AI and Global Science and Technology Assessment

Hsinchun Chen, *University of Arizona*

Many countries are aspiring to advance their economies through science and technology (S&T) innovations. Some have been more successful than others and are becoming major players in the international intellectual landscape. For example, Figure 1 shows some results from a bibliometric study that my colleagues and I conducted at the University of Arizona's AI Lab.[1] The study was based on academic nanotechnology papers published in the Thomson Science Citation Index's (SCI) Web of Sciences database from 1976 to 2004.[2] Prior to 1991, the US, Japan, Germany, France, and the UK were the major countries publishing nanotechnology research. After 1991, several additional countries joined them. By 2003, China was the second most productive country. South Korea also showed rapid development after 2000. In four years, its output exceeded Italy, Russia, and England to become the sixth most productive country in 2004.[2]

S&T strength is the foundation of a nation's economic power, so an effective, automated means of continually assessing this strength is critical to understanding a country's economic status. Such assessments require investigations in several dimensions:

- *Participants*. Who are the scientists and developers involved in R&D, and which institutions and companies ultimately benefit from these activities?
- *Processes*. What are the funding models/programs, and how are the participants linked together and organized around research initiatives?
- *Output*. What ideas, inventions, and innovations result, in which areas of technology and with what quality?
- *Benefits*. What are the economic and military advantages obtained from participation in and outputs from technology development?
- *Barriers*. Do any cultural or political factors hinder the effectiveness of a country's R&D in its quest to become a power in the global economy? (For a discussion of recent US regulations and their potential to restrict global S&T, see the "Regulatory Restrictions on Global S&T" sidebar on p. 70.)

Global S&T analytics addresses many such questions. AI, knowledge mapping, and content-analysis research contribute significantly to the answers.

In addition to analytics, global S&T assessment requires advances in several data collection and computational research areas, such as multilingual query and translation support, multimedia and unstructured data collection and management, and content analysis and visualization. These areas also benefit from AI, knowledge mapping, and content-analysis research.

For instance, Nano Mapper (http://nanomapper.eller.arizona.edu) is a knowledge mapping system that integrates the analysis of nanotechnology patents and grants into a Web-based platform. The Nano Mapper system contains nanotechnology-related patents from the US, European, and Japanese patent offices as well as information from the US National Science Foundation (NSF) grant documents. It provides simple search functionalities and a set of analysis and visualization tools that users can apply to different analytical units

1541-1672/09/$26.00 © 2009 IEEE
Published by the IEEE Computer Society

over different time periods. For example, Figure 2 shows a visualization for US Patent and Trade Organization (USPTO) patent citations from different countries and institutions over a 30-year period.

In this Issue

This issue includes five essays on global S&T assessment from distinguished experts in knowledge mapping, scientometrics, information visualization, digital libraries, and multilingual knowledge management. Each essay presents an innovative research framework, computational methods, and selected results and examples.

In the first essay, "China S&T Assessment," Ronald N. Kostoff proposes three fundamental S&T assessment metrics. "Right job" addresses the overall investment strategy. "Job right" addresses the S&T approach. "Productivity/progress" addresses the S&T output and impact. Using scientometrics techniques, Kostoff shows the strong Chinese emphasis on the physical and engineering sciences and its significant research productivity gains over the past two decades.

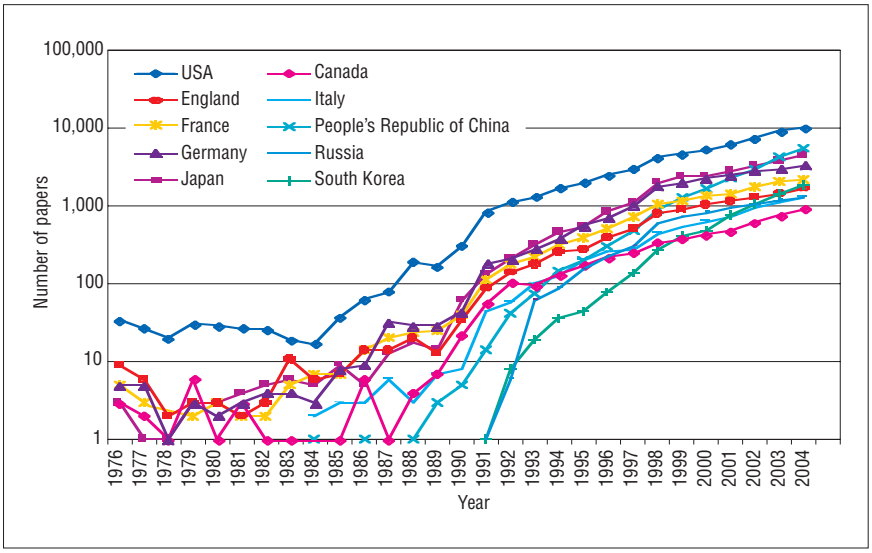


Figure 1. Top 10 countries and regions publishing nanotechnology papers (1976–2004). Although the US still publishes the most papers on this topic, the rapid growth of China's contributions after 1991 moved it to second place by 2003.

In "Mapping the Sloan Digital Sky Survey's Global Impact," Chaomei Chen, Jian Zhang, and Michael S. Vogeley adopt scientometrics and visualization techniques to study the publication and usage patterns of

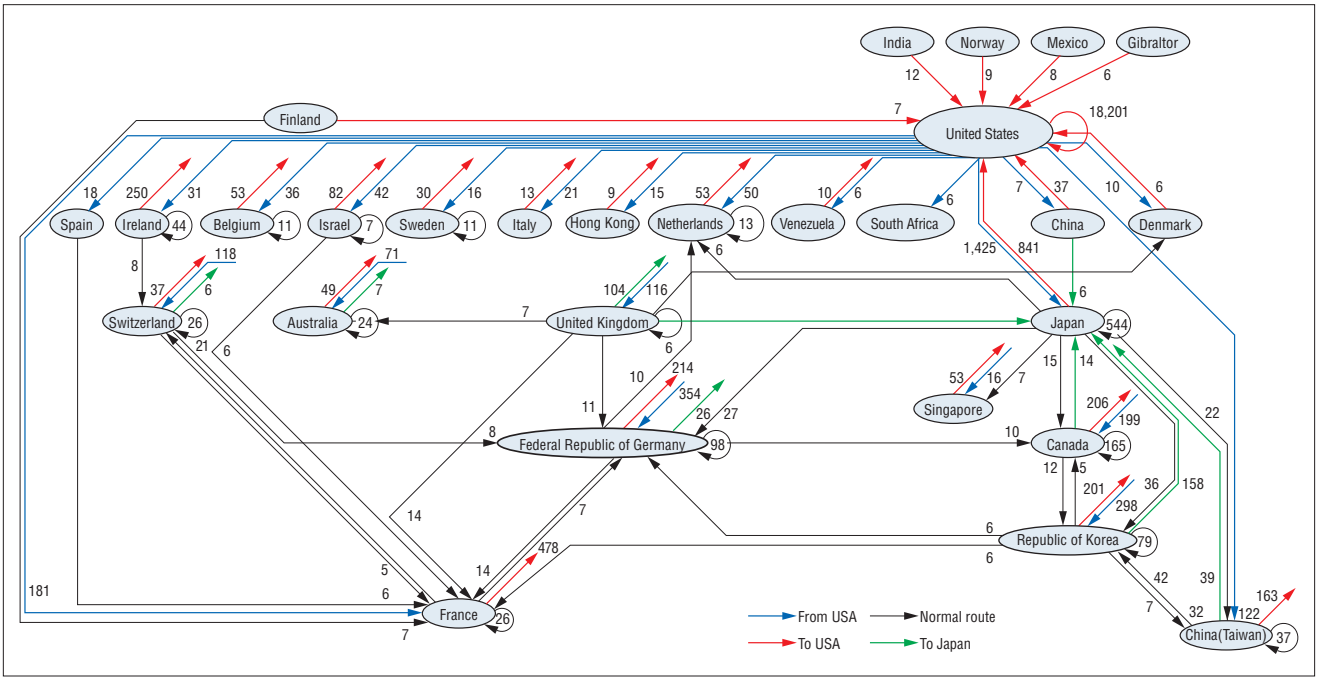


Figure 2. Country citation network of nanotechnology-related patents for the US Patent and Trade Office (1976–2006). The US appears at the top of the figure as the most-cited country. Japan, the second-most-cited country, appears below it and to the right.

Regulatory Restrictions on Global S&T

With the increasing emphasis on science and technology development by different countries and the competitive landscape of innovation and commercialization, S&T protectionism can also become a potential barrier for global knowledge diffusion. Protectionism is nothing new in modern economies. It's often administrated by different countries for strategically and socially important areas, ranging from farm products and the fishing industry, to advanced information technology and military systems. However, we've witnessed increasingly strict enforcement of regulations that control the transfer of equipment, technology, and know-how to foreign countries and nationals. Although this problem might be much more alarming on US university campuses lately, it wouldn't be surprising to see other countries adopt similar protective measures.

In the US, the two primary statues covering exports are the Arms Export Control Act and the Export Administration Act. These acts authorize two sets of regulations, the International Traffic in Arms Regulations (ITAR) and the Export Administration Regulations (EAR). ITAR covers items that are inherently military in nature. The Department of State administers these regulations, which include the Munitions List, delineating controlled types of items and technologies. EAR covers "dual-use items," which can be used for either military or civil purposes. The Department of Commerce administers these regulations, which include the Commerce Control List. The EAR defines a "deemed export" as the release to a foreign national of technology or source code subject to the EAR. These situations might include laboratory tours, joint research conducted with foreign students or professors, and even email, visual inspections, and oral exchanges.

The US regulations particularly target foreign nationals from countries such as Iran and Cuba. Research conducted by faculty and students at a university is normally considered fundamental research and is excluded from ITAR and EAR regulations. However, university-based research isn't considered fundamental research if the university or its researchers accept restrictions on the publication of the project results—for example, proprietary restrictions or a requirement for sponsor approval prior to publication.

Recently many research universities have become vigilant in informing faculty and students about these export regulations and in enforcing due process,[1] especially for defense or security-related research projects from US federal agencies. Universities frequently suggest that relevant projects develop a technology control plan to outline the procedures needed to secure controlled technology from use and observation by unlicensed non-US citizens. For more information about export control regulations and processes, many universities have begun to provide useful resources—for example, the University of Maryland (www.umresearch.umd.edu/ORAA/ecg/index.html).

Why are such measures of relevance to *IEEE Intelligent Systems* readers? What is the cost of noncompliance?

ITAR violation can result in up to $1 million per violation and 10 years of imprisonment. Professor Reece Roth in the University of Tennessee's Department of Electrical and Computer Engineering was convicted in September 2008 and faces up to 160 years in jail and $1.5 million in fines for disclosing restricted US military data about unmanned aerial vehicles to foreign nationals without first obtaining the required US government license or approval.[1] EAR violation can result in fines of $50,000 or five times the value of export, whichever is greater, per violation, and 10 years of imprisonment. Professor Thomas Butler of the Texas Tech University faces 2 years in prison for making fraudulent claims and unauthorized export (plague bacteria).

Although these might be isolated incidents, researchers in academic institutions must become more knowledgeable about such developments and watchful in future international collaborations. University boards and government agencies need to debate and evaluate the impacts of such measures in light of the unstoppable force of global S&T development, diffusion, cross-fertilization, and competition.

