JSCIRES REVIEW ARTICLE

# Text mining for science and technology – a review part I – characterization/scientometrics

**Dr. Ronald N. Kostoff**

*Research Affiliate, School of Public Policy, Georgia Institute of Technology, 13500 Tallyrand Way, Gainesville, VA 20155*

## ABSTRACT

This article is the first part of a two-part review of the author's work in developing text mining procedures. The focus of Part I is Scientometrics. Novel approaches that were used to text mine the field of nanoscience/nanotechnology and the science and technology portfolio of China are described. A unique approach to identify documents related to an application theme (e.g., military-related, intelligence-related, space-related) rather than a discipline theme is also described in some detail.

**Keywords:** Scientometrics; Bibliometrics; Text Mining; Nanotechnology; China Science and Technology; Military-Related Technology.

## OVERVIEW

Text mining is the extraction of useful information from large volumes of text. Science and technology (S&T) text mining focuses on the S&T literature, mainly in electronic form. This review will address three major text mining sub-divisions: Characterization; Seminal Literature Review (SLR); Literature-Related Discovery and Innovation (LRDI), and will be divided into two separate papers. Part I, published in this inaugural issue of The Journal of Scientometric Research, will focus on Characterization, mainly its non-citation components. Part II, published in the second issue of the Journal of Scientometric Research, will focus on the citation component of Characterization, the citation-based SLR, and the citation-enabled LRDI.

Characterization is the assignment of metrics to the technical literature of interest to identify patterns that will increase understanding of the topical matter. These metrics may include: 1) bibliometric quantities such as key authors, journals, institutions, countries; 2) the relationships among these bibliometric quantities, such as institutions whose authors frequently co-publish; 3) technical quantities such as key technical thrusts and the relationships among the thrusts. While isolated metrics may have specific uses, the key challenges are to identify 'signatures', or combinations of metrics, that provide unique insights into technology literatures or to countries' technology portfolios.

SLR is identifying the key papers that provide the intellectual heritage of a topic or discipline, and relating these documents in narrative form to display the evolution and breadth of the discipline. The Citation-Assisted Background (CAB) approach (Kostoff and Shlesinger, 2005b) was developed to assist the SLR process, but CAB and SLR require substantial human judgment to supplement the mechanistic algorithms developed.

LRDI relates disparate literatures to identify novel concepts that provide value-added to problem solving. LRDI has been mainly applied to challenging medical problems, but can be applied to the technical non-medical literature as well (Kostoff et al, 2008a; Kostoff, 2012a). Its main operational modes are: 1) start with a problem and search for a solution; 2) start with a technology and search for an application; 3) start with two problems and search for common features; 4) start with a problem and a solution, and search for the mechanism(s) by which the solution addresses the problem. Other variants are possible as well.

***Corresponding author.**

E-mail: rkostoff@gmail.com
571-248-2661

**DOI:** 10.5530/jscires.2012.1.5

## ANALYSIS

### Characterization

This review focuses on text mining methodologies developed by the author's research group. There were two main types of characterization text mining studies, based on sponsor interest. One type characterized a technical discipline/technology, and the other type characterized a country's S&T portfolio. The main difference in the actual study mechanics was in the query for retrieving the database to be analyzed.

For the country studies, the address field of the records in the database(s) was examined, and if at least one author of a document was listed as being from the country of interest, that document was included in the retrieval. Given researcher mobility among institutions, permanently or transiently, there was obviously some ambiguity introduced by this approach, but for examining large numbers of records, there were no feasible alternatives. In addition to the two types of characterization studies above, targeted scientometric studies were performed, mainly searching for messages contained in citations. Selected examples from all these studies will be presented.

For the technology studies, an iterative relevance feedback technique developed in the 90s (Kostoff et al, 1997a) was used to generate the query for retrieving documents. The iterative procedure started with postulating a simple test query, which was then inserted into the database search engine The retrieved records were examined for relevance and associated patterns, the test query was then modified with these patterns, and the process was repeated until convergence (few new relevant records retrieved with added iterations).

In both the technology and country characterization studies, similar core metrics were evaluated: key authors, institutions, journals, countries, etc, key research thrusts and the relationships among those thrusts. This review will provide a few of the more interesting findings from the 2006 nanotechnology study (Kostoff et al, 2007e, 2011a) and the 2009 China study (Chen et al, 2009). It will also provide interesting findings from a text mining study that doesn't fit neatly into any one of the categories described above, namely, the Military Relevant Technologies study that straddles scientometrics query development and LRDI (Kostoff and Bhattacharya, 2010). For details on technology characterization studies other than those mentioned above, the following references are recommended: Electrochemical Power Sources (Kostoff et al, 2002); Nonlinear Dynamics (Kostoff et al, 2004a); Fractals (Kostoff et al, 2004b); Power Sources (Kostoff et al, 2005a); Anthrax (Kostoff et al, 2008c); SARS (Kostoff and Morse, 2011b). For details on country characterization studies, the following references are recommended: Mexico (Kostoff et al, 2005c); Finland (Kostoff et al, 2006a); China/India (Kostoff et al, 2007a, 2007b, 2007c, 2007d); Brazil (Schoeneck et al, 2011).

