SCIENCE AND TECHNOLOGY TEXT MINING: GLOBAL TECHNOLOGY WATCH

BY

DR. RONALD NEIL KOSTOFF
OFFICE OF NAVAL RESEARCH
800 N. QUINCY ST.
ARLINGTON, VA 22217
PHONE: 703-696-4198
FAX: 703-696-4274
INTERNET: KOSTOFR@ONR.NAVY.MIL
http://ww2.onr.navy.mil/test/technowatch/default.htm

(THE VIEWS IN THIS ARTICLE ARE SOLELY THOSE OF THE AUTHOR AND DO NOT NECESSARILY REPRESENT THE VIEWS OF THE DEPARTMENT OF THE NAVY, OR ANY OF ITS COMPONENTS)

KEYWORDS: technology watch; text mining; science and technology; technical literature; research impact; research evaluation; research assessment; research duplication; research opportunities; research planning; research management; information technology; database; information retrieval; information extraction; information processing; innovation, taxonomies; clustering; roadmap; technology forecasting; research documentation; research dissemination; bibliometrics; scientometrics; literature-based discovery.

ABSTRACT

Global Technology Watch is the assemblage of methodologies, both human-based and computer-based, required to understand the status of science and technology (S&T) globally. Since one important dissemination avenue for S&T is its literature, analysis of technical documentation is an important component of Technology Watch.

This paper examines the role and utilization of the technical literature in S&T development and exploitation. Ready access to the results of all global research performed is required in order to:

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

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

4. TITLE AND SUBTITLE
SCIENCE AND TECHNOLOGY TEXT MINING
GLOBAL TECHNOLOGY WATCH
Unclassified

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

6. AUTHOR(S)
Kostoff, Ronald Neil ;

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

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

8. PERFORMING ORGANIZATION REPORT NUMBER

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

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

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

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Global Technology Watch is the assemblage of methodologies, both human-based and computer-based, required to understand the status of science and technology (S&T) globally. Since one important dissemination avenue for S&T is its literature, analysis of technical documentation is an important component of Technology Watch. This paper examines the role and utilization of the technical literature in S&T development and exploitation. Ready access to the results of all global research performed is required in order to: 1) Track research impacts, to help identify benefits arising from sponsored research; 2) Evaluate science and technology programs; 3) Avoid research duplication; 4) Identify promising research directions and opportunities; 5) Perform myriad oversight tasks, and, in general, 6) Support every step of a strategic research planning/ selection/ management/ evaluation process that makes optimal use of S&T investment resources. In addition, recent counter-terrorism concerns have highlighted the need for ready access to, and analysis of, databases that could link people with institutions and activities. In the S&T arena, this requires linking research performers with organizations, countries, and technical areas.

15. SUBJECT TERMS
technology watch; text mining; science and technology; technical literature; research impact; research evaluation; research assessment; research duplication; research opportunities; research planning; research management; information technology; database; information retrieval; information extraction; information processing; innovation, taxonomies; clustering; roadmap; technology forecasting; research documentation; research dissemination; bibliometrics; scientometrics; literature-based discovery.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19. NAME OF RESPONSIBLE PERSON
			Same as Report (SAR)	41	Kostoff, Ronald kostofr@onr.navy.mil
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER International Area Code Area Code Telephone Number 703696-4198 DSN -

1) Track research impacts, to help identify benefits arising from sponsored research;
2) Evaluate science and technology programs;
3) Avoid research duplication;
4) Identify promising research directions and opportunities;
5) Perform myriad oversight tasks, and, in general,
6) Support every step of a strategic research planning/ selection/ management/ evaluation process that makes optimal use of S&T investment resources.

In addition, recent counter-terrorism concerns have highlighted the need for ready access to, and analysis of, databases that could link people with institutions and activities. In the S&T arena, this requires linking research performers with organizations, countries, and technical areas.

Powerful information technology techniques now exist to identify the relevant S&T literatures and extract the required information from these literatures efficiently. We have developed and used these techniques, especially text mining (the extraction of useful information from large volumes of free unstructured text), to:

1) Substantially enhance the retrieval of useful information from global S&T databases;
2) Identify the technology infrastructure (authors, journals, organizations) of a technical domain;
3) Identify experts for innovation-enhancing technical workshops and review panels;
4) Develop site visitation strategies for assessment of prolific organizations globally;
5) Generate technical taxonomies (classification schemes) with human-based and computer-based clustering methods;
6) Estimate global levels of emphasis in targeted technical areas;
7) Provide roadmaps for tracking myriad research impacts across time and applications areas.

Text mining has also been used or proposed for discovery and innovation from disjoint and disparate literatures. This application has the potential not only to serve as a cornerstone for credible

technology forecasting, but also help predict the technology directions of global military and commercial adversaries.

There are five necessary conditions for ready access to, and exploitation of, the global technical literature:

1) Content
The research results need to be documented comprehensively;
2) Dissemination
The research results need to be easily available to a wide audience;
3) Extraction
High quality information extraction methods are required;
4) Utilization
The research results need to be integrated into the strategic S&T management process;
5) Motivation
The technical community needs to be motivated to use the global technical literature for planning, selection, management, review, and transition.

All five conditions have severe limitations today. These limitations are described, and recommendations for overcoming them are presented.

OVERVIEW

Global Technology Watch is the assemblage of methodologies, both human-based and computer-based, required to understand the status of science and technology (S&T) world-wide. Since one important dissemination avenue for S&T is its literature, analysis of technical documentation is an important component of Technology Watch. This article describes how modern information technology can help maintain awareness of global S&T by extracting useful information from large volumes of structured and unstructured S&T text. It starts by showing the importance of Global Technology Watch, then defines the information technology terms and techniques, focusing mainly on text mining. The requirements for high quality text mining are summarized next, followed by the roadblocks and barriers to achieving those requirements. At this point, approaches for overcoming the roadblocks and existing/ potential applications that

use these approaches are described in some detail. These approaches and applications draw heavily from our direct experiences with global S&T text mining. This article is targeted to the research performer/ manager/ evaluator/ administrator/ sponsor/ intelligence analyst, as well as to the information technology professional.

WHY IS GLOBAL TECHNOLOGY WATCH NEEDED, AND WHAT KIND OF INFORMATION CAN IT PROVIDE?

Science and technology form the core of modern economies and militaries. Global S&T expenditures are in the neighborhood of 500 billion dollars to a trillion dollars annually, depending on one's definition of S&T. No single organization, or even nation, can begin to cover the full spectrum of S&T development required for a modern competitive economy or military. Cooperative S&T development efforts, leveraging, exploiting, and awareness of external S&T efforts are required if an organization or nation is to remain competitive.

Global Technology Watch maintains awareness of all levels of global S&T through a combination of human-based overt and covert activities, and automated/ semi-automated approaches for analyzing and tracking the myriad outputs of S&T. These outputs include text (reports, papers, patents, etc), other media, physical products, and trained technical people. The main supporter of USA domestic research (especially fundamental research) is the Federal government, and the main supporters of most global research are the respective sovereign governments. These governments need ready access to the results of all global research performed in order to:

1) Track research impacts, to help identify benefits arising from sponsored research;
2) Evaluate science and technology programs;
3) Avoid research duplication;
4) Identify promising research directions and opportunities;
5) Perform myriad oversight tasks; and, in general,
6) Support every step of a strategic research planning/ selection/ management/ evaluation process that makes optimal use of S&T investment resources.

In addition, recent counter-terrorism concerns have highlighted the need for ready access to and analysis of databases that could link people with institutions and activities. In the S&T arena, this requires linking research performers with organizations, countries, and technical areas.