Reference

1. R. Monastersky, "Professor's Conviction on Export Violations Alerts U.S. Universities," *Chronicle of Higher Education*, 8 Sept. 2008; http://researchintegrity.asu.edu/security/documents/Export_Violations_Article-Chronicle9-8-08.pdf.

researchers in the Sloan Digital Sky Survey (SDSS) astronomy community. In addition to identifying S&T assessment challenges, the authors demonstrate the integral roles computational algorithms and advanced visualizations can play in science policy making and monitoring, in tracking the diffusion of knowledge, and in matching expertise and resources with local and global needs.

In the third essay, "Open Data and Open Code for S&T Assessment," Katy Börner, Nianli Ma, Russell J. Duhon, and Angela M. Zoss introduce "science maps" to help humans mentally organize, access, and manage complex digital library collections. The maps are based on the authors' Scholarly Database project at Indiana University. Their essay shows how S&T studies can benefit from selected free data from the NSF, National Institutes of Health, and USPTO, together with free code—namely, the Network Workbench tool.

In "Global S&T Assessment by Analysis of Large ETD Collections," Venkat Srinivasan and Edward A. Fox introduce the highly successful Networked Digital Library of Theses and Dissertations (NDLTD) project. As of March 2009, NDLTD has 663,515 electronic theses and dissertations (ETDs) from universities around the world. Using the NDLDT's Union Catalog metadata in a training set and a naïve Bayes classifier, the authors demonstrated a semiautomatic approach to topic categorization. The research can help identify emerging topics in relevant S&T collections.

The fifth and final essay, "Managing Multilingual S&T Knowledge" by Christopher C. Yang and Chih-Ping Wei, describes a research framework for cross-lingual and polylingual text categorization and category integration. They illustrate the significance of cross-lingual document retrieval and management for global S&T assessment and identify rich future research directions.

Global S&T opens the door to global cooperation as well as competition. Again, I use China as an example. Chinese researchers have published a wealth of information about S&T developments beyond nanotechnology. Except for publications in major English journals and conference proceedings, much of this material is difficult for scholars outside China to locate or access, and most of it is unknown outside a small circle of researchers. One of the most comprehensive Chinese academic databases, the Wanfang Data, contains 13,971,265 articles from 6,065 journals, 918,915 conference articles, and 1,184,412 dissertations (as of 7 July 2008), all of them in Chinese.

The breadth and depth of such material in China and other emerging economies offers insight into everything from industry and agriculture, to technology development and scientific research, to politics and military issues. Exploring these information resources can help advance economies throughout our evolving world.

Acknowledgments

Some findings reported in this essay are results of an NSF funded project, "NanoMap: Mapping Nanotechnology Development," DMI-0533749, Aug. 2005–July 2008.

References

1. H. Chen and M. Roco, *Mapping Nanotechnology Innovations and Knowledge: Global and Longitudinal Patent and Literature Analysis*, Springer, 2009.
2. Li et al., "A Longitudinal Analysis of Nanotechnology Literature: 1976–2004," *J. Nanoparticle Research*, vol. 10, suppl. 1, Dec. 2008, pp. 3–22.

Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona and director of the Artificial Intelligence Lab. He received his PhD in information systems from New York University. Contact him at hchen@eller.arizona.edu.

China S&T Assessment

Ronald N. Kostoff, *Mitre Corp.*

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

What are the central principles in conducting an S&T assessment? In my *Handbook of Research Impact Assessment*,[1] I identify three foundational S&T assessment metrics, whether for a project, a program, or a nation's total S&T output. I summarize these as right job, job right, and productivity/progress. "Right job" addresses the overall investment strategy: Are the larger S&T objectives being addressed correctly? "Job right" addresses the S&T approach: Are the best techniques being used to conduct the S&T? "Productivity/progress" addresses the S&T output and impact.

In this brief essay, I provide examples of how to use these metrics to assess the S&T of a rapidly growing country—namely, the People's Republic of China. To place the assessment in context, I compare China's metrics with those of the leader in S&T output—namely, the US. I could have used countries such as India for the baseline,[2] but my goal here is to show how far China must go to become the leader in S&T output metrics. Much more detailed exposition of the use of these metrics in assessing China's S&T output are available elsewhere.[2–5]

Right Job

S&T strategy, as reflected in published technical output in the global literature, can be inferred from different perspectives. Clustering documents by technical discipline provides one categorization approach,[4] and it's perhaps the main approach used.

A complementary approach is to show relative areas of technical emphasis among multiple countries. In 2007, my colleagues and I downloaded equal numbers of US and China research articles from basic and applied research databases and compared the occurrence frequencies of phrases.[2,4] Table 1 reflects a conceptually similar approach to compare research discipline emphases in the US and China. The Science Citation Index (SCI), the premier database of research journals, includes a subject category field for each record—that is, for each article published. This field indicates the main technical discipline for the

Table 1. Ratio of China/US articles with subject categories specified.

Subject category	China/US ratio
Crystallography	18.55
Metallurgy and metallurgical engineering	15.91
Materials science, textiles	6.21
Materials science, ceramics	6.02
Chemistry, inorganic and nuclear	5.66
Polymer science	5.62
Materials science, composites	5.55
Chemistry, applied	5.11
Physics, multidisciplinary	5.06
Electrochemistry	4.84
Mathematics, applied	4.67
Materials science, multidisciplinary	4.65
Chemistry, physical	4.23
Energy and fuels	4.21
Engineering, chemical	4.18
Sociology	0.06
Psychology, multidisciplinary	0.05
Women's studies	0.04
Law	0.04
Psychology, mathematical	0.04
Political science	0.04
Humanities, multidisciplinary	0.04
Psychology, biological	0.04
Ethnic studies	0.04
Medical ethics	0.03
History and philosophy of science	0.03
History of social sciences	0.02
Religion	0.02
Philosophy	0.02
History	0.01
Psychology, psychoanalysis	0.01

journal in which the article was published. For this essay, I examined the subject category distribution for the 100,000 most recent articles (ending 31 December 2008) published in the SCI from China and the US. I downloaded the subject categories and their frequencies. For each of almost 500 categories, I computed the ratio of China's frequency to that of the US, then sorted the list according to the China/US ratio.

The first 15 categories in Table 1 represent strong technical area emphasis by China relative to the US, and the last 15 categories are the reverse. These results, which I've replicated by other means and for other databases,[2–5] show China's strong relative emphases in the physical and engineering sciences and the US emphases in the biomedical, social, and psychological sciences. If we couple these results with China's strong production of technical graduates, then China's investment strategy is providing a solid technology-based foundation for future military and commercial competitiveness.

Job Right

The second metric addresses research quality. The only universally accepted indicator of publication quality is a panel of experts reviewing a specific document. One commonly used proxy metric for quality is the number of times other research articles cite an article. My colleagues and I examined the citation trend of China's published articles in nanotechnology, an area of strong emphasis in Chinese research.[5] The citation quality (percent of publications in the top citation tier) was low relative to that of the US, but it grew monotonically within a five-year period—from 4 percent of the US figure in 1998 to 20 percent in 2003, the latest period examined.

Another approach to assessing relative quality is to examine publication trajectories in high-quality journals. For these journals, articles must exceed a quality threshold to be accepted. I had three criteria for selecting journals to include in this assessment: high total citations, high citations per paper, and focus on specific physical science disciplines. Figure 3 compares the ratios of the number of Chinese to US articles published in two important SCI journals—namely, the *Journal of the American Chemical Society* (JACS) and the *Journal of Applied Physics* (JAP). The figure includes a comparison to total China/US article production in the SCI.

Over the past decade, the China/

US ratio for total SCI nanotechnology articles grew by about a factor of eight; the ratio for JACS articles grew by an order of magnitude, and the ratio for JAP articles grew by more than a factor of five. These quality findings reflect results from earlier studies.[3–5] However, those studies also showed many Chinese articles being published in low-impact-factor journals. From this newest study, we can conclude that a small high-quality component is achieving rates of increase that match the overall growth in Chinese technical literature.

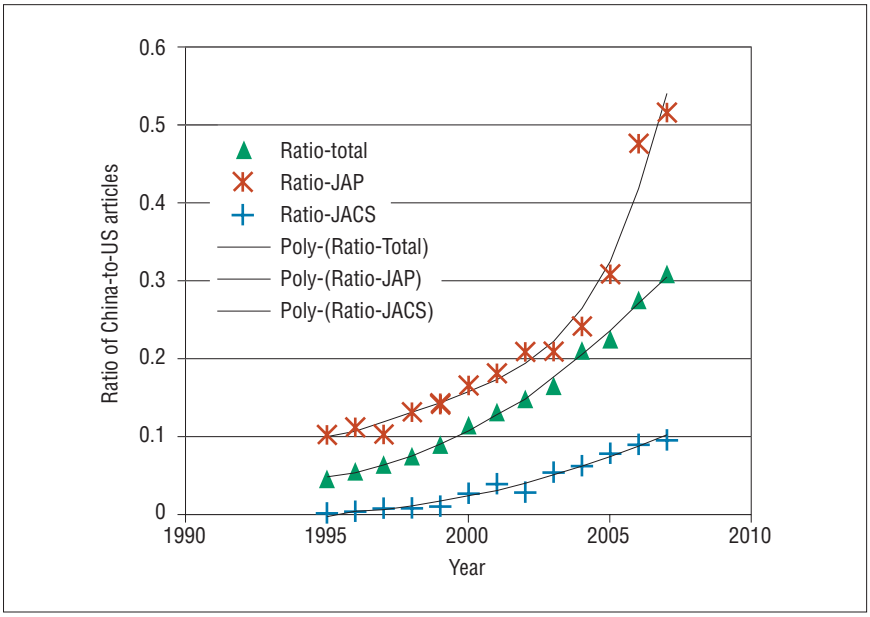


Figure 3. Ratio of China/US articles in flagship journals. The middle curve is the ratio of China/US articles for all journals. The upper curve is the China/US ratio for the *Journal of Applied Physics* (JAP), a leading physics journal, and the lower curve is the ratio of China/US articles for the *Journal of the American Chemical Society* (JACS), a leading chemistry journal.

Productivity/Progress

By any measure, China's productivity in published technical papers over the past two decades has been astounding. The bottom curve in Figure 4, reproduced from the middle curve in Figure 3 , shows outstanding relative total publication growth. The absolute publication growth numbers are equally impressive. However, I believe aggregate statistics have limited value for operational decision making. For bibliometrics, we must identify specific investment spikes to infer the true importance of an investment strategy. Table 1 addressed this issue to some extent.