#### *Nanotechnology-2006*

In the 2004-2005 time frame, the author's group conducted an initial global nanotechnology study based on 2003 Science Citation Index/Social Science Citation Index (SCI)/(SSCI) data (Kostoff et al, 2006c). Because of the heightened world-wide interest in nanotechnology at the time, and specific interest in the 2003 study initial results, the author's group was encouraged to conduct an even larger and more comprehensive study. In 2006, the 90+ term nanotechnology and nanoscience query (used in the previous nanotechnology study) was expanded to more than 300 terms. The bulk of the query (used for the 2006 study) consisted of technical terms generated by the iterative relevance feedback technique (Kostoff et al, 1997a) described above for insertion into the SCI/SSCI Topic field. Two additional fields were accessed for the remainder of the 2006 query. All journals with nano* in their title were retrieved using the Source field, and all institutions with nano* in their address were retrieved using the Address field. At the time, the full 2006 query was the most comprehensive nanotechnology query extant. This expanded query was inserted into the SCI/SSCI search engine for the year 2005, and resulted in retrieval of approximately 65000 records (Articles and Reviews only).

In the analysis, the core metrics listed above were examined, including most prolific authors, journals containing most nanotechnology papers, institutions and countries producing most nanotechnology papers, etc. In some of the cross-plots that included journals (e.g., institution-journal matrices), journal Impact Factors were incorporated. For example, the leading five institutions based on the number of nanotechnology publications were listed along with the five journals in which they published nanotechnology articles most frequently in 2005. The journals and institutions were shown with their Impact Factors and the number of articles published. Then, an average Impact Factor was calculated for each institution as a weighted average of the five Impact Factors and numbers of publications listed. This weighting approach was a first step in introducing 'quality' to the publication metrics.

*One feature that was not used in this study, but used in later non-nanotechnology studies with smaller databases, was citation searching. After retrieving articles with the text-based query, the most relevant articles were identified, and then the citation network around each article was searched for more relevant articles. Thus, for relevant article x, its references, citing papers, and an SCI field called Related Records (all records in the database that have at least one reference in common with article x) would be searched. This allowed identification of articles that had* ***concepts*** *of interest without having the* ***exact words*** *of interest.*

A substantial amount of categorization was performed. Institutions were grouped to see which ones collaborated heavily, countries were grouped for the same purpose, and records of similar technologies were grouped to understand the main technology thrusts and the relationships among the thrusts. Three main grouping approaches were used: document clustering, factor analysis, and correlation mapping. These approaches were always validated with the tried and true grouping by visual inspection. Each grouping approach provided a complementary perspective on the relationships, and all grouping approaches were included in the final narratives.

In document clustering, used mainly for identifying major technical thrusts, documents are combined into groups based on their text similarity. Document clustering yields the numbers of documents in each cluster directly, a proxy metric for level of emphasis in each taxonomy category. The specific document clustering software used (CLUTO, an abbreviation for CLUstering TOolkit) provides three different classes of clustering algorithms that operate either directly in the object's feature space or in the object's similarity space (Zhao and Karypis, 2005). These algorithms are based on the partitional, agglomerative, and graph-partitioning paradigms. A key feature in most of CLUTO's clustering algorithms is that they 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.

Two hundred and fifty-six individual clusters were chosen for the database of 65,000 nanotechnology articles and reviews retrieved from the SCI/SSCI, and are presented in detail in (Kostoff et al., 2007e). The clustering algorithm also agglomerated the 256 clusters in a hierarchical tree (taxonomy) structure, and this taxonomy (first four levels) is shown in Figure 1.

| | | | |
|---|---|---|---|
| Quantum Phenomena, Optics, Electronics, Magnetism, Tribology, and Films (32983 Rec) | Quantum Phenomena, Optics, Electronics, Magnetism, and Tribology (26077 Rec) | Quantum Phenomena (3326 Rec) | Quantum Dots (2028 Rec) |
| | | | Quantum Wells, Wires, and States (1298 Rec) |
| | | Optics, Electronics, Magnetism, and Tribology (22751 Rec) | Optics and Electronics (16432 Rec) |
| | | | Magnetism and Tribology (6319 Rec) |
| | Films (6906 Rec) | Thin Films (4760 Rec) | Properties of Thin Films (2251 Rec) |
| | | | Applications of Thin Films (2509 Rec) |
| | | Deposition of Films (2146 Rec) | Deposition of Thin Films (1752 Rec) |
| | | | Diamond Films (394 Rec) |
| Nanotubes, Nanomaterials, Nanoparticles, Polymers, Composites, Metal Complexes, and Bionanotechnol -ogy (31742 Rec) | Nanotubes (3211 Rec) | Multi-walled Nanotubes (2350 Rec) | Applications of Carbon Nanotubes (474 Rec) |
| | | | Multi-walled Nanotubes (1876 Rec) |
| | | Single-walled Nanotubes (861 Rec) | Single- and Double-walled Nanotubes (447 Rec) |
| | | | Single-walled Nanotubes (414 Rec) |
| | Nanomaterials, Nanoparticles, Polymers, Composites, Metal Complexes, and Bionanotechnology (28531 Rec) | Nanomaterials, Nanoparticles, Polymers, Composites, and Metal Complexes (22686 Rec) | Nanomaterials and Nanoparticles (14263 Rec) |
| | | | Polymers, Composites, and Metal Complexes (8423 Rec) |
| | | Bionanotechnology (5845 Rec) | DNA (775 Rec) |
| | | | Proteins and Cellular Components (5070 Rec) |

**Figure 1.** Four level hierarchical taxonomy – nanotechnology.

Figure 1 is a four level hierarchical taxonomy of the global nanoscience and nanotechnology literature. In each succeeding level, the categories are bifurcated. Categories with no shading are those in which the USA has the most publications. Categories with solid shading denote China publication lead, and categories with vertical bar shading denote Japan publication lead. Light shading means category leader has 100-125% of USA publications (denoted by 'modestly' in results below); medium shading 125-150% (denoted by 'noticeably' in results below); dark shading >150% (denoted by 'strongly' in results below).