This paper focuses on the approaches for extracting and analyzing useful information from text, collectively known as text mining. Powerful information technology techniques now exist to identify the relevant S&T literatures, and extract the required information from these literatures efficiently. We have developed and used these techniques, especially text mining, to:

1) Substantially enhance the retrieval of useful information from global S&T databases (1-10);
2) Identify the technology infrastructure (authors, journals, organizations) of a technical domain (1-10);
3) Identify experts for innovation-enhancing technical workshops (11) and review panels (12);
4) Develop site visitation strategies for assessment of prolific organizations globally;
5) Generate technical taxonomies (classification schemes) with human-based (1-10) and computer-based (6-10, 13, 14) clustering methods;
6) Estimate global levels of emphasis in targeted technical areas (4, 6-10);
7) Provide roadmaps for tracking myriad research impacts across time and applications areas (14-16).

Text mining has also been used (17-22) or proposed (11, 23, 24) for discovery and innovation from disjoint and disparate literatures. This application has the potential to serve as a cornerstone for credible technology forecasting, and help predict the technology directions of global military and commercial adversaries.

Additionally, text mining has been used to identify asymmetries and stratifications in technical databases where none were expected, potentially leading to improved understanding of system structure and dynamics (25).

However, these advanced techniques have been under-utilized by the corporate and national security intelligence communities, and by other segments of the potential user community, as well. Reasons for this lack of use differ among the diverse communities, but center around lack of awareness and understanding of available tools and techniques for information retrieval and extraction.

Any lack of global S&T awareness can waste time, people, funding, and other resources due to:

1) Pursuing approaches demonstrated to be non-productive
2) Duplicating previously performed S&T
3) Not using the latest techniques and instrumentation
4) Pursuing research in the absence of guidance available from the literature
5) Making research decisions based on incomplete information

WHAT IS TEXT MINING, AND HOW DOES IT RELATE TO GLOBAL TECHNOLOGY WATCH?

Data mining is the extraction of useful information from any type of data. In the modern context, it is the employment of sophisticated computer algorithms to extract useful information from large quantities of data. Text mining is an analogous procedure applied to large volumes of free unstructured text. S&T text mining is the application of text mining to highly detailed technical material. It is the primary technique for extracting useful information from the global technology literature.

The added complexity of text mining relative to data mining stems from the multiple meanings and interpretation of language, and their intrinsic dependence on context (e.g., Bank-repository of money, depend, edge of river, airplane maneuver). This context dependency of text interpretation renders the use of thesauri for text normalization and dimension reduction ineffective, since thesauri options do not reflect context without human selection. The complexity of S&T text mining relative to text mining of non-technical material arises from the need to generate a lexicon for each technical area mined and the need to have technical experts participate in the analysis of the

technical material. In the remainder of this paper, only S&T text mining will be addressed.

WHAT ARE THE MAJOR COMPONENTS OF TEXT MINING?

There are three major components of S&T text mining.

1) Information Retrieval
2) Information Processing
3) Information Integration

Information retrieval is the selection of relevant documents or text segments from source text databases for further processing. *Information processing* is the application of bibliometric and computational linguistics and clustering techniques to the retrieved text to typically provide ordering, classification, and quantification to the formerly unstructured material. *Information integration* combines the computer output with the human cognitive processes to produce a greater understanding of the technical areas of interest.

WHAT ARE THE DIFFERENT TEXT MINING TIME FRAMES?

Text mining is applicable to a continuum of time frames. Short time frame applications can be viewed as tactical in nature. Long time frame applications can be viewed as strategic. Tactical text mining tends to be more automated, more reactive, with less emphasis on depth of understanding. Strategic text mining requires more human cognition, is more proactive, and provides greater in-depth understanding of subject matter.

If documents are visualized as blocks, tactical mining provides real-time information about each block and its relation to other blocks. Strategic mining uses the blocks as building blocks, to obtain new information and insights unavailable from superficial block analysis.

Sample tactical text mining applications would include routing of incoming documents to appropriate categories, tracking of consumer preferences by Web site analysis, and tracking of terrorist activities by Web traffic analysis. Sample strategic text mining applications would include prediction of potential bio-warfare agents through

technical literature analysis, identification of unknown system asymmetries through literature imbalances, and determination of national technology investment strategy trends over time.

Most sponsored text mining development is for tactical applications. The development emphasizes sophisticated software, automated extraction and analysis processes, minimal human intervention, and rapid and low cost operation. These tangible products are easier to market and justify in the 'hard' technology environment of the 21st century. Unfortunately, this results in the rich ore potentially available from deep strategic mining remaining untapped.

WHAT ARE THE DIFFERENT LEVELS OF TEXT MINING COMPLEXITY, AND WHAT ARE THEIR COST AND TIME IMPLICATIONS?

S&T text mining complexity is probably the major time and cost driver of S&T text mining studies, and is perhaps the least understood and appreciated aspect of S&T text mining. It has led to more confusion about S&T text mining personnel, software, capability and cost requirements than any other issue, and is the main reason for the chasm between the expectations from text mining and the products delivered.

The complexity of S&T text mining is a proxy term for the level of detail, effort, time, and cost required for a target application. While there are many analysis parameters that affect the complexity of a particular application, there are two generic types of applications that determine the overall complexity level: sociological (overall trends) and analytical (specific details).

Sociological S&T text mining is aimed at providing a high-level understanding of a technical area for the non-expert. It provides low resolution results such as gross trends and broad categorizations. It typically involves working with high frequency phenomena (e.g., high frequency phrases, phrase combinations), requires relatively modest inputs of technical expertise, is amenable to semi-automated analysis using available software, can typically be performed in a very short time period and can be done at very low cost. It provides little new information for the technical experts in a discipline, and may be

counter-productive in providing the experts an image of S&T text mining having little new to offer.

Analytical S&T text mining is aimed at providing detailed insights to a technical area for the technical expert. It can be applied at the micro or macro discipline level. While it provides gross trends and broad categorizations to orient the customer initially, it eventually provides high resolution results such as the most detailed trends and lower level categorizations. While it involves working with high frequency phenomena initially, it evolves eventually to the examination of very low frequency phenomena. It alone has the capability to provide the technical expert with new information and insights about his/ her discipline, and the potential extrapolation of results and insights from other technical disciplines (related or disparate) to his/ her discipline of interest.

In my text mining experience, ***every type of analytical S&T text mining application has required the examination of many thousands of highly technical phrases and phrase combinations (and bibliometric items), and has required that technical judgements be made on the disposition of each of these phrases. This is the reality of analytical S&T text mining: the search for a few nuggets of information in an overwhelming sea of background clutter.***

In contrast to the sociological S&T text mining, analytical S&T text mining requires substantial inputs of technical expertise from different technical disciplines, in conjunction with information technology expertise in applying the techniques and interpreting the data. It is not overly amenable to semi-automated analysis using available software, but rather requires substantial intensive manual labor. It requires substantial time and a reasonable cost.

There is a widespread belief in the technical and possibly the intelligence communities that the combination of high speed computers with large memories, supported by intelligent agents and other artificial intelligence spin-offs, can produce automated/ semi-automated information extraction from large technical databases. As reference 26 states: “The ability to search the World Wide Web, or, in essence, millions of pages of text within seconds, has been a major

impetus for the rise in popularity of open source information." In reality, this high-tech search capability is the last in a series of complicated labor-intensive time-consuming steps required to convert copious and meaningless data into useful information. Days, weeks, or in some cases months of prior preparation are required before meaningful millisecond-level searches of "millions of pages of text" can be executed. An equal or greater amount of time is required in the post-retrieval period to translate the retrieved data into useful knowledge. Unfortunately, these precursor requirements are not advertised in the text mining software literature, and the results obtained from the consequential abbreviated analyses have not motivated the sponsor community to expand text mining efforts commensurate with both the needs of the commercial and military intelligence communities and the state of the art.