Figure 4 provides an example of what we can derive from different levels of aggregation. The bottom curve, showing the overall China/US publication ratio, indicates that China lags the US in total SCI publications by a factor of three. The middle curve (ratio of overall nanotechnology publications) shows relative growth similar to the overall relative growth pattern, albeit starting at a higher relative level due to China's emphasis on nanotechnology. By this metric, China has essentially obtained parity with the US in overall nanotechnology publication production. The top curve, for the important nanotechnology subarea of nanocomposites, shows a substantially higher (and linear) rate of ratio increase relative to the other two curves. By this metric, China is 60 percent ahead of the US in nanocomposite publication production. At this level of detail, the analyst can examine specific

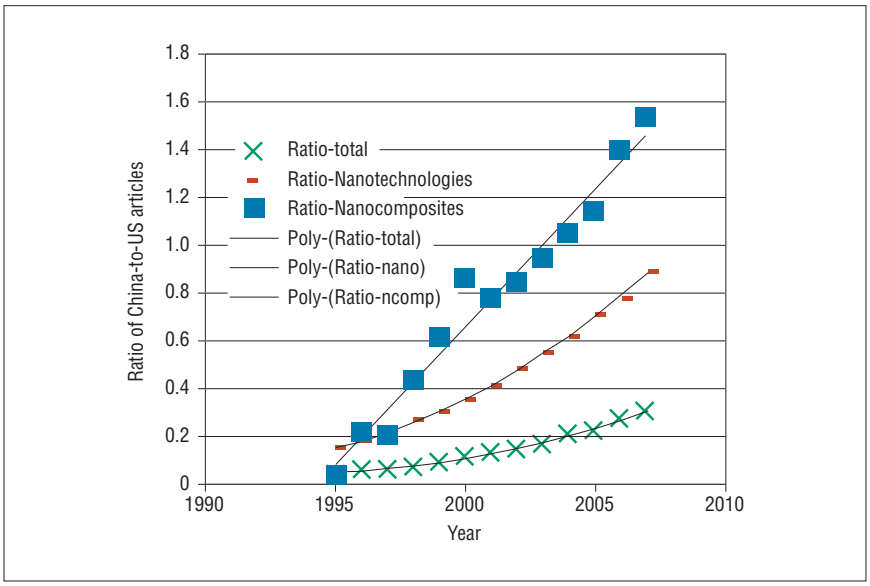


Figure 4. Ratio of China/US articles in nanotechnology at different aggregation levels. The bottom curve is the China/US ratio for all Scientific Citation Index (SCI) publications; the middle curve is the ratio for all SCI nanotechnology publications, and the upper curve is the ratio for all SCI nanocomposite publications. Nanocomposites are a subset of nanotechnology.

investment spikes, such as nanocomposites, and start to connect the dots to identify the investment strategy priorities on an integrated basis.

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

Acknowledgments

The views in this article are solely those of the author and do not necessarily represent the views of the Mitre Corp. The work contained in this article was supported in part by Mitre internal research funding.

References

1. R.N. Kostoff, *The Handbook of Research Impact Assessment*, 7th ed., DTIC tech. report ADA296021, Defense Technical Information Center, 1997.
2. R.N. Kostoff et al., "Comparisons of the Structure and Infrastructure of Chinese and Indian Science and Technology," *Technological Forecasting and Social Change*, vol. 74, no. 9, Nov. 2007, pp. 1609–1630.
3. R.N. Kostoff et al., "The Structure and Infrastructure of Chinese Science and Technology, DTIC tech. report ADA443315, Defense Technical Information Center, 2006.
4. R.N. Kostoff et al., "Chinese Science and Technology—Structure and Infrastructure," *Technological Forecasting and Social Change*, vol. 74, no. 9, Nov. 2007, pp. 1539–1573.
5. R.N. Kostoff, R.B. Barth, and C.G.Y. Lau, "Quality vs. Quantity of Publications in Nanotechnology Field from the People's Republic of China," *Chinese Science Bull.*, vol. 53, no. 8, Feb. 2008, pp. 1,272–1,280.

Ronald N. Kostoff is a researcher at the Mitre Corporation. He has a PhD in aerospace and mechanical sciences from Princeton University. Contact him at rkostoff@mitre.org.

Mapping the Sloan Digital Sky Survey's Global Impact

Chaomei Chen, Jian Zhang, and Michael S. Vogeley,
Drexel University

A country's scientific capacity is essential in today's increasingly globalized science and technology (S&T) ecosystem. Scientific capacity has four increasingly advanced capability levels: absorbing, applying, creating, and retaining scientific knowledge.[1] Moving up these levels requires more skill and training. For example, applying scientific knowledge requires more specialized skills than absorbing it. Similarly, making new discoveries requires more knowledge than applying existing procedures.

Research has shown the importance of addressing specific, local problems while tapping into globally available expertise and resources. Accessing scientific knowledge is the first step toward absorbing knowledge. Low-income countries have increased their access to scientific literature on the Internet,[2] but to what extent has this access led to more advanced scientific capacity?

Interdisciplinary and international collaboration might hold the key to creating and retaining knowledge.[3,4] For example, creative ideas tend to be associated with inspirations originating from diverse perspectives.[3] On the other hand, not all collaborations are productive. Assessing global S&T must therefore consider both successes and failures and the reasons behind them.

Sloan Digital Sky Survey

Researchers have addressed science policy issues by investigating the connection between the growth of a country's scientific publications and its economic capacity.[5] We focus on international collaborations associated with astronomy's Sloan Digital Sky Survey (SDSS; www.sdss.org) to illustrate some fundamental challenges for assessing global S&T in a rapidly growing and globalized research field.

The SDSS is the largest digital sky survey. It collects multiple types of data about stars, galaxies, quasars, and other astronomical objects in the universe. The survey has been funded by the Alfred P. Sloan Foundation, along with the participating institutions, the National Science Foundation, the US Department of Energy, NASA, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS releases the survey data to the public through the SDSS SkyServer website. Researchers and the general public can access the data directly on the Internet. SDSS I operated between 2000 and 2005. SDSS II operated between 2005 and 2008. SDSS III is operating currently.

Our analysis focuses on two data sources: the SkyServer's SQL query log and the bibliographic records of SDSS publications retrieved from the Web of Science (http://isi.knowledge.com). The query analysis aims to identify query patterns and areas of particular interest in the sky. We

use bibliometric analysis, text mining, and network visualization techniques to assess a country's capability of absorbing and applying knowledge in terms of growth in its scientific work force. We also want to reveal collaborating countries and collaboration topics and to investigate broad trends of data access, publishing, and impacts.

Rapidly Growing Scientific Capacity

Absorbing knowledge is easier than creating it: more countries access the SDSS data than contribute to the SDSS literature.

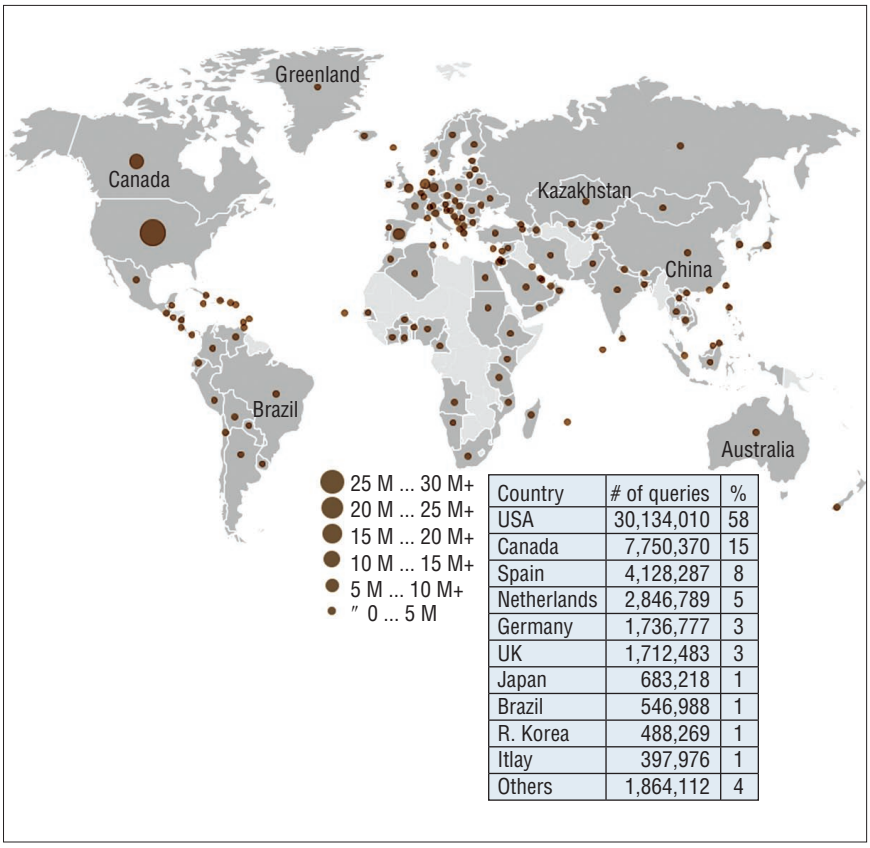
In general, we expect that the more data a country accesses, the more it publishes (see Figure 5). Determining the dynamics linking data access and publication is difficult. Reliably tracing data sources in unstructured or semistructured texts such as scientific papers remains a technical challenge for relevant fields such as natural language processing and ontology construction.

We retrieved 2,137 bibliographic records of SDSS publications between 1994 and 2008 from the Web of Science. Table 2 summarizes the statistics for these publications. We divided the last 15 years into three 5-year periods: 1994–1998, 1999–2003, and 2004–2008. Large increases are found at the country, institutional, and individual levels. Citations exceeding 300,000 in the periods 1994–1998 and 1999–2003 clearly indicate that SDSS's impact has reached far beyond the boundary of the international SDSS community. For example, the SDSS consortium from 2004–2008 has 25 participating countries. In contrast, authors who published in this period came from 51 countries.

Figure 6 depicts the dynamics of global SDSS research according to the growth rate of data access, publications, and citations received by each country. The acceleration of a given country's data access measures the growth rate from the half-life point of the accumulative data requests to the end of 2008—that is,

$$a = \frac{\frac{Q}{2}\left(\frac{1}{T_T - T_H + 1} - \frac{1}{T_H}\right)}{T_T - 2T_H + 1},$$

where Q is the total number of queries from the country and T_H and T_T



Country	# of queries	%
USA	30,134,010	58
Canada	7,750,370	15
Spain	4,128,287	8
Netherlands	2,846,789	5
Germany	1,736,777	3
UK	1,712,483	3
Japan	683,218	1
Brazil	546,988	1
R. Korea	488,269	1
Itlay	397,976	1
Others	1,864,112	4

Figure 5. Geographic distributions of SQL queries received by the SDSS SkyServer between 2003 and 2008. Many countries access the SDSS data despite considerable differences in their scientific capacity. (Source: http://manyeyes.alphaworks.ibm.com/manyeyes/visualizations/world-map-of-sdss-query-and-publicat, used by permission.)

Table 2. Statistics about SDSS publications over three 5-year periods.

Unique Items	1994–1998	1999–2003	2004–2008	Total
Countries	11	37	51	52
Institutions	54	529	2,352	2,619
Authors	135	1,103	3,879	4,372
Articles	47	369	1,722	2,137
Keywords	308	4,052	21,271	25,631
Phrases	1,019	9,838	46,801	57,658
References	955	8,265	31,647	35,999
Citations	15,828	388,638	329,008	733,474

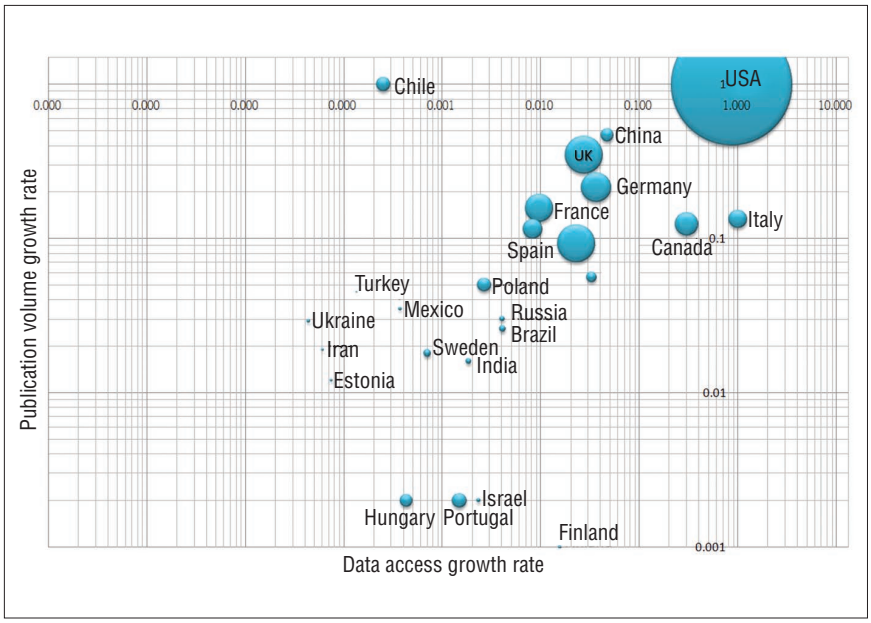


Figure 6. SDSS data access (on a logarithmic scale), publications (on a logarithmic scale), and citations (bubble size). The chart omits countries with zero or negative growths. A country's faster and larger access to data might not necessarily translate to a faster growth in publications or citations.

