OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE

FORESIGHT AND UNDERSTANDING FROM SCIENTIFIC EXPOSITION (FUSE)

Incisive Analysis Office

LEADING INTELLIGENCE INTEGRATION

Dewey Murdick
Program Manager
Dewey.Murdick@ugov.gov

2011 Graph Exploitation Symposium
August 9-10 2011

Situation

- **Technical emergence** … the process whereby innovative ideas, capabilities, applications, and even entirely new fields of study arise, are tested, mature, and if conditions are favorable, make a significant impact.
- Those able to "**scan the horizon**" for the early signs of technical emergence, and take advantage of the resulting capabilities and applications, can gain a significant competitive edge.
- The increasing **globalization of science and technology** raises the potential for high-impact technical capabilities to emerge in increasingly diverse technical, socio-economic, and geographic areas.

FUSE Program - Broad Agency Announcement (BAA)
http://www.iarpa.gov/solicitations_fuse.html

What is FUSE?

FUSE seeks to enable the early detection of real world technical emergence as found within the full-text scientific, technical, and patent literatures in English, Chinese, German, Japanese, ~~Korean~~,* Russian, and ~~Spanish~~.*

Novelty → Discover patterns of emergence and connections between technical concepts at a speed, scale, and comprehensiveness that exceeds human capacity

Usage → Alert analyst of emerging technical areas with auditable evidence to support further exploration

Impact → Provide a relevant, timely, and unbiased analytic force multiplier necessary to maintain technical vigilance, across all disciplines and multiple languages, in the face of the rapidly rising flood of publications

Complete, Continuous, Unbiased

*Status will be re-evaluated during Phase I

FUSE Approach

Today, *ad hoc* "technical horizon scanning" consumes substantial expert time, is narrowly focused on a small number of topics, and is subject to limited systematic validation.

Analysts need a reliable and transparent capability to scan continuously for signs of technical emergence.

Multiple independent research teams

Iterative prototype development in parallel with evaluation

Formal review of program by IARPA leadership every 6 months

Today	FUSE
Manual	Automatic
Limited full-text coverage (text analytics)	Comprehensive literature coverage
Updated infrequently	Updated on-demand
Months to produce (for one technical area)	24hrs to produce (for all technical areas)
Ad hoc evaluation	Formal models of emergence

Key Technical Challenges

automated detection of emerging concepts, methods, technologies …

Hypothesis: Features exist within literature that can be connected to reliably identify technical emergence

- Process multidiscipline, multilingual, and noisy full-text from scientific, technical & patent literature from around the world
  - Extract usable within-document and cross-document features *(e.g., methods, applications, infrastructure, concepts in context …)*
  - Generate meaningful Related Document Groups (RDGs)
  - Operate within a massive and rapidly growing data set
- Develop and validate indicators of technical emergence and establish models / theories of emergence
- Identify, prioritize, and nominate technical areas; provide understandable evidence of technical emergence

Why now?

- Problems to overcome:
  - Too much information to analyze, in too many languages
    - Support strategic investment
    - Facilitate discovery and innovation
  - Cannot reliably query for patterns that indicate emergence without starting with a known, named subject
- Automated analysis is likely to work because:
  - The scientific literature is now available in digital formats
    - Metadata records are well curated and ready for use
    - Exploitation of the full text of documents is now possible (although not easy)
  - Emerging text and “signal” analysis (temporal pattern) techniques are promising
    - Context-sensitive feature extraction from text
    - Unsupervised clustering
    - Machine learning
    - Statistical modeling
    - Pattern matching and analysis
    - Indicator development and validation

