fusion Energy

globally dominate the higher tech nuclear S&T performed in a smaller number of the more developed countries. To obtain a more detailed picture of the advances in nuclear power S&T, a standard DT focused analysis of the literature would need to be performed. Detailed technical terms would be used in the query, and 15 nuclear-specific journals listed in Table 11 could be added to form the total database.

4.3.5.2. Renewables. About 20% of the power systems in the query database are focused on renewables, whereas about 40% of the sources in the journal database are focused on renewables. Additionally, the emphases on specific renewables are different between the two databases. For example, in solar energy, the query database emphasizes the higher tech solar electric (especially Photovoltaics targeted at higher direct electricity conversion efficiencies). The journal database emphasizes the lower tech non-direct electricity component of solar (desalinization, distillation, heating, refrigeration). In biomass, the query database had more generic representation (biomass, solid waste, sewage sludge, vegetable oils), while the journal database had higher representation in the traditional types of biomass (firewood, rice husks, wheat straw). Wind energy had low representation in both databases. Geothermal had very low representation in the journal database, and did not even display as a cluster in the query database.

4.3.5.3. Fossil. Fossil appears in two sections of the query database taxonomy. There is a modest effort on analysis of CO_2 generation from fossil sources, and a more substantive contribution from fossil combustion techniques (catalytic combustion, engine droplet combustion). Combined, these two fossil components represent about 30% of the query database. The journal database taxonomy also represents fossil explicitly in two sections. There is a substantial section on fossil generation, and a smaller section on CO_2 emissions from vehicles. Combined, these two fossil components represent about 35% of the journal database. The main difference between the two databases relative to fossil is that the journal database emphasizes source preparation and extraction, while the query database emphasizes the higher tech fuel combustion. Also, coal seems to have a much higher representation compared to oil in the journal database, whereas the representations are about equal in the query database. Natural gas had low representation in both databases relative to coal or oil.

4.3.5.4. Conversion. Nowhere are the structural differences between the query and journal databases better illustrated than in conversion. Energy conversion is identified as a separate thematic thrust at the highest taxonomy level of the query database, consisting of almost half the database records. In the journal database, energy conversion components can be found in solar thermal, low emission replacement systems, and fossil generation. Because of the lower tech focus of the journal database, the structure is determined more by specific systems than by advanced phenomena or processes, and conversion tends to be hierarchically identified under specific systems.

In the query database, the sub-categories within the conversion category emphasize the primary conversion phenomena, such as combustion, electrochemical, and magnetic field conversion. The systems aspects of the full conversion cycle, such as the final step in the conversion of energy to electricity (e.g. turbines, power cycles), can be found within the specific power generation systems.

In the journal database, there is less emphasis on the higher tech direct conversion relative to the lower tech thermal step conversion. There is no category of magnetic field conversion, as exists in the query database. Additionally, both databases have a turbine conversion category. In the query database, the turbine conversion is closely associated with the higher tech nuclear power production category, whereas in the journal database, the turbine conversion is associated with the lower tech renewables category, most closely with the wind component. As mentioned under renewables, in the journal database, much of the solar conversion stops at the heating and cooling category, whereas in the query database, relatively more of the solar conversion is directly to electricity.

4.3.5.5. Storage. In the journal database, a separate second-level taxonomy category of batteries, containing about 15% of total database articles, is identified. Many of these battery articles, and fuel cell articles in the journal database as well, result from the inclusion of the electrochemical-dominant Journal of Power Sources in the database. The main battery focus is divided between Nickel and Lithium batteries, with somewhat less effort devoted towards the traditional Lead-Acid batteries. No other types of storage are evident in the journal database, at least down to the fourth taxonomy level of resolution.

In the query database, energy storage is identified only at the third taxonomy level. The storage function is closely associated with control of power flow in systems. While batteries receive the primary emphasis, some work is reported in capacitors, especially electrochemical, and much less reported work in mechanical storage systems. The battery work appears focused toward vehicles, in concert with some hydrogen storage efforts for hydrogen-powered vehicles as well.

5. Final observations

Advantages of using DT and bibliometrics for deriving technical intelligence from the published literature include:

- 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 more use of the rich semantic relationships among the words.