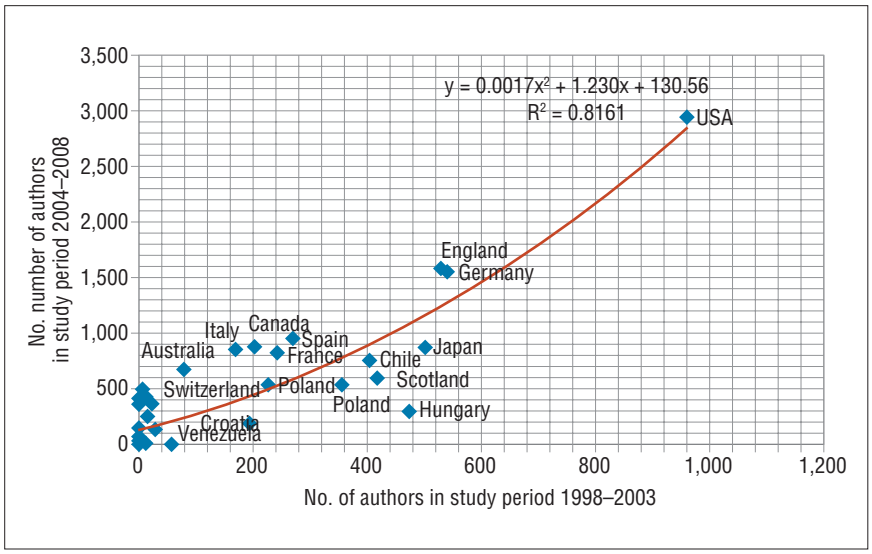


Figure 7. The growth of SDSS research workforces from the 1998–2003 study period to the 2004–2008 period. Workforces in countries above the trend line grew faster than countries below the line in the 2004–2008 period.

are the half-life and the total lifetime in months.

We compute the acceleration of a country's publications similarly. Figure 6 charts both acceleration rates logarithmically for clarity. The bubble sizes represent the citations received by the country. Chile appears as an outlier with much faster growth in publication than in data access. This observation invites further investigations into Chile's astronomical research infrastructure, its researchers' expertise, and its policies for international scientific collaboration.

In contrast, Italy appears as a different kind of outlier, with faster data access but a relatively slower publication rate. Such observations might direct global S&T assessment to focus on the role of local resources and how collaborating countries tap into shared resources and expertise.

An essential component of a country's sustainable S&T development is its population of active researchers. Growth in its scientific workforce directly reflects a country's potential. The US and Japan were the original SDSS participating countries. Germany joined in 1999. Since 2006, additional countries have participated in SDSS II, including the UK, Switzerland, and South Korea. Except in Japan, the research work forces in participating countries grew much faster than average (see Figure 7). The number of publishing authors in South Korea increased from 7 to 498 during the first two 5-year periods—that is, 1994–1998 and 1998–2003.

On the other hand, the workforces of a few nonparticipating countries grew remarkably as well. For example, a sharp increase in the number of active authors in Australia might be due in part to earlier sky surveys conducted by Australian astronomers, such as the two-degree field Galaxy Redshift Survey (2dF), which is often cited in the SDSS literature.

Global Impact

Capturing the big picture of international collaboration at a macroscopic level and simultaneously linking to subject matters at finer granularities is a long-standing challenge for computational methods.

Figure 8 illustrates how computational approaches can help improve our understanding across macro- and microscopic levels. The visualization represents two layers of information and is generated using the latest ver-

sion of CiteSpace, a freely available Java application for analyzing and visualizing scientific literature.[6] The *base layer* is a network of collaborating countries between 1994 and 2008. If researchers from different countries coauthored a published SDSS paper, the visualization connects those home countries. The *thematic layer* aggregates individual countries into clusters such that countries in the same cluster have tighter collaboration ties than those in different clusters. Each cluster reflects SDSS publications collaboratively written by researchers from these countries.

Researchers can choose cluster labels algorithmically at different abstraction levels, from the publication titles, their indexing terms, or noun phrases extracted from their abstracts. We used td*idf (term frequency × inverse document frequency) weighting to select the cluster labels in Figure 8 from indexing terms of the collaborative publications. In cluster 6, the predominant topic for collaborating researchers from Germany, England, Italy, and France is "halo," whereas in cluster 5, the primary focus of collaborations between Brazil and Argentina is likely "dark energy." Showing patterns at this level is useful not only for researchers in the trenches but also for science policy makers and evaluators.

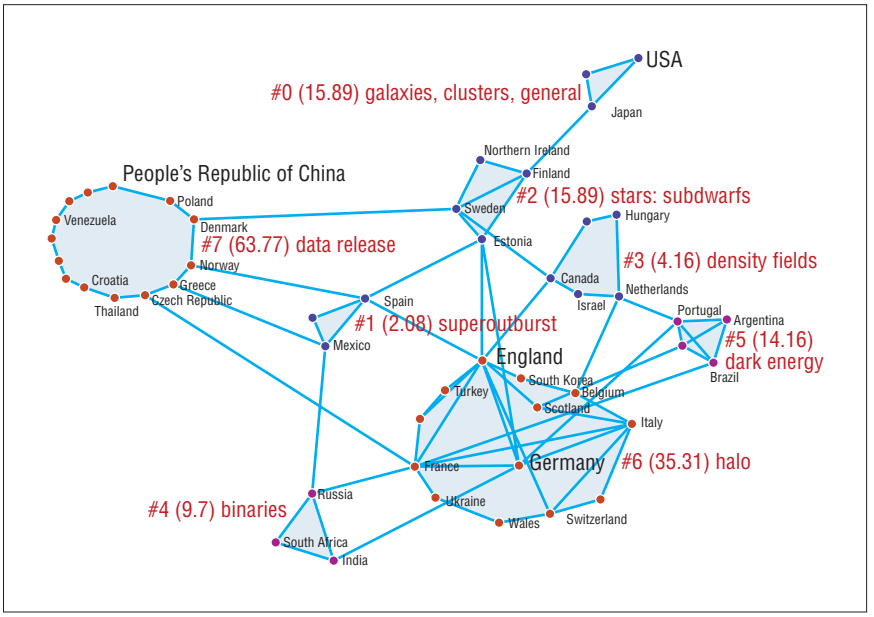


Figure 8. An international collaboration network of 51 countries and 55 collaborative ties in SDSS publications. The strength of collaborative ties identifies eight collaboration clusters. The classic tf*idf weighting scheme selects cluster labels from indexing terms in corresponding collaborative publications.

In summary, we've identified some of the challenges for assessing globalized S&T development and demonstrated some computational algorithms that can play integral roles in science policy making and monitoring. These approaches can also help in tracking how knowledge diffuses from large-scale, data-driven, cyber-enabled scientific activities and in matching complementary expertise and resources between local and global needs.

Acknowledgments

This work is supported in part by the National Science Foundation under grant IIS-0612129.

References

1. C.S. Wagner, *The New Invisible College*, Brookings Press, 2008.
2. E.A. Henneken et al., "Use of Astronomical Literature: A Report on Usage Patterns," *J. of Informetrics*, vol. 3, no. 1, 2009, pp. 1–8.
3. C. Chen et al., "Towards an Explanatory and Computational Theory of Scientific Discovery," *J. of Informetrics*, vol. 3, no. 3, 2009, pp. 191–209.
4. T. Heinze and G. Bauer, "Characterizing Creative Scientists in Nano-S&T: Productivity, Multidisciplinarity, and Network Brokerage in a Longitudinal Perspective," *Scientometrics*, vol. 70, no. 3, 2007, pp. 811–830.
5. L. Leydesdorff and C.S. Wagner, "Research Funding and Research Output: A Bibliometric Contribution to the US Federal Research Roadmap," to be published in *J. of Informetrics*, 2009; http://users.fmg.uva.nl/lleydesdorff/roadmap/roadmap.pdf
6. C. Chen, "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature," *J. Am. Soc. for Information Science and Technology*, vol. 57, no. 3, 2006, pp. 359–377.

Chaomei Chen is an associate professor in Drexel University's College of Information Science and Technology. Contact him at chaomei.chen@drexel.edu.

Jian Zhan is a doctoral student and research assistant in Drexel University's College of Information Science and Technology. Contact him at jz85@drexel.edu.

Michael S. Vogeley is an associate professor in Drexel University's Department of Physics. Contact him at vogeley@drexel.edu.

Open Data and Open Code for S&T Assessment

Katy Börner, Nianli Ma,
Russell J. Duhon,
and Angela M. Zoss,
Indiana University

There are more active science and technology (S&T) researchers today than ever before, and they either publish or perish. Some S&T areas produce more than 40,000 papers a month. Not only library buildings and storage facilities but also databases are filling up more quickly than we can build them. In addition, there are data sets, algorithms, and tools to be mastered for S&T to advance. No single person, machine, or institution can process and make sense of this enormous stream of data, information, knowledge, and expertise.

The tools we use to access, manage, and utilize our collective knowledge are primitive. Search engines are our main means of accessing everything we know collectively. This seems to work well for fact-finding, but it keeps us on the floor of confirmed and unconfirmed records. There is no "zoom out" button that provides a global view of our collectively knowledge—how it's interlinked; what patterns, trends, or outliers exist; or the context in which a specific piece of knowledge was created or can be used. Without context, intelligent data selection, prioritization, and quality judgments become extremely difficult to make. This reality leads to increasing specialization of researchers, practitioners, and other knowledge workers, a disconcerting fragmentation of science, and a world of missed opportunities for collaboration.

Recent advances in the digitization, federation, mining, and mapping of data make it possible to chart the structure and dynamics of science.[1–3] The resulting science maps serve today's explorers navigating scholarly networks and S&T results. The maps are generated through analysis of large-scale scholarly data sets in an effort to connect and make sense of bits and pieces of knowledge. Maps identify major research areas, experts, institutions, collections, grants, papers, journals, and ideas in domains of interest. They provide overviews of specific S&T fields—their homogeneity, import-export factors, and relative speed of innovation. They let us track the emergence, evolution, and disappearance of topics and identify the most promising areas of research.

Currently, many of the data sets and tools used to generate maps of science are proprietary and particular to each analyst. There are few, if any, standardized tools that can access and process appropriate data and present the results in a way that enables decision making by nonexperts. In this essay, we present open data and open code that can be freely used for S&T assessment together with sample analyses.

Linking Open Data

The Scholarly Database (SDB; http://sdb.slis.indiana.edu) at Indiana University evolved from seven years of development toward a free data source for S&T studies.[4] SDB offers three critical advantages for these studies:

- Search queries for an author, investigator, or inventor name or topic term can be run against multiple databases offering simultaneous retrieval of all funding, publications, and patents relevant for a query.
- Search results can be downloaded as complete record data dumps in an easy-to-process format.
- As query results are processed, derivative data sets such as coauthor or patent-citation tables can be downloaded as well.

