TECH MINING FOR FUTURE-ORIENTED TECHNOLOGY ANALYSES

By
Alan L. Porter

Acknowledgments

This paper draws heavily on the work of colleagues associated with development of "Technology Opportunities Analysis" at Georgia Tech. I particularly note the "TF" (technology forecasting) contributions reflected herein of Nils C. Newman and Robert J. Watts, and contributions to development of our text mining approach and software by Paul Frey, Scott Cunningham, Donghua Zhu, Douglas Porter, and Brian Minsk. Webb Myers and David Schoeneck have advanced the analytics. My long-term colleague, Fred Rossini, heavily influenced my futures perspective. I thank Joe Coates, Ted Gordon, and Peter Bishop for their reviews on the initial version.

I. HISTORY OF THE METHOD

Tech Mining is short for "text mining of science & technology information resources." The key premise underlying it is that *intelligence* (as in Central Intelligence Agency) is a prime requirement for effective technology management. Organizations operating in competitive and/or collaborative environments must track information on external technology developments.

Such technology intelligence underpins *Future-oriented Technology Analysis* ("**FTA**"). FTA is a growing community centered on the ongoing series of conferences hosted by the Institute for Prospective Technological Studies (IPST) -- //forera.jrc.ec.europa.eu/fta_2008/intro.html. It encompasses technology forecasting, assessment, foresight, and roadmapping.

So, how do we gather science & technology ("**S&T**") intelligence? We seek external information sources to mine. The earliest roots of this method lie in traditional literature reviews. "Technology monitoring" reviews the literature for a given technology of interest, usually scanning a broad range of sources. It seeks to digest sizable amounts of raw information to identify developmental patterns and key events (breakthroughs). Closely related, "environmental scanning" aims to provide early warning about important changes in various environments (e.g., competitive, policy or natural contexts; see the *Environmental Scanning* paper by Gordon and Glenn).

The advent of widely accessible, electronic information resources qualitatively changes these intelligence gathering efforts. One can now process huge quantities of information easily and effectively. "Bibliometrics" – counting publications or other bibliographic items – can help track scientific and technological developments.

Tech Mining can also be considered a form of "Content analysis." That has roots in the pre-electronic information processing era, but has blossomed with the availability of electronic datasets and analytical software. One examines patterns in term usage to infer emphases in given domains. Text mining empowers powerful variations on content analysis. This detective work is not restricted to technology analyses – for instance, intelligence agencies may analyze press releases to track foreign political leaders' shifting interests.

Data mining seeks to extract useful information from any form of data, but it is commonly used to mean analyses of numerical data (e.g., linking your credit card purchases to your demographic profile). Tech Mining exploits text and numerical data sources of various sorts. In particular, watch the distinction between unstructured text sources (e.g., Internet website content) and structured text sources (e.g., abstract records from managed databases that separate information into fields such as "author," "publication date," and "keywords"). Specialties in text understanding are rapidly evolving with fine distinctions in emphases. For our purposes, we just note emerging areas that concern text mining: "Computational Linguistics," "Natural Language Processing," and "KDD" (Knowledge Discovery in Databases).

This paper emphasizes the development of text mining tools to analyze emerging technologies. Such intelligence extraction efforts need not be restricted to a particular subject matter. However, focusing on technological change can lead to particularly useful outputs called

"innovation indicators" (discussed later and elaborated in the book, *Tech Mining*, by Porter & Cunningham – see "Selected Literature" at the end).

Who invented text mining? I don't know. Talmudic scholars have tabulated content patterns in the bible for ages. Counting scientific publication activity dates back at least to the pioneering work of Derek Price in the early 1960's (see Appendix). Content analysis dates back decades in the Social Sciences. Modern Tech Mining would count among its pioneers, Henry Small, Tony van Raan, Michel Callon, and Loet Leydesdorff (see Appendix).

What did we do before text mining? The benchmark in some regards is the traditional "literature review." Therein, a researcher summarizes the most pertinent (e.g., 20 or so) literature sources pertinent to a given study. In an age of electronic information resources, we can still search for those few "nuggets" of really key reviews, models, and developments. What is different in Tech Mining is that the *full body* of related literature is analyzed. In one study we analyzed the development of excimer lasers for machining. We located some 12,000 pertinent references. Instead of winnowing those down to find a handful of key items, Tech Mining elicits useful patterns from the entire 12,000 sources. For instance, one can track relative emphases on particular topics, over time, by key research groups.

II. DESCRIPTION OF THE METHOD

Tech Mining extracts useful intelligence from electronic text sources. This information can serve needs to know about current activities (e.g., profiling who's doing what in a target area). It can also serve FTA interests in several ways:

- identifying R&D emphases that portend future developments
- providing time series for trend extrapolation and growth modeling
- generating "innovation indicators" that speak to the prospects for successful technology applications [The *Tech Mining* book identifies some 13 enduring Management of Technology ("**MOT**") issues, that spawn about 40 questions, which some 200 innovation indicators help answer.]

Effectively answering MOT questions is why one does Tech Mining. So, the first vital step is to spell out the key questions to be addressed. Knowing those will guide choices about which data sources to use and which analyses to pursue. Focus is critical; otherwise Tech Mining can wallow in too much data, with too many intriguing – but not directly useful – analyses to pursue.

That said, let's consider how one "mines" text, particularly field-structured S&T information resources. Let's take an example case. Bob Watts was analyzing technology opportunities for the U.S. Army's Tank-Automotive Research, Development & Engineering Center (TARDEC) in the mid-1990's. He began to explore the potential for innovative applications of ceramics in automotive engines (especially tank engines). Reviewing local history uncovered earlier bad experiences. TARDEC had invested in ceramics R&D in the 1980's without significant return on investment. The Army had stopped such funding. Nonetheless, Bob saw promise and pursued his investigation.

Searching in an appropriate publication abstract database such as *EI Compendex* (also known as *Engineering Index*) yielded several thousand abstracts. In beginning the analysis, one likely wants to LILST the leading organizations contributing research articles on the topic. Knowing *WHO* is engaged is vital – e.g., it may suggest potential partners. Another list might give the prominent "keywords" descriptive of research emphases in these articles. Further analysis could then cross these two lists as a MATRIX showing which topics particular organizations mention frequently – i.e., we now add a *WHAT* dimension. That could help you spotlight particular organizations doing work of interest. For instance, we might spot that General Motors is working on structural ceramics and silicon nitride, whereas Sandia National Lab is addressing thin film ceramics and silicon carbide.

