Journal of Nanoparticle Research (2006) 8: 301–321
DOI 10.1007/s11051-005-9035-8

Perspectives

The structure and infrastructure of the global nanotechnology literature

Ronald N. Kostoff[1,*], Jesse A. Stump[1], Dustin Johnson[1,3], James S. Murday[1,4], Clifford G.Y. Lau[1,5] and William M. Tolles[2]
[1]*Office of Naval Research, 875 N. Randolph St., Arlington, VA 22217, USA;* [2]*8801 Edward Gibbs Place, Alexandria, VA 22309, USA;* [3]*Northrop Grumman TASC, 12015 Lee Jackson Highway, Fairfax, VA 22033, USA;* [4]*Chemistry Division, Code 6100, Naval Research Laboratory, Washington, DC 20375, USA;* [5]*Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA; *Author for correspondence (Tel.: +1-703-696-4198; Fax: +1-703-696-3098; E-mail: kostofr@onr.navy.mil)*

Received 21 February 2005; accepted in revised form 19 August 2005

Key words: nanotechnology, nanoscience, nanomaterials, nanoparticles, nanotubes, nanostructures, nanocomposites, nanowires, nanocrystals, nanofabrication, nanolithography, quantum dots, self-assembly, text mining, computational linguistics, bibliometrics

Abstract

Text mining is the extraction of useful information from large volumes of text. A text mining analysis of the global open nanotechnology literature was performed. Records from the Science Citation Index (SCI)/Social SCI were analyzed to provide the infrastructure of the global nanotechnology literature (prolific authors/journals/institutions/countries, most cited authors/papers/journals) and the thematic structure (taxonomy) of the global nanotechnology literature, from a science perspective. Records from the Engineering Compendex (EC) were analyzed to provide a taxonomy from a technology perspective.

- The Far Eastern countries have expanded nanotechnology publication output dramatically in the past decade.
- The Peoples Republic of China ranks second to the USA (2004 results) in nanotechnology papers published in the SCI, and has increased its nanotechnology publication output by a factor of 21 in a decade.
- Of the six most prolific (publications) nanotechnology countries, the three from the Western group (USA, Germany, France) have about eight percent more nanotechnology publications (for 2004) than the three from the Far Eastern group (China, Japan, South Korea).
- While most of the high nanotechnology publication-producing countries are also high nanotechnology patent producers in the US Patent Office (as of 2003), China is a major exception. China ranks 20th as a nanotechnology patent-producing country in the US Patent Office.

Introduction

Nanotechnology is the development and use of techniques to study physical phenomena and construct structures in the physical size range of 1–100 nm, as well as the incorporation of these structures into applications. Experiments and computer simulation have been targeted at very

small scales for decades. However, the advent of high speed and high storage capacity computers, as well as accurate instruments for measuring and manipulating at the nanoscale, have accelerated the development of nanoscale structures and devices into reality.

Public and private support for further nanotechnology development has increased dramatically. In the National Nanotechnology Initiative, established in 2001, the U. S. Federal government will contribute billions of dollars to further development by the end of the decade. Worldwide, other governments have infused substantial funding to nanotechnology programs. The private sector is heavily investing in this technology, anticipating the large size of the potential market.

Along with the growth in the tools and products of nano-science and technology (and its financial support) has come the growth in the related technical literature. For example, in the fundamental nanotechnology research literature as represented by the Science Citation Index (SCI), publications grew from 4552 articles in 1991 to 33,060 articles in 2004.

Given this voluminous literature, as well as the other voluminous literatures of Patents, Technical Reports, other large databases, and the Web, how can one gain an integrated perspective of the overall state of nanotechnology? Text mining offers one potential approach. This paper applies text mining to the SCI and Engineering Compendex (EC) nanotechnology literatures. The query to retrieve these literatures is defined operationally as follows (* denotes the wild-card character used in most search engines).

NANOPARTICLE* OR NANOTUB* OR NANOSTRUCTURE* OR NANOCOMPOSITE* OR NANOWIRE* OR NANOCRYSTAL* OR NANOFIBER* OR NANOFIBRE* OR NANOSPHERE* OR NANOROD* OR NANOTECHNOLOG* OR NANOCLUSTER* OR NANOCAPSULE* OR NANOMATERIAL* OR NANOFABRICAT* OR NANOPOR* OR NANOPARTICULATE* OR NANOPHASE OR NANOPOWDER* OR NANOLITHOGRAPHY OR NANO-PARTICLE* OR NANODEVICE* OR NANODOT* OR NANOINDENT* OR NANOLAYER* OR NANOSCIENCE OR NANOSIZE* OR NANOSCALE* OR ((NM OR NANOMETER* OR NANOMETRE*) AND (SURFACE* OR FILM* OR GRAIN* OR POWDER* OR SILICON OR DEPOSITION OR LAYER* OR DEVICE* OR CLUSTER* OR CRYSTAL* OR MATERIAL* OR ATOMIC FORCE MICROSCOP* OR TRANSMISSION ELECTRON MICROSCOP* OR SCANNING TUNNELING MICROSCOP*)) OR QUANTUM DOT* OR QUANTUM WIRE* OR ((SELF-ASSEMBL* OR SELF-ORGANIZ*) AND (MONOLAYER* OR FILM* OR NANO* OR QUANTUM* OR LAYER* OR MULTILAYER* OR ARRAY*)) OR NANOELECTROSPRAY* OR COULOMB BLOCKADE* OR MOLECULAR WIRE*

This query, generated using an iterative relevance feedback technique (Kostoff et al., 1997), is used to retrieve relevant documents from selected source databases. Then, the retrieved database is analyzed to produce the following characteristics and key features of the nanotechnology field: recent prolific nanotechnology authors; journals that contain numerous nanotechnology papers; institutions that produce numerous nanotechnology papers; keywords most frequently specified by the nanotechnology authors; authors, papers and journals cited most frequently; pervasive technical themes of the nanotechnology literature; and relationships among the pervasive themes and sub-themes.

Background

Text mining

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

Science and technology (S&T) computational linguistics (Hearst, 1999; Losiewicz et al., 2000; Zhu & Porter, 2002; Kostoff, 2003a) identifies pervasive technical themes in large databases from technical phrases that occur frequently. 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

(Greengrass, 1997; Kostoff et al., 1997; TREC, 2004)
- Potential discovery and innovation based on merging common linkages among very disparate literatures (Swanson, 1986; Swanson & Smalheiser, 1997; Gordon & Dumais, 1998; Kostoff, 2003b, 2005a)
- Uncovering unexpected asymmetries from the technical literature (Goldman et al., 1999; Kostoff, 2003c). For example, Kostoff (2003c) predicted asymmetries in recorded bilateral organ (lungs, kidneys, testes, ovaries) cancer incidence rates from the asymmetric occurrence of lateral word frequencies (left, right) in Medline case study articles.
- Estimating global levels of effort in S&T subdisciplines (Kostoff et al., 2000, 2004a; Viator & Pestorius, 2001)
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors (Kostoff et al., 2004a, b)
- Tracking myriad research impacts across time and applications areas (Davidse & Van Raan, 1997; Kostoff et al., 2001).

Evaluative bibliometrics (Narin, 1976; Garfield, 1985; Schubert et al., 1987) 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 that are producing the patents and papers, and knowledge linkage between their subject areas (Narin et al., 1994). 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; and
- Identify impacts (literature citations) of individuals, research units, organizations, and countries

Nanotechnology

Literature review overview

A comprehensive background of the seminal works in nanotechnology is contained in a companion document (Kostoff et al., 2005a), and will not be repeated here. There are numerous books (e.g., Bhushan's Handbook of Nanotechnology (Bhushan, 2004); Goddard's Handbook on Nanoscience, Engineering, and Technology (Goddard et al., 2002); Freitas' multi-volume set on nanomedicine (Freitas, 1999, 2003); see Appendix 1 of Kostoff et al. (2005b) for more complete listing of reference books), review articles (e.g., Kricka's multi-lingual survey of nanotechnology books and patents (Kricka & Fortina, 2002); Simon's review of the science and potential applications of nanotechnology (Simon, 2005)), and reports (e.g., The Royal Society's comprehensive review on nanoscience and nanotechnologies (Dowling et al., 2004); Colton's in-depth review of nanoscale measurements and manipulation (Colton, 2004)) that cover various sub-sets of nanotechnology. For the research literature, none of these published reviews have the spatial and temporal breadth of coverage of the present paper, none use a query of the extent and complexity of the present paper, and none do full text mining of the results to obtain structure and infrastructure of the nanotechnology literature. Every published research review on nanotechnology typically covers a focused technology sub-set, not the total field as was done in the present paper. For the Patent literature, (Huang et al., 2004) provides a comprehensive text mining analysis of international nanotechnology development that serves to complement the present study.