WHAT ARE THE REQUIREMENTS FOR HIGH QUALITY TEXT MINING?

The quality of a text mining study cannot exceed the quality of any of its components. For comprehensive access to the global S&T literature and maximum extraction of useful information from this literature, five primary conditions are required.

1) A large fraction of the S&T conducted globally must be documented (INFORMATION COMPREHENSIVENESS).
2) The documentation describing each S&T project must have sufficient information content and quality to satisfy the analysis requirements (INFORMATION QUALITY).
3) A large fraction of these documents must be retrieved for analysis (INFORMATION RETRIEVAL).
4) Techniques and protocols must be available for extracting useful information from the retrieved documents (INFORMATION EXTRACTION).
5) Technical domain and information technology experts must be closely involved with every step of the information retrieval and extraction processes (TECHNICAL EXPERTISE).

However, high quality text mining is a necessary, but not sufficient, component of high quality Technology Watch. In order to obtain maximum benefit from high quality text mining, it must be seamlessly

integrated with the strategic S&T management process. Yet, the decision aids into which the technical literature information is incorporated tend to be deployed on an *ad hoc* basis. Metrics, peer review, roadmaps, data and text mining, all of which depend heavily on inputs from the global technical literature, are rarely a seamless component of the strategic S&T management process. As a result, the S&T management process becomes tactical, driven by operational data. The sequence becomes Data ? Metrics ? Objectives, where the objectives of research evaluation are driven by the operational data available. This is in contradistinction to strategic S&T management, where the management sequence is Objectives ? Metrics ? Data, and the research evaluation objectives drive the data required. Proper investment strategy requires the latter approach, where the mission drives the investment as well as the data required to populate the metrics. Thus, insufficient use of the technical literature in the S&T planning cycle converts the investment strategy from top-down driven to bottom-up driven and converts planned strategic management to *ad hoc* tactical management.

Because of the above intrinsic technical literature database limitations and the consequent impact on quality of results obtained from existing technical literature analyses, there is little motivation for potential users at all levels to use the database sources during their research planning/ selection/ management/ review/ oversight activities.

WHY IS TECHNICAL EXPERTISE A HARD REQUIREMENT FOR TEXT MINING?

A 1998 text mining pilot program (27) showed conclusively that high-quality text mining requires the close involvement of information technology experts and technical domain expert(s) in the information retrieval, information processing, and information integration phases. Multiple perspectives are often needed to detect data anomalies. Presently, and for the foreseeable future, high quality S&T text mining requires intensive human labor and thought. It requires people with expertise in the specific technical disciplines being studied. Because of the intrinsic nature of high quality S&T text mining to access literatures from many different technical disciplines (and potentially different database sources), and to require judgements on the

linguistics and bibliometrics patterns in these diverse technical literatures, it requires people with understanding of these diverse literatures, as well. In fact, it has been our experience that the highest quality text mining products are obtained when the performers operate in an interdisciplinary mode, rather than a solely multi-disciplinary mode. Interdisciplinary operation carries its own unique set of problems (28), but provides a higher quality end product. However, it is an intrinsic paradox of high quality S&T text mining that technical experts are required for analysis of all technical domains accessed during the information retrieval phase, yet these domains *are not fully known before the study is initiated*.

For example, we were involved in a text mining study of bio-warfare agent prediction. The text analysis required knowledge of pathogenicity and genetic engineering of viruses (biomedical topics), as well as knowledge of aerosols and their transport through fluid media (engineering and physical science topics). Technical experts in all these areas were required. While the need for biomedical expertise was known beforehand, some of the engineering and physical science expertise requirements arose during the course of the study.

Technical experts alone are insufficient for a full spectrum S&T text mining study. People with information technology expertise who understand how the information technology capabilities and techniques should be applied to S&T text mining are required, as well. Utilizing performers with expertise in both areas would be the optimal situation.

WHAT ARE THE ROADBLOCKS AND BARRIERS TO ACHIEVING HIGH QUALITY TEXT MINING?

The approaches presently used by the majority of the technical community to address all five of these requirements have serious deficiencies.

1) *Information Comprehensiveness* is limited because there are many more disincentives than incentives for publishing S&T results (29). Except for academic researchers working on unclassified and non-

proprietary projects, the remainder of S&T performers have little motivation for documenting their output.

a) For truly breakthrough research, from which the performer would be able to profit substantially, the incentives are to conceal rather than reveal. Proprietary research with these characteristics is especially difficult to document. As industrial sponsorship of and participation in academic research becomes more pervasive, and as many academic researchers also form small companies, there is decreasing incentive from this sector of academia to publish, as well.

b) For research that aims to uncover and correct product problems, there is little motivation (from the vendor, sponsor, or developer) to advertise or amplify the mistakes made or the shortcuts taken.

c) For very focused S&T, the objective is to transition to a saleable product as quickly as possible; no rewards are forthcoming for documentation, and the time required for documentation reduces the time available for development.

d) For research of a classified or 'grey' nature, especially in today's environment of fear of terrorism, there is no motive for documentation, at least in the open literature.

Therefore, only a very modest fraction of S&T performed ever gets documented. This conclusion may sound surprising to people who have been bombarded with an 'explosion' of technical documentation. However, much of this 'explosion' may be due to a recent phenomenon known as 'paper inflation.' This is where what would have been one substantive comprehensive technical paper three or four decades ago is now sub-divided into multiple papers, each covering a portion of the parameter range of interest. Additionally, very modest variants of a given paper are published in multiple forums.

Of the performed S&T that is documented, only a very modest fraction is included in the major databases. ***The contents of these knowledge repositories are determined by the database developers, not the S&T sponsors or the potential database users.***

None of the research-sponsoring governments, including our own, appear to have control over the contents of, or interfaces with, these large S&T databases. Basically, the Federal government is footing the bill for the research that makes these large databases useful tools, but we are at the mercy of the database developers in terms of addressing our needs for database contents and operational requirements. We are heavy on data generation and light on data dissemination. The Appendix discusses some of the specific database content and access problems in more detail.

Of the documented S&T in the major databases, only a very modest fraction is realistically accessible by the users. The databases are expensive to access, not very many people know of their existence, the interface formats are not standardized, and many of the search engines are not user-friendly.

Insufficient documentation is not just an academic issue: in a variety of ways, it retards the progress of future S&T and results in duplication.

2) *Information Quality* is the product of amount of information provided and intrinsic quality of this information. Quality control is typically exerted through the peer review process, and the *pro bono* peer review process used today by the research journals has many well-known limitations (30). *Information Quality* content is limited because uniform guidelines do not exist for contents of the major text fields in database records (Abstracts, Titles, Keywords, Descriptors), and because of logic, clarity, and stylistic writing differences. The medical community has some advantage over the non-medical technical community in this area, since many medical journals require the use of Abstracts that contain a threshold number of canonical categories – Structured Abstracts (31) – while almost all non-medical technical journals do not (32).

Compatibility among the contents of all record text fields is not yet a requirement. As our studies have shown (4), this incompatibility can lead to different perspectives of a technical topic, depending on which record field is analyzed. This field consonance condition is frequently violated, because the Keyword, Title and Abstract fields are used by

their creators for different purposes. This violation can lead to confusion and inconsistency among the readers.

3) *Information Retrieval* is limited because time, cost, technical expertise, and substantial detailed technical analyses are required to retrieve the full scope of related records in a comprehensive and high relevance fraction process. Of all the roadblocks addressed in this section, this is the one that attracts probably the most attention from the Information Technology (IT) community. Because much of the IT community's focus is on selling search engine software and automating the information retrieval process, they bypass the 'elbow grease' component required to get comprehensive and high signal-to-noise retrieval.