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 promising opportunities/work that needs to be done.

Acknowledgements

The views in this paper are solely those of the authors, and do not represent the views of the Department of the Navy or any of its components, University of Minnesota, DDL-OMNI, LLC, or Noesis, Inc. In addition, the authors acknowledge the contributions of Dr Richard Carlin, Office of Naval Research, for sponsoring this effort.

References

[1] Kostoff RN. Text mining for global technology watch. In: Drake M, editor. Encyclopedia of library and information science, vol. 4. 2nd ed. New York: Marcel Dekker 2003. p. 2789–99.
[2] Hearst MA. Untangling text data mining. In: Dale R, Church K, editors. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland. New York: ACM Press; June 20–26, 1999, p. 3–10.
[3] Zhu DH, Porter AL. Automated extraction and visualization of information for technological intelligence and forecasting. Technol Forecast Social Change 2002;69(5):495–506.

[4] Losiewicz P, Oard D, Kostoff RN. Textual data mining to support science and technology management. J Intell Inform Syst 2000;15:99–119.
[5] Kostoff RN, Eberhart HJ, Toothman DR. Database tomography for information retrieval. J Inform Sci 1997;23(4):301–11.
[6] Greengrass E. Information retrieval: an overview. Report No. TR-R52-02-96. Fort George G. Meade, MD: National Security Agency; 1997.
[7] TREC (Text Retrieval Conference), Home Page. http://trec.nist.gov/
[8] Swanson DR. Fish oil, Raynauds syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30(1):7–18.
[9] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell 1997;91(2):183–203.
[10] Kostoff RN. Stimulating innovation. In: Larisa V. Shavinina, editor. International handbook of innovation. Oxford: Elsevier 2003:388–400.
[11] Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. J Am Soc Inform Sci 1998; 49(8):674–85.
[12] Goldman JA, Chu WW, Parker DS, Goldman RM. Term domain distribution analysis: a data mining tool for text databases. Meth Inform Med 1999;38:96–101.
[13] Kostoff RN. Bilateral asymmetry prediction. Med Hypotheses 2003;61(2):265–6.
[14] Kostoff RN, Green KA, Toothman DR, Humenik JA. Database tomography applied to an aircraft science and technology investment strategy. Technical Report No. TR NAWCAD PAX/RTR-200/84. Naval Air Warfare Center, Aircraft Division, Patuxent River, MD, USA. 2000.
[15] Kostoff RN, Shlesinger M, Malpohl G. Fractals roadmaps using bibliometrics and database tomography. Fractals 2004; 12(1):1–16.
[16] Viator JA, Pestorius FM. Investigating trends in acoustics research from 1970–1999. J Acoust Soc Am 2001; 109(5):1779–83.
[17] Kostoff RN, Shlesinger M, Tshiteya R. Nonlinear dynamics roadmaps using bibliometrics and Database tomography. Int J Bifurcat Chaos 2004;14(1):61–92.
[18] Davidse RJ, Van Raan AFJ. Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. Scientometrics 1997;40(2):171–93.
[19] Kostoff RN, Del Rio JA, García EO, Ramírez AM, Humenik JA. Citation mining: integrating text mining and bibliometrics for research user profiling. J Am Soc Inform Sci Technol 2001;52(13):1148–56.
[20] Narin F. Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity. Report No. NSF C-637. Arlington, VA: National Science Foundation; 1976.
[21] Garfield E. History of citation indexes for chemistry—a brief review. J Chem Inform Computer Sci 1985;25(3):170–4.
[22] Schubert A, Glanzel W, Braun T. Subject field characteristic citation scores and scales for assessing research performance. Scientometrics 1987;12(5/6):267–91.
[23] Narin F, Olivastro D, Stevens KA. Bibliometrics theory, practice and problems. Eval Rev 1994;18(1):65–76.
[24] Kostoff RN, Eberhart, HJ, Miles, DA. System and method for database tomography. US Patent Number 5,440,481; 1995.
[25] Kostoff RN, Eberhart HJ, Toothman DR. Database tomography for information retrieval. J Inform Sci 1997;23(4):301–11.
[26] Kostoff RN, Eberhart HJ, Toothman DR. Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. JASIS 1999;50(5):427–47.
[27] Kostoff RN. Database tomography for technical intelligence. Competitive Intell Rev 1993;4(1):38–43.
[28] Kostoff RN, Eberhart HJ, Toothman DR. Database tomography for technical intelligence: a roadmap of the near-earth space science and technology literature. Inform Process Mgmt 1998;34(1):69–85.
[29] Kostoff RN, Eberhart HJ, Toothman DR, Pellenbarg R. Database tomography for technical intelligence: comparative roadmaps of the research impact assessment literature and the journal of the American chemical society. Scientometrics 1997;40(1):103–38.
[30] Kostoff RN, Braun T, Schubert A, Toothman DR, Humenik JA. Fullerene roadmaps using bibliometrics and database tomography. J Chem Inform Computer Sci 2000;40(1):19–39.
[31] Kostoff RN, Green KA, Toothman DR, Humenik J. Database tomography applied to an aircraft science and technology investment strategy. J Aircraft 2000;37(4):727–30.
[32] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA. Electrochemical power source roadmaps using bibliometrics and database tomography. J Power Sources 2002;110(1):163–76.