Database generation

The first step in database generation is query development. The iterative relevance feedback technique of Simulated Nucleation (Kostoff et al., 1997) is used to develop the query as follows. A test query is generated (e.g., 'nanotechnology'); records are retrieved from the SCI using this

query; the retrieved records are divided into relevant and non-relevant categories; the phrase patterns of each category are analyzed using the TextDicer software; and the query is modified by inclusion of selected phrase patterns. The process is repeated until convergence occurs. During the iterative query development process, clustering of the retrieval is performed at least one time, to identify the main technical categories and insure that the query includes adequate technical terms that represent each of the main technical thrusts (Kostoff, 2005c).

For the final retrieval, the query shown in the Introduction was inserted into the SCI (2005) search engine to retrieve relevant records from the source SCI database published in 2003 only. Due to SCI downloading limitations at the time the data were taken, records had to be downloaded separately from the top 350 journals containing the most nanotechnology papers. These 21,474 downloaded papers were used for the computational linguistics and most of the bibliometrics. The institution and country bibliometrics were obtained from direct query of the SCI. These downloaded records were current at the time of the study, and the numbers of records retrieved provided an adequate sampling of the literature. The SCI-retrieved 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 nanotechnology articles. The prolific institution and country bibliometrics were updated to 2004, especially to highlight the rapid advance made by a number of countries in recent years. The query was also inserted into the Engineering Compendex to retrieve relevant records.

Results

The results from the publications bibliometric analyses are presented in section 'Publication statistics on authors…', followed by the results from the citations bibliometrics analysis in section 'Citation statistics on authors…'. Results from the computational linguistics analyses are shown in section 'Taxonomy results'. 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.

Publication statistics on authors, journals, organizations, countries

The output and productivity metrics of paper counts are presented initially.

Author frequency results

There were 50,969 authors listed in the 21,474 downloaded records. Table 1 contains the names of the 20 most prolific of these authors. All the names of the 20 most prolific authors appear to be of Asian origin, and the number of their publications listed for 2003 appeared quite high. While some of the names applied to multiple authors, in some cases an author listed as most prolific was indeed one person. YT Qian, for example, published 106 *research articles* in SCI-accessed journals in 2003 (based on direct query of the SCI), and 346 articles in SCI-accessed journals over a four-year period.

As the first author's previous text mining studies have shown, one characteristic of prolific authors is that they tend not to be first authors on many of their articles. For example, the 20 most prolific authors from 2000 to 2003 in the two

Table 1. Most prolific authors – 2003

Author	#Papers
Zhang, Y	84
Li, J	63
Qian, YT	62
Wang, J	62
Wang, Y	62
Lee, JH	59
Liu, Y	58
Zhang, LD	58
Chen, Y	56
Bando, Y	52
Chen, J	52
Wang, X	52
Zhang, J	51
Gao, L	50
Wang, H	47
Kim, JH	46
Li, Y	45
Kim, J	44
Zhang, H	44
Wang, L	41

nanotechnology-focused journals Nano Letters and Nanotechnology authored/co-authored 193 papers in the two journals since the journals' inception. They were the first authors on just 14 of the 193 papers, or about 7%. Eleven of the authors had zero first authorships, five of the authors had one first authorship, three of the authors had two first authorships, and one author had three first authorships. Interestingly, three of the four authors who had more than one first authorship work at national laboratories. The practical implications of prolific authors tending not to be first authors, in terms of Bibliometrics impact, will be addressed in the Most Cited Authors section.

Journals containing most nanotechnology papers

The 20 journals containing the most nanotechnology papers (Table 2) from the 21,474 downloaded records tend to be in the technical disciplines of Physics, Chemistry, and Materials, with an emphasis on surface science. The top tier in volume of nanotech-related articles had three physics journals (*Applied Physics Letters*, *Physical Review*, and *Journal of Applied Physics*).

Table 2. Journals containing most nanotechnology research articles – 2003

Journal	#Papers
Applied Physics Letters	1240
Physical Review B	899
Journal of Applied Physics	875
Langmuir	690
Journal of Physical Chemistry B	558
Japanese Journal of Applied Physics Part 1-Regular Papers Short Notes & Review Papers	435
Journal of the American Chemical Society	408
Chemical Physics Letters	390
Physical Review Letters	353
Nano Letters	346
Chemistry of Materials	319
Applied Surface Science	291
Physica E-Low-Dimensional Systems & Nanostructures	278
Thin Solid Films	260
Inorganic Chemistry	254
Journal of Magnetism and Magnetic Materials	247
Journal of Materials Chemistry	243
Macromolecules	243
Advanced Materials	239
Journal of Vacuum Science & Technology B	229

Institutions producing most nanotechnology papers/patents

Table 3 contains the most prolific paper-publishing institutions in 2003, based on direct query of the SCI. Also in this table are the most prolific patent-publishing institutions (for nanotechnology patents granted by the US Patent Office) in 2003, according to (Huang et al., 2004). Seven of the prolific paper-publishing institutions are research centers, and the rest are universities. No industry institutions are listed. Nineteen of the prolific patent-producing institutions are industry, and the other two are universities. The two universities listed (MIT, Univ of Cal) are common to both lists.

The relatively high fraction of paper-publishing research centers (1/3) compared to the first author's previous technology text mining studies suggests a more applied focus to the research. The near-orthogonality of the two lists suggests that the organizations/people who publish prolifically are not the same as those who patent prolifically. It might be useful for management researchers to study the MIT example, to ascertain how MIT maintains a prolific publishing-patenting balance, and what (if any) advantages accrue from maintaining such a balance.

Table 4 contains the most prolific institutions, updated for 2004. Of the 20 most prolific institutions, 13 are universities, and the remaining 7 are government laboratories. Thirteen are from the Far East, three are from the USA, three are from Western Europe, and one is from Eastern Europe. There are no major changes at the top of the list between 2003 and 2004.

Countries producing most nanotechnology papers/patents

Table 5 contains the most prolific paper-producing countries for 2003, based on direct query of the SCI. Also in this table are the most prolific countries for nanotechnology patents granted by the US Patent Office in 2003, according to (Huang et al., 2004). The most striking difference is that of China, which effectively is tied for second on the paper-producing list and tied for 20th on the patent-producing list. Canada, Netherlands, and Israel have a paper performance that outproduces the patent performance by a substantial margin.

Table 6 shows the most prolific paper-producing countries, updated for 2004. At the top, there is little difference in the rankings compared to 2003,

Table 3. Most prolific institutions – SCI; USPTO Patents – 2003

Institution – SCI	#Pap	Institution – Patents	#Pat
Chinese Acad Sci	1303	IBM	198
CNRS	1198	Micron Technology	129
Russian Acad Sci	687	Advanced Micro Devices	128
Tsing Hua Univ	454	INTEL	90
Univ Tokyo	429	Regents, Univ Of California	89
Tohoku Univ	352	MMM	79
Osaka Univ	345	Motorola	72
Natl Inst Adv Ind Sci & Technol	341	Hitachi	68
Univ Sci & Technol China	297	Xerox	68
Nanjing Univ	288	Canon Kabushiki Kaisha	64
Natl Inst Mat Sci	287	Eastman Kodak	64
Tokyo Inst Technol	283	NEC	57
CNR	275	Corning	50
CSIC	268	Applied Materials	47
Univ Illinois	245	Fuji Photo Film	42
Peking Univ	245	Matsushita Electric	41
Seoul Natl Univ	244	Lucent Technologies	37
Univ Texas	230	Texas Instruments	37
Univ Cambridge	229	Genentech	36
MIT	226	Kabushiki Kaisha Toshiba	36
Univ Calif Berkeley	210	MIT	36

with the exception that China has clearly moved into second place. There were 101 countries listed. In 2004, three countries dominate: USA, China, and Japan; Germany is a strong contributor as well. In the top six countries, the three from the Western group (USA, Germany, France) have about 8% more publications than the three from the Far Eastern group (China, Japan, South Korea).