4) *Information Extraction* is limited because the automated phrase extraction algorithms required to convert the free text to phrases and frequencies of occurrence as a necessary first step in the text mining process leave much to be desired. This is especially true for S&T free text, which the computer views as essentially a foreign language due to the extensive use of technical jargon. Both a lexicon and technical experts from many diverse disciplines are required for credible information extraction.

Poor performance by the automated phrase extraction algorithms can result in:

- lost candidate query terms for semi-automated information retrieval;
- lost new concepts for literature-based discovery;
- generation of incomplete taxonomies for classifying the technical discipline of interest; and
- incorrect concept clustering.

For clustering in particular, the non-retrieval of critical technical phrases by the phrase extractor will result in artificial cluster fragmentation. Conversely, the retention of non-technical phrases by the phrase extractor will result in the generation of artificial mega-clusters.

Detailed labor-intensive manual cleanup is therefore crucial to success. Thousands of phrases must be examined and culled by technical experts to insure that the appropriate high technical content phrases are generated in usable form. This level of human effort required is not advertised by the software vendor community, and as a result, many users are disappointed by the results produced from the software alone.

5) Two types of *Technical Expertise* are required for a credible text mining study, text mining technology expertise and technical (and related) domain expertise. Text mining technology *Technical Expertise* is limited because the intrinsic complexity of text mining has not been appreciated by the technical community, and resources have not been made available for the development of text mining experts (33). In contrast, target domain and related technical expertise exist, but their use in text mining studies has been limited both by tradition and by lack of understanding of the role of technical domain experts in high quality text mining. Because much information retrieval in the past and present has been performed by non-technical domain expert library support staff, the need and cost for higher priced technical experts to participate in the text mining studies is viewed as a non-essential expenditure. In addition, the developers of text mining software promote the concept that intelligent agents and smart algorithms can substitute for human experts.

An on-going text mining literature survey shows that there are in fact very few people actually developing the true text mining processes globally and increasing the understanding of what text mining can offer. For example, the only group actually publishing the results from the literature-based discovery text mining application is Swanson and Smalheiser (17-22). Perhaps a couple of other people, including ourselves, have written concept papers about literature-based discovery (11, 23). The literature-based discovery experience mirrors that of the other S&T text mining applications, as well. The research impact road-mapping application (14, 15) is being addressed by only one group (ourselves). There is a major mis-match between the potentially substantial benefits of these myriad S&T text mining approaches and the number of researchers and developers who understand, advance, and apply them.

WHAT ARE THE EXISTING AND POTENTIAL APPROACHES USED TO OVERCOME THESE BARRIERS, AND WHAT ARE SOME OF THEIR EXISTING AND POTENTIAL APPLICATIONS?

The approaches and their text mining applications tend to be integrated, and will be presented as such. Most of the applications center around information retrieval and extraction; the other three areas (information comprehensiveness and quality, and technical expertise), serve as enablers.

1) Information Comprehensiveness

Progress in information comprehensiveness requires that more of the S&T performed be documented, more of the documentation be placed in the larger databases, and more of these larger databases be made accessible to the wider user community. There are no efforts of any significance (of which I am aware) requiring increased S&T documentation. Since the science component of S&T is funded mainly by governments, a cooperative agreement among governments is needed requiring comprehensive documentation of performed research.

Some of the larger databases, such as the Science Citation Index (SCI), are expanding journal coverage. While the SCI covers the premier research journals, there is still valuable information existing in the lower tier journals, as our studies have shown (4). The major databases need to expand their raw data coverage, and the storage and access capabilities now exist to support this added requirement.

In the past couple of years, there has been a noticeable trend toward increased user friendliness in the major database search engines. However, the diversity of search engine protocols still presents a significant barrier to seamless utilization for the majority of part-time users. Some community-wide standards and uniformity in these protocols would go a long way toward making database access transparent to the user.

2) Information Quality

In the documentation of S&T, canonical category requirements are needed for the full text papers and for the Abstracts. Consonance of coverage is needed for all record text fields.

A major canonical requirements improvement occurred in the late 1980s with the introduction of Structured Abstracts in some medical journals (31). While the specific canonical requirements differ among journals, the categories tend to include Background, Objectives, Approach, Results, Conclusions. Because of the increased specialization that exists in most technical communities today, awareness and use of these Structured Abstracts have not impacted the total medical journal community, or almost any of the non-medical technical journal community. We have disseminated recommendations for Structured Abstracts and consonance of coverage among all record text fields (34) to editors of the premier non-medical technical journals, as well as editors of premier medical journals not using Structured Abstracts. Additionally, we have placed these recommendations in the open literature (32, 35). Insufficient time has passed to assess the changes that resulted.

We know of no efforts to require consonance of contents among all text fields in a record (albeit at a different level of description), other than our own described above.

3) Information Retrieval

S&T text mining can be used to enhance the retrieval of information from global S&T databases (1-10). Our group used all the S&T text mining components listed above to extract very comprehensive and highly relevant S&T information from global/ national semi-structured (free and structured text) S&T databases such as:

a) Science Citation Index (compendium of 5300 journals addressing basic research)
b) Engineering Compendex (compendium of 2600 journals addressing applied research and technology development)
c) MEDLINE (journal medical literature covering basic and applied research)
d) National Technical Information Service (reports from U. S. government-sponsored basic research to advanced development)

e) INSPEC (journal and conference proceedings covering basic research to technology development in physics, electronics, computing)
f) RADIUS (narratives of U. S. government agency R&D programs)
g) IBM and USPTO Patents (patent database)

WHY IS INFORMATION RETRIEVAL COMPLEX? WHAT IS WRONG WITH THE TRADITIONAL METHOD OF PROVIDING A FEW KEY WORDS TO YOUR LIBRARIAN?

Our group has been developing information retrieval techniques using an iterative relevance feedback approach. The source database queries result in retrieval of very comprehensive source database records that encompass direct and supporting literatures with very high ratios of desired/ undesired records. Some of the queries consist of hundreds of terms (4), in stark contrast to the handful of phrases used in typical information retrieval. In many cases, large queries are necessary to achieve the retrieval comprehensiveness and 'signal-to-noise' ratio required. Queries of a specific size are not a query development target of our group; rather, the query development process produces a query of sufficient magnitude to achieve the target objectives of comprehensiveness and high relevance ratio.

In each iterative step of our information retrieval process (5), a sample group of records retrieved with a previously modified query is classified into two categories: relevant to the central theme of interest, and non-relevant. Bibliometric and linguistic patterns of each category of the sampled records are examined, to generate terms that can be used to modify the query in order to increase relevant records (addition terms) and decrease non-relevant records (negation terms). The underlying assumption is that records in the source database that have the same linguistic patterns as the relevant records from the sample will have a high probability of being relevant, and records in the source database having the same linguistic patterns as the non-relevant records from the sample will also be non-relevant. Selection of such terms from the many thousands of candidate terms is a daunting task, and is extremely complex and time consuming.

To expand the relevant records retrieved, a phrase from the sample records should be added to the query if it:

1) appears predominately in the relevant record category;
2) has a high marginal utility (will retrieve a large ratio of relevant to non-relevant records) based on the sample;
3) has reasons for its appearance in the relevant records that are understood well; and
4) *IS PROJECTED TO RETRIEVE ADDITIONAL RECORDS FROM THE SOURCE DATABASE (E.G., SCI) MAINLY RELEVANT TO THE SCOPE OF THE STUDY*.

