Mining Conference Proceedings for Corporate Technology Knowledge Management

ROBERT J. WATTS
U.S. Army Tank and Automotive Research, Development & Engineering Center, USA:
wattsb@tacom.army.mil

ALAN L. PORTER
Search Technology, Inc., Norcross, Georgia, and Technology Policy & Assessment Center, Georgia Tech, Atlanta, Georgia, USA: alan.porter@isye.gatech.edu

Abstract
An organization's knowledge gained through technical conference attendance is generally isolated to the individual(s) attending the event. The aggregate corporate knowledge is extremely limited, unless the organization institutes a process to document and transfer that knowledge to the organization. Even if such a process exists, the knowledge gains are limited to the experiences and communication skills of the individuals attending the conference. Many conference proceedings are now published and provided to attendees in electronic format, such as on CD-ROM and/or published on the internet, such as IEEE conference proceedings listed at http://www.computer.org/proceedings/proceed_a-h.htm .

These proceedings provide a rich repository that can be mined. Paper abstract compilations reflect "hot topics," as defined by the researchers in the field, and delineate the technical approaches being applied. R&D profiling can more fully exploit recorded conference proceedings' research to enhance corporate knowledge. This paper illustrates the potential in profiling conference proceedings through use of WebQL information retrieval and TechOasis (VantagePoint) text mining software. It shows how tracking research patterns and changes over a sequence of conferences can illuminate R&D trends, map dominant issues, and spotlight key research organizations.

Introduction

How does one keep up with R&D? Information is spewing forth at tremendous rates. Multiple electronic access modes now bring this information “to your fingertips.” New tools enable powerful analyses and information visualizations to help exploit these information resources (Borner et al., 2003; Steinbach et al., 2000; Salton et al., 1994]. We have been particularly interested in ways to exploit compilations of technical text records (usually patent or publication abstracts) (c.f., Watts et al., 2004: Zhu and Porter, 2001). Furthermore, these can be directly pointed to answer one’s pressing management of technology (MOT) questions (Kostoff and Geisler, 1999; Losiewicz et al., 2000; Porter and Cunningham, 2005; Teichert and Mittermayer, 2002; Watts and Porter, 2002).

One key venue for exchange of fast-breaking research developments is “the conference.” This paper targets management of technology issues (originally presented at PICMET, 2005). We presented another paper at that conference as well, called “Mining PICMET: 1997-2005 Papers Helps You Track Management of Technology Developments” (Porter et al., 2005). This built

Preprint of an article submitted for consideration in International Journal of Innovation and Technology Management © 2007 copyright World Scientific Publishing Company, www.worldscinet.com/ijitm

upon an earlier conference compilation (Porter et al., 2003). Those papers accompanied a Reader version of text mining software[a] to enable the PICMET attendees to mine the 2005 conference contents "live." This can help attendees identify papers, people, and institutions doing research which intersects their own interests to interact while they are present at the conference (as well as upon their return home).

The notion of providing the set of a conference's papers over time for ready analysis offers an additional dimension. Namely, one can track the emergence of issues and research streams in that field. In many research areas, conferences lead (and sometimes dominate) journal papers in setting forth research frontiers. Some leading research databases only cover journal articles (e.g., MEDLINE, Web of Science). By having access to a body of conference papers, one can readily generate trends and plot these to identify topics that are of emerging (and declining) interest. One can also see which researchers and which research organizations have been pursuing particular topics within the overall domain, over time. In this way, one can spotlight colleagues with whom one wants to share ideas (either at the conference or later). By analyzing topic congruence (i.e., via forms of clustering or factor analyses), one can focus on particular themes and see how these come to intersect each other. Such text mining can help one identify opportunities of greatest promise in tailoring one's own lines of inquiry.

This paper extends the other paper (Porter et al., 2005) that keys on exploiting the compilation of PICMET research over time. Just as PICMET exposes one to the latest explorations in MOT, other conferences address manifold technical issues. In this paper we illustrate how to gain value-added information from conferences. We explore alternative data access modes and what these can offer technology managers. We compare what it takes to obtain useful MOT intelligence from a) free web versions of the data vs. b) obtaining the proceedings abstract records via pay databases such as EI Compendex and INSPEC. This is one instance of a broader set of opportunities. Many data compilations lack structured indexing of the textual contents (e.g., most web searches, many databases). The approach demonstrated here illustrates a) identification of topical themes from one information resource; b) searching for thematic content in another information resource with structured indexing; and c) then applying that external structure to understand patterns and relationships within the initial information resource.

Technical Approach

Having access to the IEEE website, and professional interest in certain of its topics, we began our investigation there. We searched the IEEE list of conference proceedings for specific topics (e.g., noun phrases within the full conference name) to locate those covering topics of special interest. For this application, we selected a particular conference – IEEE International Workshops on Database and Expert Systems Applications (DESA). We are interested in their coverage from 2001-04 (four conferences). We used WebQL [http://www.ql2.com/] to mine the IEEE Conference Proceedings site [http://www.computer.org/proceedings/proceed_a-h.htm]. WebQL, from QL2 Software, is a software tool enabling quick development and easy deployment of software agents to extract data from the World Wide Web and many other unstructured data sources. We thus identified the conference proceedings of interest and corresponding web links to be mined (using a second WebQL script). We focused our conference listing search on expert systems and discovered the IEEE conferences, "International

[a] This used the *VantagePoint Reader*; see www.theVantagePoint.com for further information on this software.

Workshops on Database and Expert Systems Applications (DESA)." The second WebQL script searched for and retrieved specific conference proceedings web link information and compiled it in Excel format. Figure 1 presents a screen capture of the IEEE web site page that provided the links mined to retrieve the IEEE abstracts analyzed here.

Once extracted, WebQL can structure the data into standard output file formats – HTML, PDF, DOC, etc. We formatted the retrieved conference listings for Excel file analysis. This WebQL output file could have simply been viewed and searched in Excel. However, we imported the file into Tech OASIS[b] to facilitate richer analyses. Each paper's summary information included: conference name, conference date, conference location, paper title, authors name(s), author(s') affiliation(s) and the paper abstract.