Many other combinations could provide insights into the emergence of a technology. Just to give one other example, one could cross "keyword usage" by "publication year" to suggest which topics are "hot" (increased emphasis recently). In this way we are combining *WHAT & WHEN* aspects – perhaps to help the Army set its R&D priorities. One might add a third dimension to the mining effort. For instance, in our text mining software, *VantagePoint*, you could pop open a detail window to see in which countries (*WHERE*) particular topics are being pursued recently. [We'll pick up this case story later.]

Which information resources you choose to mine determines what sort of intelligence you can elicit. Obviously someone checking levels of public concern over an emerging technology's possible environmental hazard would select different resources to tap than someone investigating ceramic R&D.

I find it useful to distinguish 6 main types of information resources (Table 1). Many technology managers fail to take advantage of rows A and B; they rely almost exclusively on tacit judgment. This is folly. Those who obtain empirically based knowledge, in addition to using human expertise, will dominate those who don't do so. Others equate electronic information resources with the internet (row B). That also misses a vital resource – row A. Tech Mining emphasizes extracting useful intelligence from databases (both publicly available and organization-confidential). Databases represent someone having compiled, filtered, and formatted text (or other) data, making this of higher value than the unfiltered mess of the internet.

Table 1. Six Information Types

Medium \ Message	**1) Technology**	**2) Context**
A) Databases	Research funding, publication & patent abstracts, citations	Business, market, policy, popular opinion
B) Internet	Technical content sites	Company sites, blogs, etc.
C) People	Technical experts	Business experts

Analyzing a spectrum of databases that span different stages along a technological innovation can help benchmark how mature the technology is. One expects S&T activity to migrate along

this continuum, over time, as an innovation progresses, but the reader is reminded that technological innovation rarely progresses linearly:

- Fundamental research [reflected by research project funding (e.g., NSF Award or NIH CRISP database) and publication abstract databases (e.g., Science Citation Index; PubMed)]
- Applied research & development [reflected, for instance, by engineering oriented publication abstract databases (INSPEC, EI Compendex)]
- Invention [consider patent databases such as Derwent World Patent Index and PatStat (from the European Patent Office for academic use]
- Commercial Application [e.g., new products databases, Business Index, marketing data sources]
- Broader Contextual Factors and Implications [e.g., Lexis-Nexis, Factiva, Congressional Record]

Having commended such databases as extremely valuable repositories, I need to say that they don't tell the whole story. Database content lags – it takes time for most sources to make information available (e.g., publication processes may take a year or more; patents can take many years), so the internet is more current. Not all sources are covered by databases; so one should seek enriching internet content as well. And database access – especially unlimited use licenses vital for Tech Mining – is costly (with relatively few exceptions – e.g., PubMed is free). We like to analyze database content (typically using one or a few key sources on a given study) first; then augment with internet searches (e.g., to check websites of key research organizations). A notable frontier for Tech Mining is the improving software able to find, compile, and format internet content (c.f., www.**ql2**.com/). And, above all, engaging experts to review the empirical findings and help interpret those (Table 1, Row C) is critical to generate credible technical intelligence and subsequent FTA.

III. How To Do It

Here's an 8-step approach to Tech Mining.

1. Spell out the focal MOT questions and decide how to answer them
2. Get suitable data
3. Search (iterate)
4. Import into text mining software (e.g., VantagePoint)
5. Clean the data
6. Analyze & interpret
7. Represent the information well – **communicate!**
8. Standardize and semi-automate where possible

I'd like to illustrate via two case examples: A) The Army's analysis of the potential new use of thin-film ceramics in engines, and B) National benchmarking of R&D initiatives. We've introduced (A) already, so let's step through the Tech Mining method using that, then later we can enrich our perspective with (B).

Step 1 – The Army wanted to assess whether thin-film ceramics held potential for its tank engine applications. If so, then it sought guidance on how best to pursue such development. In particular, might they want to partner with particular R&D organizations?

Step 2 – We've introduced a few of the hundreds of available databases. The Army (TARDEC) had previously determined which of these to license for its technical intelligence and research needs.

Step 3 – Search entails several key elements. One must formulate the inquiry – what do you want to know, about what topic? Properly bounding the inquiry is vital – e.g., profiling all R&D relating to "ceramics" would be a herculean task, unlikely to yield useful insights. At the opposite extreme, our interests are very different from traditional librarian efforts to find a handful of "just right" publications. We want a middle ground of a large swath of related research – e.g., ceramic development associated with engine applications.

We often find it effective to conduct a "quick and dirty" initial search; download a sample of resulting abstract records; and quickly analyze those using our text mining software (*VantagePoint*). By browsing the resulting term lists (i.e., keywords), preferably with benefit of expert review, we can sharpen the search phrasing. Then we redo the search and retrieve the actual record set to be analyzed in depth. This process can be done in as little as three minutes given desktop access to the searchable database(s) via network or CD.

Step 4 – In the TARDEC ceramics work, Bob Watts downloaded his search results into *TechOASIS* software (see the Appendix) for further analyses. Such software is tuned to analyze field-structured text records.

Step 5 – Data cleaning might be called "clumping." We want to consolidate author or organization name variations (especially if we're combining search results from multiple databases). *VantagePoint* (and other) software tools use fuzzy logic to clean up terminology (e.g., to combine singular and plural forms of a term; overcome small name discrepancies). General purpose and special thesauri are then used to combine related items. One handy thesaurus aggregates organizations by types: universities; companies; government or non-governmental organizations; and hospitals. Then, a handy macro (a little software script) shows the extent of industrial R&D on a topic in a pie-chart as a useful innovation indicator.

Step 6 – Analyses can take many paths and forms. These should remain mindful of Step 1 – i.e., they should help answer the essential MOT questions driving the analysis. Text mining software does basic operations, such as counting which author organizations show up most frequently in your set of records. It then allows easy exploration – e.g., via a detail window to examine a given organization's topical emphases over time. The software can also perform more advanced functions – e.g., extracting noun phrases from abstracts, clustering related entities, and mapping research networks (based on co-authoring, co-citation, or whatever).

