Journal of Information Science
http://jis.sagepub.com/

Clustering methodologies for identifying country core competencies

Ronald N. Kostoff, J. Antonio del Río, Héctor D. Cortés, Charles Smith, Andrew Smith, Caroline Wagner, Loet Leydesdorff, George Karypis, Guido Malpohl and Rene Tshiteya

Journal of Information Science 2007 33: 21
DOI: 10.1177/0165551506067124

The online version of this article can be found at:
http://jis.sagepub.com/content/33/1/21

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
cilip
Chartered Institute of Library and Information Professionals

JIS

Clustering methodologies for identifying country core competencies

Ronald N. Kostoff

Office of Naval Research, Arlington, VA 22217 USA

J. Antonio del Río and Héctor D. Cortés

Centro de Investigación en Energía, UNAM, Temixco, Mor. México

Charles Smith

Booz-Allen Hamilton, Bethesda, MD 20852, USA

Andrew Smith

University of Queensland, Queensland, Australia

Caroline Wagner and Loet Leydesdorff

University of Ámsterdam, Amsterdam, The Netherlands

George Karypis

University of Minnesota, Minneapolis, MN 55455, USA

Guido Malpohl

University of Karlsruhe, Postfach 6980, 76128 Karlsruhe, Germany

Rene Tshiteya

DDL-OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600, Mclean, VA 22102, USA

Received 28 November 2005
Revised 23 January 2006

Abstract.

The technical structure of the Mexican science and technology literature was determined. A representative database of technical articles was extracted from the Science Citation Index for the year 2002, with each article

Correspondence to: Ronald N. Kostoff, Office of Naval Research, 875 N. Randolph St, Arlington, VA 22217, USA. E-mail: kostofr@onr.navy.mil

containing at least one author with a Mexican address. Many different manual and statistical clustering methods were used to identify the structure of the technical literature (especially the science and technology core competencies), and to evaluate the strengths and weaknesses of each technique. Each method is summarized, and its results presented.

Keywords: Mexico; science and technology; bibliometrics; computational linguistics; core competencies; research evaluation; factor analysis; concept clustering; document clustering; data compression; network analysis; Leximancer; CLUTO; greedy string tiling

1. Background and research objectives

1.1. Country technology assessments

National science and technology (S&T) core competencies represent a country's strategic capabilities in S&T. Knowledge of country core competencies is important for myriad reasons, including:

(1) Assignment of priority technical areas for joint commercial or military ventures.

(2) Assessment of a country's military potential.

(3) Knowledge of emerging areas to avoid commercial or military surprise.

Obtaining such global technical awareness, especially from the literature, is difficult for multiple reasons, including:

(1) Much science and technology performed is not documented.

(2) Much documented science and technology is not widely available.

(3) Much available documented science and technology is expensive and difficult to acquire.

(4) Few credible techniques exist for extracting useful information from large amounts of science and technology documentation [1].

Most credible country technology assessments are based on a combination of personal visitations to the country of interest, supplemented by copious reading of technology reports from that country. Such processes tend to be laborious, slow, expensive, and accompanied by large gaps in the knowledge available. The more credible and complete evaluation processes will focus on selected technologies from a particular country, and provide in-depth analysis.

In the past half century, driven mainly by the Cold War, a large number of country technology assessments were performed [2–14]. The last two decades have seen an expansion in focus to technologies of major economic competitors. Over the past two decades, some of the most credible of these country technology assessments have come from two organizations: the World Technology Evaluation Center (WTEC – Loyola University) and the Foreign Applied Sciences Assessment Center (FASAC – SAIC). In conducting their studies, both of these organizations would gather topical literature from the country of interest, assemble teams of experts in the topical area, have the teams review the literature as well as conduct site visitations, and have the teams brief their findings and write a final report. The studies performed by these groups remain seminal approaches to country technology assessments.