The Tech OASIS Excel quick import engine/filter was edited to provide natural language processed (NLP) text fields for both the paper title and abstract fields. Use of NLP-parsed terms and phrases provides a way to mine the actual content of the abstracts. Through NLP text profiling we can get at the topics researchers are pursuing. Our targets include knowledge about the entire research domain of interest, including:

* *what* – what are the hot topics?
* *who* – who are the research leaders on particular topics?
* *where* – where are the centers of knowledge?
* *when* – what are the trends in research?

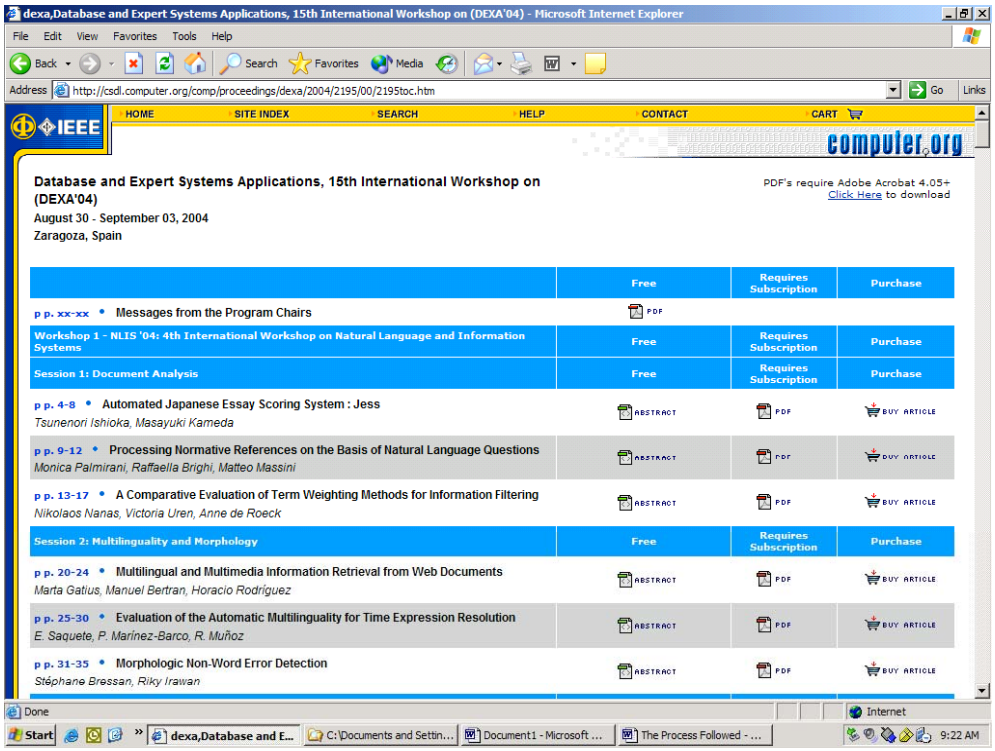

Figure 1- IEEE web site mined by WebQL software

[b] Tech OASIS is for U.S. Government use. The commercial versions of this software are *VantagePoint* [www.theVantagePoint.com] and *Thomson Data Analyzer* [scientific.**thomson**.com/ts/products/tda]

NLP algorithms capture useful chunks of text within the free-text portion of the abstracts. We have found that certain text-processing-aids greatly improve the quality of the information available. By borrowing keywords from indexed databases we assure that domain-specific terms and phrases are captured in the free text. For instance, if we are interested in "expert systems," we don't want the NLP parser to separate the terms into "expert" and "systems."

To identify informative terms in our analyses, we used a 3-step process. First we examined a limited set of abstracts containing research domain terms and phrases within the conference proceedings. Domain-specific terms and phrases from that source formed the basis for a search strategy for a second source -- indexed databases (EI Compendex and INSPEC). The descriptors and identifiers (i.e. the index terms or keywords) from the indexed databases were compiled to create an improved set of terms and phrases for the domain under study. In the third step, these terms and phrases were tagged in the conference proceedings' abstracts files and extracted (i.e., protected during NLP processing on import into Tech OASIS). This resulted in a contextually rich set of entities on which to profile the conference proceedings. Put another way, we "borrowed" the index terms from EI Compendex and INSPEC to help analyze the version of the conference proceedings gathered directly from the website (that lacks index terms).

We began this process by looking for clues. The 2001-04 International Workshops on Database and Expert Systems Applications cover many topics, so devising a suitable search strategy to retrieve corresponding information from huge databases was not trivial. Our "Rosetta Stone" appeared in the 2002 DESA proceedings in the form of a sequence of messages from session chairpersons. These consisted of descriptive abstracts by the co-chairs about the sub-workshops on: holonic and multi-agent systems (HoloMAS**)**, electronic business hubs (WEBH**),** trust and privacy in digital business **(**TrustBus**)**, negotiations in electronic markets (e-Neg**)**, mobility in databases and distributed systems (MDDS), theory and applications of knowledge management (TAKMA), management of information on the web (MIW), web based collaboration (WBC), natural language and information systems (NLIS), web semantics (WebS), and very large data warehouses (VLDWH). The text from these co-chairs' messages was manually scanned and the terms and phrases in the search strategy, Table 1, were identified. This search strategy uses Boolean logic to search EI Compendex and INSPEC. Closed parentheses mean that the terms are required to be adjoining. The question mark indicates wild card character(s).

Had we not found this set of messaging telling about the workshop themes, we would have considered two other ways to generate search terms to use in the databases. One approach is to list the NLP title phrases and highlight defining terms therein. Another is to locate abstracts within the workshop whose titles and/or texts suggest over-viewing – e.g., "forecast of," "technology assessment, "new trends," and so forth.