Table 4. Prolific institutions – SCI papers – 2004

Institution	Country	#Papers
Chinese Acad Sci	China	1533
CNRS	France	1241
Russian Acad Sci	Russia	641
Tsing Hua Univ	China	504
Univ Tokyo	Japan	444
Osaka Univ	Japan	373
Tohoku Univ	Japan	363
CSIC	Spain	345
Univ Sci & Technol China	China	342
Nanjing Univ	China	333
Natl Inst Adv Ind Sci & Technol	Japan	311
CNR	Italy	311
Tokyo Inst Technol	Japan	308
Seoul Natl Univ	S. Korea	296
MIT	USA	284
Univ Illinois	USA	283
Natl Univ Singapore	Singapore	277
Univ Texas	USA	273
Natl Inst Mat Sci	Japan	272
Peking Univ	China	262

Table 5. Most prolific countries – SCI; USPTO patents – 2003

Country – SCI	#Pap	Country – Patents	#Pat
USA	7512	USA	5228
Japan	4431	Japan	926
Peoples R China	4417	Germany	684
Germany	3099	Canada	244
France	1900	France	183
South Korea	1592	South Korea	84
United Kingdom	1520	Netherlands	81
Russia	1293	United Kingdom	78
Italy	1015	Taiwan	77
India	830	Israel	68
Spain	727	Switzerland	56
Taiwan	706	Australia	53
Canada	690	Sweden	39
Poland	515	Italy	31
Switzerland	498	Belgium	28
Netherlands	492	Denmark	23
Brazil	455	Singapore	20
Sweden	435	Finland	17
Australia	434	Ireland	10
Singapore	372	Austria	8
Israel	347	Peoples R China	8

Table 6. Prolific countries – 2004

Country	Ratio							
	2004			1994			2004/1994	
	Nano Pap	Tot Pap	Nanpap/Totpap	Nano Pap	Tot Pap	Nanpap/Totpap	Nanpap	Totpap
USA	8037	294,762	0.027266	2388	283,530	0.008422	3.365578	1.039615
China	5644	54,024	0.104472	271	8976	0.030192	20.82657	6.018717
Japan	4617	71,411	0.064654	1346	49,524	0.027179	3.430163	1.441947
Germany	3120	65,358	0.047737	928	45,686	0.020313	3.362069	1.430591
France	1954	46,647	0.041889	519	35,346	0.014683	3.764933	1.319725
South Korea	1912	22,284	0.085801	77	3450	0.022319	24.83117	6.45913
England	1465	57,134	0.025641	467	43,254	0.010797	3.137045	1.320895
Russia	1300	23,992	0.054185	249	24,737	0.010066	5.220884	0.969883
Italy	1115	35,561	0.031355	204	21,054	0.009689	5.465686	1.689038
India	1025	21,117	0.048539	115	12,129	0.009481	8.913043	1.741034
Taiwan	941	13,456	0.069932	73	5244	0.013921	12.89041	2.56598
Spain	829	26,302	0.031519	114	12,548	0.009085	7.27193	2.096111
Canada	785	35,630	0.022032	246	29,200	0.008425	3.191057	1.220205
Switzerland	598	14,552	0.041094	175	9882	0.017709	3.417143	1.472576
Netherlands	584	20,176	0.028945	207	14,376	0.014399	2.821256	1.40345
Poland	582	12,968	0.04488	67	5878	0.011398	8.686567	2.206193
Singapore	527	5348	0.098542	14	1378	0.01016	37.64286	3.880987
Sweden	471	15,021	0.031356	128	11,167	0.011462	3.679688	1.345124
Brazil	462	14,631	0.031577	47	4368	0.01076	9.829787	3.349588
Australia	462	22,789	0.020273	101	14,392	0.007018	4.574257	1.583449

However, studies have shown an English language bias for the SCI (Winkmann et al., 2002), and these Far Eastern publication numbers based solely on the SCI should be viewed as an under-estimate.

In addition, trends are very important. In Table 6, the first column on the left (Country) represents the country, the next column (2004 Nano Pap) contains the nanotechnology papers published by each country in 2004, the next column (2004 Tot Pap) contains the total papers published by each country in 2004, the next column is the ratio of the 2004 nano papers to total papers, the next three columns are the same type of data for 1994, and the final two columns are the ratios of 2004/1994 nanotechnology papers and total papers, respectively. Three important observations follow.

First, the 2004/1994 ratio of nanotechnology papers is in double digits for the Far Eastern countries only (Peoples R China, South Korea, Taiwan, and Singapore). Figure 1 shows this trend more dramatically, where the short bar for the countries depicted represents the 1994 nanotechnology papers, and the long bar represents the 2004 nanotechnology papers. Second, the 2004/1994 ratio of total SCI papers is above ~4 for Far Eastern countries only (Peoples R China, South Korea, Singapore). Third, the fractions of nanotechnology papers to total papers for 2004 above 8% are for Far Eastern countries only (Peoples R China, South Korea, Singapore). Thus, in the past decade, these Far Eastern countries have shown substantial growth in total SCI papers, in nanotechnology papers, and in the ratio of nanotechnology papers to total papers.

Citation statistics on authors, papers, and journals

The second group of metrics presented is counts of citations to papers published by different entities. 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 listed in order of decreasing frequency. While citations are ordinarily used as impact or quality metrics (Garfield, 1985), 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 (MacRoberts & MacRoberts, 1996; Kostoff, 1998).

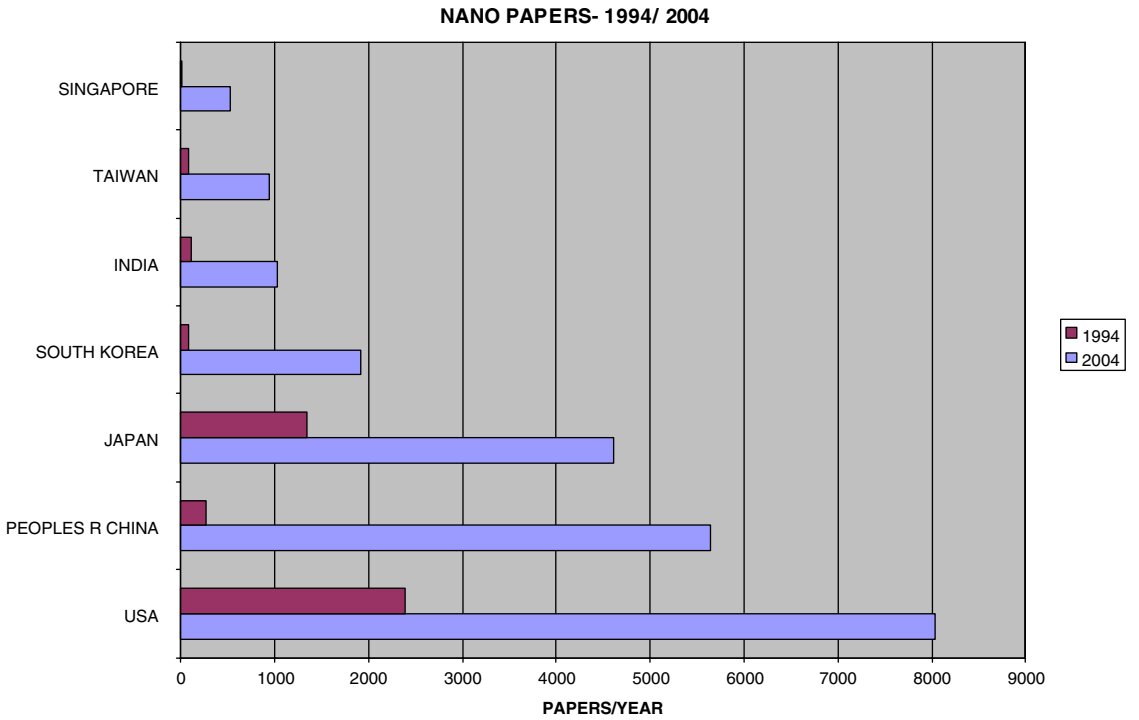

Figure 1. 2004 and 1994 papers published for select countries.

Most cited authors

The most cited first authors were examined initially, since only the first authors are shown on the references downloaded en masse from the SCI. There were 134,906 cited first authors listed. About half of the most cited first authors are from the Far East, with most of the remainder being from the USA (Table 7).

The citation data for authors and journals represent citations generated only by the 21,474 specific records extracted from the SCI database for this study. They do 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 (as will be shown in the Most Cited Documents section), and by papers published in other years.