Many of these operations can be done without text mining software. To begin, many website database search engines enable you to tally occurrences of particular terms, slice search sets by year, and so forth (c.f., SciFinder; EI Village). Once you've downloaded search results, you can

use common literature tools (e.g., EndNote) to handle the records; MS Excel to help count up frequencies; and so forth. However, if you frequently want to perform such analyses, or want to get at more intricate innovation indicators, I'd recommend you investigate software aids (see Appendix).

Step 7 – Representing your findings effectively begins by knowing how your key users like to receive technical intelligence. Accordingly, you may well want to judiciously combine:

- Oral and written reporting
- Electronic and paper reporting
- Text, tables, and figures – possibly with video or animations
- Interactive exchange (e.g., workshop)

Sometimes, as analysts, we neglect to explain. We wallow in the data so deeply that we forget to point out the implications of our tallies and figures. And, to paraphrase a Russian colleague, "decision-makers don't want pretty pictures, they want answers."

Step 8 – Sacrificing some flexibility of the strategy in which every study starts from Ground Zero is advisable. We advocate an "RTIP" – Rapid Technology Intelligence Process – premised on standardizing and automating Tech mining analyses. There are multiple advantages to this:

- Technology managers and professionals become familiar with the Tech Mining outputs. They learn how to get the answers they need from them, thereby making better decisions informed by empirical knowledge.
- Analyses are done much faster. One company told us how a series of "quarterly" competitive technical intelligence reports could now be done in about 3 days. Suddenly, that makes data-based intelligence accessible to pressing decisions.
- Analyses are cheaper. By semi-automating data cleaning and analysis steps, analysts can generate more value, but use fewer resources.

IV. Strengths and Weaknesses of the Method

If you noticed, I mentioned how Tech Mining can answer four types of questions – who, what, where, and when? The other two of the so-called "reporter's questions" – how and why? – almost always require expert opinion to infer processes (how?) and reasons (why?).

No FTA method stands well alone. This is certainly true for Tech Mining. Basic tabulations require thoughtful interpretation. "Innovation indicators" require complementary expert opinion to test observations and relationships generated from the information resources.

Do you need expert help to mine technological information resources? Ron Kostoff, a leader in technological text mining (see Appendix), argues that heavy involvement of substantive experts in the process is essential. This yields advantages in assuring sensible interpretation, proper treatment of terminology, and "buy-in" from those experts who become actively engaged.

The information products generated can be no better than the sources mined. So, if vital elements of technological development are not treated in your sources, you will not uncover

them. For example, if you base your analysis on patents on an emerging technology for which the developers do not generally seek a patent, you will be misled.

Certain analyses can become complicated. "Transparency" enables users to understand results. Representations of relationships discovered in the text are usually based on co-occurrences. These entail terms appearing in the same documents more frequently than expected from their overall popularity. It is important to recognize that clustering and mapping of such relationships generally do not have a single, "right" way to go. Representations, unless invalid (i.e., misrepresentations due to incorrect statistical analyses or such), prove their value by being used – that is, by depicting relationships so that a user gains new insights.

Tech Mining is still new to technology managers and professionals, as well as futurists. As such it faces credibility issues. If one makes decisions relying heavily on derived knowledge from text mining, whom can you blame if the decision goes awry? Many decision makers feel more comfortable leaning on expert opinion, even if it is parochial. We suggest combining Tech Mining results with expert opinion to take advantage of the strengths of each.

A good way to get started at Tech Mining is to contract for a couple analyses. If those usefully inform MOT decision-making, then consider developing in-house capabilities. To do so, one first needs access to suitable electronic information resources. One then wants the tools to search, retrieve, and analyze selected records. The appendix notes several software options.

Training is needed to perform Tech Mining effectively. Analysts familiar with treating "text as data" (e.g., the "Internet generation" who have been students since about 1995) can quickly grasp the concepts and generate useful findings with a couple of hours of training. Less analytically oriented persons, or those residing on "the wrong side of the digital divide," may never get it. Training workshops, supported by ongoing access to advice, offer a reasonable middle ground for motivated analysts to pick up how best to use given software.

Costs are driven by database charges. Many universities already license a number of R&D databases useful for technology foresight, so there is no additional cost involved. Simple analyses can be performed using the basic analytical capabilities of the search engines (e.g., *Web of Knowledge*). To license the software illustrated in this paper, *VantagePoint*, costs some thousands of dollars (but varies by organizational type and needs).

Putting information products into comfortable, easily grasped form is critical. We recommend blending numerical tabulations with graphical depictions and text interpretations – tailored to your audience's preferences. An interesting option is to combine a report with a CD containing the raw abstracts and the mining software. That can enable users (who are willing to learn to use the software) to dig down into the data themselves as the report triggers questions. For instance, imagine you're reading a profile of ceramic engine developments and note that Sandia National Lab is researching thin film ceramics – you can instantly browse through their abstracts to see what they are doing. Note that to get to the full articles requires either a full text database, live links, or a visit to the library.

In observing what makes for effective utilization of text mining outputs, we distinguish users (and organizations) who prefer reports from those who prefer answers to their questions. For instance, a governmental agency (e.g., the Army) interested in ceramic engine applications, would most likely want a well-substantiated report. In contrast, we have found that many company technology managers want to know that well-founded analyses underlie results obtained, but they just want their immediate question answered directly (no lengthy report).

Tailoring information products carefully to the key users' needs is critical. An interesting approach is "packetizing" findings to avoid information overload. For instance, instead of holding off until a 100-page report is ready, dole out small amounts on particular aspects that you feel the user wants and can readily digest. This can also help engage the user in the analytical process, greatly facilitating buy-in.

Building familiarity with text mining outputs and how to use them effectively takes real effort. Relationships between analysts and users must be strong. Factors that enhance prospects include:

- facilitating direct links between the analysts and users (interposing additional people to perform information searches, qualify statistical analyses, review results, etc., is deadly)
- engaging multiple analysts and multiple technology decision makers to develop a robust, learning network
- directing attention to successes so that the organization appreciates the value gained.

V. USE (IN COMBINATION WITH OTHER METHODS)

This section illustrates how Tech Mining can be used to inform Future-oriented Technology Analyses. Two case studies are offered. Throughout, keep in mind that text mining needs to be combined with other methods, particularly expert opinion.