Our second step applied the Table 1 search strategy to retrieve 3067 and 1344 abstracts, respectively, from the INSPEC and EI Compendex databases (as licensed from Dialog, Inc., a database provider). A combined list of descriptors and identifiers was compiled from the two Dialog search files.

[c] Tech OASIS is for U.S. Government use. The commercial versions of this software are *VantagePoint* [www.theVantagePoint.com] and *Derwent Analytics* [http://www.derwent.com/products/dapt/derwentanalytics/].

Table 1 - Databases and Expert Systems Search Strategy

Set	Items	Description
S1	1297161	PY>2000
S2	3150	S1 AND (EXPERT()SYSTEM?)
S3	5566	S1 AND (MULTI-AGENT?)
S4	3073	S1 AND (DISTRIBUTED()SYSTEM?)
S5	48	S1 AND ((ELECTRONIC()MARKET?) AND NEGOTIATION?)
S6	1519	S1 AND ((COLLABORATIVE OR GRID)()COMPUTING)
S7	2927	S1 AND (KNOWLEDGE()MANAGEMENT)
S8	4391	S1 AND (SOFTWARE()AGENT?)
S9	501	S1 AND (TEXT()(MINING OR SUMMARIZATION OR CATEGORIZATION))
S10	19216	S2 OR S3 OR S4 OR S5 OR S6 OR S7 OR S8 OR S9
S11	41752	S1 AND (INTERNET OR WWW OR (WORLD()WIDE()WEB))
S12	1372	S1 AND (WEB(2N)INFORMATION)
S13	278	S1 AND (WEB(2N)COLLABORAT?)
S14	967	S1 AND (WEB(2N)SEMANTIC?)
S15	42743	S11 OR S12 OR S13 OR S14
S16	3067	S10 AND S15

For our third step, this list was used to extract domain-specific terms via another import of the web-sourced IEEE conference proceedings. Compared to files compiled using the standard Tech OASIS import engine, the resulting abstract files had more than triple the number of abstract NLP terms and phrases available for cluster analyses. This demonstrates the utility of applying index terms (keywords) from outside sources. It also shows the value in protecting those terms during natural language parsing. The results were

- The 2001 proceeding abstracts' extracted NLP lists had 335 terms with record frequencies greater than 2 (208 were descriptor/identifier domain specific entities) vs. 91 terms compiled by the standard NLP processed list.
- The 2002 proceeding abstracts' had 454 such terms (263 entities) vs. 114 for the standard NLP import
- The 2003 proceedings had 316 (195 entities) vs. 81 and
- The 2004 proceedings had 336 (207 entities) vs. 102.

We next describe how these enriched terms were used to profile the IEEE DESA Proceedings. This results in qualitatively richer understanding of the content of these conferences. It enables users to understand overall research emphases, as well as to pinpoint papers of particular interest.

Results

The WebQL web crawler software retrieved 148, 152, 157 and 173 abstracts, respectively, for the 2001, 2002, 2003 and 2004 IEEE Databases and Expert Systems Application (DESA)

conference proceedings. We analyzed the four annual proceedings separately and combined. Managers can gain insights on research "hot topics" by analyzing the individual proceedings. The combined proceedings file provides information on topical trends and regular attendees.

Research Profiling

Research profiling (Porter et al., 2002) extracts information about a research domain by identifying patterns from collections of research outputs. A particular interest lies in tracking changes in research emphases over time periods (Watts, 2003). Table 2 shows the leading organizations vs. conference dates. Such a compilation provides knowledge about who regularly presents at these conferences. This can point us toward cutting edge researchers. For instance, were we planning to send someone to attend the next DESA workshop, we might expressly point them to make contact with the Czech Technical University and University of Greenwich researchers. Observing Table 2, one also observes that foreign sources dominate publication of research at this forum. Is this observation true for the broader field of research?

Table 3 shows the topical emphases of the leading organizations at DESA. These reflect term clusters (or factors – we apply principal components analysis to the NLP extracted entities, terms and phrases). This provides information on research focus areas of each organization. The leading conference presenter (Czech Technical University) concentrates on three primary areas: heritage, interoperability and multi-agents. The Open University's abstracts primarily cluster in only one area -- the heritage factor. Five of six of Hewlett-Packard's abstracts fall in the business factor group. This intelligence would support decisions on who might make attractive collaborative partners. Interesting observation -- four factor groups (authentication, evolution, electronic commerce and e-government) have only one lead organization with more than one abstract. So, the technology manager seeking expertise at this venue has clear targets.

Factor Map Cluster Groups

Although the messages from the co-chairs of the 2002 conference were most useful in developing our database search strategy, they appeared to bias the clustering of the conference proceeding abstracts. Therefore, the co-chair messages were removed from the files before Table 3 was derived. However, Table 4 shows in what groups the co-chair messages clustered during initial analyses. This knowledge helps verify the process for using the NLP-extracted entities to derive the factor groupings.

During the first iteration, eleven factor groups were derived, as shown in the 2nd and 5th columns of Table 4 and preceded by "Map:". The remaining terms in the 2nd and 5th columns are the other terms of the respective factor group. For example, the factor group, Map: University, consists of all abstracts containing terms: technology, research, messaging, university, topic or exchange. All of the co-chair messages clustered together in this University factor group. Additionally, the University factor group contained 190 of the 630 published abstracts. This large number influenced the decision that the co-chair messages were biasing the factor analysis.