The number of citations for first authors underrepresent the actual (total) citations for those authors for the following reasons. When references in a document are downloaded as a group from the SCI, they list the first author only. Thus, any reference on which any of its authors was not first author will not count as a citation for the author. As was shown by the prolific author example for the two nanojournals, the most prolific authors tend not to be first authors. Therefore, they will not receive citation credit for most of their papers. The more accurate citation numbers could be obtained, but only for references that are themselves in the SCI, and then only with a laborious manual filtering process where each reference is downloaded individually (example shown later in this section). In essentially all the text mining studies performed by the first author, it was found that the lists of most prolific authors and most cited first authors were almost disjoint. While there are a number of reasons for this phenomenon, the lack of first authorship by most prolific authors is clearly an important reason.

In the results for most cited first authors, some of the researchers acknowledged to be major contributors to the development of nanotechnology (e.g., Smalley, Lieber) did not appear. One potential reason, under-representation due to limited first authorship, has been described above. To test this hypothesis, the following experiment was performed. The 232 most cited papers in the

Table 7. Most cited first authors

Author	Institution	Country	#Cites
Iijima, S	Nec Corp Ltd	Japan	1048
Dresselhaus, MS	MIT	USA	529
Wang, J	Nanjing Univ Technology	China	465
Ulman, A	Polytechnical University	USA	456
Saito, R	Tohoku Univ	Japan	455
Alivisatos, AP	Univ Cal Berkeley	USA	449
Chen, J	Nankai Univ	China	395
Murray, CB	IBM Corp	USA	392
Caruso, F	Univ Melbourne	Australia	380
Vaia, RA	AFRL	USA	367
Ajayan, PM	Rensselaer Poly	USA	357
Decher, G	Univ Strasbourg	France	318
Kong, J	Acad Sinica	China	314
Tans, SJ	Fom Inst Atom & Mol Phys	Netherlands	310
Huang, MH	Univ Cal Berkeley	USA	309
Inoue, A	Tohoku Univ	Japan	307
Nakamura, S	Teikyou Univ	Japan	303
Perdew, JP	Tulane Univ	USA	302
Chen, Y	Hefei Univ Technology	China	297
Zhang, Y	Peking Univ	China	295

retrieved database were downloaded individually from the SCI. This allowed all the authors to be represented on the download. Then, the most prolific authors in this database of 232 highly cited references were identified, and listed in frequency order. The top 20 authors on this list are shown in Table 8. Smalley, Lieber, and others who did not appear in the most cited first authors list now are listed in Table 8. This confirms the limited first authorship hypothesis, and shows that these acknowledged leaders participated in many highly cited papers, although not as first authors.

Not all the prolific authors of the 232 most cited papers were independent. Some of these authors functioned as groups, for multiple papers. A clustering analysis of the authors' relationships was performed. For example, Smalley, Rinzler, Colbert, and Nikolaev form a moderately close knit unit, as evidenced by their clustering. More detailed examination of their publication data shows significant co-publication, with Smalley being the central figure on many of the publications.

Of these 232 most cited papers, 66 were published in *Science*, 44 in *Nature*, 15 in *Physical Review Letters*, 10 in *Applied Physics Letters*, 10 in *Chemical Reviews*, and 9 in *Physical Review B*.

Most cited documents

There were 308,961 references listed in the retrieved papers. Essentially all the top tier most cited

Table 8. Authors of most highly cited papers

Author	#Papers	Author	#Papers
Smalley, RE	17	Vaia, RA	6
Lieber, CM	16	Dresselhaus, MS	5
Giannelis, EP	10	Eklund, PC	5
Rinzler, AG	9	Mirkin, CA	5
Alivisatos, AP	8	Duan, XF	4
Colbert, DT	8	Bawendi, MG	4
Thess, A	7	Murray, CB	4
Dai, HJ	7	Okada, A	4
Dekker, C	7	Peng, XG	4
Nikolaev, P	6	Gratzel, M	4

documents were published within the last decade, showing the dynamic nature of this discipline. These are the most recent references of any discipline examined in the first author's previous text mining studies. Additionally, only one of the authors in this tier, SS Fan (24th in the ranking), was listed at a Chinese institution. Thus, while the prolific author, institution, country lists, and most cited first author list show a substantial Chinese (country) representation, the most cited document list shows a minor Chinese (country) representation. While there are a number of reasons for this difference, one possible reason is that the citations received by Chinese authors are spread over many documents. A re-examination of the most cited documents in 3 or 4 years should show whether the large number of Chinese documents published currently are accompanied by adequate quality (as measured by citations).

Table 9 lists the ten highest frequency references in the 21,474 retrieved papers. The fields for each record, starting from left, are: Author (first author); Year (year of publication); Source (name of journal or book); Vol (volume number); Page (page number); #Cites-2003-Ref (frequency of reference in the 2003 retrieved database); #Cites – SCI-Tot (number of total citations listed in the Times Cited field of the SCI record). The narrative line (in bold italics) following the field listings for each record contains the record title (in parenthesis).

Seven of the ten references had first authors from the USA. *Science* and *Nature* journals accounted for eight of the first ten. Three articles focused on nanotubes, two on nanowires, two on nanocrystallites/quantum dots, and the remainder on surface-dominated applications (molecular sieves, self-assembled monolayers, and solar cells). The articles as a unit focused on demonstration of growth, fabrication, synthesis, and some small-scale device integration. Two authors were from industry, and the remainder from universities.

Most cited journals

There were 31,321 journals cited in the 21,474 retrieved papers. Table 10 contains a list of most cited journals. At the very top were *Phys Rev B* and *Appl Phys Letters*. On average, the most cited journals appear more fundamental than the most prolific journals, a trend that has been observed in other text mining studies as well. The distribution of journal disciplines is about the same in both the most prolific and most cited journals, focusing on Physics, Chemistry, and Materials, in that order. Eleven of the journals are in common between the two lists. There are no Chinese journals on either

Table 9. Most cited references

First author	Country	Year	Source	Vol	Page	#Cites 2003-Ref	#Cites SCI-Tot
Iijima, S	Japan	1991	Nature	V354	P56	730	4079
(Helical Microtubules of Graphitic Carbon)							
Alivisatos, AP	USA	1996	Science	V271	P933	249	1538
(Semiconductor Clusters, Nanocrystals, and Quantum Dots)							
Kresge, CT	USA	1992	Nature	V359	P710	213	3801
(Ordered Mesoporous Molecular-Sieves Synthesized by a Liquid-Crystal Template Mechanism)							
Thess, A	USA	1996	Science	V273	P483	196	1601
(Crystalline Ropes of Metallic Carbon Nanotubes)							
Murray, CB	USA	1993	JACS	V115	P8706	194	1317
(Synthesis and Characterization of Nearly Monodisperse (E=S, SE, TE) Semiconductor Nanocrystallites)							
Ulman, A	USA	1996	Chem Rev	V96	P1533	191	1534
(Formation and Structure of Self-Assembled Monolayers)							
Morales, AM	USA	1998	Science	V279	P208	177	772
(A Laser Ablation Method for the Synthesis of Crystalline Semiconductor Nanowires)							
Tans, SJ	Netherlands	1998	Nature	V393	P49	174	968
(Room-Temperature Transistor Based on a Single Carbon Nanotube)							
Oregan, B	Switzerland	1991	Nature	V353	P737	173	1878
(A Low-Cost, High-Efficiency Solar-Cell Based on Dye-Sensitized Colloidal TIO_2 Films)							
Huang, MH	USA	2001	Science	V292	P1897	170	529
(Room-Temperature Ultraviolet Nanowire Nanolasers)							

Table 10. Most cited journals

Journal	#Cites
Phys Rev B	27,936
Appl Phys Lett	27,281
Phys Rev Lett	20,000
J Am Chem Soc	17,127
Science	16,154
J Appl Phys	13,620
Nature	13,429
Langmuir	13,280
J Phys Chem B	10,038
Chem Mater	8415
J Chem Phys	7956
Macromolecules	7683
Adv Mater	7623
J Phys Chem-Us	6188
Chem Phys Lett	6133
Thin Solid Films	4804
Angew Chem Int Edit	4537
J Electrochem Soc	4501
Surf Sci	4024
Anal Chem	3608

list, implying that many Chinese authors are publishing in the more recognized international journals, where they are more likely to receive higher citations.