1.2. Text mining technology assessments

The first author's group has been developing text mining approaches to extract useful information from the global science and technology literature for the past decade [15–26]. These studies have typically focused on a technical discipline, and have examined global S&T efforts in this discipline. It is believed that such approaches, with slight modification, could be adapted to identifying the core S&T competencies in selected countries or regions, including estimation of the relative levels

of effort in each of the core technology areas. It is also believed that coupling of the text mining approach with WTEC and FASAC approaches would amplify the strengths of each approach and reduce the limitations. The text mining component would be performed initially to identify:

- Key core competencies and technology thrusts in the country of interest.
- Key interdisciplinary thrusts.
- Approximate levels of efforts in technology-specific competency areas and in interdisciplinary areas.
- Highly productive researchers.
- Highly productive centers of excellence, including those not well known.
- Highly cited researchers.

Once the key technologies, researchers, and centers of excellence had been identified, then site visitation strategies could be developed. The second phase of the effort would be the actual site visitations. A key step in this hybrid process would be demonstration of the ability of text mining to identify the targets of interest with reasonable precision in a timely manner at an acceptable cost. These three driving parameters (performance, time, cost) could be traded-off against each other to provide a balance acceptable and tailored to a variety of potential customers.

1.3. Research objectives

- Evaluate approaches for identifying the technology core competencies of the Mexican research literature, and for assessing levels of effort/emphasis in these core competencies.
- Include both manual and statistical approaches.
- Identify unique capabilities of each approach.
- Focus on clustering approaches whose categories will be determined by the data and algorithms, rather than using pre-determined categories.
- Include network-based approaches as well, especially for identifying the relationships among categories.
- Compare results from the different core competency identification approaches.

2. Overview of approaches and databases used

2.1. Overview

Two major types of information are required for a country S&T core competency assessment. One is technical infrastructure, which encompasses the prolific performers, the journals that contain many of the papers, the prolific institutions, and the most cited papers/authors/journals. The other is technology thrusts, and the relationship among the thrusts. This study focused on obtaining multiple approaches for identifying the S&T thrusts and their relationships.

Section 2.2 describes the database used for the taxonomy analyses. Based on the sampled set of 4529 retrieved papers representing Mexico's total research, two types of taxonomies are presented, manual and statistical. The manual taxonomies require mainly hand-classification of abstracts, journals, and keywords into categories, whereas the statistical approaches use more computer-based pre-classification. In both approaches, strong human input is required for final categorization Section 3 presents the manual taxonomy approaches and results, Sections 4–6 present the statistical taxonomy approaches and results, and Section 7 presents taxonomy comparisons.

There are five manual taxonomy results presented (Section 3), and three major classes of statistical taxonomy approaches presented (concept clustering (Section 4), document clustering (Section 5), and network mapping (Section 6)). Concept clustering is the grouping of words or phrases based

on their co-occurrence in the same text unit. In the present paper, concept clustering techniques include factor matrix-based clustering and multi-link hierarchical aggregation clustering.

In document clustering, documents are clustered based on their overall text similarity. In the present paper, document clustering techniques include greedy string tiling (Section 5.1), entropy-based data compression (Section 5.2), partitional (Section 5.3), journal (Section 5.4), and latent semantic (Section 5.5).

Network mapping presents analysis of Mexico's technology capabilities using network analysis of word co-occurrence to reveal patterns within the data. These patterns can provide information that would not be evident from a visual examination of the data.

The reader interested in detailed results on any of the techniques mentioned above should see reference [27].

2.2. Databases and information retrieval approach

For the present study, the Science Citation Index database was used as the record source. At the time the final data was extracted for the present paper (fall 2002), the version of the SCI used accessed about 5600 journals (mainly in physical, environmental, engineering, and life sciences basic research). The retrieved database used for analysis consisted of selected journal records (including the fields of authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the web version of the SCI for articles that contained at least one author with a Mexico address.

3. Manual taxonomies

Five manual categorization techniques were compared: article titles, journal titles, keywords, full abstracts, journals. Table 1 compares the different manual categorizations of articles into technical disciplines. If manual categorization of the full abstracts is taken as the benchmark, then manual characterization of the article titles is the best approximation, and keyword and journal title counts are poorer approximations.