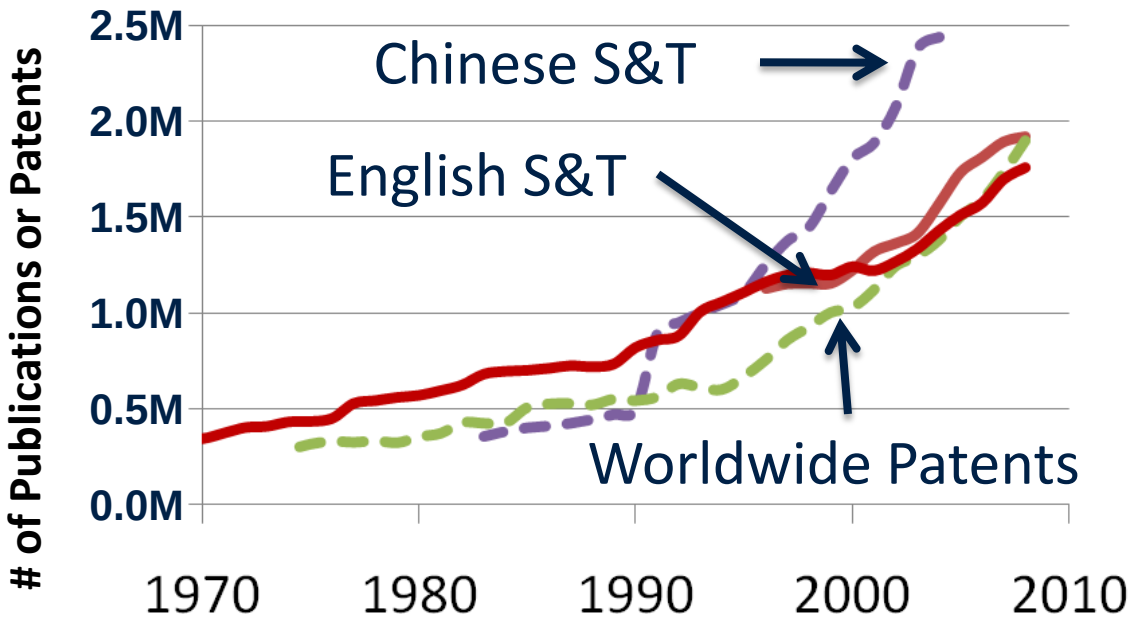
Worldwide Scientific, Technical & Patent Literature

Top Languages
(English Language Indices*)

Language	Pubs/Patents
English	~55M / 6M
German	~1.5M / 3M
French	~1M / 1M
Russian	~1M / 1.5M
Chinese	~650k / 2M
Japanese	~350k / 8M
Spanish	~300k / 500k

*Many additional non-English collections

Publications and Patents (by Source)



Source: Thomson Reuters Web of Science® (>10k journals & >100k conference proceedings, 1900-present) and Derwent World Patents Index® (41 patent issuing authorities, 1970s-present) , Elsevier B.V. Scopus® (18k journals & 3.6M conference papers, 1996-present), Chinese National Knowledge Infrastructure (1986-2008, data source includes broader range)

Document Repository (Phase 1)

- Contains **scientific literature** and **patents**
- Scientific and Technical Literature
  - Broad disciplinary coverage, but case study emphasis
  - Commercial Metadata (e.g., WoS, Scopus)
  - Commercial Full-text from journals and conferences
  - Open Access Full-text (e.g., PubMed Central)
  - *Acquisition will continue (more content, multiple languages)*
- Applied and Granted Patents, Utility Models
  - European Patent Office DOC DB Metadata
  - Broad coverage of patent offices, full-text in multiple languages
  - *Acquisition will continue (full-text, multiple languages)*

FUSE Validation / Metrics

- Validated theories and indicators of emergence
  - *"Emergence Theory Peer Review"*
- Effective identification, prioritization and nomination of technical areas as compared to real world (e.g., experts, case studies, present day tests for both positive / negative examples)
  - *"Nomination Quality"*
- Evidence provided in a clear and humanly usable form
  - *"Evidence Quality"*
- System to perform at scale across multiple languages
  - *"Computational Efficiency" and "Multilingual Performance"*
- Control experiment to ensure full-text features are leveraged by models (not just metadata); develop environment for RDG generation and evidence explanation
  - *"FUSE Lite"*

Scientific, Technical, Patent Literature -> Graphs

Interesting Attributes	Interesting Relationships	Interesting Graphs
• Author / Inventor • Organization • Emails • Geo-location • Funding org / contract # • Subject categories • Controlled vocab / keywords • International Patent Classification • **Technical methods** • **Equipment** • **Infrastructure** • **Applications**	• Co-occurrence Persons Organizations Co-location Papers Patents Cross-corpus citations (often unresolved) Clusters • Co-citation • Semantic relationships "Bag-of-words" Metadata **Zone of full-text** **Rhetorical stance** **Sentiment link type for citations**	• Co-authorship graphs • Co-citation graphs • Geo-centric graphs • Graphs to enhance entity resolution • Multigraphs & hypergraphs • **Lots of room to explore**

Bold: New features to be explored

Graph Related Challenges

- Large datasets, millions of documents
- Time-dependent analysis of networks
- Lack of a rigorous probabilistic framework for evolving and noisy data
  - Node uncertainty
  - Link uncertainty
- Multigraph and hypergraph analysis of networks (with time domain)

Persistent Issues … Can Graphs Help?