For multi-discipline source databases, application of these conditions can be complex. A recent example from the query development in a text mining study on the discipline of text mining illustrates this point. The phrase IR (an abbreviation for 'information retrieval' used in many SCI Abstracts) and its frequency of occurrence were extracted from a large group of retrieved records. It was characteristic of predominantly relevant sample records, had a very high absolute frequency of occurrence in the sample, and had a high marginal utility based on the sample. However, it was **not** 'projected to retrieve additional records from the source database mainly relevant to the scope of the study' when used as an SCI query term. A test query of IR in the SCI source database showed that it occurred in 65740 records dating back to 1973. Examination of only the first thirty of these records showed that IR is used in science and technology as an abbreviation for InfraRed (physics), Immuno-Reactivity (biology), Ischemia-Reperfusion (medicine), current(I) x resistance(R) (electronics), and Isovolume Relaxation (medical imaging). IR occurs as an abbreviation for information retrieval in probably one percent of the total records retrieved containing IR, or less. As a result, the phrase IR was not selected as a stand-alone query modification candidate.

Consider the implications of this real-world example. Assume a query consists of 200 terms. Assume 199 of these terms are selected correctly, according to the guidelines above. If the 200th term were like IR above, then the query developer would have been swamped with an overwhelming deluge of unrelated records. ONE

MISTAKE IN QUERY SELECTION JUDGEMENT can be fatal for a high signal-to-noise product.

In recent studies with our iterative information retrieval technique, about three iterations were typically required to obtain convergence. The resulting queries ranged in size from a dozen terms to a couple of hundred, depending on the relationship between the objectives of the study and the contents of the database. In the most recent studies, a marginal utility approach was applied to the query. The number of additional relevant records retrieved with the use of each additional query term was plotted, and the query length was terminated when the slope became small. This method resulted in substantial efficiencies in number of query terms.

As an example of a pre-marginal utility query, one of the studies focused on S&T for aircraft platforms (4). The query philosophy was to start with the term AIRCRAFT, then add terms to the query that would increase the number of relevant S&T papers (e.g., VSTOL, HELICOPTER, LANDING GEAR, FUSELAGE) or would eliminate papers not relevant to S&T (e.g., CROP SPRAYING, BUFFALO TRACKING). The resulting information retrieval process for the SCI required three iterations and produced 207 terms, while the process converged after a single iteration for the Engineering Compendex (EC) with 13 terms.

Because of the technology focus of the EC, most of the papers retrieved using AIRCRAFT, HELICOPTER, or similar query terms focused on the S&T of the Aircraft platform itself, and were aligned with the study goals. The research focus of the SCI differed. Many of the retrieved papers focused on the science that could be performed from an Aircraft platform, rather than the S&T of the Aircraft platform itself. Those papers were not aligned with the study goals. Therefore, no adjustments were required to the EC query, whereas many terms were added to the SCI query using Boolean negation to eliminate non-relevant papers. The Web, with its intrinsic lack of structure relative to the SCI or EC, would compound the difficulty of relating the study's objectives with any database focus.

Thus, the relation of the candidate query term to:

1) the objectives of the study, and
2) the contents and scope of the total records in the full source database (i.e., all the records in the SCI, not just those retrieved by the test query),

must be considered in query term selection. The quality of this selection procedure will depend upon the expert(s)' understanding of both

1) the scope of the study, and
2) the different possible meanings of the candidate query term across many different areas of R&D.

This strong dependence of the query term selection process on the overall study context and scope makes the 'automatic' query term selection processes reported in the published literature very suspect. These conclusions were reached on the basis of our experience with the homogeneous high quality databases such as the SCI or EC or Medline. The problem is even more difficult for the Web, where the intrinsic heterogeneity and expanded contexts further increase the number of meanings and interpretations available for each candidate query term.

4) Information Extraction

a) Infrastructure Identification

S&T text mining can be used to identify the technology infrastructure (authors, journals, organizations) of a technical area (1-10). This infrastructure includes the authors (if known), journals, performing organizations, and countries. Such information is valuable for identifying experts for technical workshops and review panels, and for planning site evaluation visits. This information becomes critical for intelligence studies, where tracking of people and institutions and analyzing time trends is a central component of the analysis. For technical journal-based databases such as the SCI or EC, once the information retrieval process has been completed, extraction of the technical infrastructure is straight-forward.

b) Literature-Based Discovery

S&T text mining can be used to discover new concepts or new relationships from literature, especially extrapolated from disparate literatures (11, 17-22). Such information can be used to identify promising research or technology opportunities and promising new directions for research. For intelligence applications, this approach can forecast the emergence of new unforeseen capabilities, based entirely on cutting-edge findings from global research laboratories.

In my estimation, this is an area of investigation that has completely fallen through the cracks. As stated previously, there is essentially one group that has published completed studies using this general technique, and these published studies have focused solely on the medical literature. Far more efforts are needed to test the applicability of competing techniques and the utility of different databases.

In another variant of literature-based discovery, a recently published study showed that unexpected asymmetries in systems could be identified through technical literature text mining alone (25). While the specific application was to disease asymmetries in bilateral body organs, and the specific literature examined was the Medline medical database, in principle the technique is applicable to any system and any literature. Knowledge of these asymmetries can target investigations into the underlying mechanisms that drive the asymmetries, and could lead to a deeper understanding of system structure and dynamics.

HOW CAN TEXT MINING BE USED FOR TECHNOLOGY FORECASTING?

Two of the credible major approaches to technology forecasting are:

1) the use of expert workshops for group dynamic approaches, and
2) the use of experts for literature-based innovation and discovery.

For the latter approach, some of the most revolutionary discoveries from text mining/ information retrieval have occurred in the medical field (17-22), resulting from linking disparate literatures to the primary target literature.

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

A series of papers (11, 24, 28) recommended combining these two approaches to eliminate their individual weaknesses and exploit their synergies. In this tandem approach, literature-based innovation and discovery using text mining would be performed initially. Based on the results of this initial step, a workshop would then be assembled using the linked disciplines from the literature-based study for the structure, and the experts identified from the literature-based study as the participants. The 1999 paper provides an example of the tandem process for Autonomous Flying Systems, although the literature-based component had insufficient time to operationalize all the concepts listed in the theoretical section of the paper.

c) Theme Identification

S&T text mining can be used for identifying the main technical themes or sub-themes in a large body of technical literature. Visual categorization of words/ phrases allows technical taxonomies (classification schemes) to be generated (1-10). Statistical clustering of words/ phrases (concept clustering), using factor analyses or partitional or hierarchical aggregation clustering schemes, also allows technical taxonomies to be generated (6-10). By categorizing phrases and counting frequencies (6-8), or by using statistical (or manual) document clustering (8, 10), S&T text mining can also be used to estimate global levels of emphasis in technical areas or sub-areas. These results can be used as the basis for S&T adequacy or deficiency judgements (4).

ARE THERE ANY PITFALLS WITH USING PHRASE FREQUENCY ANALYSIS FOR THEME IDENTIFICATION?

i) Level of Emphasis Studies

Our phrase frequency analysis approach proceeds as follows. Single word, adjacent double word, and adjacent triple word phrases are extracted from the Abstracts of the retrieved papers, and their frequencies of occurrence in the text are computed. A taxonomy is generated (top-down based on experience, bottom-up based on natural groupings of high frequency multi-word technical phrases, or some hybrid of top-down and bottom-up), and the appropriate technical phrases and their associated occurrence frequencies are placed in the appropriate categories. The frequencies in each category are summed, thereby providing an estimate of the level of technical emphasis of that category. For example, in the Aircraft S&T study (4), a taxonomy consisting of categories such as Structures, Aeromechanics, and Flight Dynamics was defined using a hybrid top-down and bottom-up approach. The sum of the frequencies of the phrases assigned by the technical expert to each of these categories was then computed.