In the first level (leftmost column), the total retrieved records are divided into two technical categories. One category (Quantum Phenomena, Optics, Electronics, Magnetism, Tribology, and Films) focuses mainly on physical phenomena, whereas the other category (Nanotubes, Nanomaterials, Nanoparticles, Polymers, Composites, Metal Complexes, and Bionanotechnology) focuses on materials and structures. The two categories are about the same size.

The primarily phenomena category sub-divides into two categories, with the larger category (phenomena) being roughly four times the size of the smaller category (films). The materials and structures category likewise divides into two asymmetric categories, with the smaller sub-category focusing on nanotubes and the nine times larger category focusing on all other structures and materials. China has a modest publications lead in this latter category.

At the fourth level, China out-publishes the USA in:

- Properties of Thin Films (modestly, 2251 rec)
- Diamond Films (modestly, 394 rec)
- Applications of Carbon Nanotubes (strongly, 474 rec)
- Multi-Walled Nanotubes (modestly, 1876 rec)
- Nanomaterials and Nanoparticles (noticeably, 14263 rec)
- Polymers, Composites, and Metal Complexes (noticeably, 8423 rec)

Also at this level, Japan out-publishes the USA in Deposition of Thin Films. Note that identification of these islands of strength of USA competitors required accessing more detailed levels of the taxonomy.

In addition, fuzzy document clustering was run to identify medical applications. This fuzzy version allowed multi-theme records to be assigned to multiple categories. This is a particularly valuable feature for records that contain both research and applications themes, but is useful for any types of records that contain more than one thrust area. The detailed medical applications can be found in (Kostoff et al, 2007e; Kostoff et al, 2011a; Kostoff et al, 2008a).

Factor analysis of a database aims to reduce the number of variables in a system, and to detect structure in the relationships among variables. Correlations among variables are computed, and highly correlated groups (factors) are identified. The relationships of these variables to the resultant factors are displayed clearly in the factor matrix, whose rows are variables and columns are factors. In the factor matrix, the matrix elements Mij are the factor loadings, or the contribution of variable i (in row i) to the theme of factor j (in column j).

The theme of each factor is determined by those variables 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 typically dominates in terms of absolute value magnitude. This dominant tail is used to determine the central theme of each factor. Factor analysis was used to quantify word or phrase, institution, and country collaborations. It was used extensively in this nanotechnology study, and in most of the characterization studies on other topics as well.

An autocorrelation function describes the correlation between a random function and a copy of itself shifted by some 'lag' distance. One can produce a map showing terms that commonly occur together. For example, an autocorrelation map of institutions shows teams of institutions that publish together. A cross-correlation map shows relationships among items in a list based on the values in another list. A cross-correlation map of institutions and phrases can show groups of organizations that write about the same things i.e., publish on the same technical themes. A cross-correlation map of countries and phrases can show groups of nations that write about the same things. Both types of maps were used extensively in the nanotechnology study, and in some of the later characterization studies on other topics.

There were two advances of note in this study; both will be illustrated.

*Identifying new categories*

The various grouping approaches listed above tend to group by disciplines, emphasizing those determined by the algorithms. But, some categories of interest may be applications-oriented, consisting of many disciplines (e.g., military relevant technologies), or may be non-applications non-discipline oriented (e.g., instrumentation),

and these categories may not be readily obtainable by the grouping approaches above.

After the 65000 records were retrieved, and some initial grouping analyses had been performed, a list of categories that was deemed important but could not be identified sharply using the above grouping approaches was generated (e.g., instrumentation, materials, properties, phenomena, nanostructures, non-medical applications). To populate each category with relevant phrases, the following approach was taken. One of the options in the TechOasis text analysis software used was identification of all single, double, triple, and quadruple adjacent word phrases in the text being analyzed (all phrases beginning or ending with 'stop-words' from a pre-determined list were eliminated), along with the frequency of each phrase's appearance in the text. The list of phrases and their frequencies was generated (phrase frequency analysis), and the 60,000 highest frequency phrases were inspected visually. Phrases deemed relevant to any of the pre-selected categories were assigned manually to that category. While this procedure was manually intensive and time-consuming, the resulting database of categories and their phrases was unmatched by any other in the field. Having such a database allowed cross-plots of this data with each other, with existing technical categories, and with existing bibliometric categories. Because of modifications that had been made to the CLUTO document clustering software, phrases for any of the defined categories of interest (e.g., instrumentation) could be entered by themselves only, and clustering could be performed for only those records that addressed e.g. instrumentation. Figure 2 shows a four-level hierarchical taxonomy for nanotechnology instrumentation (Kostoff et al, 2007f). As can be seen from Level 1 (left-most column), there were only ~27,500 records related to nanotechnology instrumentation out of the 65,000 total. For all the cross-plots and cross-correlation maps that were made with these and other categories, the interested reader is advised to see (Kostoff et al, 2011a; Kostoff et al, 2007e).