A. Investigating the Potential of an Emerging Technology (thin-film ceramics) to meet Army Engine Needs

I earlier introduced the analytical challenges prompting this series of analyses on ceramic R&D. Figure 1 gave the key innovation indicator that alerted to a key advance in the maturation of this technology. The message in Figure 1's time-slices has several parts. Consider, first, the back row showing "Number of Publications." In the 1987-88 time slice, this reached 200 or more, and then crashed down. In the most recent time period, 1993-95, this has begun to recover slightly. The second row, "Discrete Sources," shows a similar pattern. In essence, following an initial burst of ceramics enthusiasm, funding and research effort had cooled.

The front row, "Number of Keywords," tells a markedly different story. This indicates that the richness of R&D discourse about ceramic engine applications had sharply accelerated in the recent time period. That was the signal for action.

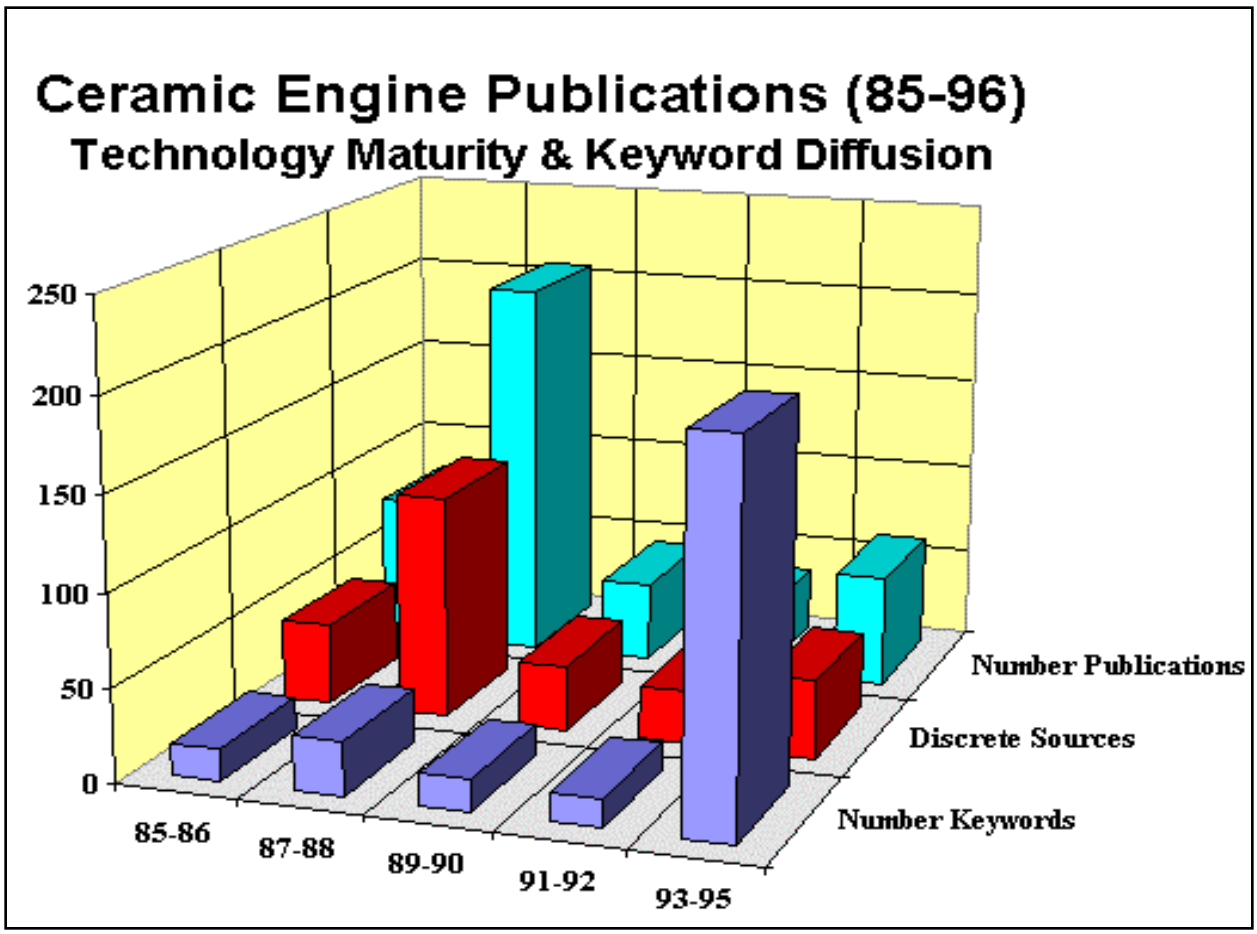


Figure 1. Keyword Behavior Change as an Innovation Indicator (from Watts & Porter, 1997)

Watts then presented these results to ceramics experts who confirmed that, indeed, the research arena had gotten more specialized and was moving toward practical applications. Emerging key terms implied serious research underway on particular materials (e.g., silicon nitride); special properties (e.g., work on fatigue testing); modeling; and, in particular, on thin film features. TARDEC senior management was convinced by this storyline that thin-film ceramics now could help the Army meet its engine needs.

The second Tech Mining phase pursued the prospects for thin film ceramic coating applications in the automotive context. TARDEC sought to identify the research leaders. Rather to the surprise of the mechanical engineers working on engine improvements, they were not from the structural ceramics community. Instead, it was the semiconductors sector doing the cutting edge R&D on thin ceramic films for chip design and development. Bob's Tech Mining work was able to span these traditional bounds by searching in EI Compendex. He identified particularly exciting R&D activity underway at Sandia National Lab and a company. Those researchers had not considered coating tank engine parts!

To wrap up the story, TARDEC made contact with those research organizations and funded two large projects to explore coating applications. Adaptation of vapor deposition to pistons and

turbine blades proved workable. In 2004, a major production plant opened to coat used Abrams tank turbine blades for extended life. The ceramic coating offered additional advantages in terms of higher operating temperatures and reduced air pollution. So, Tech Mining had provided two key insights: 1) this research domain had significantly matured to offer real promise, and 2) it identified potential development partners from a distant sector with which TARDEC mechanical engineers did not routinely interact.

B. Benchmarking National Research Initiatives

This Tech Mining case also involves the public sector (our private sector work is almost always in confidence). Let's use this to illustrate variations in the "8-step" approach.

Step 1) Here, a Canadian Government group sought to benchmark a number of "research cluster" initiatives in other countries as comparisons to assess their own initiatives – i.e., a research evaluation issue. I draw illustrations here from one that looked at photonics research in the U.S.

Step 2) We searched in Science Citation Index (SCI)
[a companion study investigated patenting, using the Derwent World Patent Index.]