This proved to be a very useful approach for estimating levels of emphasis. When coupled with information about the desired level of emphasis for selected categories, judgments of adequacy or deficiency could then be made. Either a requirements-driven methodology could be used to relate what is being done to what is needed, or an opportunity-driven methodology could be used to relate what is being done to what the state-of-the-art will allow.
For the specific areas studied, phrase frequency analyses of requirements/ guidance documents were performed to obtain requirements-driven quantitative estimates of the desired levels of emphasis, and the phrase frequency results from the S&T documents were then compared with the phrase frequency results from the requirements/ guidance documents. Opportunity-driven desired levels of emphasis were estimated based on the intuition and judgement of technical experts, and compared with the phrase frequency results from the S&T documents. The two comparisons were used together to arrive at overall judgements regarding adequacy or deficiency (4).

More recent studies (8, 10) have incorporated document clustering approaches, where documents are categorized into groups, rather than the phrase or word categorizations described above. Levels of emphasis can now be estimated by simply counting the number of

documents assigned to each technical category. This categorization can be done manually, or with the different statistical clustering packages available. While manual document, or even word/ phrase, clustering may seem outmoded in the present age of computers, it offers a level of insight and training unmatched by the sterile computerized approaches. In some studies, we have used the manual clustering to benchmark the computer-defined clusters. In some cases, the computer-driven categories will provide innovative structures not considered beforehand.

ii) Taxonomy Generation

A deeper taxonomy will naturally lead to greater resolution among subcategories and thereby to greater specificity in the judgments of adequacy and deficiency that can be made. For example, if the lowest level materials category in a taxonomy of ship subsystems is MATERIALS, then a gross judgement of adequacy or deficiency of technical emphasis in the very broad category of MATERIALS is all that can be made. This does not help guide decisions because of the lack of specificity. If, however, the lowest level materials category in the taxonomy includes subcategories such as WELDED TITANIUM ALLOYS, then judgements as to the adequacy or deficiency of technical emphasis in WELDED TITANIUM ALLOYS can be made. The more detailed the sub-category, the more useful the result from a programmatic viewpoint, and the greater are the numbers of adequacy or deficiency judgements that can be made. However, the greater the number or level of sub-categories, the lower the frequencies of the phrases required for statistical significance of each sub-category, the greater is the amount of work required, and the more expensive and time consuming is the study. Thus, a tradeoff must be made between the study time and costs and the quality of results required.

iii) Multiple Field Perspectives Required

It was also found useful to apply phrase frequency analysis to different database record fields to gain different perspectives. The Keyword, Title and Abstract fields are used by their creators for different purposes, and the phrase frequency results can provide a different picture of the overall discipline studied based on which field was examined. For example, in the Aircraft study (4), a group of high frequency Keywords was concentrated on longevity and

maintenance; this view of the Aircraft literature was not evident from the high frequency phrases from the Abstract field, where lower frequency phrases had to be examined to identify thrusts in this mature technology area.

The contents of the Keyword field reflect summary judgements of the main focus of the paper's contents by the author or indexer, and represent a higher level description of the contents than the actual words in the paper or abstract. Thus, one explanation for the difference between the conclusions from the high frequency Keywords and Abstract phrases is that the body of non-maintenance Abstract phrases, when considered in aggregate from a gestalt viewpoint, is perceived by the author or indexer as oriented towards maintenance or longevity. For example, the presence of the material category phrase CORROSION in the Abstract could be viewed by the indexer as indicative of a maintenance-focused paper, since many maintenance problems are due to the presence of corrosion. Another explanation is that maintenance and longevity issues are receiving increased attention, and the authors/ indexers may be applying (consciously or subconsciously) this 'spin' to attract more reader interest.

As another example, the Abstract phrases from the Aircraft study contained heavy emphasis on laboratory and flight test phenomena, whereas there was a noticeable absence of any test facilities and testing phenomena in the Keyword field. Again, the indexers may view much of the testing as a means to an end, rather than the end itself, and their Keywords reflect the ultimate objectives or applications rather than detailed approaches for reaching these objectives. However, there was also emphasis on high performance in the Abstract phrases, a category conspicuously absent from the Keywords. The presence of descriptors from the mature technology or longevity categories in the Keywords, coupled with the absence of descriptors from the high performance category, provide a very different picture of the Aircraft research literature than do the presence of high performance descriptors and the lack of longevity and maintenance descriptors in the Abstract phrases.

Consider the implications of this finding for Web-based analyses. The Web does not have a field structure, and effectively combines

Abstracts, Titles, Full Text, Keywords, and every other field. *What meaning can be ascribed to phrase frequency analyses of this agglomeration, and how can the results be interpreted meaningfully?*

iv) Theme Relationship Identification

S&T text mining can be used to identify the relationships between technical themes and between technical themes and infrastructure components (1-10). Much of our present effort is focused on understanding and developing clustering approaches that will sharpen the groupings of common themes and identify theme linkages from a variety of perspectives. We pioneered the use of multi-word phrase clustering for S&T databases (37-39), and are presently using the more sophisticated clustering software in parallel with clustering technique development to link major technical themes, major technical themes with supporting technical themes, and technical themes with infrastructure components. Such linkages are important not only in the development of science and technology, but have important corporate and defense intelligence applications, and can provide direction for program and organizational restructuring based on technical content. Further, S&T text mining can generate taxonomies from the bottom-up, removing human subjectivity from the process to some extent.

The clustering appears to offer much promise in information retrieval, literature-based discovery, taxonomy development, and a host of other application areas that cannot be mentioned here. It is foundational to S&T text mining. We are using manual and statistical clustering, with both concept clustering (words/ phrases) and document clustering being subsumed under the statistical clustering. Most of the published text clustering literature addresses new statistical clustering techniques, mainly in the context of document clustering. While statistical clustering technique development is important, it represents the tip of the clustering roadblock iceberg. It is not the major clustering barrier.

Extensive technical expertise and labor are required to understand optimal phrase groupings into clusters and to determine each cluster's central theme(s). Again, thousands of detailed technical phrases must be examined and interpreted by technical experts to understand their inter-relationships and the rationale for allocating

them to specific clusters. These requirements for extensive human involvement, labor, time, and technical expertise in the clustering process (and all other S&T text mining processes) are not reflected in the published literature.

v) Impact Roadmaps

S&T text mining can provide roadmaps of myriad research impacts (14, 15), to the degree that citations are assumed to bear some relation to impact. Such information is useful for impact tracking and subsequent S&T sponsor presentations. It provides performer organizations the ability to determine if the audience reached is the target audience. The Citation Mining approach we developed for these applications includes both bibliometric profiling and text mining. It is a sub-set of the more general trans-citation analysis, and has important consequences for Web-based corporate and national security intelligence information extraction.

5) Technical Expertise

In 1999, we developed a strategy for implementing text mining in the Federal government (33). Central to the strategy was development of information technology experts who understood the mechanics of text mining, and how it should be applied under different scenarios. We emphasized the need for multiple domain experts to work with the information technology experts. The development process proposed was neither fast nor cheap. Little has been done nationally to start implementation of such a strategy.

Since that time, it has become clear that efficiency and quality of a text mining study are improved further when the information technology and technical domain experts increase their understanding of each other's specialty area. It is insufficient to have each group conduct its share of the study without input from the other group. Our studies now require each group to participate in the other's spheres of expertise. In other words, our present focus is interdisciplinary research.

LESSONS LEARNED

From a long-range strategic view, the main output from the text mining studies conducted over the last decade was technical experts

who had their horizons and perspectives broadened substantially through participation in the data mining process. The text mining tools, techniques and tangible products, while important, were of secondary importance relative to the expert with advanced capabilities who could be a long-term asset to an organization.

There was a steep learning curve required to integrate the domain expert with the computational tools that required substantial training time. The operational mechanics were not the problem; the major roadblock was the time required for the expert to understand how the tools should be applied to address the study's specific objectives, and how their products should be analyzed and interpreted. The problem stems from the fact that text mining requires additional skills beyond traditional science and engineering training and experience. Technical domain experts do not necessarily develop such skills in a traditional technical specialty career.