Table 2 - Leading Affiliations at IEEE Databases and Expert Systems Applications

	# Records	173	157	152	148
# Records	**Affiliations (Cleaned 2)**	**2004**	**2003**	**2002**	**2001**
15	Czech Technical University, Prague, Czech Republic	3	4	6	2
12	University of Greenwich, London, UK	8	4		
9	University of Vienna, Austria	3	3	2	1
8	Vienna University of Technology		2	1	5
8	Open University, Milton Keynes, UK	3	2	3	
8	Poznan University	1	3	2	2
7	Nanyang Technological University		2	2	3
6	Tohoku University, Japan		2	2	2
6	University of Linz, Austria	1	1	1	3
6	University of Pittsburgh, PA	2	1	2	1
6	Hewlett-Packard Corporation	2		2	2
6	Tokyo Denki University, Japan	2	2	1	1
6	Fraunhofer AIS, Germany	3		3	
5	Middlesex University	1		1	3
5	Università di Milano, Italy	3		2	
5	Yamagata University, Japan	2		1	2
5	National Technical University of Athens			1	4
5	University of Calgary			3	2
5	Toyo University, Japan	2	1	1	1
5	University of Zaragoza, Spain	3			2
5	Monash University		1	2	2
5	Iwate Prefectural University	2	1		2
5	Johannes Kepler University Linz, Austria	2	1	1	1
4	Fukuoka Institute of Technology (FIT), Japan	3	1		
4	University of Alberta, Edmonton, Canada	1	1	1	1
4	University of Oklahoma		2	1	1
4	University of Missouri-Kansas City, Kansas City	1	2		1
4	Università di Brescia, Italy	1	1	2	
4	Imperial College London	3	1		
4	University of Tokyo			2	2
4	City University of Hong Kong		2	1	1
4	University of Montreal			2	2
4	Univ. de Castilla la Mancha, Spain	4			
4	University of Malaga, Spain	2	2		

Table 3 - Leading Affiliations vs. Factor Map Groups

		ABSTRACT (NLP) C:Entities (factors)																	
	# Records	136	118	113	105	83	83	68	64	59	51	43	42	41	36	32	28	26	20
# Records	Affiliations (Cleaned 2)	datasets	retrieval	interoperability	traffic	query	video	Business	authentication	multi-agent	learning	distributed	mobile	knowledge management	evolution	real-time	electronic commerce	Heritage	e-government
15	Czech Technical University, Prague, Czech Republic		3	5			2			4				3	2			6	
12	University of Greenwich, London, UK	3		2			2	2				2		2					
9	University of Vienna, Austria	3	3		3			2			2	4		2					
8	Vienna University of Technology	5	2	2		2													
8	Open University, Milton Keynes, UK		2				2											6	
8	Poznan University	3		2	3	3		2											
7	Nanyang Technological University		2				3										2		
6	Tohoku University, Japan	2			3			2											
6	University of Linz, Austria		2																
6	University of Pittsburgh, PA													3					
6	Hewlett-Packard Corporation							5											
6	Tokyo Denki University, Japan				3														
6	Fraunhofer AIS, Germany				3														
5	Middlesex University	2	2								3								
5	Università di Milano, Italy	2		2	3														
5	Yamagata University, Japan				4														
5	National Technical University of Athens						2	3						2					2
5	University of Calgary				2					5						3			
5	Toyo University, Japan		4																
5	University of Zaragoza, Spain												2						
5	Monash University												3						
5	Iwate Prefectural University		2		2		2												
5	Johannes Kepler University Linz, Austria											2							
4	Fukuoka Institute of Technology (FIT), Japan		3		3														
4	University of Alberta, Edmonton, Canada	3																	
4	University of Oklahoma				2								2			2			
4	University of Missouri-Kansas City, Kansas City																		
4	Università di Brescia, Italy	2		3															
4	Imperial College London				2														
4	University of Tokyo						2												
4	City University of Hong Kong		2		2						2								
4	University of Montreal																		
4	Univ. de Castilla la Mancha, Spain																		
4	University of Malaga, Spain								2										

Table 4 - Factor Map Group Defining Terms (Combined 2001-04 IEEE Proceedings)

# Records	Map: University	Hi-Load (1.63)	# Records	Map: commerce	Hi-Load (0.54)
101	technology	WBC (1.63); MDDS (1.36)	31	commerce	WEBH(0.4)
92	research	NLIS (1.02): MIW (0.92)	14	electronic commerce	e-Neg(0.34); TrustBus(0.28)
25	messaging	HoloMAS(0.9); WebS(0.64)	**# Records**	**Map: XML**	**Hi-Load (-0.51)**
21	University	TAKMA(0.61); e-Neg(0.59)	49	XML	MDDS(-0.5); WEBH(-0.5)
19	topic	TrustBus(0.56)	38	standard	HoloMAS(-0.46)
14	exchange	VLDWH(0.49); WEBH(0.45)	20	data model	WebS(-0.41)
# Records	**Map: trust**	**Hi-Load (-0.69)**	**# Records**	**Map: Heritage**	**Hi-Load (-0.85)**
45	control	TrustBus(-0.61)	64	Ontology	WebS(-0.26)
37	security		17	Heritage	
18	privacy		13	exploration	
18	trust		12	cultural heritage	
14	access control		**# Records**	**Map: datasets**	**Hi-Load (0.99)**
13	authentication		49	analysis	VLDWH(0.58)
# Records	**Map: agent**	**Hi-Load (-0.6)**	16	storage	WBC(0.18)
38	mobile	WBC(-0.34)	15	cluster	
33	agent	MDDS(-0.25)	14	grids	
20	server		13	data warehouse	
16	mobile agents		12	data mining	
# Records	**Map: multi-agent**	**Hi-Load (1.79)**	12	datasets	
48	agents	HoloMAS(1.79)	**# Records**	**Map: information retrieval**	**Hi-Load (-0.60)**
18	agent system	WBC(0.66)	57	documents	NLIS(-0.43)
18	multi-agent		48	search	
14	manufacturing		37	retrieval	
# Records	**Map: traffic**	**Hi-Load (0.94)**	17	information retrieval	
31	algorithms	MDDS(0.94)	**# Records**	**Map: learning**	**Hi-Load (0.40)**
19	real-time	HoloMAS(0.62)	41	learning	MIW(0.30); WebS(0.29)
15	traffic		14	e-learning	NLIS(0.16)

However, observing the clustering of the messages in the other ten factor groups helps validate the NLP entities extraction and standard factor map process used to cluster the abstracts. Viewing Table 4, the "trust" factor, defined by the terms: control, security, privacy, trust, access control and authentication, had the highest loading abstract (i.e., Hi-Load (-0.69)). The co-chair message for the Trust and Privacy in Digital Business working group (TrustBus), the only co-chair message to be clustered in the trust factor group, had a loading coefficient of -0.61, and thus appears appropriately grouped. Two co-chair messages, Web Based Collaboration (WBC) and Mobility in Databases and Distributed Systems (MDDS) had loading coefficients of -0.34 and -0.25 in the agent factor group, which had a highest abstract loading coefficient of -0.6. A level of confidence in the proceedings analysis process can be gained by comparing the factor defining terms and the high loading co-chair messages shown in Table 4.