### *Optimal display frequency ranges*

This is an area that has received little attention in scientometric analysis. In many cross-plots and network maps, so much information is presented as to make the display almost unintelligible. In the nanotechnology study, it was found that, by using selected portions of the frequency spectrum for plotting, much sharper and intelligible

| | | | |
|---|---|---|---|
| **AFM, NMR, Calorimetry (8423)** | **NMR, RS, Calorimetry (4684)** | NMR, Complexes, Compounds (1546) | **NMR Spectroscopy (306)** |
| | | | NMR, Complexes, Compounds (1240) |
| | | **RS, Calorimetry (3138)** | DSC (1138) |
| | | | **Raman Scattering, RS, AFM (2000)** |
| | **AFM (3739)** | **AFM, Films, Tip, Imaging (2003)** | **AFM, Film, Tip, Imaging (1055)** |
| | | | **AFM, Film, Substrate, Deposit (948)** |
| | | **AFM, Films, Deposition, Growth, Substrate (1736)** | **AFM, Film, Deposit, Substrate, Growth (1511)** |
| | | | **AFM, Magnetic (226)** |
| EM, XRD (19090) | **EM (4492)** | **TEM (2545)** | HRTEM (296) |
| | | | **TEM (2249)** |
| | | **SEM, Films, Composites, Particles, Cells (1947)** | **SEM, Film, Particle, Cell (1652)** |
| | | | **SEM, IS (295)** |
| | XRD, Films (14598) | SEM, XRD, Films, Coatings, Composites (3634) | SEM, XRD (1451) |
| | | | SEM, Film, Coating, Deposit, XRD (2183) |
| | | XRD, TEM, Thin Films (10964) | TEM, Film, Particle, Nanoparticle, STM (5986) |
| | | | Film, XRD, XPS (4978) |

**Figure 2.** Four level taxonomy - nanotechnology instrumentation.

displays were possible, and much more information could be transmitted as well. One example will suffice.

Figure 3 is a cross-correlation map of institutions and cited journals. It shows how institutions are linked indirectly through their direct linkages with journals their researchers cite in publications. The thirty institutions with the most publications were cross-correlated with the 500 journals cited most frequently, and the resultant 'spaghetti' map on Figure 3 made detailed relationships difficult to ascertain. The readers may want to contemplate how many network plots they have seen that are almost unintelligible due to the density of data presented.

Figure 4 is a cross-correlation map of the top 30 institutions with the next 500 cited journals, and it provides a much clearer picture of linkages that exist. Figure 4 shows four clusters based on nationality: one Chinese, one Japanese, one American, and one European. Another point to note from Figure 4 is that the Chinese group is isolated from the other institutions, whereas the Japanese and the American institutions link to the European research centers, both through CNRS. Figure 3 is essentially a demonstration in visual terms that all the institutions are citing the same most cited journals heavily (almost by definition), and because of the citation level and generality of these highly cited journals, this portion of the cited journal frequency spectrum is not optimal for helping to identify institutional relationships. Figure 4 shows much better discrimination possible when operating at a mid-frequency spectrum, where more focused but less cited journals help to emphasize institutional relationships based on common technologies.

Other cross-correlation variables were examined, such as phrases, and similar results to those above were found. There is obviously a role for using the highest frequency cited journals, phrases, etc, in the text mining analyses, but there is also a role for mid and perhaps even low frequency ranges of these variables.