Finally, it became clear that in order to have maximal impact from the text mining studies, they had to be conducted as an integral part of the organization's strategic management process (33). Otherwise, their influence on organizational strategic, and even tactical, decision-making would be minimal.

SUMMARY AND CONCLUSIONS

There is much technical information of value to military and corporate intelligence, and to the advancement of S&T in general, from global Technology Watch. The documentation of this technical information exists in many forms, ranging from the high quality peer-reviewed literature and its associated databases, to unstructured memos, reports, and the Web. Extraction of this information meaningfully is complex even for the most ordered and structured of technical databases. For the Web, technical information extraction is orders of magnitude more time consuming and complex. In either case, we have only begun to scratch the surface in developing information extraction techniques. Far more concerted efforts are required if global Technology Watch is to assume its rightful role as a useful source of national security and corporate technological information.

SUGGESTED FURTHER READING

1) Information Retrieval
There has been a substantial amount of work done to improve information retrieval, especially since the advent of large-scale use of the Web search engines. The proceedings of two major annual conferences (40, 41) are highly recommended, as is a survey of information retrieval techniques (42). While credible papers on information retrieval are published in many information sciences journals, three journals that appear to be of particular use are the *Journal of the American Society for Information Science and Technology*, *Information Processing and Management*, and the *Journal of Information Science*. The interested reader should peruse recent editions of these volumes.

2) Science Indicators
Tracking global S&T trends is an important component of Global Technology Watch. In order to generate a concise picture of global technology dynamics, some quantification of technology activity, inputs, outputs, and outcomes is required. The discipline of science indicators, or scientometrics, has expanded to address the required quantification. Three references are recommended. Every issue of the journal *Scientometrics* has articles devoted to the quantification of S&T. A 2000 book by Professor Geisler offers a comprehensive review of S&T metrics (43, 44). Finally, a biennial report by the National Science Foundation uses myriad S&T indicators to provide a comprehensive picture of the status and trends of institutional, sector, national, and international S&T (45).

REFERENCES

1. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature". Information Processing and Management. 34:1, 1998.
2. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography". Journal of the American Society for Information Science. 15 April 1999.

3. Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. Jan-Feb 2000.
4. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". Journal of Aircraft, 37:4, July-August 2000. Also, see Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. A., "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy", TR NAWCAD PAX/RTR-2000/84, Naval Air Warfare Center, Aircraft Division, Patuxent River, MD.
5. Kostoff, R. N., "Database Tomography for Technical Intelligence: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society, Scientometrics, 40:1, 1997.
6. Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography". Journal of Power Sources. 110:1. 163-176. 2002.
7. Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography". International Journal of Bifurcation and Chaos. In Press.
8. Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". Fractals. December 2003.
9. Kostoff, R. N., Andrews, J., Buchtel, H., Pfeil, K., Tshiteya, R., and Humenik, J. A. "Text Mining and Bibliometrics of the Journal Cortex". Cortex. Under Review.
10. Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". Under Review.
11. Kostoff, R. N. "Science and Technology Innovation". Technovation. 19, October 1999.
12. Kostoff, R. N., "The Principles and Practices of Peer Review", in: Stamps, A. E., (ed.), Science and Engineering Ethics, Special Issue on Peer Review, 3:1, 1997. Also, Kostoff, R. N., "Research Program Peer Review: Principles and Practices", http://www.dtic.mil/dtic/kostoff/index.html, 1997.

13. Kostoff, R. N., and DeMarco, R. A., “Science and Technology Text Mining”, Analytical Chemistry, June 2001.
14. Kostoff, R. N., Del Rio, J. A., Humenik, J. A., “Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling”, JASIS. 52:13. 1148-1156. 52:13. November 2001.
15. Del Rio, J. A., Kostoff, R. N., Garcia, E. O., Ramirez, A. M., and Humenik, J. A. “Phenomenological Approach to Profile Impact of Scientific Research: Citation Mining.” Advances in Complex Systems. 5:1. 19-42. 2002.
16. Kostoff, R.N., Del Rio, J. A., Bedford, C.W., Garcia, E.O., and Ramirez, A.M. “Macromolecule Mass Spectrometry-Citation Mining of User Documents”. Under Review.
17. Smalheiser, N.R., Swanson, D.R., “Assessing a Gap in the Biomedical Literature – Magnesium – Deficiency and Neurologic Disease”, Neurosci Res Commun, 15: (1), 1994.
18. Smalheiser, N.R., Swanson, D.R., “Calcium-Independent Phospholipase A (2) and Schizophrenia”. Arch Gen Psychiat, 55 (8), 1998.
19. Smalheiser, N.R., Swanson, D.R., “Using ARROWSMITH: A Computer Assisted Approach to Formulating and Assessing Scientific Hypotheses”, Comput Meth Prog Bio, 57: (3), 1998.
20. Swanson, D.R., “Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge”, Perspect Biol Med, .30: (1), 1986.
21. Swanson, D.R., Smalheiser, N.R., “An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery”. Artif Intell, 91 (2), 1997.
22. Swanson, D.R., “Computer – Assisted Search for Novel Implicit Connections in Text Databases”. Abstr Pap Am Chem S, 217, 1999.
23. Hearst, M. A., “Untangling Text Data Mining”, Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
24. Kostoff, R. N. “Stimulating Innovation”. International Handbook of Innovation. In Press.
25. Kostoff, R. N. “Bilateral Asymmetry Prediction”. Medical Hypotheses. August 2003.
26. Kimery, A. L., “The Truth is Out There”, Military Information Technology, 5(2), 2001.
27. Losiewicz, P., Oard, D., and Kostoff, R. N. "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15, 2000.

28. Kostoff, R. N. "Overcoming Specialization." BioScience. 52:10. 937-941. 2002.
29. Kostoff, R. N., "The Underpublishing of Science and Technology Results", The Scientist, 1 May 2000.
30. Rennie, D. "Fourth International Congress On Peer Review In Biomedical Publication." *JAMA* 2002; 287: 2759-2760
31. Ad Hoc Working Group for Critical Appraisal of Medical Literature. (1987) "A Proposal For More Informative Abstracts Of Clinical Articles". Ann. Intern. Med. 1987; 106: 598-604.
32. Kostoff, R. N., and Hartley, J., "Structured Abstracts for Technical Journals", Science, 11 May 2001.
33. Kostoff, R. N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations". Technology Analysis and Strategic Management. 11:4. 1999.
34. Kostoff, R. N., and Hartley, J., "Structured Abstracts For Technical Journal Articles", letter dated 12 May 2001.
35. Kostoff, R. N., and Hartley J. "Structured Abstracts for Technical Journals". Journal of Information Science. 28:3. 257-261. 2002.
36. Anon, "Managing Innovation: Combine Workshops And Text Mining To Speed Discovery", Inside R&D Alert, 12 January 2001.
37. Kostoff, R. N., "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Proceedings: Portland International Conference on Management of Engineering and Technology, October 27-31, 1991.
38. Kostoff, R. N., "Research Impact Assessment," Proceedings: Third International Conference on Management of Technology, Miami, FL, February 17-21, 1992. Larger text available from author.
39. Kostoff, R. N., "Database Tomography for Technical Intelligence," Competitive Intelligence Review, 4:1, Spring 1993.
40. TREC (Text Retrieval Conference), Home Page, http://trec.nist.gov/
41. Annual ACM Conference on Research and Development in Information Retrieval, Home Page, http://www.acm.org/pubs/contents/proceedings/series/sigir/
42. Greengrass, E., "Information Retrieval: An Overview", National Security Agency, TR-R52-02-96, 28 February 1997.
43. Geisler, E., "The Metrics of Science and Technology", Quorum Books, Westport, CT, 2000.