Currently, SDB provides access to four data sets:

- 17,764,826 Medline papers provided by the National Library of Medicine (NLM),
- 1,043,804 funding awards from the National Institutes of Health (NIH),
- 174,835 funding awards from the National Science Foundation (NSF), and
- 3,875,694 patents from the US Patent and Trademark Office (USPTO).

Information regarding data provenance, system architecture, table schemas, and search functionality is available on SDB's "About" page.

Any researcher or layperson can register to search approximately 23 million records. Currently, the system has over 150 registered users from four continents and over 60 institutions in academia, industry, and government.

Sharing Free Code

The Network Workbench (NWB, http://nwb.slis.indiana.edu) is a tool that supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks. As of June 2009, the tool contains more than 110 plug-ins for the preprocessing, analysis, modeling, and visualization of networks. About 40 of the plug-ins can be applied to or were specifically designed for S&T studies.

The NWB tool comes with an associated community wiki (https://nwb.slis.indiana.edu/community), extensive documentation of algorithms, and sample data sets. The tool has

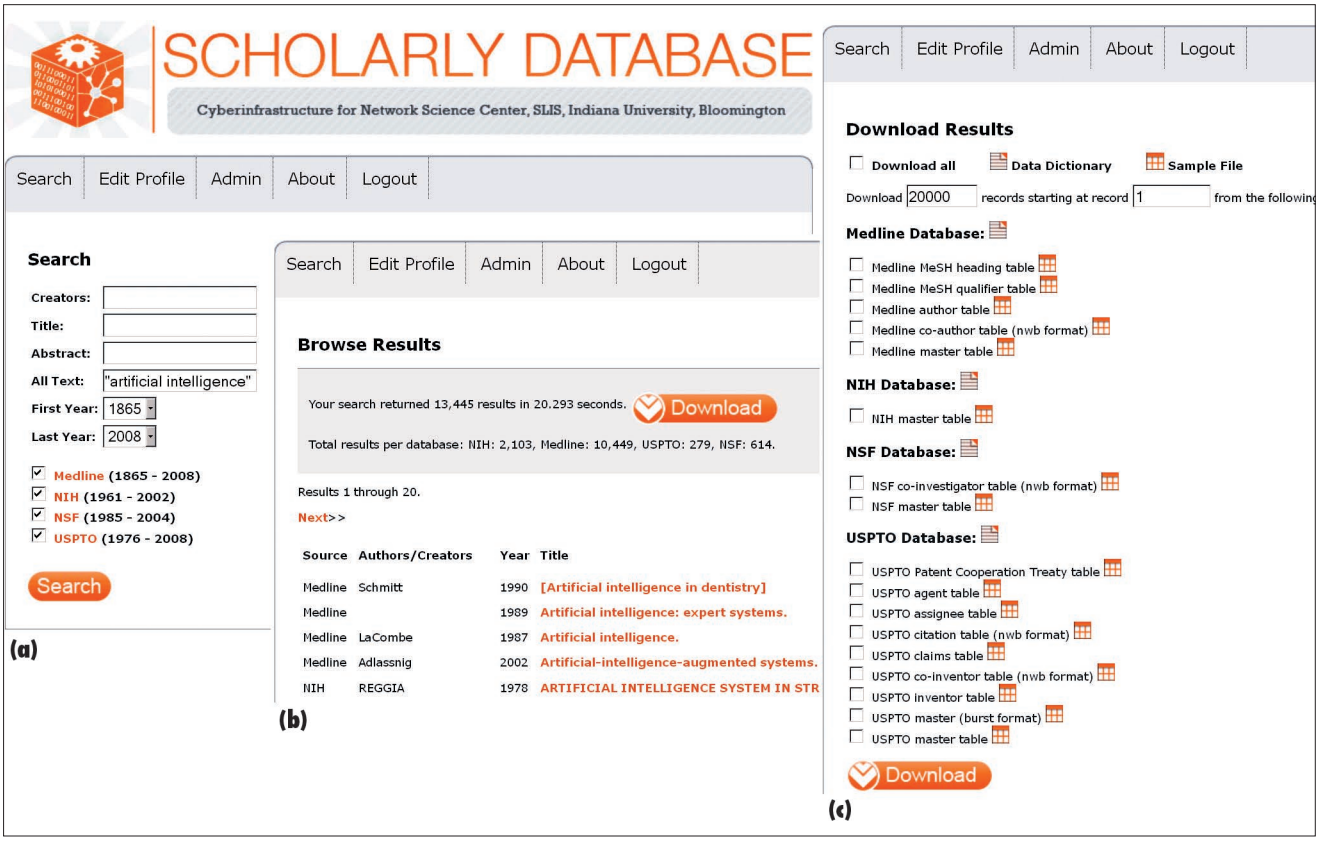


Figure 9. SDB interfaces for (a) search, (b) browsing results, and (c) downloading results. These interfaces guide users through the search of multiple data sets and the download of results in different combinations and formats. (Screen shots courtesy of the Cyberinfrastructure for Network Science Center, Indiana University, Bloomington.)

been downloaded more than 22,000 times since December 2006.

S&T Studies That Anyone Can Replicate

Users can combine the SDB with the NWB tool to study S&T data sets professionally in a manner that anyone can easily replicate. The process involves three steps: data set retrieval and download using SDB, data analysis and visualization using the NWB tool, and interpretation of results.

Data Acquisition

Figure 9a shows a query for "artificial intelligence" in the "All Text" field over all data sets available in SDB. The browse results page comprises 13,445 records—10,449 Medline papers, 2,103 NIH awards, 614 NSF awards, and 279 USPTO patents. The top-five highest scoring records are five Medline papers (see Figure 9b). Clicking on the record title opens a page showing the abstract and other information associated with the record.

Users can select different data types from the download results (see Figure 9c). For example, the Medline database offers a master table with general information, an author table that provides paper-author associations, a coauthor table that stores the coauthor network in a format compatible with the NWB tool, as well as several other tables. The icons next to each table link to data dictionaries for each database and sample data sets as well.

Medline Coauthorship Network

The Medline master table lists all paper records for the AI query. The five most frequently occurring journals are *IEEE Transactions on Pattern Analysis and Machine Intelligence* with 761 papers, *IEEE Transactions on Image Processing* (526), *Bioinformatics* (469), *IEEE Transactions on Systems, Man, and Cybernetics – Part B, Cybernetics* (456), and Springer's *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention* (443).

A user can load the Medline coauthor table into the NWB tool. The table then appears in the tool's Data Manager window (see Figure 10a).[5] With plug-ins specific to scientometrics research, NWB can be used to extract the coauthorship network. A network-analysis toolkit computes basic properties. The network has 26,206 author nodes and 59,140 coauthor edges. Exactly 944 authors are isolates (that is, unconnected). The

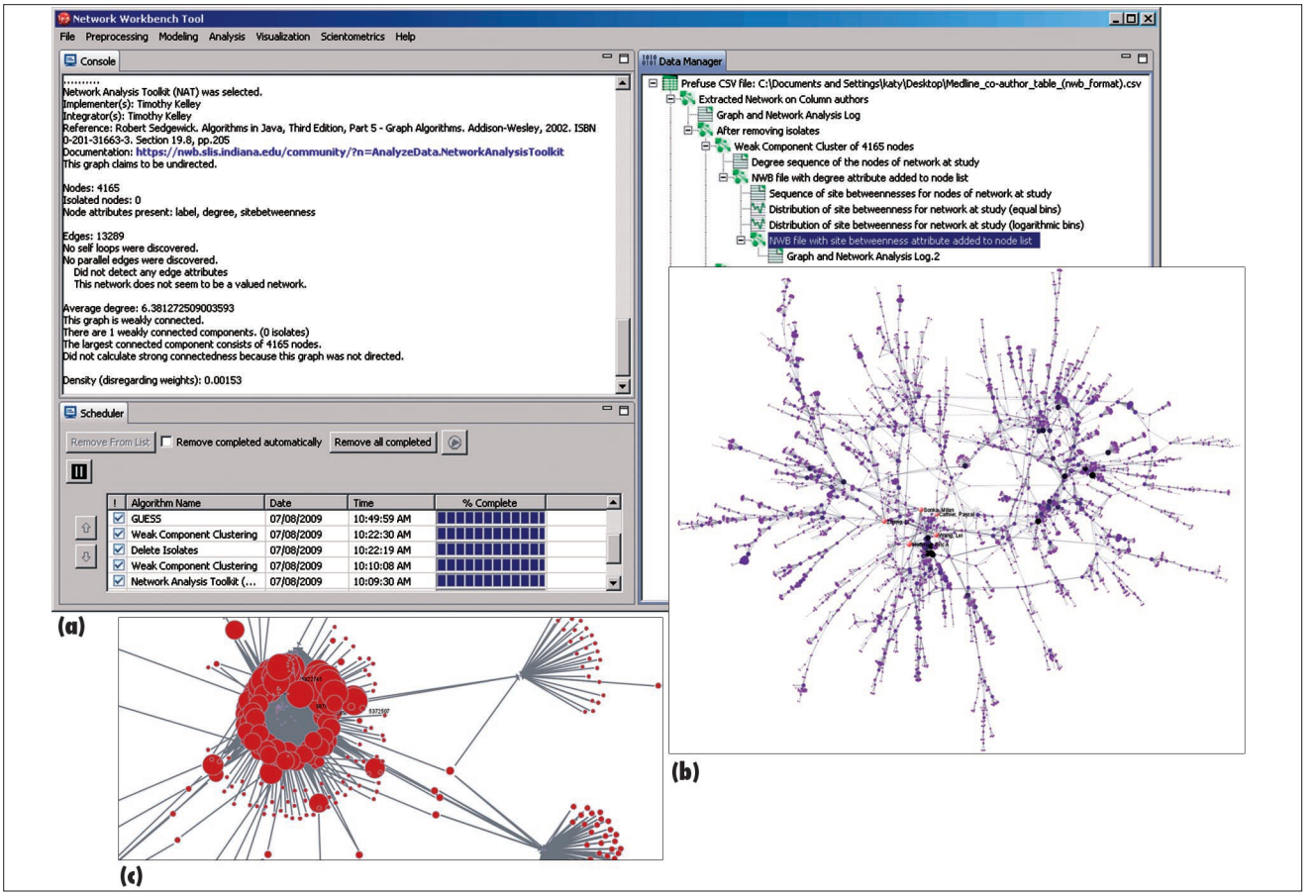


Figure 10. Interfaces to (a) the NWB tool with its console, which records the number of algorithms run, and its data manager, which lists loaded and computed data sets; (b) a Guess layout of the Medline coauthorship network's largest component; and (c) a Guess zoom feature showing details of the patent-citation network. (Screen shots courtesy of the Cyberinfrastructure for Network Science Center, Indiana University, Bloomington.)

number of clusters is almost 5,000. Using the weak component clustering algorithm, the user can extract the largest component, which has 4,165 nodes and 13,289 edges. Subsequently, a node-degree analysis computes each node's degree—that is, its number of distinct edges.

For each node, the betweenness-centrality (BC) algorithm determines the fraction of shortest paths between node pairs that pass through the node of interest. The Guess graph exploration tool (http://sourceforge.net/projects/guess), available under the NWB tool's visualization menu, visualizes the resulting network.

Figure 10b shows the coauthor network with author node area sizes and color-coding according to their degree—that is, the number of distinct coauthors. The five nodes with the highest BC value are labeled and appear in pink. The highest BC node is "Zhang, Li," the author of 10 papers from the Medline AI search results. His papers have been published in journals with Institute for Scientific Information subject categories varying from "computer science, hardware and architecture" to "endocrinology and metabolism." This diversity is mirrored in his coauthorship connections to researchers from many different clusters in the network. Medline contains little computer science research—primarily work within the biomedical sciences. Consequently, the network features major experts that apply AI techniques to biomedical research and practice.