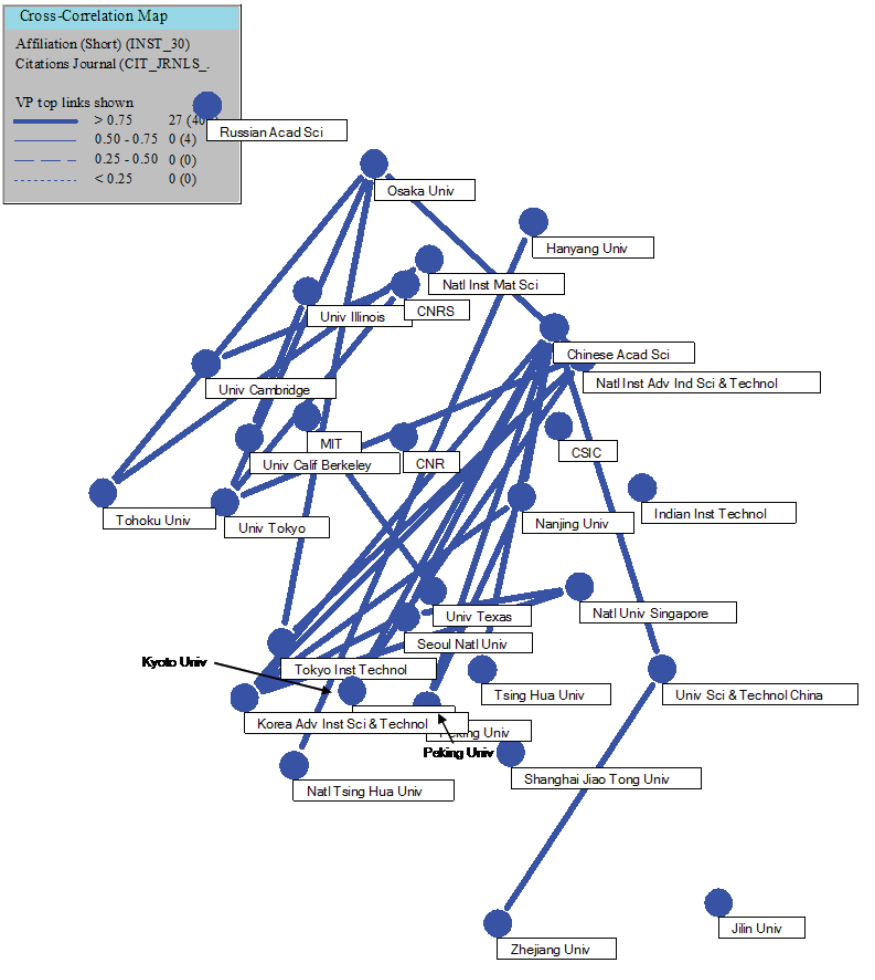


**Figure 3.** Institution-cited journal cross-correlation map (2005) (cited journals 1-501).

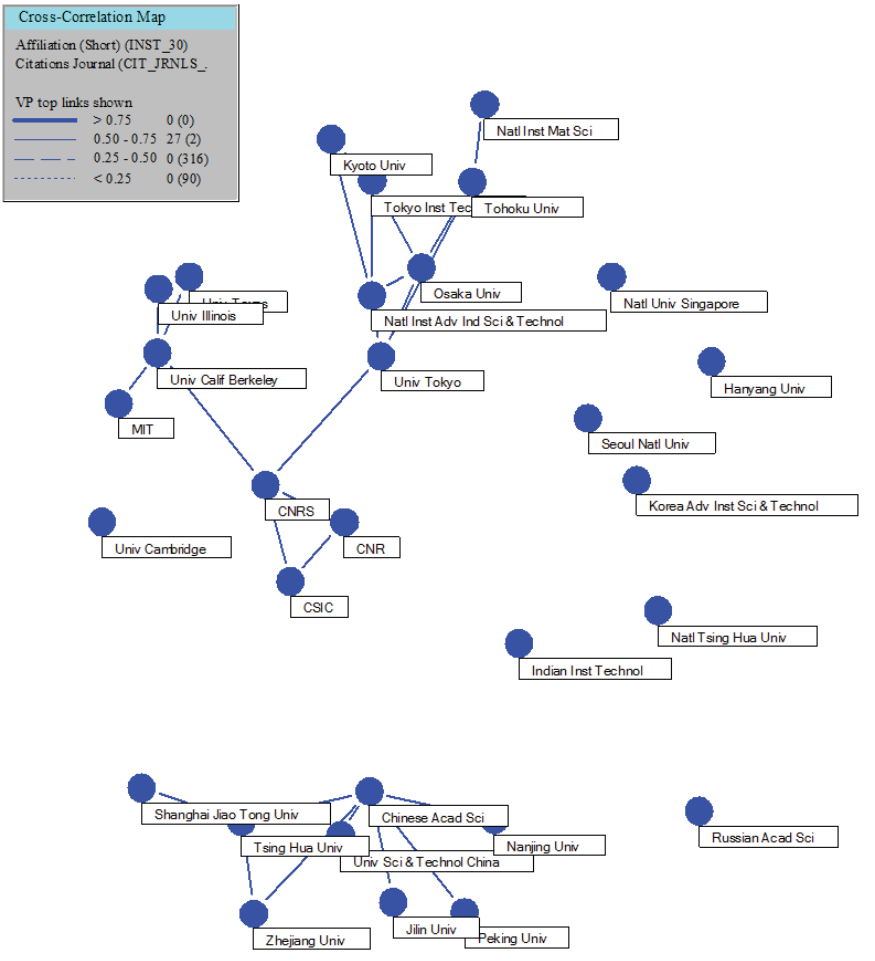


**Figure 4.** Institution-cited journal cross-correlation map (2005) (cited journals 502-1003).

## *Comparison of China's S&T Program to USA*

S&T assessment at the nation-state level is important from many perspectives. It can provide some understanding of a nation's military potential, which is useful for defense planning. It can also provide understanding of a nation's commercial potential, which is useful for competitiveness. Finally, it can identify areas of S&T that can be leveraged and coordinated for mutual benefit.

What are the central principles in conducting an S&T assessment? In The Handbook of Research Impact Assessment (Kostoff, 1997b), three foundational S&T assessment metrics are identified, whether for a project, a program, or a nation's total S&T output. These three metrics can be summarized as right job, job right, and productivity/progress. 'Right job' addresses the overall investment strategy: Are the larger S&T objectives being addressed correctly? 'Job right' addresses the S&T approach: Are the best techniques being used to conduct the S&T? 'Productivity/progress' addresses the S&T output and impact.

In this high level S&T assessment, examples were provided of how to use these metrics to rapidly assess the S&T of a scientifically growing country—namely, the People's Republic of China. To place the assessment in context, China's metrics were compared with those of the leader in S&T output, namely, the USA. Much more detailed expositions of the use of these metrics in assessing China's S&T output are available elsewhere (Kostoff et al, 2006b, 2007a, 2007c, 2008d).

### *Right job*

S&T strategy, as reflected in published technical output in the global literature, can be inferred from different perspectives. Clustering documents by technical discipline provides one categorization approach, and it is perhaps the main approach used.

A complementary approach is to show relative areas of technical emphasis among multiple countries. The SCI includes a Subject Category field for each record—that is, for each article published. This field indicates the main technical discipline for the journal in which the article was published. In 2009, the Subject Category distribution for the 100,000 most recent articles (ending 31 December 2008) published in the SCI from China and the US was examined. The Subject Categories and their frequencies were downloaded. For each of almost 500 categories, the ratio of China's frequency to that of the US was computed, and then the list was sorted according to the China/US ratio.