Step 3) Formulating a viable search posed a challenge because there is so much R&D relating to photonics. After an initial trial, with expert feedback to refine what should be included, we settled on a multi-modular strategy: we retrieved abstract records that included 1) certain target terms (e.g., photonic, optoelectronic); we also retrieved records that had 2) both a "photo or optical" term stem AND a term relating to electronics, telecommunications, instrumentation, materials, coatings, or lenses. Given the timeframe of the cluster initiatives of interest, we retrieved data from 1999-2008. We considered using classification information too – SCI Subject Categories – but did not like the results as well.

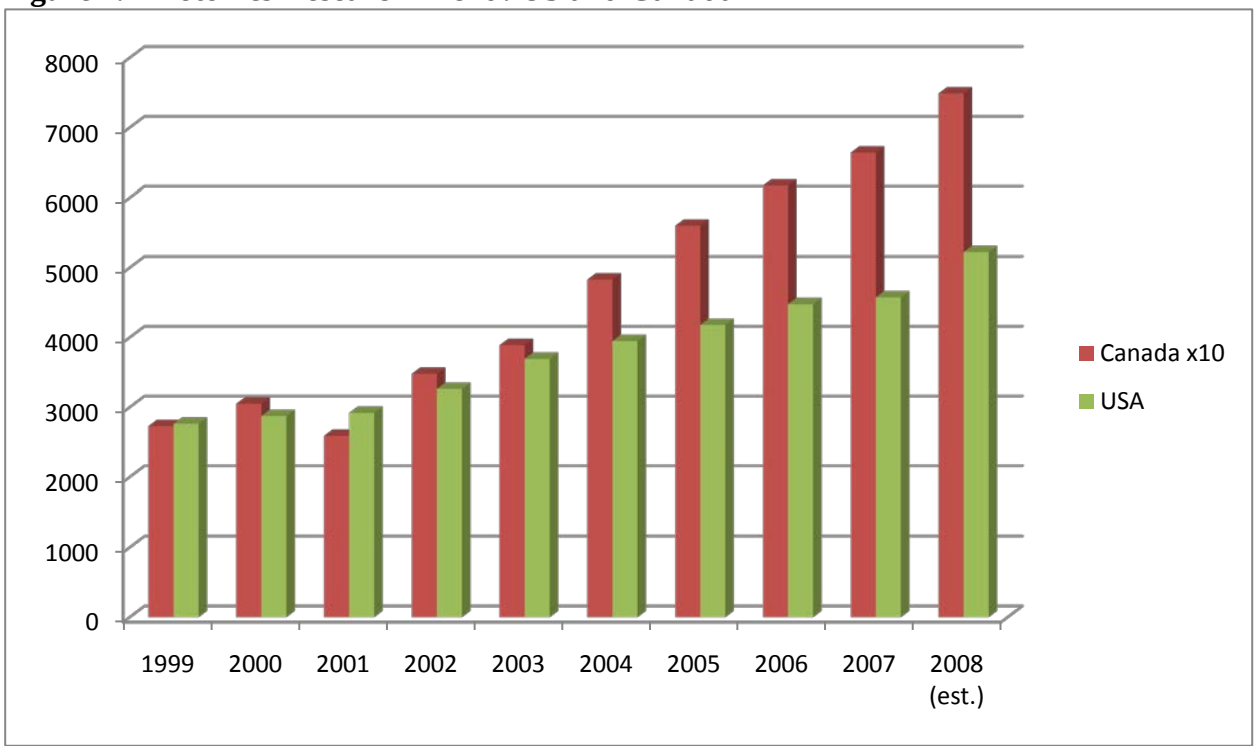
Steps 4 & 5) We imported the data into our software and extracted the desired fields for further analysis (e.g., authors, author organizations, key terms, title phrases, publication year, times cited). We cleaned and consolidated these fields.

Steps 6 & 7) Analyses – we elaborate on these below.

Step 8) We take advantage of a number of existing macros to generate these results.

Figure 2 shows the overall research trend, comparing Canada to the US. Of special interest for the benchmarking comparison is that Canadian photonics research increases relative to the US over this decade.

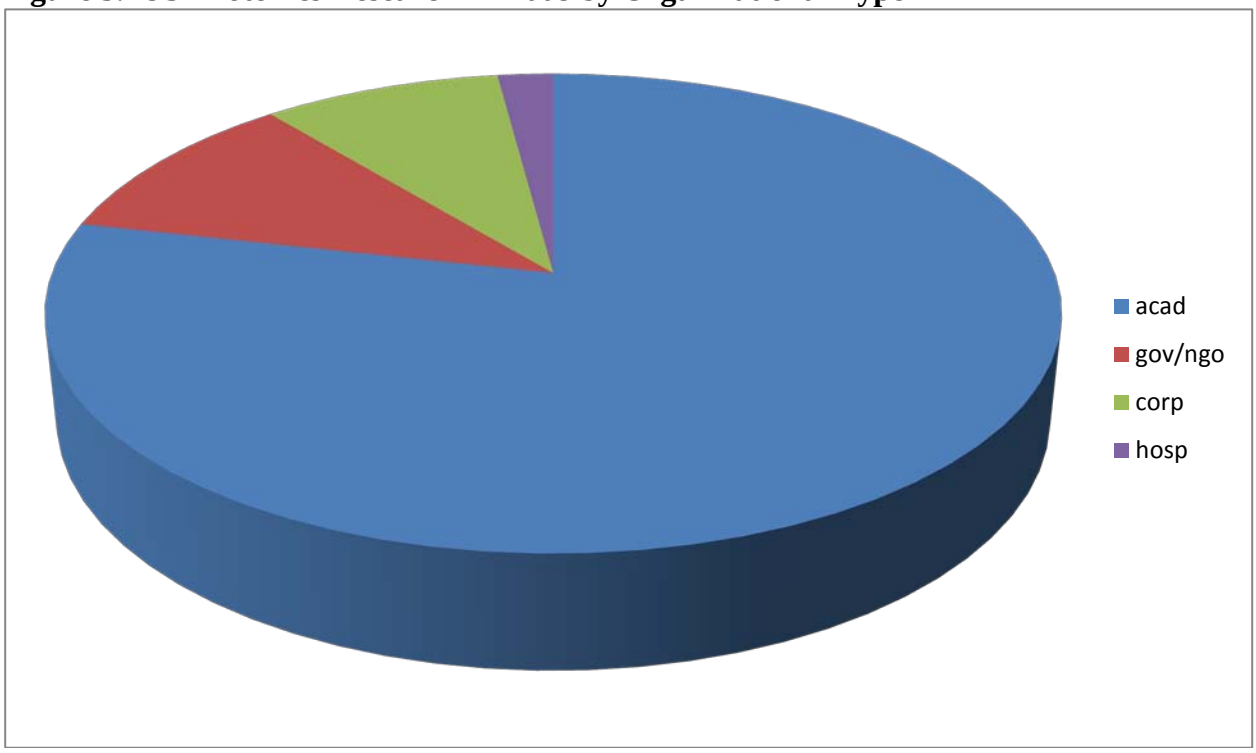
Figure 2. Photonics Research Trend: US and Canada



Note: In that SCI continues indexing additional publications well after the calendar year ends, we multiplied the 2008 tally by 1.2 to get the 2008 (est.) values.