- Lack of truth / insufficient truth for technical emergence
  - How does it occur?
  - How do the processes vary across disciplines and communities of practice?
  - How does one prioritize which technical area is more emergent than another?
  - *Many more questions will arise…*
- Models of background and foreground behavior
  - We don't always know what we are looking for…

Anticipated Impact

- **Scientific & Technical Intelligence Analysis Impact**
  - Relevant, timely, and bias-controlled analytic force multiplier to maintain technical vigilance, across all disciplines and multiple languages
  - Discover previously unknown emergence signals of interest at speed, scale, and comprehensiveness that exceeds human capacity
- **Technical Impact**
  - Generalized and validated theories of technical emergence
  - New cross-document conceptual feature extraction technologies
  - Significant progress in computer-generated evidence representations for human use
- **Secondary Impact**
  - Improved priority filter for USG investment strategies and policy
  - Technology applies to additional genres

Questions

Program Structure

Phase (Period)	Length (months)	Primary English and *Multilingual* Goals
Phase 1 (Base Period)	18	Demonstrate that full-text literature can be the source for robust indicators of technical emergence within a consistent theoretical construct. Automatically prioritize a small number of provided Related Document Groups (RDGs), each representing a single technical area. Nominate those RDGs that exhibit technical emergence. *Demonstrate proof-of-concept functionality in at least two languages in addition to English.*
Phase 2 (Option Periods 1 & 2)	30 (15 & 15)	Demonstrate automatic generation and nomination of those RDGs that exhibit single technical area emergence, from a collection of millions of full-text documents. *For at least two languages in addition to English, automatically prioritize provided RDGs, each representing a single technical area. Nominate those RDGs that exhibit technical emergence.*
Phase 3 (Option Period 3)	12	Demonstrate automatic generation and nomination of those RDGs that exhibit technical emergence across disparate technical areas, from a collection of millions of full-text documents. *For at least two languages in addition to English, demonstrate automatic generation and nomination of those RDGs that exhibit single technical area emergence, from a collection of full-text documents.*

Case Studies

- Drawn from many areas of scientific inquiry & application:
  - Biological Sciences / Biotechnology; Computer Science / Information Science; Earth Science; Engineering; Mathematics / Statistics; Medical / Clinical / Infectious Disease / Health Services; Physical Sciences; Social Sciences; …
- Technical emergence measured from literature & “real world” views
- Specific topics will start with DNA Microarrays & Genetic Algorithms
  - Multiple case studies to be produced quarterly; some are held back for evaluation
  - Expect about 8+ to be released in Phase I
- Case studies are representative but not comprehensive
  - Insufficient for machine learning solutions to train technical emergence classifiers
  - Limited examples of emergence & non-emergence over 5 years of the program (~60)
  - Reference baseline will have limited temporal resolution (~5 year blocks)

Case Study: Genetic Algorithms Example

"Genetic algorithms are evolutionary inspired techniques used in computing to find exact or approximate solutions to optimization and search problems by using inheritance, mutation, selection, and crossover."

- **Is there a capability development trigger?**
  - 1950s-1960s: 1st articles in evolution-inspired algorithms appear (little follow-up)
  - 1962: Crossover and recombination operators first emerge (Holland et al.)
  - 1966: Evolutionary programming concepts introduced (Fogel et al.)
  - 1975: "Adaptation in Natural and Artificial Systems" published (Holland) and dissertation shows wide variety of functionality (De Jong)

Source: http://www.talkorigins.org/faqs/genalg/genalg.html

Genetic Algorithms Example (Continued)

- **Is there evidence of capability maturation and impact?**
  - 1985: First Int'l Conference on Genetic Algorithms and Applications
  - 1988: Machine Learning special double issue
  - 1989: Goldberg. "Genetic Algorithms in Search, Optimization, and Machine Learning" book helps pave way for rapid growth in application of methods
  - July 1992: Scientific American article; excitement about capability
  - 1980s-1990s (enabling conditions): Increase in computing power
  - Increasing usage trend in technical papers as successful method
- **Is there evidence of the application of a capability?**
  - 1980s and beyond: Applied to a broad range of subjects
    - stock market prediction and portfolio planning
    - aerospace engineering
    - microchip design
    - biochemistry and molecular biology
    - scheduling at airports and assembly lines

Capability emerged from within one technical area and is applied to many

FUSEnet – Computational Environment

- Government system hosted by Oak Ridge National Laboratory (ORNL); a protected unclassified system with remote access for performers, test and evaluation team, and transition partners (prototype)

- **FUSEnet Specifications**
  - 770 gigaFLOPS* of maximum performance (can double)
  - 16 blade servers, each with 6 cores, totaling 192 processors
  - 96 GBytes of RAM per server for a total of 1,536 GBytes
  - 250 TBytes of storage utilizing a scalable virtualized storage pool
  - iSCSI 10 Gigabit connectivity
  - Virtualized computing space through VMware
  - Access to Document Repository (DR)
  - Functional aspects exposed in a Service Oriented Architecture (SOA)
  - Access and control policies are enforced by ORNL
  - Help line provided

* **FL**oating point **OP**erations per **S**econd