The results, which have been replicated by other means and for other databases, show China's strong relative emphases in the physical and engineering sciences and the US emphases in the biomedical, social, and psychological sciences (Chen et al, 2009). If these results are coupled with China's strong production of technical graduates, then China's investment strategy is providing a solid technology-based foundation for future military and commercial competitiveness.

### *Job right*

The second metric addresses research quality. The only universally accepted indicator of publication quality is a panel of experts reviewing a specific document. One commonly used proxy metric for quality is the number of times other research articles cite an article. The citation trend of China's published articles in nanotechnology, an area of strong emphasis in Chinese research, was examined (Kostoff et al, 2008d). The citation quality (percent of publications in the top citation tier) was low relative to that of the US, but it grew monotonically within a five-year period—from 4 percent of the US figure in 1998 to 20 percent in 2003, the latest period examined.

Another approach to assessing relative quality is to examine publication trajectories in high-quality journals. For these journals, articles must exceed a quality threshold to be accepted. There were three criteria used for selecting journals to include in this assessment: high total citations, high citations per paper, and focus on specific physical science disciplines. The ratios of the number of Chinese to USA articles published in two important SCI/SSCI journals—namely, the Journal of the American Chemical Society (JACS) and the Journal of Applied Physics (JAP), were compared (Chen et al, 2009).

Over the past decade, the China/USA ratio for total SCI/SSCI nanotechnology articles grew by about a factor of eight; the ratio for JACS articles grew by an order of magnitude, and the ratio for JAP articles grew by more than a factor of five. These quality findings reflect results from earlier studies (Kostoff et al, 2006b, 2007a, 2007c, 2008d). However, those studies also showed many Chinese articles being published in low-Impact-Factor journals. From the 2009 study (Chen et al), it was concluded that a small high-quality component is paralleling rates of increase that match the overall growth in Chinese technical literature.

### *Productivity/progress*

By any measure, China's productivity in published technical papers over the past two decades has been astounding. The bottom curve in Figure 5 (Chen et al, 2009) shows outstanding relative total publication growth. The absolute publication growth numbers are equally impressive. However, aggregate statistics have limited value for operational decision making. For bibliometrics, specific investment spikes must be identified to infer the true importance of an investment strategy.

Figure 5 provides an example of what can be derived from different levels of aggregation. The bottom curve, showing the overall China/USA publication ratio, indicates that China lags the USA in total SCI publications by a factor of three. The middle curve (ratio of overall nanotechnology publications) shows relative growth similar to the overall relative growth pattern, albeit starting at a higher relative level due to China's emphasis on nanotechnology. By this metric, China has essentially obtained parity with the USA in overall nanotechnology publication production (in fact, a recent analysis shows China to be about twenty percent ahead of the USA in nanotechnology publications at the end of 2011 (Kostoff, 2012b)). The top curve, for the important nanotechnology sub-area of

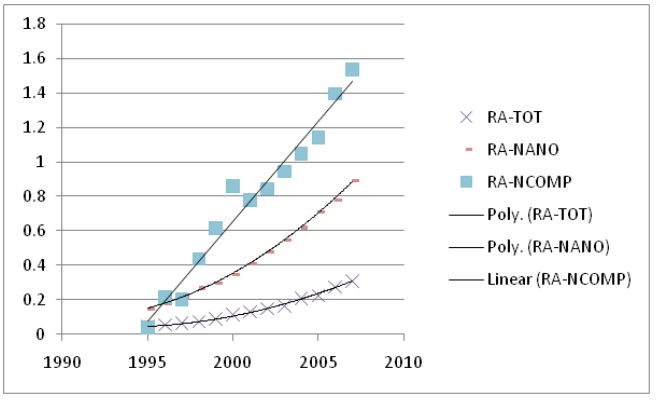


**Figure 5.** Ratio of China/USA articles in nanotechnology at different aggregation levels.

nanocomposites, shows a substantially higher (and linear) rate of ratio increase relative to the other two curves. By this metric, China is 60 percent ahead of the USA in nanocomposites publication production (increased to 130% ahead of the USA in nanocomposites publications by the end of 2011 (Kostoff, 2012b)). At this level of detail, the analyst can examine specific investment spikes, such as nanocomposites, and start to connect the dots to identify the investment strategy priorities on an integrated basis.

S&T assessment at the project, program, or nation level can be very valuable. However, the analyst must be judicious in selecting the appropriate metrics to evaluate the investment strategy, research approach, and productivity, and the appropriate level of aggregation.

### *Identification of Military-Related Technologies*

The goal of this study was to identify the portion of a country's S&T portfolio that could be considered military relevant. Military relevant technologies cannot typically be identified from a large group of technologies using document clustering or factor analysis. These grouping methods tend to provide ***discipline*** breakdowns, whereas military-relevant, space-relevant, intelligence-relevant, etc., have ***application*** orientation.

In the first step of a two-step approach, an iterative relevance feedback approach (Kostoff et al, 1997a) was taken to generate a query to retrieve military identifiable documents. These are documents that are unmistakably military-oriented, containing terms like ***artillery, aircraft carrier, fighter aircraft, weapons of mass destruction***, etc. In the second step, an iterative relevance feedback approach was taken to generate a query to retrieve strongly military-related documents. These are documents that may or may not contain the military-identifiable terms, but the technologies within these documents are those that are strongly associated with the military-identifiable documents.