The remaining illustrations give a current profile based on US photonics research in 2008. As mentioned, we find that approximating the percentage of R&D from industry is a telling innovation indicator (albeit one with important nuances). Figure 3 shows the sector breakout. Interestingly, for the benchmarking, the Canadian corporate publication rate (7%) is considerably less than that of the US (12%). We also probed further to investigate the extent of cross-sector collaboration. In the US, for instance, 11% of the papers with an academic author also had an author with a corporate affiliation. Such evidence of research networking could be important for cluster development.

Figure 3. US Photonics Research in 2008 by Organizational Type



In investigating the effects of the photonics initiatives, we also broke out research publications to enable comparisons by region and by organization. These bibliometric analyses often enable one to set up quasi-experimental design comparisons to help infer cause and effect. For instance, if one wants to assess the impacts of university U setting up a major new photonics research center in Year X, you can design a set of comparisons (e.g., of research publications and citations received):

- comparing U time series before and after Year X (probably setting aside Year X+1 to allow for effects to show up) – does research activity escalate after the center is formed?
- comparing U with other similar universities before (hypothesizing similarity) and after (hypothesizing difference)

We also wanted to compare topical emphases between the US and Canada. To do that, we explored information at different levels of detail. For instance we extracted the prominent key words and title phrases prevalent in sets of the articles. We also pulled out information on the Subject Categories into which SCI classifies the publication journals. We then aggregated those into Macro-disciplines and compared the US and Canadian publications for 2007-08. Table 2 shows the leading ones, and we do see some interesting differences (e.g., relatively more Canadian photonics work in Physics).

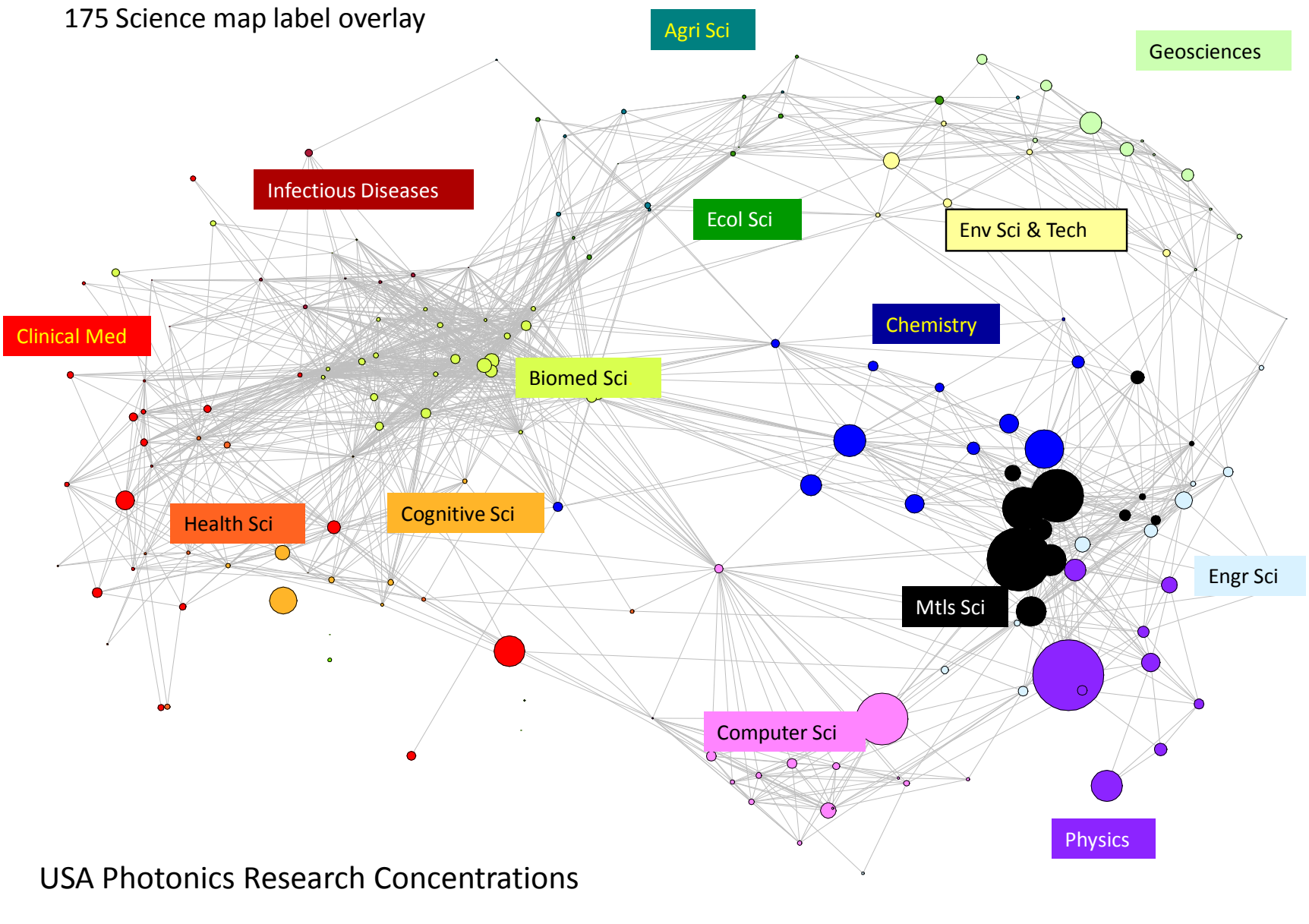
Table 2. National Photonics Publication Macro-Disciplines, SCI, 2007-08

	# Records	8715	1251
#	**Macro-Disciplines**	**USA**	**Canada**
3844	Physics	38.4%	44.9%
3722	Mtls Sci	38.2%	33.3%
1489	Computer Sci	14.4%	19.5%
1467	Chemistry	15.2%	12.8%
843	Clinical Med	8.8%	7.1%

To give a visual feel for the research fields engaged in this photonics work, Figure 5 presents a science overlay map. Against a background of intersecting arcs to represent the 175 Subject Categories covered by SCI, we overlay the photonics publication activity.[1] Here we quickly grasp that photonics research touches into a range of fields. So, if someone wants to bolster R&D leading to economic returns, funding and other support should not be concentrated in a single field. Furthermore, one may want to take action to promote cross-field collaboration. Figure 5 suggests involving Materials Scientists, Physicists, Chemists and Computer Scientists particularly.