Taxonomy results

The first author's past text mining studies have used a variety of approaches to identify the main technical themes in the database. These include extracting key phrases and manually assigning them to categories; extracting key phrases and assigning them with statistical computer algorithm, using factor analyses and multi-link clustering; and grouping documents based on text similarity.

Both factor analysis and document clustering were used for the present study. Appendix 1 contains the factor analysis results. In document clustering, documents are combined into groups based on their text similarity. Document clustering yields numbers of documents in each cluster directly, a proxy metric for level of emphasis in each taxonomy category. For both the total SCI and EC databases, document clustering was performed using the Abstracts text only.

The clustering approach presented in this section is based on a partitional clustering algorithm (Zhao & Karypis, 2004; Karypis, 2005) 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.

CLUTO requires specification of the number of clusters desired. Cluster runs (using the retrieved 21,474 SCI records) ranging from 64 to 1024 clusters were generated, providing thematic details at different levels of specificity (resolution). CLUTO also agglommerated the 64 cluster results into a hierarchical tree (taxonomy) structure. This taxonomy is presented in some detail in the next sections. Appendix 3 of Kostoff et al. (2005b) contains the details of each cluster's contents. A 256 cluster run of the EC database was made, and a schematic of the EC taxonomy is shown in Figure 2, following the SCI taxonomy description in the next section.

Nanotechnology taxonomy

SCI

Based on the CLUTO output, a multi-level hierarchical taxonomy was generated of the 21,474 SCI retrieved records for 2003. Because a number of the component technologies of nanotechnology are quite different from each other, the hierarchical structure was converted to the following flat taxonomy (all categories at the same level) by modifying the fourth level of the hierarchical taxonomy slightly (combining some categories, splitting others to fifth level).

Each category is defined by its component themes, in bullets. The number preceding each category heading is the number of records in that category. The bullets listed under each category are the major themes within the category, and represent the themes of the elemental clusters within the category.

(2127) Polymers/Nanocomposites

- clays, emphasizing production of polymer-layered silicate nanocomposites from organoclays and montmorillonite-derived clays using melt intercalation

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
Carbon Nanotubes (CNTs) 1225	Structural Properties of CNTs (1044)	General Single-Wall CNT Theory/Measurements (159)	
		CNT Wall Properties-Chiral/Zigzag/Armchair (885)	Single-Wall CNT Wall Properties (679)
			Multi-Wall CNT Wall Properties (206)
	Growth & Alignment of CNTs (181)	CNT Alignment (by Plasma-Induced Fields) (72)	
		Growth of CNTs on Patterned Catalysts (109)	
Other Nanostructures (11411)	Nano-Photonics, Thin Film Deposition (6417)	Nano-Photonics/Nano-Optics (4573)	Quantum Dots (2541)
			ZnO Laser Optics (2032)
		Thin Film Deposition (1844)	Non-Carbon Thin Film Deposition (1293)
			Carbon Layer Film Deposition (551)
	Nanostructures for Informational, Biological, and Material Reinforcement Applications (4994)	Nanostructures with Electronic/Magnetic Application (IT) (2370)	Magnetic Properties of Nanostructures (708)
			Nanowire Properties/Fabrication (1662)
		Nanoparticles with Material Reinforcement or Biological Application (2624)	Colloidal Au/Ag Nanoparticles (Biological Application) (988)
			Polymer Nanocomposites (1636)

Figure 2. Four level hierarchical taxonomy – EC

- nanocomposites, mainly polymer, including fiber composites as well as nanoparticles embedded in matrices
- addition of block copolymers, or polymeric micelles, to promote self-assembly and improve material properties and structures

- polymers, especially on the molecular chain structures, and the structures and molecular weights of polymer aggregates in solution, especially water-based
- bonds and ligands among groups in complexes and compounds, with some emphasis on hydrogen bonds
- structure of crystals, emphasizing space group parameters

(1713) Particles/Nanoparticles

- nanoparticles, with primary emphasis divided between gold/noble metal nanoparticle mixtures and magnetic nanoparticles in magnetic fluids, and secondary emphasis on ZnO nanoparticles. Also addresses production of nanoparticles or nanobubbles by core-shell separation
- silver, especially nanoparticles (especially with core-shell nanostructures), colloids, particles, and determination of their structural, chemical, and electrical properties
- particles in fluids, especially colloids, typically a particle core with surfactant shell, and use of emulsions and microemulsions polymerization to generate these particles
- particles, especially nanoparticles, their size distribution, and properties of particle aggregates, especially magnetic

(2641) Nanowires, powders, and catalysts

- TiO2 (including titania colloids), especially for its photocatalytic activity, and examines electronic and metallurgical properties resulting from different fabrication techniques, including conversion of the anatase phase into rutile phase as a function of annealing temperature
- catalysts using very small particles, especially their deposition on carbon supports, and the nature of reactions at these small particle sizes
- porous materials, especially mesoporous silica structures generated with nanomaterial templates, and emphasizes pore size distribution of activated meso-carbon-microbeads
- nanowires, especially the fabrication and synthesis of nanowire arrays, and on evaluation of the geometric, structural, and electronic properties of these nanowires as a function of fabrication technique and parameters
- growth and fabrication of ZnO nanomaterials and nanostructures, especially nanobelts and nanorods, emphasizing structural determination with transmission electron microscopy
- nanorod and nanocrystal production through chemical reaction synthesis routes, and determination of the structural properties by transmission electron microscopy and x-ray diffraction
- powders, emphasizing sol–gel synthesis processes with different precursors for optimal growth, and parameterizing the effect of temperature on growth during the calcination process
- nanomaterial structures with emphasis on implants, emphasizing phases of crystals and amorphous materials, and especially their variation with thermal factors, such as annealing, growth, implantation, and synthesis temperatures

(1171) Materials

- alloys, especially relation of phase composition to magnetic properties of nanocrystalline alloys and amorphous alloys, and the tailoring of these properties by annealing
- high-energy ball milling to produce alloy powders, including effects on particle structure and phase of mill time, material composition, and annealing temperature
- grains, especially their size and boundaries, and how bulk crystalline properties depend on grain size, especially at nanometer levels
- coatings, and the effect of sintering on their properties, especially for Al_2O_3 and SiC powders and other structures, and Al_2O_3–SiC composites
- indentation, especially nanoindentation, and plastic deformation to measure mechanical properties of nanostructures, including stress–strain relationships, tensile strength, shear, ductility, and fracture

The following three segments deal with Surfaces, Films and Layers:
(2200) Thin films

- films, both thick and thin, and the variation of properties with film thickness, especially magnetic and dielectric properties
- thin films, emphasizing PZT films for application to high-density ferroelectric random access memory, and TiO_2 films for application to high efficiency solar cells, and further emphasizing films created by the sol–gel process
- films, especially thin films and their deposition on substrates, and parameters that affect their properties such as annealing

(1194) Self-assembled monolayers and gold electrodes

- self-assembly, emphasizing thiols because of their capability to form self-assembled monolayers (SAM) on noble and semi-noble metals, and examining the adsorption properties of thiols with various terminal groups
- monolayers, especially self-assembled surface monolayers, with some emphasis on alkyl monolayers, gold monolayers or gold substrates, and molecular chains in ordered and disordered monolayers
- surface adsorption, emphasizing proteins, monolayers, and molecules, and the use of scanning tunneling microscopy to characterize the adsorption process
- gold electrodes in electrochemical systems, typically coated with SAM for enhanced electrochemical performance, as well as deposition of gold nanoparticle films on surfaces for detection/sensing purposes

(1794) Surface layer modification

- layers, especially multi-layer oxides/SiO_2 on silicon-based substrates, emphasizing thick layers/coatings, factors affecting their deposition, and characterization of their interface properties
- ion bombardment, irradiation, and implantation of surfaces, examines the effects as a function of energy levels, dose, depth of penetration, fluence, and annealing
- growth of surface layers on substrates, including GaN layers, emphasizing epitaxial deposition, and the formation of islands and their parameter-dependent clustering
- etching of surface patterns, especially silicon-based films or crystals/wafers, and the relationship, and control, of surface roughness to increase etching resolution. Also focuses on AFM for both measuring surface roughness and wear, as well as performing the etching process
- proximal probe tip properties and dynamics, including cantilever dynamics and fabrication complexities, and the use of electron beam lithography for mask fabrication