In the first step, terms like 'military' were used to query the Abstract and Keyword fields (in some cases, controlled vocabulary fields as well), records were retrieved, phrases and phrase combinations that were unmistakably military (e.g., artillery, aircraft carrier, etc) were extracted, added to the query, and the process was repeated until convergence. Initially, organization names (e.g., XXX military academy) were queried as well, but many of the records retrieved were not sufficiently military-identifiable, so the organization/address field was not accessed further.

The second step proved to be much more difficult. The objective was to generate a query to retrieve strongly military-related documents. The approach selected was to identify technology text patterns that occurred in the military-identifiable records from the first step, and to add these text patterns to a query that would retrieve records containing military-relevant technologies. Unfortunately, there were many types of technology text patterns that could be extracted from the military-identifiable records, and these different text patterns resulted in different strengths of relationships among the technologies in the records retrieved and their linkages to the military application.

For example, the simplest pattern was a list of technology phrases extracted from the military-identifiable records by straight-forward phrase frequency analysis. This list was obtained by visual inspection of all phrases contained within the military-identifiable records, above a pre-selected threshold frequency of occurrence. Selected phrases were entered into the Ei Compendex database search engine, and the records retrieved were sampled for military relevance.

In many cases (e.g., information technology, neural networks, signal processing), only a small fraction of the retrieved records had a direct relationship to the military application, although one could argue that e.g. 'signal processing' expertise focused on a non-military application could be readily transferable to a military application. The retrieval volume was quite large, given the broad coverage of the technologies identified.

The next simplest pattern was a variant of the first, whereby more detailed stand-alone phrases were extracted. Thus, instead of extracting 'signal processing' as above, a specific variant like 'radar signal processing' that might be more targeted to a military application was extracted. A query spanning six pages in length was generated, consisting of ~1600 words or approximately 500-600 long multiword phrases. Again, hundreds of thousands of records were retrieved, with perhaps a slightly higher fraction of military-relevant documents, but still relatively low.

The final two patterns examined were phrase combinations. The first combination approach was inspired by the LRDI methodology (Kostoff, 2012a, 2012b), whereby document clustering or factor analysis was performed on the military-identifiable documents, and combinatorials of the key technologies in each cluster or factor were used as a query. This approach yielded records somewhat closer to what was desired, but still of insufficient military specificity.

The second combination approach (which was used in the study) was derived as follows: A list of the key technologies in the military-identifiable records was generated by visual inspection of the phrases in these records (e.g., 'signal processing', 'information technology', 'synthetic aperture radar', 'neural network', 'wireless networks', etc). A second list of desired functions or actions of more specific technologies was generated by visual inspection of the phrases in these records (e.g., 'target identification', 'intrusion detection', 'obstacle avoidance', 'target recognition', 'active jamming', etc). The two lists were matrixed against each other, and the potential combinations reflected by the contents of each cell were broad technologies focused on achieving a military-driven mission (e.g., 'signal processing' AND 'target detection', 'genetic algorithms' AND 'feature extraction', etc). This approach narrowed the retrieval considerably to strongly military-relevant technologies. The methodology was applied to the Indian S&T literature, and provided some interesting insights on the structure of the Indian defense S&T establishment (Kostoff and Bhattacharya, 2010).

Thus, it appears that classes of technology document retrievals need to be defined with different degrees of military-relevance. The study was limited to retrieving the military-identifiable and strongly military-relevant technologies. To identify military-relevant research, the process could be repeated, starting with the military relevant technologies and identifying linkages with military-related research.

## DISCUSSION AND CONCLUSIONS

The three examples presented in this review are a microcosm of some of the challenges faced in using scientometrics to solve problems. The nanotechnology example showed the importance of selecting variables that will directly address the questions of interest, and in selecting ranges of those variables to enhance the communication of the research results. The China S&T example showed the importance of selecting a minimum set of the most important evaluation criteria for assessing the value of a country's S&T portfolio, and the military relevant technologies example showed the value of a methodology for extracting technologies of interest from a sea of random technologies.

In the more general case, the scientometrics analyst is presented with questions of interest (usually to sponsors or stakeholders of one type or another), and has at his/her disposal an almost infinite amount of multi-dimensional multi-media data of very differing levels of quality with which to address the questions. The challenging decisions required in practice are: 1) which datasets to include; 2) how can one distinguish between reliable and unreliable data in the datasets selected for analysis; 3) which variables or combinations of variables should be selected to populate with data in order to answer the questions being raised; 4) how should data be extracted from the data sources selected to comprehensively and precisely populate the variables; and 5) how should the results of this massive data retrieval and data analysis be presented most clearly to the sponsors/stakeholders?

The author reviews for a number of technical and social sciences journals. Scientometrics articles invariably tend to be heavy on presentation of data and light on the type of analysis described above. This lack of real analysis, and especially lack of combining the data vectors into insightful 'signatures', limits the utility of these scientometrics analysis, and raises the barriers to acceptance of scientometrics techniques by the scientific community. Much more research is required in identifying key variables to extract from massive blocks of raw data, and how these variables can be combined into meaningful 'signatures' that will provide profound insights into superficially random data.

This review has only scratched the surface of what is possible in identifying variables and variable ranges of interest in a large body of data, and in extracting data of interest from large amounts of raw data. Hopefully, it will incentivize the community to generate far more complex protocols for achieving these ends.

## REFERENCES

Chen H, Kostoff RN, Chen C, Zhang J, Vogeley MSE, Börner K, Ma N, Duhon RJ, Zoss A, Srinivasan V, Fox EA, Yang CC, Wei CP. (2009). AI and global science and technology assessment. IEEE Intelligent Systems. 24(4). 68–88.