4. Concept clustering

Two statistically based concept clustering methods were used to develop taxonomies, factor matrix clustering and multi-link clustering. Both offer different perspectives on taxonomy category structure from the document clustering approach described later. None of the clustering approaches included here is inherently superior.

In this section, a synergistic combination of factor matrix and multi-link clustering is described that offers substantial improvement in the quality of the resultant clusters. Once the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant

Table 1
Comparison of manual categorization techniques

Manual categorization comparisons	Article titles	Journal titles	Keywords	Full abstracts	Journals
Physics	29.90%	37.50%	26.00%	23.10%	20.40%
Biological and medical sciences	33.20%	31%	57.60%	34.70%	39.90%
Chemistry	16.50%	11.90%	10.10%	12.90%	10.30%
Other topics	7.10%	6.40%	2.90%	10.50%	11.80%
Agriculture	4.70%	3.60%	1.80%	4.90%	3.70%
Mathematical and computer science	3.60%	3.60%	0.40%	6.30%	5.30%
Earth sciences and oceanography	2.50%	2.60%	0.60%	5.10%	4.70%
Material science	2.50%	3.50%	0.60%	2.40%	3.80%

technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (e.g. a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc.) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context [28, 29].

The remainder of this Section presents the multi-link clustering only. See reference [27] for factor matrix details.

4.1. Multi-link hierarchical word clustering

4.1.1. Multi-link clustering approach A symmetrical co-occurrence matrix of the highest frequency high technical content words/phrases was generated. The matrix elements were normalized using the equivalence index $E_{ij} = C_{ij}^2/C_i{*}C$, where C_i is the total occurrence frequency of the ith word/phrase, and C_j is the total occurrence frequency of the jth word/phrase, for the matrix element ij), and a multi-link clustering analysis was performed using the WINSTAT statistical package. The complete linkage hierarchical aggregation method was used. A detailed description of the final word dendrogram (a hierarchical tree-like structure), and the aggregation of its branches into a taxonomy of categories, are shown in reference [27]. A summary description now follows.

4.1.2. Multi-link word clustering results The top level clusters form a flat set. Some of the clusters have a distinct hierarchical structure into sub-clusters, where a technology area can be divided into its specific sub-technologies.

The 249 words in the dendrogram are grouped into top level clusters. At this level, five broad topics (categories) can be discerned from visual inspection of the types of words in each cluster. These include biology, medicine, physics, chemistry, and environment. Each of these highest level clusters is then divided into smaller clusters by the technical experts, who evaluate the mix of words in each smaller cluster, and then assign a theme to each cluster.

Category 1 – biology

There are four main groupings: membrane biology/cell–cell recognition; microbial molecular biology/gene expression; recombinant DNA biology; plant population genetics.

Category 2 – medicine

There are five main groupings: cardiopulmonary; reproductive; liver damage; immunology; chronic disease treatment.

Category 3 – physics

There are four main groupings: quantum and dynamical systems; accelerator physics; solid-state; astrophysics.

Category 4 – chemistry

There are three main groupings: polymers; molecular characterization; thin films.

Category 5 – environment

There are four main groupings: forest and agriculture; oceanography and geophysics; heavy metals in sediments; fish growth.

These thematic areas coincide with the major thematic areas listed in Table 1, especially those determined by manual categorization of the full abstracts. In Table 1, agriculture and earth sciences and oceanography were listed as separate themes, whereas the present taxonomy lists them under environment.

5. Document clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [30–37]. Five approaches were examined in this paper: greedy string tiling,

entropy-based data compression, partitional clustering, automatic journal categorization, and latent semantic clustering.

5.1. Greedy string tiling

5.1.1. Greedy string tiling approach The approach presented in this section is based on a greedy string tiling (GST) text matching algorithm [38, 39]. Basically, GST clustering forms groups of documents based on the cumulative sum of shared strings of words. Each group is termed a cluster, and the number of records in each cluster, and the highest frequency technical keywords in each cluster, are two outputs central to this analysis.