The factor analyses of the NLP extracted entities were redone, excluding the co-chair message abstracts. The factor map of the combined file (2001-04) of proceedings' abstracts is shown in Figure 2. Each factor, represented as a node, has a drop-down box containing the group-defining terms. When viewed together, these hi-loading terms help provide a better understanding of the concepts documented in the grouped abstracts. Links between nodes show factors that relate more closely to each other.

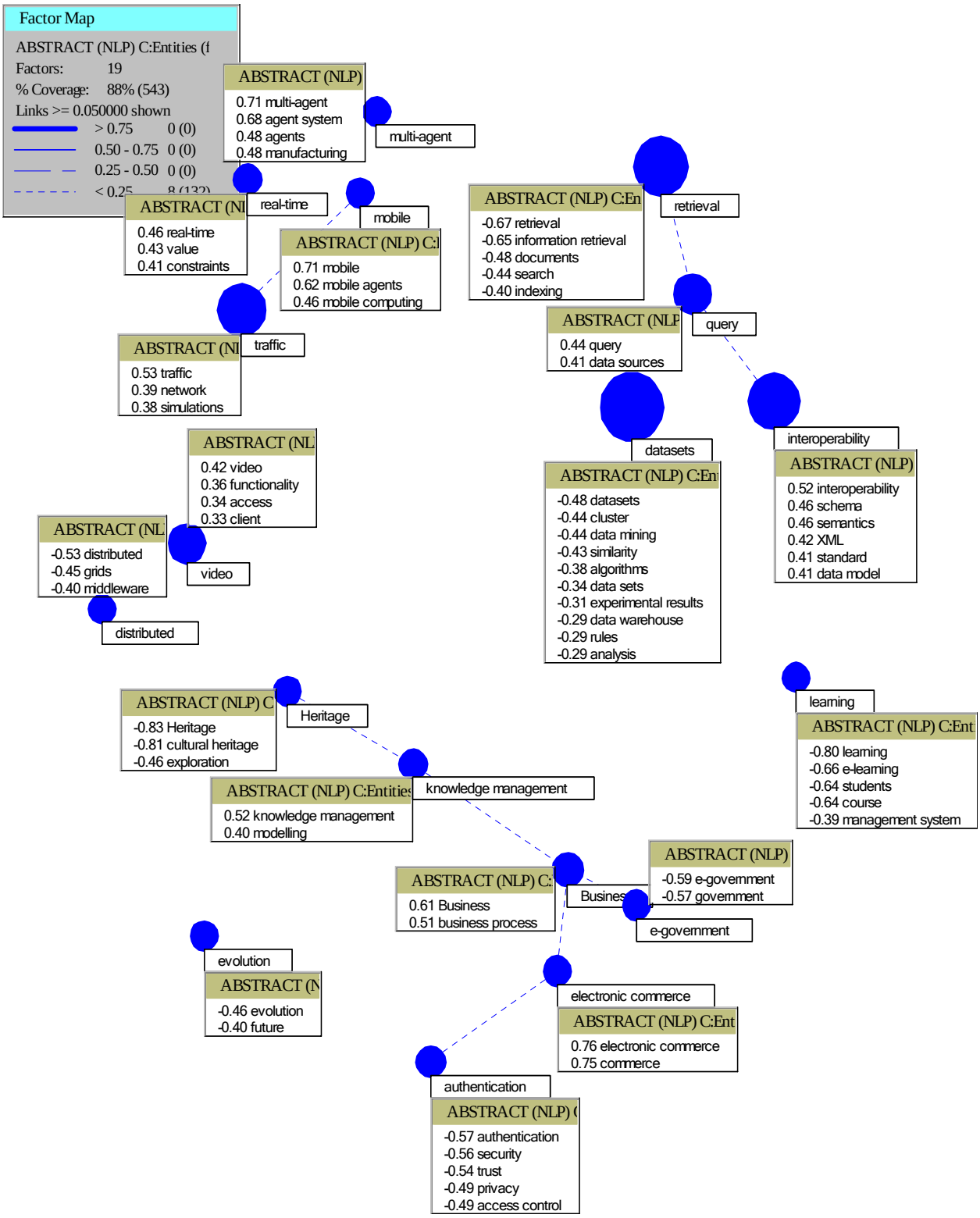

Figure 2 - Factor Map of Abstract NLP Entities – IEEE Databases and Expert System Applications Conference Proceedings 2001-2004 (No 2002 Messages)

Figure 3 provides the histograms for each of the Figure 2 factor groups and the number of abstracts presented annually. Such charts can provide managers intelligence on which sub-disciplines dominate the conference subject matter and which categories of research are declining or rising. For example, publications in e-government, electronic commerce and the business factor groups have declined over the four-year period. Experts in the field could best explain the reasons for the declining research; perhaps, applications have increased (technology matured) and need for research declined. Similarly, one can observe that the four most active areas of research in the 2004 conference were retrieval, interoperability, traffic and query.