Kostoff RN, Eberhart HJ and Toothman DR. (1997a). Database Tomography for information retrieval. Journal of Information Science. 23(4). 301–11.

Kostoff RN. (1997b). The Handbook of Research Impact Assessment. Seventh Edition. Defense Technical Information Center. Fort Belvoir, VA. DTIC Report Number ADA296021. (http://www.dtic.mil/).

Kostoff RN, Tshiteya R, Pfeil KM and Humenik JA. (2002). Electrochemical power source roadmaps using bibliometrics and Database Tomography. Journal of Power Sources. 110(1). 163–76.

Kostoff RN, Shlesinger M and Tshiteya R. (2004a). Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 14(1). 61–92.

Kostoff RN, Shlesinger M and Malpohl G. (2004b). Fractals roadmaps using bibliometrics and Database Tomography. Fractals. 12(1). 1–16.

Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA and Karypis G. (2005a). Power source roadmaps using Database Tomography and bibliometrics. Energy. 30(5). 709–30.

Kostoff RN and Shlesinger MF. (2005b). CAB-citation-assisted background. Scientometrics. 62(2). 199–212.

Kostoff RN, Del Rio JA, Smith C, Smith A, Wagner CS, Malpohl G, Karypis G and Tshiteya R. (2005c). The Structure and infrastructure of Mexico's science and technology. Technological Forecasting and Social Change. 72(7).

Kostoff RN, Tshiteya R, Bowles CA and Tuunanen T. (2006a). The structure and infrastructure of the Finnish research literature. Technology Analysis and Strategic Management. 18(2). 187–220.

Kostoff RN, Briggs M, Rushenberg R, Bowles C and Pecht M. (2006b). The structure and infrastructure of Chinese science and technology. DTIC Technical Report Number ADA443315. (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA.

Kostoff RN, Stump JA, Johnson D, Murday J, Lau C and Tolles W. (2006c). The structure and infrastructure of the global nanotechnology literature. Journal of Nanoparticle Research. 8 (3-4). 301–21.

Kostoff RN, Bhattacharya S, Pecht M. (2007a). Assessment of China's and India's science and technology literature – introduction, background, and approach. Technological Forecasting and Social Change. 74(9). 1519–38.

Kostoff RN, Johnson D, Bowles CA, Bhattacharaya S, Icenhour AS, Nikodym KF, Barth RB, and Dodbele S. (2007b). Assessment of India's research literature. Technological Forecasting and Social Change. 74(9). 1574–1608.

Kostoff RN, Briggs M, Rushenberg R, Bowles CA, Icenhour AS, Nikodym KF, Barth RB and Pecht M. (2007c). Chinese science and technology — Structure and infrastructure. Technological Forecasting and Social Change. 74(9). 1539–73.

Kostoff RN, Briggs M, Rushenberg R, Johnson D, Bowles CA, Bhattacharaya S, Icenhour AS, Nikodym KF, Barth RB, Dodbele S, Pecht M. (2007d). Comparisons of the structure and infrastructure of Chinese and Indian science and technology. Technological Forecasting and Social Change. 74(9). 1609–30.

Kostoff RN, Koytcheff R and Lau CGY. (2007e). Structure of the global nanoscience and nanotechnology research literature. DTIC Technical Report Number ADA461930 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA.

Kostoff RN, Koytcheff R and Lau CGY. (2007f). Nanotechnology instrumentation and its measurements. Current Nanoscience. 3(2). 135–54.

Kostoff RN, Koytcheff R and Lau CGY. (2008a). Structure of the nanoscience and nanotechnology applications literature. The Journal of Technology Transfer. 33(5). 472–484. DOI: 10.1007/s10961-007-9042-2.

Kostoff RN, Block JA, Solka JA, Briggs MB, Rushenberg RL, Stump J.A, Johnson D, Wyatt JR. (2008b). Literature-Related Discovery. ARIST. 43. 243–85.

Kostoff RN, Morse S and Oncu S. (2008c). Text mining of the Anthrax literature. Defense Science Journal. 58(5). 678–85.

Kostoff RN, Barth RB and Lau CGY. (2008d). Quality vs quantity of publications in nanotechnology field from the Peoples Republic of China. Chinese Science Bulletin. 53(8). 1272–80.

Kostoff RN and Bhattacharya S. (2010). Identification of militarily-relevant science and technology. Defence Science Journal. 60(3). 259–70.

Kostoff RN, Koytcheff RG and Lau CGY. (2011a). Structure of the global nanoscience and nanotechnology research literature. Encyclopedia of Nanoscience and Nanotechnology. American Scientific Publishers. Vol. 23. 417–86.

Kostoff RN, Morse SA. (2011b). Structure and infrastructure of infectious agent research literature: SARS. Scientometrics. 86(1). 195–209.

Kostoff RN. (2012a). Literature-related discovery and innovation — update. Technological Forecasting and Social Change. doi:10.1016/j.techfore.2012.02.002.

Kostoff RN. (2012b). China/USA nanotechnology research output comparison—2011 update. Technological Forecasting and Social Change. doi:10.1016/j.techfore.2012.01.007.

Kostoff RN. (2012b). Text mining for science and technology - a review: Part II - citation and discovery. Journal of Scientometric Research. In Press. .

Schoeneck DJ, Porter AL, Kostoff RN, Berger EM. (2011). Assessment of Brazil's research literature. TASM. 23(6). 601–21.

Zhao Y, Karypis G. (2005). Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery. 10(2). 141–68.