[33] SCI. Science Citation Index. Institute for Scientific Information 2003. Philadelphia, PA.
[34] Kostoff RN. The underpublishing of science and technology results. The Scientist 2000;14(9):6–6.
[35] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA, Karypis G. Science and technology text mining: electric power sources. DTIC Technical Report No. ADA421789, 2004. Springfield, VA: National Technical Information Service.
[36] Garfield E. History of citation indexes for chemistry—a brief review. J Chem Inform Computer Sci 1985;25(3):170–4.
[37] Kostoff RN. The use and misuse of citation analysis in research evaluation. Scientometrics 1998;43(1):27–43.
[38] MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics 1996;36(3):435–44.
[39] Cutting DR, Karger DR, Pedersen JO, Tukey JW. Scatter/gather: a cluster-based approach to browsing large document collections. In: Belkin NJ, Ingwersen P, Pejtersen AM, editors. Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), Copenhagen, Denmark. New York: ACM Press; 1992, p. 318–29.
[40] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. In: Haas LM, Tiwary A, editors. Proceedings of the ACM-SIGMOD International Conference on Management of Data (SIGMOD'98), Seattle, WA. New York: ACM Press; 1998, p. 73–84.
[41] Hearst MA. The use of categories and clusters in information access interfaces. In: Strzalkowski T, editor. Natural language information retrieval, 1999. Dordrecht: Kluwer. p. 333–74.
[42] Karypis G, Han EH, Kumar V. Chameleon: a hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Spl Issue Data Anal Mining 1999;32(8):68–75.
[43] Prechelt L, Malpohl G, Philippsen M. Finding plagiarisms among a set of programs with JPlag. J Universal Computer Sci 2002;8(11):1016–38.
[44] Rasmussen E. Clustering algorithms. In: Frakes WB, Baeza-Yates R, editors. Information retrieval data structures and algorithms, 1992. Englewood Cliffs: Prentice-Hall. p. 419–42.
[45] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Report No. 00-034. Minneapolis, MN: Department of Computer Science and Engineering, University of Minnesota; 2000.
[46] Willet P. Recent trends in hierarchical document clustering: a critical review. Inform Process Mgmt 1988;24:577–97.
[47] Wise MJ. String similarity via greedy string tiling and running Karb-Rabin matching. Department of Computer Science, University of Sidney; 1992. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps
[48] Zamir O, Etzioni O. Web document clustering: a feasibility demonstration. In: Croft WB, Moffat A, Van Rijsbergen CJ, Wilkinson R, Zobel J, editors. Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), Melbourne, Australia. New York: ACM Press; 1998, p. 46–54.
[49] Karypis G. CLUTO-A clustering toolkit. http://www.cs.umn.edu/~cluto
[50] Zhao Y, Karypis G. Empirical and Theoretical comparisons of selected criterion functions for document clustering. Machine Learning 2004;55(3):311–31.