USPTO Patent Citation Network

The AI search results generate a USPTO citation network that has 3,614 nodes, 8,393 edges, and 107 components. NWB users can load the USPTO citation table and apply the scientometrics-specific extract-directed network algorithm to extract a patent-citation network.

The network shows many network components connected by weak linkages. The 20 nodes with the highest outdegree—that is, the highest number of citations within the set—are labeled by patent number. Figure 10c shows a zoom into the set of most-

cited patents. Among them are patent number 5597312, entitled "Intelligent tutoring method and system"; number 5372507, describing a "Machine-aided tutorial method"; and number 5696885, an "Expert system and method employing hierarchical knowledge base, and interactive multimedia/hypermedia applications."

The availability of open data and open code will make S&T assessment more available and potentially more powerful. Over time, more data sets will become available via the SDB. At the core of the NWB tool is the Cyberinfrastructure Shell (CIShell, http://cishell.org), which makes it easy to plug-and-play new algorithms and to bundle sets of algorithms into custom branded tools. CIShell builds on and extends industry-developed code by the OSGi Alliance (http://osgi.org), reducing time-to-market and development costs by letting developers exploit many pre-built and pre-tested modules.

Other work currently under way will make it possible to create high-quality visualizations that support insight from raw data, at the push of a button, including geographic maps and hierarchical community visualizations.

Acknowledgments

We thank the NWB team and Kevin W. Boyack for stimulating discussions that helped shape the material reported here. This work was partially supported by the NSF under grants SBE-0738111, CBET-0831636, and IIS-0750993 and by the James S. McDonnell Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. K. Börner, C. Chen, and K.W. Boyack, "Visualizing Knowledge Domains," *Ann. Rev. Information Science & Technology*, vol. 37, 2003, pp. 179–255.
2. C. Chen, Mapping Scientific Frontiers: *The Quest for Knowledge Visualization*, Springer, 2003.
3. R. Shiffrin and K. Börner, "Mapping Knowledge Domains," *Proc. Nat'l Academy of Sciences*, vol. 101 (suppl. 1), 2004, pp. 5,183–5,185.
4. G. La Rowe et al., "The Scholarly Database and Its Utility for Scientometrics Research," *Scientometrics*, vol. 79, no. 2, 2009, pp. 219–234; http://ivl.slis.indiana.edu/km/pub/2008-larowe-sdb-scientometrics.pdf
5. *Network Workbench Tool: User Manual*, 1.0.0 beta, Cyberinfrastructure for Network Science Center, 2009; http://nwb.slis.indiana.edu/Docs/NWB-manual-1.0.0beta.pdf.

Katy Börner is the Victor H. Yngve Professor of Information Science at the School of Library and Information Science, adjunct associate professor in the School of Informatics, core faculty of cognitive science, research affiliate of the Biocomplexity Institute, fellow of the Center for Research on Learning and Technology, member of the Advanced Visualization Laboratory, and founding director of the Cyberinfrastructure for Network Science Center at Indiana University. She has a PhD in computer science from the University of Kaiserslautern. Contact her at katy@indiana.edu.

Nianli Ma is senior systems analyst and database administrator at Indiana University's Cyberinfrastructure. She has a master's degree in computer science from Carnegie Mellon University. Contact her at nianma@indiana.edu.

Russell J. Duhon is a software developer at the Cyberinfrastructure for Network Science Center at Indiana University and is finishing his bachelor of science degree in Informatics. His recent projects include mapping the collaborations of the Chinese Academy of Science and designing a S&T analysis and mapping tool through a contract with the NSF. Contact him at rduhon@indiana.edu

Angela Zoss is a doctoral student and research assistant at Indiana University's Cyberinfrastructure for Network Science Center. She has an MS in communication from Cornell University. Contact her at amzoss@indiana.edu.

Global S&T Assessment by Analysis of Large ETD Collections

Venkat Srinivasan
and Edward A. Fox,
Virginia Polytechnic Institute & State University

Electronic theses and dissertations (ETDs) are a key part of global scholarship. If we can determine the distribution of ETDs in broad topical areas, such as science, technology, engineering, and mathematics (STEM), for each region around the world, we can gain critical insights into prevailing research trends.

In this essay, we present a technique for identifying STEM dissertations from a large ETD collection. We derived our testbed ETD collection from the Networked Digital Library of Theses and Dissertations (NDLTD; www.ndltd.org),[1] which has members from more than 80 universities (or university consortia) around the world (see Figure 11). Hence our results can be used to gauge global interest in STEM areas, particularly since the mid-1990s.

Background

ETDs form an important part of the open access scholarly literature but

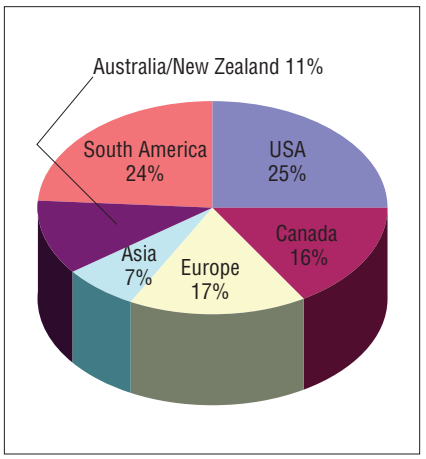


Figure 11. Major regions contributing to the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog. The NDLTD has members from more than 80 universities worldwide.

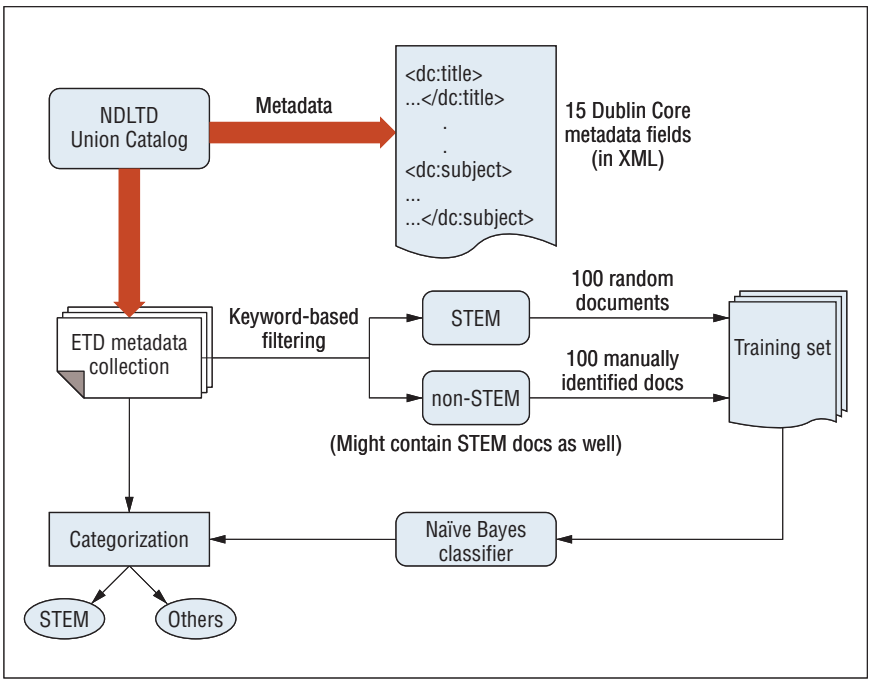


Figure 12. Categorization pipeline. Electronic thesis and dissertation (ETD) metadata is used as features to the naïve Bayes classifier, which then distinguishes the science, technology, engineering, and mathematics (STEM) ETDs from others.

are largely underutilized. Though there are many analyses of the scholarly literature—for example, studies of research trends, citation networks, or clickstream data—to the best of our knowledge, very few of the studies consider ETDs.

Yet ETDs are a valuable resource in and of themselves. They have broad topical coverage, include comprehensive and up-to-date literature surveys with pointers to related papers, and, importantly, also have quality control, in that dissertations are reviewed by a committee of experts. Easier access to ETDs would therefore be a valuable aid to scholarly activities.

In a larger effort here at Virginia Tech, we're working on developing techniques for performing information retrieval in large documents, such as books, ETDs, and patent documents. As part of our preliminary studies, we've worked on categorizing ETDs into topical areas, and we'll subsequently provide an appropriate search and browse interface. Here, we present results from a pilot study.

Our work's implications go beyond just providing an approach to tag ETDs into STEM and non-STEM areas. In recent years, in the US in particular, concern has increased about declining interest in STEM areas among the student population. The problem is particularly accentuated at the undergraduate and graduate levels, where such a declining interest could lead to reduced competitiveness in the global technology environment. Electronic dissertations can be very good indicators in this regard, as there is a direct correlation between the number of dissertations produced and the number of students graduating in the corresponding area.

Study Methods

The NDLTD Union Catalog is an effort that started in the mid-1990s to aid the preparation and wider dissemination of ETDs. As of March 2009, the catalog has 663,515 ETDs from universities around the world. It provides 15 Dublin Core metadata fields (see http://dublincore.org) relevant to a dissertation (title, subject, abstract, year, publisher, and so on) plus a link to the dissertation itself at the corresponding university. While compiling the list of NTLTD for a particular region, we considered only those universities that contributed more than 2,000 dissertations for our experiments.

We also confined our pilot studies to English language dissertations. Many NDLTD-affiliated universities have dissertations in languages other than English—mostly in Portuguese, Spanish, or Chinese—so ETDs from the US are significantly overrepresented in our sample. Another major issue with the NDLTD Union Catalog is the amount of noise present in various metadata fields. For example, although you might expect the "date" field to hold the year in which the dissertation was published, it sometimes holds author or university information or other unrelated data instead. Because our study includes a timeline analysis, we considered only those dissertations that have the Dublin Core "date" field set correctly. Fortunately, this is the case for many universities, especially for those from the US and Australia.

Figure 12 describes our categoriza-

Table 3. Science, technology, engineering, and mathematics (STEM) dissertations for some major contributors to the NDLTD Union Catalog.

	NDLTD Source (University)	No. of ETDs in NDLTD	No. of STEM EDTs identified by the classifier
USA	Massachusetts Institute of Technology	29,804	23,157
	Virginia Polytechnic Institute & State University	11,976	6,776
	Ohiolink (Ohio universities)	8,020	5,467
	North Carolina State University	5,026	4,179
	California Institute of Technology	4,774	4,596
	Georgia Institute of Technology	3,582	2,628
	Total	63,182	46,803
Rest of the World	Australasian Digital Theses	37,958	15,121
	NSYSU (Taiwan)	11,087	5,407
	University of Manitoba (Canada)	24,989	1,647
	Middle Eastern Technical University (Turkey)	2,247	1,659
	University of Waterloo (Canada)	1,396	584
	University of Auckland (New Zealand)	1,176	821
	Total	56,362	25,239

tion pipeline. We use only metadata information—specifically, only the Dublin Core title, subject, and abstract metadata fields. The first step is to build a good-quality training set to use in training a classifier to distinguish between STEM dissertations and others. To do this, we filter the dissertations according to keywords occurring in the Dublin Core subject field. For science categories, we check for words such as biology, chemistry, or math. For the technology category, we check for the keyword "engineering" in the subject field.

We selected dissertations from three universities: Massachusetts Institute of Technology (MIT), California Institute of Technology (Caltech), and Virginia Tech. We filtered these dissertations on the basis of keywords. We then used 50 science and 50 technology dissertations (selected at random) as a training set for STEM areas, and 100 "other" dissertations (identified manually) to form our non-STEM training set.