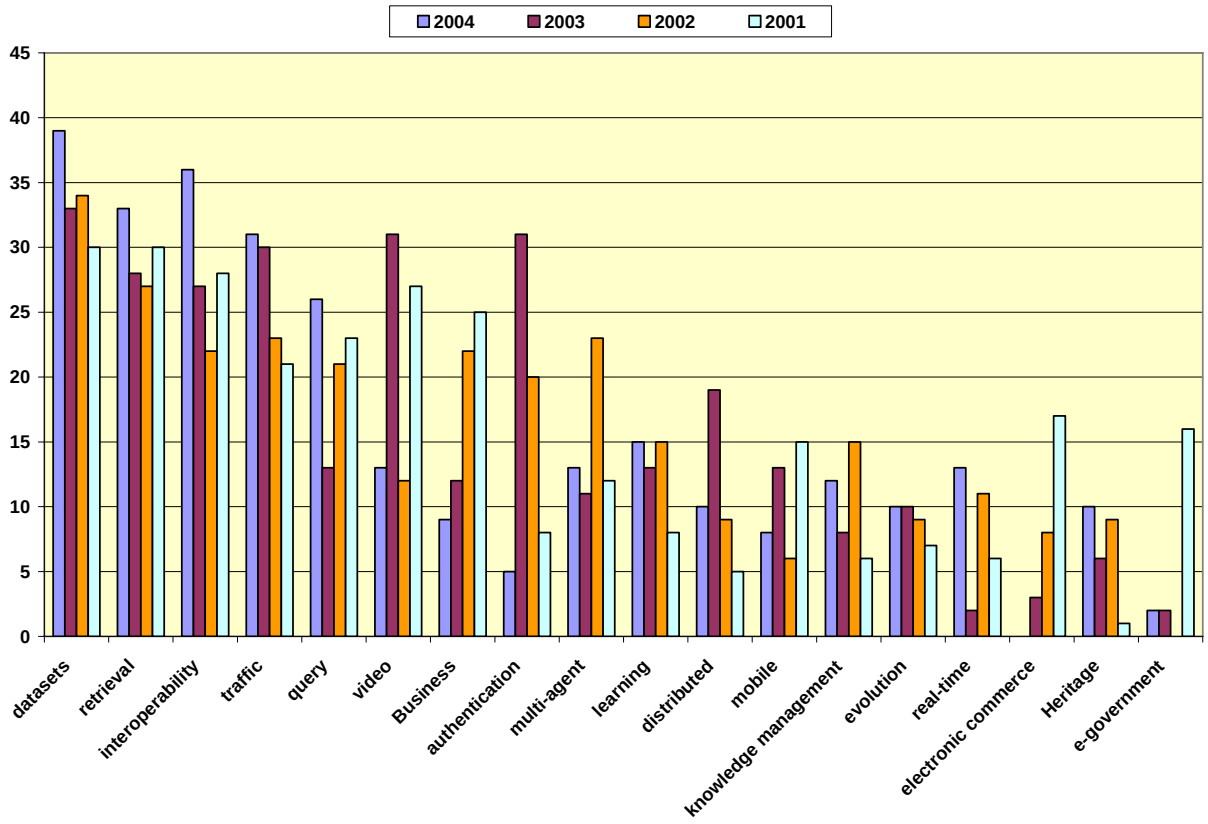

Figure 3 - IEEE Databases and Expert System Applications Factor Groups' chronologies

Figure 4 depicts the factor map for the 12% outliers – the abstracts that were *not* clustered in the factor mapping (Figure 2 depicts 88% of the abstracts). Factor map groups represent consensus term usage. Abstracts not using these consensus terms may represent new research topics.

Let's explore Figure 4 further. The term "autonomic computing" appears in two factor groups. Autonomic computing occurred first in 2003 in 7 abstracts and then in 4 abstracts in 2004. In a 2003 paper, Constantinescu states "Systems which are autonomic, capable of managing themselves are required" in "Towards an Autonomic Distributed Computing System." In a 2003 paper, Sterritt et al. claim autonomic computing aims to (i) increase reliability by designing

systems to be self-protecting and self-healing; and (ii) increase autonomy and performance by enabling systems to adapt to changing circumstances, using self-configuring and self-optimizing mechanisms. This field, autonomic computing, appears to fit the definition of an emerging area of research.

By mining down to individual abstracts that have been self-organized into topical groups, managers can quickly gain insights on the "hot topics." Through such mining in Autonomic Computing, we find that an application needs to be aware of its environment. In the 2004 paper, "Simulation Model for Self-Adaptive Applications in Pervasive Computing," Huebscher et al. state "While the term "environment" is not normally understood as being a physical environment, in Pervasive Computing many applications do actually need to monitor the physical environment in which they are deployed." The profiled conference proceedings can, thus, provide both a "meta-perspective" – a bird's eye view (e.g., who are the leading publishers, what are the central research focus areas, etc.), and targeted access to specific information.

Discussion

This paper analyzes sets of conference proceedings retrieved using a web crawler. Its interests lie in both the content and the mode of accessing it. The content – conference proceedings – is of special interest as indicative of leading research in many domains (not all). This content is not covered as fully as are journal publications by many of the leading science and technology abstracting databases. Thus, we are interested in ways to get at this information for in-depth probing of trends and patterns that indicate areas of emerging promise and the particular research groups pursuing topics of special interest to us. For instance, tracking the emergence of particular topics in the PICMET conferences over a decade is a fruitful way to track the evolution of our field (Porter et al., 2005).

Mining conference proceedings also offers practical management of technology options. The cost of sending a member of an organization to a conference becomes a sunk cost; so CD-ROMs or web links offered to attendees to access the proceedings offer "free" knowledge mining opportunities. The cost for accessing journal articles from licensed databases generally runs on the order of $2.00 to $3.00 per abstract downloaded. Costs for full papers, often available as conference proceedings, would further justify an organization's attention to mining these. Then the time-aspect of the information makes mining conference proceedings additionally compelling. Typically, journal articles from licensed database providers do not become available on the database for one-two years after initial writing (due to reviewing, revision, publication lags, and indexing lags). Conference proceedings are often handed out on CD-ROM at the conference or released on a web link shortly thereafter. Yet, such proceedings offer special challenges in that they may not be as well-indexed as records in the technical databases.