(2352) Optics/Spectroscopy

- optics, especially nonlinear optical materials, and material refractive indices, especially for photonic crystals
- optical waveguides, including their gratings and optical fibers
- laser power and output, especially second harmonic generation from diode and optically pumped lasers
- pulsed lasers, emphasizing beam properties, and their use in characterizing optical properties of materials, nanofabrication of materials, and on materials for solid-state lasers
- luminescent and fluorescent emissions from excited energy states, emphasizing intensity, emission and absorption spectra, emission peaks, and photoluminescence
- molecular dynamics, emphasizing calculations of excited state energies, dissociation spectra, molecular energy transfer, electron vibrational energy and transitions, photon energy absorption, and molecular bonds
- radiation interaction with nanomaterials, emphasizing spectral bands, absorption bands, band gaps, especially at Raman and optical frequencies

(1255) Quantum dots

- exciton (electron–hole pair) states, especially in quantum dots
- quantum dots, emphasizing electronic states and energy levels, and growth mechanisms
- quantum dots, especially InAs, GaAs, CdSe QDs, emphasizing growth techniques, self-assembled layers, and photoluminescent properties
- InAs and GaAs, especially InAs quantum dots grown by molecular beam epitaxy on GaAs substrates

(1079) Magnetics

- tunneling, in tunneling junctions, especially magnetic tunnel junctions in magnetoresistance devices, with emphasis on Kondo states and Coulomb blockades
- spin, including spin-dependent electron scattering, spin–orbit interactions, spin channels, quantum dot spin states, quantum dot spin polarization, and electron spin resonance
- behavior of magnetic nanostructures in magnetic fields, including effect on spin, domain structures, and optical, magnetic, and mechanical anisotropies
- magnetic properties of nanomaterials and nanostructures, and the variation of these properties

with growth and treatment parameters, such as annealing

(1518) Solid state electronic structure/Properties

- GaN for light emitting diodes, and also includes AlGaN, InGaN, and AlN
- electroluminescent emitters and fabrication of light-emitting devices/diodes, with strong emphasis on determining and increasing efficiency
- gates for transistors and other electronic devices
- electrical properties and characteristics of nanomaterial structures, including voltage–current plots, electric fields, field emission, electrical conductivity, and electronic devices
- quantum wires, emphasizing energy states, and electrical conductivity and transport in one dimensional systems

(1624) Nanotubes

- single wall nanotubes, especially carbon, including growth of bundles and ropes, and determination of composition using Raman Scattering, as well as adsorption properties
- nanotubes, especially single-wall carbon nanotubes, and addresses properties of bundles, emphasizing zigzag and armchair nanotubes
- nanotubes, mainly carbon but including carbon nanotube composites and other nanotube materials as well. Emphasizes multi-wall nanotubes, their alignment, and their use as field emission devices
- multi-wall nanotubes, especially carbon, especially vertically aligned catalyzed chemical vapor deposition grown films, including use as glassy-coated film electrode
- carbon nanotubes, especially vertically aligned catalytically activated plasma-assisted chemical vapor deposited grown CNT, and examines their applications to field emission devices and field-effect transistors

(806) Nano-Bio-technology

- detection of proteins and inhibitors, emphasizing their active binding sites
- artificial and biological membranes, including their structure determination, and formation of the artificial membranes as well. Some emphasis was placed on nanoscopic structures using hydrated single lipids and lipid mixtures, where the nanostructures formed by these extruded vesicles/liposomes ranged from isolated unilamellar vesicles to flat sheet membranes.
- animal and solar cells, emphasizing the use of indicator dyes to enhance the photosensitivity of these cells, and both increase the efficiency of solar cells and use the luminescence as detectors for animal cells
- DNA, emphasizing oligonucleotides used in hybridization studies in order to detect and study specific nucleic acid fragments, such as single or double-strand DNA

EC

For the purpose of comparing the EC technical nanotechnology structure with that of the SCI, a hierarchical taxonomy of the EC nanotechnology literature was generated. The first four levels of the taxonomy are shown in Figure 2.

The first level of the EC taxonomy bears similarity to the second level of the SCI taxonomy. In both cases, Carbon Nanotubes form a separate major category, and are about 10% of the other nanostructure records (EC, 10.7%; SCI, 8.2%). At the fourth taxonomy level, the categories are quite similar. The EC has moderately more emphasis on fabrication, while the SCI has more emphasis of the fundamental areas such as excited emissions, band absorption, energy states, self-assembly, and DNA proteins.

Taxonomy observations

SCI

Relative to the other categories, Nano-Bio-Technology appears to be under-represented. This may be a real effect, or it may result from use of a query terminology different from that used by the biology research authors. This observation is supported by the absence of any biology journals in the top 20 most cited journals or top 20 journals containing most nanotechnology papers.

Also, the focal point of the total database is research and development to develop products using technology at the nanometer scale. This is essentially a *technology production database*, focused on the nanotechnology front end. There is almost no research on health effects (animal or human), environmental/climate impacts, security issues, vulnerability, synergistic effects from coupling with other new technologies, etc. Of the 64

Table 11. Potential applications

- Catalysts (Photo, Electro, Auto, Methane-Reforming, Heterogeneous/Immobilized Enzyme, Direct Bio-Electro, Bimetallic);
- Electrodes (Gold, Silver, Platinum);
- Semiconductors (Metal-Oxide, Cuprate, Amorphous, Polymer, Single Nanocrystal, Diluted Magnetic);
- Lithography (Soft, Etching, Dry Etching, Plasma Etching, Electrochemical Etching, Selective Etching, Electron Beam, Optical, Dip-Pen Nano, X-Ray);
- Storage (Hydrogen, Oxygen, Optical);
- Field Emission (Thermionics);
- Drugs (Delivery, Release);
- Switching;
- Solar (Cells, Photovoltaics);
- Superconducting (Nanowires, Tapes, Thin Films, High Temperature, Microbridges/Ultrafast Switches, Coulomb-Blockade Electrometers, Quantum Logic Gates);
- Recording (Magnetic Media, Heads);
- Waveguide (Optical);
- Transistor (Field-Effect, Mosfet, Single Electron, Thin Film, Organic, Quantum Dot);
- Capacitors (Mos, Super, Double-Layer);
- Detectors;
- Printing;
- Piezoelectric;
- Gene Delivery;
- Electrolyte;
- Wires (Quantum, Nanowire Arrays);
- Displays (Nematic Liquid Crystal, Flat Panel);
- Filters (Add-Drop, Chromatic Dispersion Reduction, Wavelength Division Multiplexing, Optical, Resonant Grating, Holographic Interference, Thermal Wavelength Tuning, Molecular Sieves);
- Insulators (Gate, Low K);
- Blood (Vessel Engineering, Serum Testing, Flowmeter, Catecholamine Monitoring);
- Holograph (Diffraction Gratings, Recording, Data Storage);
- Tribology (Lubrication, Solid Lubricants, Wear Rate/ Resistance, Friction Coefficient, Durability);
- Methanol;
- Ferroelectric;
- Lasers;
- Ceramics;
- Diodes (Light-Emitting);
- Resists (Photo);
- Sensing (Antibody, Bio);
- Circuits;
- Corrosion (Resistance);
- Enzymes (DNA Damage Detection, Glucose Sensing);
- Batteries (Rechargeable Lithium);
- Gate (Oxides, Logic, Mosfet);
- Fuel Cells;
- Membranes;
- Electrolytes (Polymer);
- Shape Memory;
- Quantum Computer;
- Memory (Random Access);
- Molecular Devices (Diodes, Wires, Memory, Switches, Data Storage)
- Optical Fibers;
- Magnets (Permanent, Ferro);
- Bone (Tissue Engineering, Implants, Fracture Repair);
- Environmental Protection (Waste Water Treatment, Air Purification)

elemental document clusters examined, none had themes or even critical phrases that addressed these important issues.

EC

The comments on the SCI taxonomy from the previous section apply here as well, especially as pertaining to the focus of the database. In both cases, more research on the back-end of nanotechnology production would add balance to the overall science and technology effort.

Potential applications

A taxonomy of potential applications was also generated manually. The keywords and phrases from the SCI Abstracts were inspected visually, and those relating to potential applications were extracted. They were categorized by visual inspection. Table 11 contains the applications categories, and examples of applications phrases within each category, for those categories that include phrases other than the category name. The technique of Citation Mining (Kostoff et al., 2001) would provide supplementary information on potential applications by examining papers that cite nanotechnology papers, and would complement the present approach.