44. Kostoff, R. N., “The Metrics of Science and Technology”, Book Review, Scientometrics, 50:2, 2001.
45. National Science Board, “Science and Engineering Indicators 2000”, Arlington, VA, National Science Foundation, 2000, NSB-00-1.

APPENDIX - TECHNICAL LITERATURE DATABASE PROBLEMS

General

In the information age, data has become a strategic resource. In particular, technical data has assumed primary importance. At the same time, terrorism has become more of a national threat. Modern-day terrorism is fueled by advanced technology, which leverages and amplifies the impact of individuals and small cells strongly. Twenty people with bows and arrows would have had a miniscule fraction of the devastating impact of twenty people commandeering three fully-fueled jet liners into the World Trade Center and the Pentagon, which in itself is miniscule compared to what twenty people could do with selected bio-warfare agents. To help counter terrorism, even more advanced technology is required, especially related to surveillance, detection, and prediction.

Maximal technology advancement and exploitation are limited by the availability and awareness of technical data. The most comprehensive dissemination channel for fundamental technical data is the global technical literature. The necessary conditions for ready access to, and exploitation of, the global technical literature, have been discussed in the main text, and a more detailed description of these necessary conditions follows in the specific section of this Appendix.

The foundational condition is documentation and integration of primary data into comprehensive databases. In particular, technology advancement and exploitation are very dependent on having tailored comprehensive technical databases that address the objectives of the applications.

There exist many databases today covering the full spectrum of technology, ranging from government-sponsored and maintained databases to privately-controlled databases. Many of these databases are one-of-a-kind in terms of capabilities. In total, these databases constitute a strategic resource for the information age and technology development in general, and for counter-terrorism in particular. Like any strategic resource, they should be readily

available for national use, in the form that will produce the right data for specific applications.

Yet many of the databases have little coupling to application requirements. Especially for privately-sponsored databases, the contents are determined by the database developers, mainly for profitability and not for user requirements. To me, this is as serious as limitations on any other strategic resource, and requires a proactive remediation by national management. The government needs to take an active role in either sponsoring more of the unique and critical databases, or at least infusing sufficient funds to insure that the databases contain the appropriate material for national security applications. Because of the relative intangibility of data compared to more tangible strategic resources such as energy or materials, data as a strategic resource has not been elevated to the position of importance it deserves. That situation needs to be corrected now.

Specific

The large most widely used technical literature databases (e.g., SCI, EC, Medline) contain records from a variety of journals, conferences, workshops, and organization reports. Any of these large technical literature databases can be characterized by a number of objective parameters. These include, but are certainly not limited to:

1) Number of sources accessed (e.g., number of journals, number of conferences),
2) Time frame of source publications,
5) Fields available (e.g., title, author, abstract),
6) Fields entered,
7) Quality of entries,
8) Search engine capabilities,
9) Search engine interface,
10) Search engine download capabilities.

Each of the parameters listed above differs, in varying degrees, for the different large databases. Some of the difference is due to difference in technical focus, such as the SCI's focus on basic research, EC's focus on applied research and technology, and

Medline's focus on medicine. Much of the difference is due to specific database developer interests and motivations. Alignment of the capabilities provided by these databases with user needs, especially those of the main S&T government sponsors, appears to be serendipitous rather than the result of joint planning. Changes in database capabilities can be arbitrarily made by the developer, in some cases degrading the capabilities of the database, and the user has no influence on this direction.

Some of the issues associated with the above parameters include:

1) Number of sources accessed (e.g., number of journals, number of conferences)
While the leading journals are represented in the major databases, many more could be added. The Web version of the SCI, for example, accesses 5600 journals presently. While this appears comprehensive, there are many good technical journals that are excluded. Many more credible journals could and should be added. Journals that contain high quality technical information of particular interest to the intelligence community, for example, are not necessarily those cited most highly. There is little motivation to advertise those technical studies that straddle the gray area between classified and unclassified in the high circulation global literature. In addition, technology journals should be added, so that better research-technology linkages could be obtained from the text. This could provide database users a better idea of the potential applications resulting from specific research areas. A development target should be to determine the realistic limits to number of journals accessed, and incorporate as many of these as is possible.

2) Time frame of source publications
All these databases have different starting points in time for their records. Especially for studies that cross different technical disciplines and development categories, it would be valuable to have some consistency of starting points. For studies such as literature-based discovery, it is imperative to insure that potential discovery concepts have not been addressed in the technical literature previously (prior art). In these cases, earlier literatures are extremely valuable.

3) Fields available (e.g., title, author, abstract)
While some fields are common among the databases, not all are common. For example, the SCI does not contain a field with sponsor information. This deficiency becomes very important when an organization is attempting to ascertain the impact of its research from the literature. Also, there is not a comprehensive taxonomy structure in the SCI, parallel to the MESH capability contained in Medline. The closest field in SCI is Keywords Plus. It does not have the breadth or structure of MESH, it is not always populated, and its quality varies considerably.

4) Fields entered
Even though space is reserved for specific fields in the different databases, not all fields have entries. For example, I have been conducting a literature-based discovery study using Medline. My focus has been on records from the mid-early 1980s. Even though an Abstract field exists for those records, perhaps 30-40% of the records don't have Abstracts. In the SCI, Abstracts were not included until 1991. For many types of studies, especially those based on modern-day information technology techniques, records without Abstracts are essentially records missing! Proper database management would require that all fields be entered before the record could be accepted for the database.

5) Quality of entries
There is substantial variation in the quality of data entered. I have read thousands of Abstracts from different disciplines in the different databases. Some are accurate and informative. Many have substantial spelling and grammatical errors, introduced in the conversion from original source data. Especially for the Abstracts, there is a wide variation in the content. Some journals, like many from the medical literature, require Structured Abstracts. These include canonical criteria that must be addressed before the source paper can be published, and they tend to be very informative for readers from other technical disciplines. Other Abstracts have no structure requirements, and many of these are essentially useless. There have been four International Congresses on Peer Review in Biomedical Publication (http://www.ama-assn.org/public/peer/peerhome.htm) that have addressed the relationship between enhanced peer review quality and improved paper content. Some of the results of these

conferences could be implemented for improved paper content. Proper database management would require criteria for the major fields that would have to be addressed before the record could be included in the database.

6) Search engine capabilities
Search engine capabilities differ widely among the databases. Some search engines, such as Medline's OVID, have adjacency searching (e.g., word A within five words of word B), while others don't. Some, such as Medline's PubMed, can't do phrase searching, unless the phrase is on some pre-determined list. Some search engines can retrieve records from multiple fields using the OR boolean operator. Others can only retrieve records from multiple fields using the AND operator, a much more restrictive condition. A development target should be that all the major databases have uniform state-of-the-art search engine capabilities.

7) Search engine interface
All the search engines have different user interfaces, of widely varying quality and ease. For the occasional search engine user, this is a major obstacle, analogous to having to learn a new language whenever a different database is used. A development target should be uniformity of user interfaces among search engines, using state-of-the-art capabilities.

8) Search engine download capabilities
All the search engines have different download capabilities, with different levels of ease and quality. For example, PubMed can download 10000 records at a time, while OVID can download 200 max. A development target should be uniformity of download capability at the highest end.

The above are only a few of the existing problems, and improvements possible, with the large scale technical literature databases. Many more improvements would be surfaced at an initial meeting of government users, and others involved in the data generation and use cycle. I have heard Congressional hearings where databases to support counter-terrorism have been proposed (e.g., INS, FBI, etc), and the proposals have been received very enthusiastically. I believe there is a new attitude in the country for expansion and enhancement

of databases that could, as part of their function, support Federal security functions.