We both demonstrate and begin validating a process to profile non-indexed free-text information (i.e., web-accessible conference proceedings abstracts). We first compile sets of research domain-specific terms from indexed databases (e.g., EI Compendex, INSPEC) from other literature. We then tag and protect those terms or phrases in building lists from the abstract records' free text. The tagged entities are extracted during Natural Language Processing (NLP) parsing of the abstracts to compile a contextually rich set of terms and phrases on which to

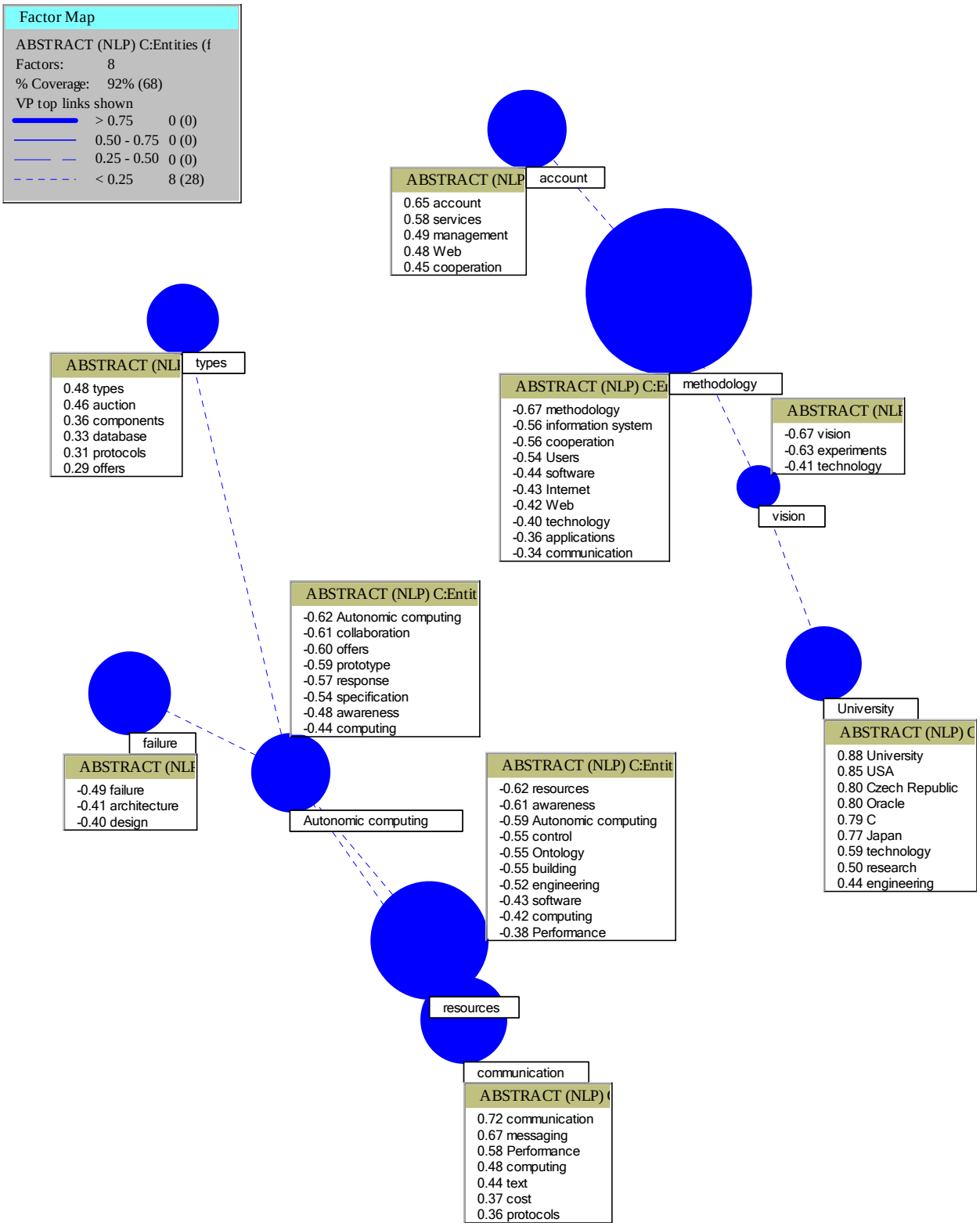

Figure 4 - Factor Map of NLP Entities for Non-factored Abstracts – IEEE Databases and Expert System Applications Conference Proceedings 2001-2004 (No 2002 Messages)

profile the free-text documents. To accomplish this process, we briefly introduce and use WebQL information retrieval and TechOasis (VantagePoint) text mining software.

We suggest analytical approaches to further validate the demonstrated analysis approach. The standard PCA factor analysis process uses a metric comprised of the population percentage clustered and cluster quality measures (entropy, F-measure and cohesiveness) [10]. Further research should compare cluster quality measures for the factor groups of alternative approaches. Cluster analysis strives to create "highly internally homogenous groups, the members of which are similar to one another, and highly externally heterogeneous groups, members of which are dissimilar to those of other groups" (Borner et al., 2003). Steinbach, et al., (2000) discuss and apply measures of cluster quality, both internal and external measures of "goodness."

For this example, we observed lower entropy and F-measures for the factor groups derived from the indexed database for the 2004 conference proceedings than obtained for the NLP entity extracted factor groups (Watts and Porter, 2003). This implies that analysis of indexed data provides better factor groups; but that indexing using external information takes time and resources. However, the NLP entity extraction process clustered the same percentage, 97%, of the 2004 abstracts into factor groups. In contrast, the factor groups, created by the standard NLP abstract terms analysis approach, clustered only 66% of the 2004 abstracts. In addition, the 2004 proceedings abstracts' yielded 336 terms with record frequencies greater than 2 (208 were descriptor/identifier domain specific entities) vs. 102 terms compiled by the standard NLP processed list and 149 available for the indexed terms database, EI Village. The NLP entity extraction process, in this case study, provided the greater number of terms for the factor group analysis. One could argue that it is difficult to make all-inclusive assignments of indexed terms for the abstracts and having self-assignment through entity extraction provides the more thorough approach. Further research should assess this claim.