We trained a naïve Bayes classifier to distinguish between STEM and non-STEM dissertations. We chose this classifier for its simplicity, low training time, and effectiveness in performing binary classification.[2] We concatenated the Dublin Core title, subject, and description metadata fields and used them to train the classifier, after some parsing (to remove special characters, mathematical equations, and so on), stopword removal, and stemming. The features provided to the naïve Bayes classifier are thus the word stems occurring in the three metadata fields. We used an open source implementation of a naïve Bayes classifier in Perl available through the Comprehensive Perl Archive Network (see http://search.cpan.org/~kwilliams/Algorithm-NaiveBayes-0.04/lib/Algorithm/NaiveBayes.pm).

Results

We determined the classifier's average precision and recall values on the training set of 200 documents by performing 10-fold cross-validation; the values are 0.94 and 0.70, respectively. We measure precision as the ratio of true STEM dissertations and the dissertations identified as STEM by the classifier, and recall as the fraction of true STEM dissertations that the classifier identified correctly as STEM. We compute precision and recall values during each of the 10 folds, and average them to obtain the overall precision and recall measures.

We used the classifier to identify STEM dissertation for universities that have dissertations in English in NDLTD. Table 3 presents the ETD sources and detailed results. We also performed a timeline analysis on this ETD collection, where we measured the STEM output over time (see Figure 13). While the percentage of ETDs that are in STEM areas, as opposed to all topical areas, seems in most of the world to be relatively constant, it appears that the US percentage is declining, which many would view as a matter of concern.

Our pilot study results indicate that information in the Dublin Core metadata fields is by itself sufficient to do the initial categorization into STEM and non-STEM areas. As part

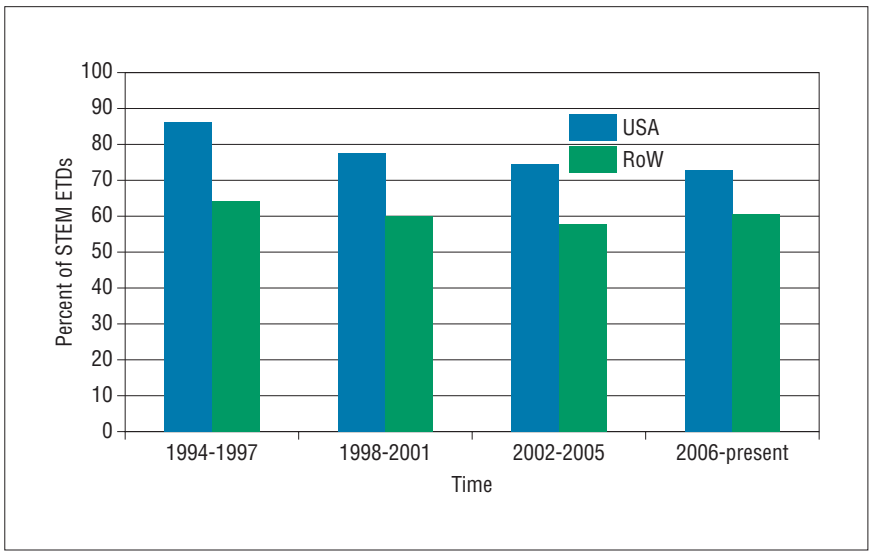


Figure 13. STEM ETDs. The distribution over the years for the US and the rest of the world (RoW) indicate that US STEM ETD output appears to be declining slightly.

of our future work, we want to do more specific categorization based on an ontology such as the Open Directory Project (www.dmoz.org). We also want to provide search and browse services.

There has been no drastic change in STEM output over the years, which should to some extent address concerns regarding declining interest in STEM areas among students, particularly in the US. However, US universities tend to have a sizeable population of international students, especially at the graduate level. Hence, we need additional metrics to identify STEM productivity for American students. Toward this goal, we have collected commonly occurring American surnames from census data.[3] Using this information and the Dublin Core "contributor" metadata field, we will filter out the STEM dissertations and do timeline analysis to get a more realistic picture of STEM productivity among American students.

On a broader note, an important future goal is to expand our work to include ETD collections of universities beyond the NDLTD Union Catalog. Lessons learned during this pilot study will help immensely in analyzing a larger collection.

Acknowledgments

The work on this project was funded by a grant from Google, which we gratefully acknowledge.

References

1. E. Fox et al., "Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources," *D-Lib Magazine*, vol. 3, no. 8, Sept. 1997; www.dlib.org/dlib/september97/theses/09fox.html.
2. F. Sebastiani and C.N.D. Ricerche, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, vol. 34, 2002, pp 1–47.
3. *Frequently Occurring Surnames from Census 2000,* US Census Bureau; www.census.gov/genealogy/freqnames2k.html.

Venkat Srinivasan is a PhD candidate in the Computer Science Department at Virginia Polytechnic Institute & State University. He has an MS in computer science from the University of Delhi. Contact him at svenkat@vt.edu.

Edward A. Fox is a professor in the Computer Science Department at Virginia Polytechnic Institute & State University. He has a PhD in computer science from Cornell University. Contact him at fox@vt.edu.

Managing Multilingual S&T Knowledge

Christopher C. Yang,
Drexel University
Chih-Ping Wei,
National Tsing Hua University

To keep pace with rapid global advances in science and technology (S&T), organizations must constantly analyze the latest scientific discoveries or technological breakthroughs and then develop effective strategies to create and sustain market advantages in increasingly competitive business environments. Effective search and management of relevant S&T documents is a critical first step in technology trend analysis, competitive intelligence surveillance, and technology roadmapping.[1,2]

Such documents can include scientific articles, patent documents, and business newswires from various sources. They are often created and maintained in heterogeneous language environments. Although substantial efforts have gone into facilitating cross-lingual information retrieval, little prior research examines the use of text mining to support effective multilingual knowledge (document) management. In this essay, we explore exciting research opportunities and important challenges in multilingual text mining for global S&T knowledge management.

An Illustrative Scenario

Tom, a senior fuel cell technology analyst, downloads thousands of US patents (in English) from the US Patent and Trademark Office (USPTO) website and organizes them into technological topics (categories). He also collects patent documents (in Chinese) and wants to classify them on the basis of his existing categories. Tom thus faces a cross-lingual text

categorization (CLTC) challenge: automated learning from a training set of preclassified documents in one language (L_1), followed by classification of other documents available in a different language (L_2).

Tom's patent repository now is polylingual—that is, his categories contain some patent documents in English and others in Chinese. Subsequently, when Tom accesses new patents, in either English or Chinese, and assigns them to his patent repository, he's performing a polylingual text categorization (PLTC) task. This task entails automated learning from a training set of preclassified polylingual documents (some in L_1 and some in L_2) and assigning unclassified documents available in L_1 or L_2 into the appropriate categories.

In addition to patent documents, Tom gathers scientific articles (in English) about fuel cell technology and maintains them using a preferred classification scheme, which might differ from what he uses to maintain the patent repository. His colleague, Jennifer, does the same thing, but she focuses on Chinese scientific articles and organizes them according to her preferred categories. Tom hopes to integrate Jennifer's Chinese repository into his English repository through cross-lingual category integration (CLCI). However, to do so, he must address the challenge of integrating different categorization schemes. Essentially, CLCI integrates a category set (the source catalog) that contains documents in L_2 into another category set (the master catalog) that contains documents in L_1.

Tom now has two repositories, both containing polylingual documents. To perform comprehensive technology intelligence analyses that identify important technological threats and opportunities, he needs support for effective polylingual category integration (PLCI). Formally, PLCI addresses the challenge of integrating a source catalog into a master catalog when both catalogs contain polylingual documents.

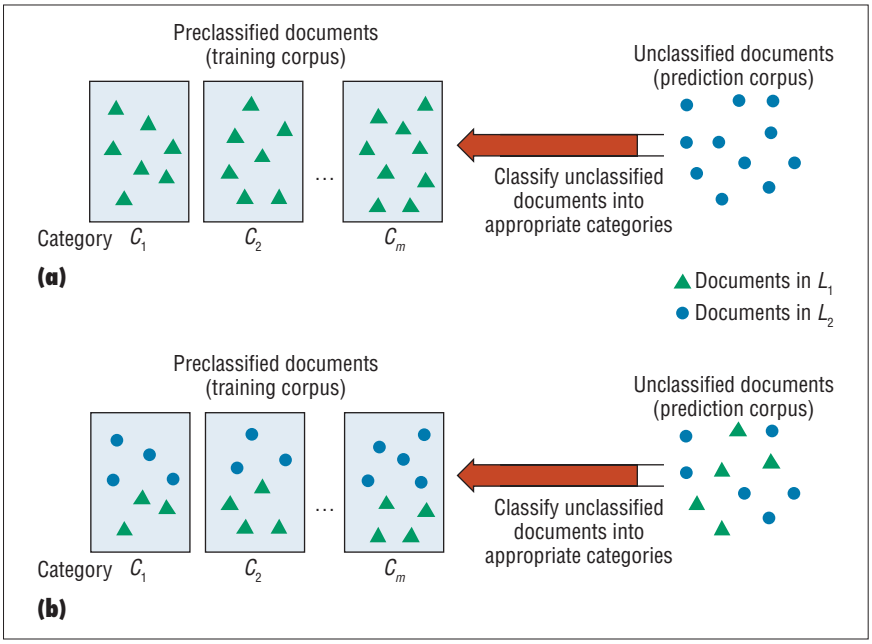


Figure 14. Illustrations of cross-lingual text categorization (CLTC) and polylingual text categorization (PLTC). (a) CLTC uses the categories established through the training corpus in one language to classify documents in another language. (b) PLTC uses the categories established through a polylingual training corpus to classify polylingual documents.

Cross-Lingual Text Categorization

As Figure 14a illustrates, CLTC deals with learning from a set of preclassified documents (the training corpus) in L_1 and then classifying unclassified documents (the prediction corpus) in L_2. A major CLTC challenge is providing cross-lingual semantic interoperability—that is, establishing a connection between representations of the training corpus in one language and representations of the prediction corpus in the other language. Mitigating the language barrier requires some form of translation, which involves two fundamental design issues.

Translation Mechanism

Previous CLTC studies have examined several translation mechanisms, including bilingual dictionaries, machine translation, and a bilingual thesaurus.[3,4]

Bilingual dictionary translation can be proprietary, costly, and intolerant of novel terms and proper nouns commonly found in S&T documents. Machine translation uses a system that translates a document from one language to another automatically, though the effectiveness of existing systems often isn't satisfactory, particularly for documents that require greater contextual information for accurate translations. A bilingual thesaurus relies on the assumption that associated terms often co-occur in documents,[5,6] and it can be constructed automatically from a parallel or comparable corpus.

Despite noise in the statistical nature of a bilingual thesaurus, it offers desirable constructability, maintainability, and capability with respect to novel terms and proper nouns. These properties make it relatively appealing. Prior studies concentrate

primarily on employing one translation mechanism when developing their respective CLTC techniques. Examining the effects of different translation mechanisms on CLTC effectiveness in the context of global S&T knowledge management is essential but has received little investigation attention.

Translation Strategy

A translation can be performed on the training corpus (that is, translate training documents from L_1 to L_2)[3] or on the prediction corpus (that is, translate unclassified documents from L_2 to L_1).[4] However, prior research lacks theoretical justifications or empirical evidence regarding which strategy is more effective. This fundamental question requires thorough examination.