[1] This mapping process categorizes articles indexed in Web of Science according to the journals in which they appear (Rafols and Meyer, forthcoming; Leydesdorff and Rafols, 2009). The Subject Categories are then grouped into "macro-disciplines" using a form of factor analysis (Principal Components Analysis) based on degree of association. Those Macro-Disciplines become the labels in the figure. The photonics research concentrations appear as nodes on these maps. Larger nodes reflect more publications.

Figure 2. US Photonics Concentrations – Overlay on a Base Map of Science



This investigation also sought to compare particular institutions to see how those engaged in research clusters have fared. Tech Mining allows us easily to profile numbers of organizations by applying a macro. We indicate which fields of information we want to see, for which institutions. For fun, Table 3 profiles photonics research activity for three leading US "institutes of technology."

Table 9. Profiles of 3 Universities: 2007-08 Photonics Research (indexed in SCI)

Organization	Macro-Disciplines	Select Terms	Countries	Collaborating Organizations	h
MIT [273]	Physics [131] Mtls Sci [110] Chemistry [48] Computer Sci [46]	crystal [60] lithograph* [30] nano [29] quantum [25] fiber [24]	Germany [18] UK [16] Japan [14] Singapore [12]	Harvard Univ [15] CALTECH [14] Stanford Univ [9]	14
Caltech [253]	Physics [145] Mtls Sci [43] Computer Sci [32] Geosciences [31]	crystal [31] lens [21] quantum [21] nano [17] solar cell [17]	France [42] UK [31] Germany [25] Canada [14]	NASA [27] Univ Arizona [22] UCLA [15] MIT [14]	14
Georgia Tech [194]	Mtls Sci [113] Physics [68] Computer Sci [49] Chemistry [27]	nano [48] crystal [44] lithograph* [25] fiber [23]	China [13] France [11] Germany [9] Taiwan [6]	NEC Labs Amer [12] Emory Univ [10]	13

This research profile breaks out for each organization:

- Top Macro-Disciplines reflected by their publications – this can help spot differences in emphases of their photonics research – contrast Caltech's physics emphasis with Georgia Tech's materials science.
- Top Key Terms – this may help locate who is researching a particular topic of interest.
- Countries – We see that Caltech seems more internationally networked than is MIT, and that Caltech's collaboration is heavily with France, the UK, and Germany, with Canada fourth.
- Collaborating Organizations – MIT shows strong ties with neighboring Harvard; Caltech, with NASA (Jet Propulsion Lab is a joint endeavor); & Georgia Tech, with a firm, NEC. This may give leads on potential interest in partnering.
- The h-index is a measure of the extent of citation of the organization's publications.[2]

C. Uses

Effective uses for Tech Mining are still being discovered. This section aims to stimulate your thinking of how this tool set could help meet your information needs.

[2] "Times Cited" provides a rough measure of paper impact (or quality) – i.e., how often do others cite the given paper. The distribution of citations is highly skewed – i.e., a very few papers receive many citations; most papers receive few or none. We examine citation intensity for the selected organizations' papers using the "h-index" – the number of papers (N) with at least N citations each. This metric is useful because it discounts the disproportionate weight of highly cited papers. The h-index was developed by J.E. Hirsch and published in *Proceedings of the National Academy of Sciences of the United States of America* 102 (46): 16569-16572 November 15, 2005.

Competitive Technological Intelligence ("**CTI**") has grown aggressively over the recent two decades, as companies see the need to know what others (companies, universities, government agencies) are doing to develop particular technological capabilities. Text mining of results from topical searches in extensive R&D databases is an essential part of CTI. This is becoming ever more critical in an era of "Open Innovation" in which R&D is no longer used exclusively in-house (see References).

Technology Roadmapping (see Appendix) also is growing in use as a vital FTA method. Industry associations [c.f., www.sia-online.org/], government agencies [c.f., http://www1.eere.energy.gov/industry/aluminum/partnerships.html], and individual companies seek to lay out prospects for development of families of related technologies (systems and components) over the next several generations. Particularly compelling are efforts by companies, such as Motorola, to map expected developments in technologies together with product development trajectories. Tech Mining can be a key contributor to understanding external competitive environments [c.f., http://mati.ncms.org/].

Research evaluation is essentially retrospective in focus, but conveys value to foresight activities by helping to manage R&D programs. Tech Mining has a role to play in comparing and assessing the research outputs of various organizational groups. Certainly, one would not base R&D management or policy solely on counting publications (bibliometrics) or patents, but innovation indicators can provide deeper insights into developmental paths and prospects. Another handy application is to use rapid text mining to inform proposal evaluation processes. Confronted with an unfamiliar proposed research project, a program manager can quickly (and cheaply) identify how much research is ongoing on the topic, what research groups are most prolific, and see how the proposer compares. The manager can also get contact information on potential reviewers knowledgeable on the topic in question.

National foresight studies can use Tech Mining to compare national R&D outputs with those of benchmark nations, or with their country's economic development targets. This can help identify gaps that may merit attention. Our photonics case exemplifies this sort of analysis.

You may have noted the heavy emphasis on utilization issues in this chapter. I choose to raise these because of our experiences in doing Tech Mining – awareness of these hurdles is a first step toward overcoming them. Furthermore, many of these same issues pertain to futures research in general (c.f., Gordon & Glenn, 2000). We probably all need to work as much on effective presentation of results as on futures research per se.