Summary and conclusions

A text mining analysis of the nanotechnology literature was performed, consisting of a bibliometrics component for obtaining the infrastructure, and a computational linguistics component for obtaining the technical themes and their taxonomy structure. Abstracts as they appear in SCI were used to represent the basic research literature, and Engineering Compendex Abstracts were used to represent the technology/engineering literature.

Bibliometrics

There appear to be a large number of prolific authors with Asian names, far larger than in any of the first author's previous text mining studies, reflecting the large contributions from the Far East Asian countries (e.g., Japan, China, South Korea). The 20 journals containing the most nanotechnology papers tend to be in the technical disciplines of Physics, Chemistry, and Materials, with an emphasis on surface science. The top tier in volume of nanotech-related articles had three physics journals (*Applied Physics Letters*, *Physical Review*, and *Journal of Applied Physics*). Conspicuously absent are the biology journals.

Of the 20 most prolific institutions, 13 are universities, and the remaining 7 are government laboratories. Thirteen are from the Far East (corresponding to the large number of prolific authors from that region), three are from the USA, three are from Western Europe, and one is from Eastern Europe.

In 2004, three countries dominate in production of research papers: USA, China, and Japan; Germany is a strong contributor as well. In the top six countries, the three from the Western group (USA, Germany, France) have about 8% more publications than the three from the Far Eastern group (China, Japan, South Korea). However, studies have shown an English language bias for the SCI, and these Far Eastern publication numbers should be viewed as an under-estimate.

Overall trends between 1994 and 2004 were tabulated. The 2004/1994 ratio of nanotechnology papers is in double digits for the Far Eastern countries only (Peoples R China, South Korea, Taiwan, and Singapore). The 2004/1994 ratio of total SCI papers is above $\sim$4 for Far Eastern Asian countries only (Peoples R China, South Korea, Singapore), showing tremendous research interest and growth in nanotechnology in Far East Asia. The fractions of nanotechnology papers to total papers for 2004 above 8% are for Far Eastern countries only (Peoples R China, South Korea, Singapore). Thus, in the past decade, these Far Eastern countries have shown substantial growth in total SCI papers, in nanotechnology papers, and in the ratio of nanotechnology papers to total papers.

About half of the most cited first authors are from Far East, with most of the remainder being from the USA. Of the 232 most cited papers, 66 were published in *Science*, 44 in *Nature*, 15 in *Physical Review Letters*, 10 in *Applied Physics Letters*, 10 in *Chemical Reviews*, and 9 in *Physical Review B*.

Essentially all the top tier most cited documents were published within the last decade, showing the dynamic nature of this discipline. These are the most recent references of any discipline examined in the first author's previous text mining studies. Additionally, only one of the authors in this tier, SS Fan (24th in the ranking), was listed at a

Chinese institution. Thus, while the prolific author, institution, and country lists show a substantial Chinese (country) representation, the top tier cited document list shows a minor Chinese (country) representation.

Seven of the ten most cited references had first authors from the USA. *Science* and *Nature* journals accounted for eight of the first ten. Three articles focused on nanotubes, two on nanowires, two on nanocrystallites/quantum dots, and the remainder on surface-dominated applications (molecular sieves, SAMs, and solar cells). The articles as a unit focused on demonstration of growth, fabrication, synthesis, and some small-scale device integration. Two authors were from industry, and the remainder from universities.

The top tier of the most cited journals contained *Phys Rev B* and *Appl Phys Letters*. On average, the most cited journals appear more fundamental than the most prolific journals, a trend that has been observed in other text mining studies as well. The distribution of journal disciplines is about the same in both the most prolific and most cited journals, focusing on Physics, Chemistry, and Materials, in that order. Eleven of the journals are in common between the two lists. There are no Chinese journals on either list, implying that many Chinese authors are publishing in the more recognized international journals, where they are more likely to receive higher citations.

Computational linguistics

Two taxonomies (technology categorizations) of the SCI nanotechnology literature were generated: a hierarchical taxonomy for displaying the high level literature structure, and a flat taxonomy for displaying the detailed thrusts in each category. The flat taxonomy of the SCI nanotechnology literature contains the following categories: Polymers/Nanocomposites; Particles/Nanoparticles; Nanowires, Powders, and Catalysts; Materials; Thin films; Self-assembled monolayers and gold electrodes; Surface layer modification; Optics/Spectroscopy; Quantum dots; Magnetics; Solid state electronic structure/Properties; Nanotubes; Nano-Bio-Technology. A hierarchical taxonomy of the EC nanotechnology literature was generated. The first level of the hierarchical EC taxonomy bears similarity to the second level of the hierarchical SCI taxonomy. In both cases, Carbon Nanotubes form a separate major category, and are about 10% of the other nanostructure records. At the fourth taxonomy level, the categories are quite similar. The EC has moderately more emphasis on fabrication, while the SCI has more emphasis of the fundamental areas such as excited emissions, band absorption, energy states, self-assembly, and DNA proteins.

For the SCI taxonomy, relative to the other categories, Nano-Bio-Technology appears to be under-represented. This may be a real effect, or it may result from use of a query terminology different from that used by the biology research authors. This observation is supported by the absence of any biology journals in the top 20 most cited journals or top 20 journals containing the most nanotechnology papers.

Based on the SCI nanotechnology literature taxonomy, the focal point of the total database is research and development to develop products using technology at the nanometer scale. This is essentially a *technology production database*, focused on the nanotechnology front end. There is almost no research on health effects (animal or human), environmental/climate impacts, security issues, vulnerability, synergistic effects from coupling with other new technologies, etc. Of the 64 elemental document clusters examined, none had themes or even critical phrases that addressed these important issues.

For the EC taxonomy, the comments on the SCI taxonomy above apply here as well, especially as pertaining to the focus of the database. In both cases, more research on the back-end of nanotechnology production would add balance to the overall science and technology effort.

Finally, none of the published nanotechnology research literature surveys offer the background, infrastructure and technology structure of the nanotechnology literature, as described in the present paper. This additional information to the traditional literature survey/review offers a perspective on nanotechnology beyond what any individual or team of individuals can offer. Future literature surveys should contain both the traditional approach and the text mining approach. *Additionally, future nanotechnology text mining studies should include (see Kostoff et al.*, 2005b *for a more detailed description): (1) Expanded databases (e.g., Medline, DTIC Technical Reports, RADIUS, and Federal agency award databases); (2) Expanded*

queries (more phrases, broader phrases, other fields; see Kostoff, 2005a *for more details); (3) Discovery analyses (literature-based discovery, literature-assisted discovery; see Kostoff, 2005a for more details); (4) Expanded time frames for detailed trend analyses; (5) Larger numbers of clusters for finer resolution; and (6) Citation mining for identifying and tracking the myriad impacts and applications of nanotechnology research (Kostoff et al., 2001).*

Appendix 1 – EC and SCI factor analysis

Factor analysis of a text database aims to reduce the number of words/phrases (variables) in a system, and to detect structure in the relationships among words/phrases. Word/phrase correlations are computed, and highly correlated groups (factors) are identified. The relationships of these words/phrases to the resultant factors are displayed clearly in the factor matrix, whose rows are words/phrases and columns are factors. In the factor matrix, the matrix elements Mij are the factor loadings, or the contribution of word/phrase i (in row i) to the theme of factor j (in column j). The theme of each factor is determined by those words/phrases that have the largest values of factor loading. Each factor has a positive value tail and negative value tail. For each factor, one of the tails dominates in terms of absolute value magnitude. This dominant tail is used to determine the central theme of each factor.

Factor analyses were performed on the EC and SCI retrievals. Factor matrices ranging from 2 to 32 factors were generated, the main themes identified, and the themes were manually categorized into a hierarchical taxonomy. The SCI taxonomy is presented first, followed by the EC taxonomy.