Other Research Issues

Besides these design issues, two research questions also warrant investigation. First, most previous CLTC studies assign each unclassified document to a category individually. However, the well-known word-mismatch problem can make category assignments based on individual documents ineffective. One solution is to group similar unclassified documents into clusters using a document-clustering technique. We could then translate each cluster into another language (if employing a prediction-corpus translation strategy) and, finally, assign all documents in each cluster to the same category. Developing and empirically evaluating a proper CLTC cluster-based category-assignment method represents an interesting research direction.

Second, prior CLTC research doesn't consider translation quality with regard to learning a classification model or classifier in the training-corpus translation strategy or assigning translated documents to categories in the prediction-corpus translation strategy. The translated terms in each document can vary considerably in quality. This means the translated documents can differ in quality as well. We therefore need to design appropriate methods for estimating translation quality at both term and document levels and to develop effective learning algorithms or category assignment methods that can reveal the quality of translated training or prediction documents.

> We need to design appropriate methods for estimating translation quality at both term and document levels.

Polylingual Text Categorization

As Figure 14b shows, PLTC differs from CLTC in that it constructs classifiers from a training corpus available in different languages and classifies unclassified documents in any of those languages. Because training documents exist in each language, we can simply consider PLTC as multiple independent monolingual text-categorization problems. That is, we can construct a classifier for each language on the basis of the training documents available in that language. When a new document in a specific language becomes available, we use the corresponding classifier for category assignment.

However, this naïve approach employs the training documents in only one language to construct each monolingual classifier. Hence, it can't take advantage of important categorization information available in the training documents of the other language.

We propose a feature reinforcement-based PLTC (FR-PLTC) technique that takes the training documents of all languages into account when constructing a monolingual classifier for each specific language.[7] Specifically, we first measure the discriminatory power of all features (terms) in each language's training documents. Then we reassess the discriminatory power of each feature in one language by considering its related features in another language, using a bilingual thesaurus. With such cross-language checking, if a feature in L_1 and its related features in L_2 possess high discriminatory power, the feature is likely to possess greater discriminatory power. However, inconsistent assessments between two languages reduce confidence in the resulting discriminatory power. Accordingly, we select a set of features with the greatest reassessed discriminatory power for each language. On the basis of the selected features for each language, we can then construct a monolingual classifier using the training documents available in that language.

Our empirical evaluation shows that FR-PLTC significantly outperforms the naïve PLTC approach in terms of classification accuracy. It achieves a 5.42 percent improvement with tf∗idf (term frequency × inverse document frequency) as the representation scheme and a support vector machine as the underlying learning algorithm.

PLTC has received far less research attention than CLTC, and several important research issues remain open. For example, when constructing a

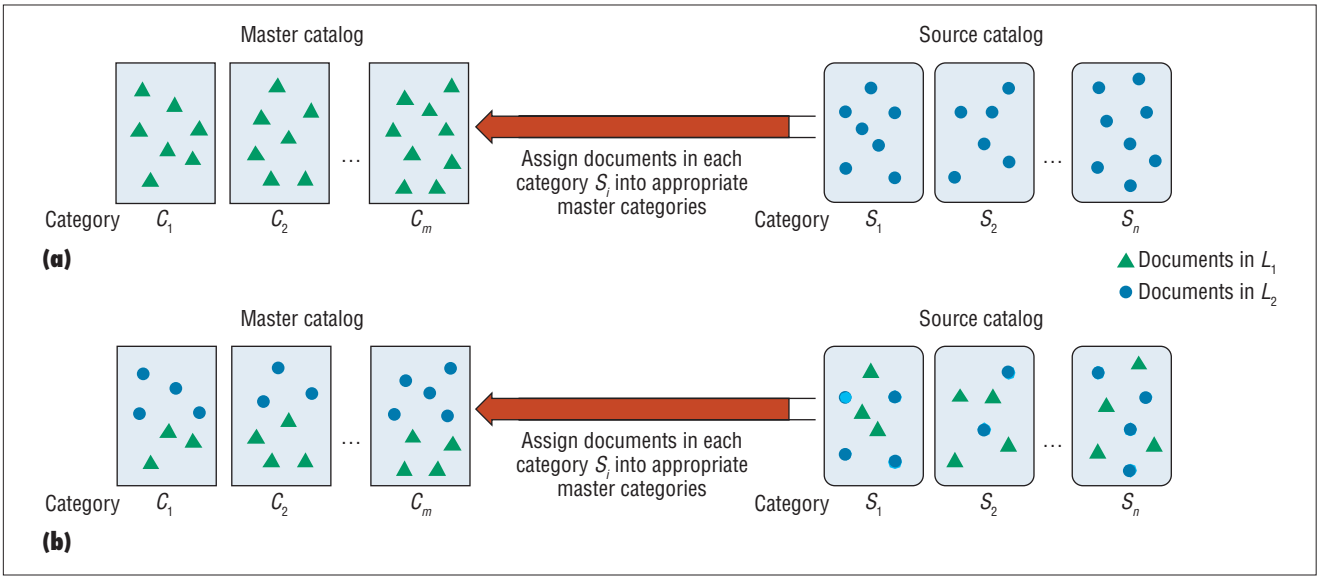


Figure 15. Illustrations of (a) cross-lingual category integration (CLCI) and (b) polylingual category integration (PLCI). CLCI finds a category in the master catalog in one language for each document in the source catalog in another language. PLCI integrates a source catalog into a master catalog when both catalogs consist of documents available in various languages.

monolingual classifier for a specific language, the FR-PLTC technique doesn't employ the training documents available in another language to expand the training sample size. We might further improve its effectiveness by including translated training documents, originally available in another language, into the target language's training corpus.

Second, as mentioned earlier, translation quality issues must be addressed to fully realize the potential utilities of the suggested PLTC solution.

Third, further research might consider developing a PLTC technique that can construct a single language-independent classifier, perhaps by employing latent semantic indexing (LSI) to build a language-independent space on the basis of a parallel or comparable corpus. In this case, all polylingual training documents would be mapped onto an LSI space, which would allow for the construction of a single classifier from the mapped training documents. When classifying unclassified documents in any of those languages, we would first need to map the documents onto the LSI space, then use the language-independent classifier for the category assignment. It would be interesting to investigate the conditions that favor the use of a PLTC technique employing a single, language-independent classifier compared with techniques involving multiple language-specific classifiers, such as FR-PLTC.

Cross-Lingual Category Integration

As Figure 15a depicts, a major CLCI objective is to find an appropriate category in the master catalog for each document in the source catalog when documents in both catalogs are available in different languages. Several category integration techniques have been proposed in the literature,[8,9] but they all target a monolingual environment.

As with CLTC, the major challenge of CLCI is overcoming the language barrier between catalogs. By properly translating the documents in one catalog, we transform the challenging CLCI task into a common monolingual category-integration problem, which we can address with an appropriate existing category-integration technique. With this approach, we must therefore address the two design issues inherent to CLTC when developing a CLCI technique to support global S&T knowledge management. Specifically, we must select the most effective translation mechanism and identify a translation strategy (master catalog or source catalog) that seems likely to improve integration effectiveness. Moreover, we might need to extend existing category-integration techniques to account for the varying quality of translated documents.

Polylingual Category Integration

Figure 15b illustrates the PLCI problem of integrating a source catalog into a master catalog when both catalogs consist of documents available in various languages. PLCI can be simplified as multiple, independent monolingual category integration (MnCI) problems—one for L_1 and one for L_2. However, similarly to PLTC, this naïve approach considers only master and source documents in one language during each MnCI task. It ignores documents in another language, thus likely compromising the integration effectiveness.

To exploit the opportunities offered by polylingual documents in both catalogs, we see several directions worth pursuing. For example, the FR mechanism proposed in the FR-PLTC technique might be interesting. Specifically, for each language, we could incorporate the FR mechanism to select more representative features from the documents in the master catalog, then use an existing MnCI technique to integrate its corresponding source catalog into the master catalog in that language.

Another possible design would take a cross-lingual approach to address PLCI. For example, to conduct category integration for L_1, we could translate those documents originally available in L_2 in the master catalog into L_1, then perform MnCI for L_1 by integrating the documents that appeared in L_1 in the source catalog into the master catalog, which currently contains documents originally in L_1 and those translated from L_2. Likewise, we would perform this process for L_2. Because each MnCI task uses a larger master catalog, the resulting integration effectiveness likely improves. If we further consider the quality of the translated documents when performing MnCI tasks, the effectiveness might improve further.

Addressing the research opportunities we've identified could substantially broaden the spectrum of multilingual text-mining and its practicality for supporting global S&T knowledge management. These opportunities also share a common set of challenges that deserve further attention. For example, competitive intelligence surveillance, which allows organizations to understand their current and potential competitors better, often requires the extraction of names of different organizations, technologies, or products from various S&T documents. When dealing with multilingual documents, adequate cross-lingual entity-resolution mechanisms are essential for effective global S&T analysis. Furthermore, some S&T documents are scientific or technologically oriented, whereas others have a predominantly business orientation. This increases the chance of different documents using different terms in referring to identical or similar concepts. Establishing cross-domain interoperability is essential, especially in multilingual environments.

For multilingual documents, adequate cross-lingual entity-resolution mechanisms are essential for effective global S&T analysis.

References

1. R.N. Kostoff and R.R. Schaller, "Science and Technology Roadmaps," *IEEE Trans. Eng. Management*, vol. 48, no. 2, 2001, pp. 132–143.
2. A.L. Porter and S.W. Cunningham, *Tech Mining: Exploiting New Technologies for Competitive Advantage*, Wiley-Interscience, 2005.
3. J.S. Olsson, D.W. Oard, and J. Hajic, "Cross-Language Text Classification," *Proc. 28th Ann. Int'l ACM Conf. Research and Development in Information Retrieval* (SIGIR 05), ACM Press, 2005, pp. 645–646.
4. L. Rigutini, M. Maggini, and B. Liu, "An EM Based Training Algorithm for Cross-Language Text Categorization," *Proc. 2005 IEEE/WIC/ACM Int'l Conf. Web Intelligence* (WI 05), IEEE CS Press, 2005, pp. 529–535.
5. K.W. Li and C.C. Yang, "Automatic Cross-Lingual Thesaurus Generated from the Hong Kong SAR Police Department Web Corpus for Crime Analysis," *J. Am. Soc. for Information Science and Technology*, vol. 56, no.3, Feb. 2005, pp. 272–282.
6. C.C. Yang and K.W. Li, "An Associate Constraint Network Approach to Extract Multi-lingual Information for Crime Analysis," *Decision Support Systems*, vol. 43, no.4, 2007, pp. 1,348–1,361.
7. C. Wei, H. Shi, and C.C. Yang, "Feature Reinforcement Approach to Poly-lingual Text Categorization," *Proc. Int'l Conf. Asia Digital Libraries*, Springer, 2007, pp. 99–108.
8. R. Agrawal and R. Srikant, "On Integrating Catalogs," *Proc. 10th Int'l Conf. World Wide Web*, ACM Press, 2001, pp. 603–612.
9. T.H. Cheng and C. Wei, "A Clustering-Based Approach for Integrating Document-Category Hierarchies," *IEEE Trans. Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 38, no. 2, Mar. 2008, pp. 410–424.

Christopher C. Yang is an associate professor in Drexel University's College of Information Science and Technology. He has a PhD in computer engineering from the University of Arizona. Contact him at chris.yang@drexel.edu.

Chih-Ping Wei is a professor in the Institute of Service Science at National Tsing Hua University in Taiwan, R.O.C. He received a PhD in management information systems from the University of Arizona. Contact him at cpwei@mx.nthu.edu.tw.