"Who" should perform Tech Mining? Today, this is probably best done by those with a serious, ongoing stake in generating empirical technology intelligence or foresight. These include patent analysts, CTI specialists, tech forecasters, and information specialists who see the need to provide analyses, not just deliver source materials. I believe this "who should do it" answer is expanding. I advocate "research profiling" – i.e., limited Tech Mining to understand one's research context – be done by "everyone." That is, every researcher ought to study the body of R&D publication (patents, etc.) relating to his or her endeavor. For instance, suppose you are analyzing "fuel cells" for a college paper (or for your organization's product development). It behooves you to overview who is doing what research on various types of fuel cells and related

technologies. Furthermore, using simple search engine capabilities you can determine trends. For example, which of the five types of fuel cells are being most actively pursued (and, consequently, have the greatest prospects for continuing advancement)? You should also look for contextual influences apt to support, or impede, further fuel cell innovation. By tabulating "hits" on particular topics within the fuel cell literature, you can gain useful clues. For instance, are newspapers "screaming" about a certain fuel cell attribute destroying the environment? Might these pressures portend regulation?

I encourage all of you to use the tools at hand to profile the technological context. Those tools will be advancing considerably over the coming few years. Check out what profiling/mining capabilities you might have already – e.g., searching *Chem Abstracts* using their *SciFinder* software allows you to list and count instances by author, keyword, date, etc. Or, even more basically, you can tabulate interesting information right as you search with simple search engines. For instance, if you are searching *INSPEC* for "fuel cells," you could successively search on "fuel cells and 1995," "fuel cells and 1996," etc., to then plot the R&D activity trend. By having the idea of mining these resources, you'll be amazed at what intelligence you can generate using your available tools.

VI. Frontiers of the Method

Tech Mining is a young method. Most promising is the advance in interpretable innovation indicators that benchmark technological progress. Our *VantagePoint* software was commercially introduced in 2000; improvements in statistical algorithms and representation modes continue at a strong pace. Most promising is the advent of "macros" (Visual Basic scripts) that control sequences of steps in both *VantagePoint* and Microsoft software (such as *MS Excel* and *Powerpoint*). Such macros enable automated production of preferred information products in desired form on the desktop.

An intriguing development entails applying a series of text mining analyses to discover inherent cross-relationships not obvious to researchers in individual research domains (see Appendix for Swanson and Kostoff notes). Such Literature-Based Discovery can leverage the expansive repertoire of R&D results to enhance one's own research and new product development.

Appendix

Software resources pertinent to Tech Mining evolve rapidly. Here's one comparative review of many software tools, in terms of doing patent analyses: Yang, Y.Y., Akers, L., Klose, T., and Yang, C.B. (2008), Text Mining and Visualization Tools – Impressions of Emerging Capabilities, *World Patent Information*, Vol. 30 (#4), December, 280-293.

We are seeing increasing involvement of database vendors in provision of analytical tools to use with their data. For instance, Thomson Reuters now offers several of these software aids (Aureka, ClearForest, Thomson Data Analyzer).

Here's a short list of software especially helpful in extracting intelligence from structured text resources:

Anavist works with Chem Abstracts, and is offered by STN: www.cas.org/products/anavist/index.html.

Aureka offers particularly attractive visualizations of relationships among patents: http://www.micropatent.com/0/aureka_online.html.

ClearForest based on the work of R. Feldman of Bar-Ilan University, Israel, See http://www.clearforest.com.

DB2 Intelligent Miner, from IBM, is not end-user software, but provides a promising platform: http://www-01.ibm.com/software/data/iminer.

Matheo software from France: http://www.matheo-software.com.

Temis is European-based text analytics: http://www.temis.com.

VantagePoint is presented at http://theVantagePoint.com.
[Closely related versions of the software are available – *TechOASIS*, for U.S. Government use; and *Thomson Data Analyzer* from Thomson Reuters.]

Selected literature and pertinent Tech Mining approaches:

The Georgia Tech Technology Policy and Assessment Center's approach, called "Technology Opportunities Analysis," is treated on http://tpac.gatech.edu. Additional Tech Mining papers are available at: //theVantagePoint.com.

M. Brenner, Technology intelligence at Air Products: Leveraging analysis and collection techniques, *Competitive Intelligence Magazine*, Vol. 8 (3), 6-19, 2005.
[a great description of how a major company integrates multiple tools to deliver a range of technical intelligence results]

Chesbrough, H.W. *Open Innovation: The New Imperative for Creating and Profiting from Technology*, Harvard Business School, Cambridge, MA (paperback edition), 2006.

Cunningham, S.W., Porter, A.L., and Newman, N.C. Tech Mining Special Issue, *Technology Forecasting and Social Change*,Vol. 73 (8), 2006.

T. J. Gordon and J. C. Glenn, *Factors Required for Successful Implementation of Futures Research in Decision Making*, Army Environmental Policy Institute, Alexandria, VA.

L. Huston, and N. Sakkab, Connect and Develop, *Harvard Business Review*, March, 58-66, 2006.
[a compelling success story for Open Innovation at P&G]

Ron Kostoff – "Database Tomography" and extensive literature reviews on S&T analyses, including Literature-Based Discovery and Technology roadmapping processes: http://www.dtic.mil/dtic/search/tr/ [guided search for KOSTOFF retrieves ~50 of his studies]

A.L. Porter, A.T. Roper, T.W. Mason, F.A. Rossini, and J. Banks, *Forecasting and Management of Technology*, New York: John Wiley, 1991 (second edition under development as of 2009).

A.L. Porter, QTIP: Quick Technology Intelligence Processes, *Technological Forecasting and Social Change*, Vol. 72, No. 9, 1070-1081, 2005.

A.L. Porter, and S.W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage*, Wiley, New York, 2005.

A.L. Porter, A. Kongthon, J-C. Lu, Research Profiling: Improving the Literature Review, *Scientometrics*, 53, 351-370, 2002.

Anthony van Raan and colleagues at the University of Leiden – bibliometric mapping: http://sahara.fsw.leidenuniv.nl/cwts/

R.J. Watts, R.J., and A.L. Porter, "Innovation Forecasting," *Technological Forecasting and Social Change*, Vol. 56, p. 25-47, 1997.

L. Leydesdorff, and I. Rafols, A Global Map of Science Based on the ISI Subject Categories. *Journal of the American Society for Information Science and Technology*, 60(2), 348-362, 2009.

I. RAFOLS, and M. MEYER, Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, forthcoming.