SCI taxonomy

Level 1

- Instruments (XRD-TEM-SEM)
- Phenomena/Properties (Crystal structure)

Level 2

- Instruments (XRD-TEM-SEM; Differential calorimetry)
- Phenomena/Properties (Crystal structure; Surface adsorption (SAM/Film deposition))

Level 3

- Instruments (XRD-TEM-SEM; Differential calorimetry; AFM)
- Phenomena/Properties (Crystal structure; Surface adsorption (SAM/Film deposition); Photoluminescence (Quantum dots); Catalysis

EC taxonomy

For a two factor analysis, the main thrusts are:

(1) Films
(2) Nanocomposites–Clay/Differential calorimetry

For a four-factor analysis, the main thrusts are:

(1) Films (hardness, mechanical properties)
(2) Nanocomposites–Clay/Differential calorimetry
(3) Nanoparticle formation/reaction/catalysis
(4) Microstructure (Ni/Zr/C/B)

For an eight-factor analysis, the main thrusts are:

(1) Differential calorimetry/Nanocomposites–Clay
(2) Films (temperature/thickness/deposition)
(3) XRD/TEM (size, catalysis)
(4) Ni/Cu (alloys, Fe, Co)
(5) Hardness/Mechanical properties
(6) CNT
(7) SAMs
(8) Crystal structure

These results contrast the differences between the SCI and EC databases from the factor matrix perspective, as well as the differences between document clustering-based taxonomies and factor matrix-based taxonomies. The document clustering taxonomies are categorized essentially by structures (e.g., nanowires, nanotubes, nanoparticles, films) and phenomena (optics, magnetics). The SCI factor matrix taxonomies are characterized by instruments (XRD, TEM, SEM, AFM, differential calorimetry) and the quantities they measure (crystal structure, surface adsorption, photoluminescence). The EC factor matrix taxonomies are characterized by structures (films, nanocomposites, nanoparticles, microstructures).

At the first level of the factor matrix taxonomies, the science focus of the SCI, which concentrates on instrumentation and basic scientific

phenomena (crystal structure), is clearly seen. The technology focus of the EC, which concentrates on structures and materials (films, nano-composites-clay) is also evident.

At the second level, the science focus of the SCI remains the same, with additional instrumentation and measured phenomena shown. The EC focus continues on particles and microstructure. At the third level, the focus of the EC on structures and materials continues (CNT, SAMs, alloys, mechanical properties), but some of the applied research aspects begin to emerge (XRD/TEM, crystal structure).

References

Bhushan B., 2004. Springer Handbook of Nanotechnology. Springer.

Colton R.J., 2004. Nanoscale measurements and manipulation. J. Vac. Sci. Technol. B 22(4), 1609–1635.

Davidse R.J. & A.F.J. Van Raan, 1997. Out of particles: impact of CERN, DESY and SLAC research to fields other than physics. Scientometrics 40(2), 171–193.

Dowling A. et al., 2004. Nanoscience and Nanotechnologies: Opportunities and Uncertainties. The Royal Society and the Royal Academy of Engineering. 29 July.

Freitas R.A., 1999. Nanomedicine, Vol. 1: Basic Capabilities. Landes Bioscience.

Freitas R.A., 2003. Nanomedicine, Vol. 2: Biocompatibility. Landes Bioscience.

Garfield E., 1985. History of citation indexes for chemistry – a brief review. JCICS 25(3), 170–174.

Goddard W.A., D.W. Brenner, S.E. Lyshevski & G.J. Iafrate, 2002. Handbook of Nanoscience, Engineering, and Technology. CRC Press.

Goldman J.A., W.W. Chu, D.S. Parker & R.M. Goldman, 1999. Term domain distribution analysis: a data mining tool for text databases. Methods Inform. Med. 38, 96–101.

Gordon M.D. & S. Dumais, 1998. Using latent semantic indexing for literature based discovery. J. Am. Soc. Inform. Sci. 49(8), 674–685.

Greengrass E., 1997. Information Retrieval: An Overview. National Security Agency. TR-R52–02–96.

Hearst M.A., 1999. Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20–26.

Huang Z., H. Chen, Z.K. Chen & M.C. Roco, 2004. International Nanotechnology Development in 2003; country, institution, and technology field analysis based on USPTO patent database. J. Nanopart. Res. 6, 325–354.

Karypis G., 2005. CLUTO – A Clustering Toolkit. http://www.cs.umn.edu/~cluto.

Kostoff R.N., 1998. The use and misuse of citation analysis in research evaluation. Scientometrics 43(1), 27–43.

Kostoff R.N., 2003a. Text mining for global technology watch. In: Drake M. ed. Encyclopedia of Library and Information Science. 2nd edn. 4 Marcel Dekker, Inc., New York, NY, pp. 2789–2799.

Kostoff R.N., 2003b Stimulating innovation. In: Larisa V. Shavinina (ed.). International Handbook of Innovation. Elsevier Social and Behavioral Sciences, Oxford, U.K., pp. 388–400.

Kostoff R.N., 2003c. Bilateral asymmetry prediction. Med. Hypotheses 61(2), 265–266.

Kostoff R.N., 2005c. Systematic Acceleration of Radical Discovery and Innovation in Science and Technology. Fort Belvoir, VA: Defense Technical Information CenterDTIC Technical Report Number ADA430720 (http://www.dtic.mil/).

Kostoff R.N., J.A. Del Rio, E.O. García, A.M. Ramírez & J.A. Humenik, 2001. Citation mining: integrating text mining and bibliometrics for research user profiling. J. Am. Soc. Inform. Sci. Technol. 52(13), 1148–1156.

Kostoff R.N., H.J. Eberhart & D.R. Toothman, 1997. Database tomography for information retrieval. J. Inform. Sci. 23(4), 301–311.

Kostoff R.N., K.A. Green, D.R. Toothman & J.A. Humenik, 2000. Database tomography applied to an aircraft science and technology investment strategy. J. Aircraft 37(4), 727–730.

Kostoff R.N., J.S. Murday, C.G.Y. Lau & W.M. Tolles, 2005a. The Seminal Literature of Nanotechnology Research. DTIC Technical Report Number ADA435986 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. Also, an abridged version is published in this issue.

Kostoff R.N., M. Shlesinger & G. Malpohl, 2004b. Fractals roadmaps using bibliometrics and database tomography. Fractals 12(1), 1–16.

Kostoff R.N., M. Shlesinger & R. Tshiteya, 2004a. Nonlinear dynamics roadmaps using bibliometrics and database tomography. Int. J. Bifurcat. Chaos 14(1), 61–92.

Kostoff R.N., J.A. Stump, D. Johnson, J.S. Murday, C.G.Y. Lau & W.M. Tolles,. 2005b. The structure and infrastructure of the global nanotechnology literature. Fort Belvoir, VA: Defense Technical Information CenterDTIC Technical Report Number ADA435984 (http://www.dtic.mil/).

Kricka L.J. & P. Fortina, 2002. Nanotechnology and applications: an all-language literature survey including books and patents. Clin. Chem. 48(4), 662–665.

Losiewicz P., D. Oard & R.N. Kostoff, 2000. Textual data mining to support science and technology management. J. Intell. Inform. Syst. 15, 99–119.

MacRoberts M. & B. MacRoberts, 1996. Problems of citation analysis. Scientometrics 36(3), 435–444.

Narin F., 1976. Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.

Narin F., D. Olivastro & K.A. Stevens, 1994. Bibliometrics theory, practice and problems. Evaluat. Rev. 18(1), 65–76.

Schubert A., W. Glanzel & T. Braun, 1987. Subject field characteristic citation scores and scales for assessing research performance. Scientometrics 12(5–6), 267–291.

SCI. 2005. Science Citation Index. Phila., PA: Institute for Scientific Information.

Simon J., 2005. Micro- and nano-technologies: dullish electrons and smart molecules. Comp. Rendus Chim. 8(5), 893–902.

Swanson D.R., 1986. Fish oil, Raynauds syndrome, and undiscovered public knowledge. Perspect Biol. Med. 30(1), 7–18.

Swanson D.R. & N.R. Smalheiser, 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 91(2), 183–203.

TREC (Text Retrieval Conference), 2004. Home Page, http://trec.nist.gov/.

Viator J.A. & F.M. Pestorius, 2001. Investigating trends in acoustics research from 1970 to 1999. J. Acoust. Soc. Am. 109(5), 1779–1783Part 1.

Winkmann G., S. Schlutius & H.G. Schweim, 2002. Citation rates of medical German-language journals in English-language papers – do they correlate with the Impact Factor, and who cites?. Klin. Monats. Augenheilkunde 219(1–2), 72–78.

Zhao Y. & G. Karypis, 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. Machine Learn. 55(3), 311–331.

Zhu D.H. & A.L. Porter, 2002. Automated extraction and visualization of information for technological intelligence and forecasting. Technol. Forecast. Soc. Change 69(5), 495–506.