



## Overview

Governments and industry worldwide rely on advances in science and technology (S&T) to maintain a competitive advantage. To this end, they need ready access to the results of global research to:

- Track the impact of research to help identify benefits
- Evaluate science and technology programs
- Avoid research duplication
- Identify promising research directions and opportunities
- Perform myriad oversight tasks
- Support every step of a strategic research process that makes optimal use of S&T investment resources

In addition, recent counterterrorism concerns highlight the need for ready access to information that links people, technology and organizations together to stop the threat of terrorist activities. To combat this threat, more advanced technology is required, especially in the areas of surveillance, detection and prediction.

Since science and technology are global enterprises, with expenditures approaching \$1 trillion dollars annually, (depending on one's definition of S&T), no single organization or nation, can begin to research and develop the full spectrum of S&T required for a modern competitive economy or military. There must be cooperative development efforts including identifying, leveraging and exploiting

external efforts — if an organization or nation is to remain competitive.

Global Technology Watch maintains awareness at all levels of global S&T through a combination of human-based overt and covert activities, and automated approaches for analyzing and tracking the myriad S&T outputs. These outputs include text (reports, papers, patents, etc.), other media, physical products and technically trained people.

This article describes how information technology can help an organization maintain awareness of global S&T efforts by extracting useful data from large volumes of structured and unstructured S&T text. It is targeted to the researcher, intelligence analyst and information technology professional.

Powerful information technology techniques, such as text mining, now exist to identify and extract relevant data from the global S&T literature. Text mining is especially useful in making sense out of disjointed and disparate data. At the Office of Naval Research, we have developed and used these techniques to substantially enhance the retrieval of useful information from global S&T databases, such as the following.

- **Science Citation Index (SCI)** – current and retrospective bibliographic information, author abstracts and cited references found in 5,600 of the world's leading

scholarly science and technical journals covering more than 150 disciplines. The Web-based Science Citation Index Expanded, used at the Office of Naval Research, has 2,100 more journals than the CD-ROM version.

- **Engineering Compendex** – a compendium of more than 5,000 journals, conference proceedings, technical reports and foreign translations addressing applied research and technology development.

- **MEDLINE** – published by the National Library of Medicine and the National Institutes of Health, containing medical data covering basic and applied research.

- **National Technical Information Service (NTIS)** – the largest central resource for government-funded scientific, technical, engineering and business related information available today with more than 600,000 information products covering over 350 subject areas from over 200 federal agencies, including the Defense Technical Information Center Technical Reports. The technical reports and other DTIC databases are easily accessible on the DTIC Web site at <http://www.dtic.mil/>.

- **Inspec** – published by the IEE, is an English-language bibliographic information service providing access to the world's scientific and technical literature in physics, electrical engineering, electronics, communications, control engineering, computers, computing, information technology, manufacturing and production engineering.

- **RADIUS** – created by the Rand Corp., in cooperation with the National Science Foundation, contains narratives of U.S. government agency research and development programs at five hierarchical levels, ranging from 24 narratives at level 1 (reflecting overall descriptions of the research and development activities of the 24 major R&D sponsoring agencies) to 592,000 narratives at level 5 (award levels from these 24 agencies).

- **U.S. Patent and Trademark Office** – patent database.

The extracted data is used to identify the technology infrastructure, including authors, journals, organizations, etc., of

a technical domain and the experts for innovation-enhancing technical workshops and review panels. It is also used to:

- Develop site visit strategies to assess organizations globally using bibliometrics (e.g., counts of publications, patents, citations and unpublished data) and other science and technology indicators.
- Generate technical taxonomies (classification schemes) using clustering methods.
- Provide roadmaps for tracking innumerable research impacts across time and applications areas based on text mining. This has important consequences for Web-based corporate and national security intelligence.

Text mining has the potential to serve as a cornerstone for credible technology forecasting. It helps predict the technology directions of global military and commercial adversaries. Text mining has also been used to identify asymmetries and stratifications in technical databases where none were expected, potentially leading to an improved understanding of system structure and dynamics.

## Components of S&T Text Mining

There are three major components:

- Information Retrieval – the selection of relevant documents or text segments from source text databases for further processing.
- Information Processing – the application of bibliometrics, computational linguistics and clustering techniques to retrieved text to provide ordering, classification and quantification to formerly unstructured material.
- Information Integration – the combination of computer-generated output with human cognitive processes to produce a greater understanding of technical areas of interest.

## Steps in a Text Mining Study

A typical text mining study (by our group), without the literature-based discovery component, includes the following steps:

- ✓ Identify the technical scope of the problem.

✓ Develop a query to retrieve published records comprehensively and accurately. This involves high recall and precision.

✓ Select appropriate source databases for analysis.

✓ Retrieve records from databases.

✓ Generate publication bibliometrics.

✓ Generate citation bibliometrics.

✓ Generate background section, whose content is based on contribution of seminal papers.

✓ Generate taxonomy of retrieved literature to identify technical structure, including themes and relationships, using manual and/or statistical clustering. Include phrases and words, document clustering and hierarchical and/or flat taxonomies.

✓ Determine adequacy or deficiency of levels of effort (based on numbers of publications) in each category of taxonomy.

## Conclusions

The confluence of comprehensive technical databases, sophisticated information extraction algorithms and advanced text-mining processes offers the capability of substantially increasing awareness of global S&T. Expanded awareness fits in with the requirement for maximal technology advancement to combat terrorism and to ensure a competitive economy. Successful global S&T text mining requires an intrinsically interdisciplinary approach, incorporating information technology and technology-specific expertise.

For further information, I suggest the following resources, which are available through technical libraries.

Kostoff, R. N. "Text Mining for Global Technology Watch." In the Encyclopedia of Library and Information Science, edited by M. Drake. Second Edition. Vol. 4. New York: Marcel Dekker, Inc., 2003: 2789-2799.

Kostoff, R. N. "Stimulating Innovation." In The International Handbook of Innovation, edited by Larisa V. Shavinina. Oxford, UK: Pergamon Press, 2003.

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 1999.

Zhu D.H. and A.L. Porter. "Automated extraction and visualization of information for technological intelligence and forecasting." Technological Forecasting and Social Change, 2002. 69 (5): 495-506.

Swanson D.R. and N.R. Smalheiser. "An interactive system for finding complementary literatures: a stimulus to scientific discovery." Artificial Intelligence, Vol. 91, 1997.(2): 183-203.



*Kostoff received a doctorate in aerospace and mechanical sciences from Princeton University in 1967. At Bell Labs, 1966 to 1975, he performed technical studies in support of the NASA Office of Manned Space Flight and economic and financial studies for AT&T. At the Department of Energy, 1975 to 1983, he managed the Nuclear Applied Technology Development Division, the Fusion Systems Studies Program and the Advanced Technology Program. He joined the Office of Naval Research in 1983 as the Director of Technical Assessment for 10 years. He invented and patented (1995) the Database Tomography process, a computer-based textual data mining approach that extracts relational information from large text databases.*

*After managing the Navy Laboratory Independent Research Program for five years, he established a new effort in textual data mining. He recently received a full-spectrum text mining system patent application, called TextTosterone. He has written many papers on his research and is listed in Who's Who in America, 56th Edition (2002), Who's Who in America, Science and Engineering, 6th Edition (2002) and 2000 Outstanding Intellectuals of the 21st Century, 2nd Edition, (2003). See also [http://www.onr.navy.mil/sci\\_tech/special/technowatch/](http://www.onr.navy.mil/sci_tech/special/technowatch/).*