Again, one form of such technical information is the conference proceedings. Many of these are not covered by major technical databases. We assert that special focus on one or more conferences can be a valuable way to generate technical intelligence. That is, a particular conference may represent cutting-edge research in one or more domains of special interest. That makes it appealing to access and analyze that set of records. In our companion paper (Porter et al., 2005) that analyzes the sequence of papers over five successive PICMET conferences, we demonstrate the potential to identify key thrusts in Management of Technology. One can then probe further. For instance, one might identify the leading research organizations working on a particular topical area; then, one could profile the trajectory of themes they are pursuing, identify the key researchers, and possibly pursue collaboration.

Most importantly, we demonstrate how the profiled conference proceedings can be used by technology managers. Specifically, intelligence about the conference research domain can be derived, including:

* *what* – what are the hot topics?
* *who* – who are the research leaders on particular topics?
* *where* – where are the centers of knowledge?
* *when* – what are the trends in research?

Even today, various technology managers make considerably less use of such empirical knowledge than do their counterparts in production, finance, marketing, and even sports management.

Today, the alert technology manager is aware of three key types of information resource:

- Databases (repositories of filtered R&D information on publications, projects, patents, etc.)
- Internet resources (e.g., "Googling" the web to identify active sites on the topic of interest)
- Human expertise (e.g., those with valuable tacit technical and/or business knowledge).

Mining the ill-structured internet resources offers a special challenge. We note an advantage of using WebQL to retrieve the information to be analyzed. Using WebQL, we could tailor the information, both content and format, to meet our analysis needs. Licensed database suppliers, on the other hand, must provide a set of standard data formats to meet the majority of customer information processing needs. The tailored retrieved information required less cleaning and provided more on-target field lists summaries.

We note the IEEE Databases and Expert System Applications Proceedings contained mostly foreign-sourced research and wondered whether this was true for the broader field. This question begs further research. In a complimentary and more general vane, research on how to gauge conferences as to how well, statistically, they reflect the broader field of research might be of value to technology managers.

In closing, we note that information profiling can support other technology management issues to allow a manager to:

- Assess another organization's strengths and weaknesses (e.g., to refine decisions on merger and acquisition)
- Assess one's own organization's gaps and strengths (then suggest vectors to pursue accordingly)
- Assess an emerging technology to determine its likely development trajectory (especially commercialization)
- Help determine "so what?" as to how that emerging technology fits our organization's plans (road-mapping technologies and products)
- Help manage R&D processes – prioritize programs and projects better by providing empirical bases for decisions
- Inform IP-based strategic choices – help figure out "why?" a competitor is pursuing particular technologies and patenting strategies
- Improve other MOT decisions – technology insertion, national foresight, …

We commend these information resources and analytical tools to organizations with a need to know about emerging technologies.

References

1. Borner, K., Chen, C., and Boyack, K.W. (2003). Visualizing Knowledge Domains, *Annual Review of Information Science and Technology*, **37**: 179-255.
2. Kostoff, R., & Geisler, E. (1999). Strategic Management and Implementation of Textual Data Mining in Government Organizations. *Technology Analysis & Strategic Management*, **11**: 493-525.
3. Losiewicz, P., Oard, D.W., and Kostoff, R.N. (2000). Textual data mining to support science and technology management, *Journal of Intelligent Information Systems*, **15**(2), 99-119.
4. Porter, A.L., and Cunningham, S.W. (2005). *Tech Mining: Exploiting New Technologies for Competitive Advantage*, Wiley, New York.
5. Porter, A.L., Kongthon, A., Lu, J-C., "Research Profiling: Improving the Literature Review," *Scientometrics*, Vol. 53, p. 351-370, 2002.
6. Porter, A. L., Watts, R. J., and Anderson, T. R., (2003). Mining PICMET: 1997-2003 Papers Help You Track Management of Technology Developments, *Proceedings, Portland International Conference on Management of Engineering and Technology* (PICMET), Portland, OR, USA, July.
7. Porter, A. L., Watts, R. J., and Anderson, T. R. (2005). Mining PICMET: 1997-2005 Papers Help You Track Management of Technology Developments, *Proceedings, Portland International Conference on Management of Engineering and Technology* (PICMET), Portland, OR, USA, July.
8. Salton, G., Allan, J., Buckley, C., and Singhal, A. (1994). Automatic Analysis, Theme Generation and Summarization of Machine-Readable Texts, *Science*, **264**: 1421-1426
9. Steinbach, M., Karypis, G., and Kumar, V. (2000). A Comparison of Document Clustering Techniques University of Minnesota, Technical Report #00-034. http://www.cs.umn.edu/tech_reports/
10. Teichert, T., and Mittermayer, M. A. (2003). Text Mining for Technology Monitoring, *IEEE IEMC 2002*: 596-601.
11. Watts, R. J. (2003). Research Evolution in Robotics Fuzzy Control Technologies, *Association of Unmanned Vehicle Systems International* (AUVSI) conference, Baltimore, MD, July.
12. Watts, R. J., and Porter, A. L. (2003). R&D Cluster Quality Measures and Technology Maturity, *Technology Forecasting & Social Change*, **70**: 735-758.
13. Watts, R. J. and Porter, A. L. (2002). Tracking the Evolution of Management of Technology (MOT), *International Association for Management of Technology* (IAMOT) 2002 Conference.
14. Watts, R. J., Porter, A. L., and Minsk, B. (2004). Automated Text Mining Comparison of Japanese and USA Multi-Robot Research, *Data Mining 2004*, Malaga, Spain, September.
15. Zhu, D. and Porter, A.L. (2001). Automated Extraction and Visualization of Information for Technological Intelligence and Forecasting, *21st International Symposium on Forecasting*, Pine Mountain, Georgia, June 17-20.