5.1.2. Greedy string tiling results A 5% similarity threshold produced a total of 1072 clusters. Ninety-three percent of the clusters contained eight abstracts or less. The 64 largest clusters (containing 804 abstracts) were extracted.

The taxonomy defined by the word clustering algorithms was used to categorize the 64 clusters generated by the greedy string tiling approach. Each cluster was assigned to the most appropriate category in the taxonomy defined by the WINSTAT-generated dendrogram of the last section, based on the theme suggested by the highest frequency technical keywords. The number of records in each taxonomy category from all the clusters in the category was calculated, and is shown in Table 2.

Compared to the full abstracts results of Table 1, the present GST categorization provides reasonable agreement in biology and medicine (30 vs 34%), modest agreement in physics (23 vs 33%), and poor agreement in chemistry (13 vs 23%).

5.2. Data compression clustering

5.2.1. Data compression clustering approach The compression algorithm approach [40] of this section assumes that the entropy of a string can be measured when this string is zipped (compressed). The main idea is that when one compresses two strings sequentially, the compression rate will increase if the second string is similar to the first one, and then the zipped string will have less disorder (entropy) than the previous two strings. The entropy is defined as

$$\text{(A)}\quad \text{Entropy} = (\text{Length}(\text{zip}(A + b)) - \text{Length}(\text{zip}(A)) - \text{Length}(\text{zip}(b + b)) + \text{Length}(\text{zip}(b)))/\text{Length}(b)$$

where A is the patron text, b is the abstract to be analyzed, and zip indicates the zipped function. The fundamental objective is to automate the classification of records into pre-defined categories, such as the Defense Technical Information Center (DTIC) themes. The complete abstract of each record is then compared against the patron text for each pre-determined DTIC theme, and then each record is assigned to an area that provides the best match.

Nineteen patron texts or lexicons for 19 DTIC themes are defined. With these 19 DTIC theme dictionaries, the 4529 abstracts are compressed. Then, using the best compression rate, the corresponding first level categorization theme for each abstract is selected.

Two other variants of the entropy formula are used:

$$\text{(B)}\quad \text{Entropy} = (\text{Length}(\text{zipL}(A + b)) - \text{Length}(\text{zipL}(A)) - \text{Length}(\text{zipL}(b + b)) + \text{Length}(\text{zipL}(b)))/\text{Length}(b)$$

where zipL indicates a zipping process with the lexicon as parameter. This variant allows shorter calculation time.

$$\text{(C)}\quad \text{Entropy} = (\text{Length}(\text{zipL}(L + b)) - \text{Length}(\text{zipL}(L)) - \text{Length}(\text{zipL}(b + b)) + \text{Length}(\text{zipL}(b)))/\text{Length}(b)$$

where the difference is that the Lexicon has been used as a patron text. The computational time is reduced of the order of 6 to 3 hours from the (A) to (C) entropy measurement.

Table 2
Assignment of GST clusters to categories

Cluster number	Biology	Medicine	Physics	Chemistry	Environment
1			75		
2					26
3		25			
4				19	
5				17	
6				17	
7			16		
8			16		
9		15			
10		15			
11		15			
12			13		
13		13			
14	13				
15					13
16			12		
17					12
18			12		
19					12
20		12			
21				12	
22		11			
23				11	
24			11		
25			11		
26			11		
27		11			
28				11	
29				11	
30		11			
31					11
32				11	
33			10		
34		10			
35			10		
36	10				
37				10	
38					10
39				10	
40		10			
41				10	
42			10		
43			10		
44			10		
45		10			
46				10	
47			10		
48	9				
49			9		
50			9		
51	9				
52	9				
53	9				
54			9		
55				9	

(continued)

Table 2 *(Continued)*

Cluster number	Biology	Medicine	Physics	Chemistry	Environment
56		9			
57	9				
58				9	
59				9	
60					9
61					9
62				9	
63					9
64		9			
SUM	68	176	264	185	111
SUM (NORM)	0.08457711	0.21890547	0.32835821	0.2300995	0.1380597

5.2.2. Data compression clustering results Here, it is important to note that with this method it is possible to analyze all abstracts. The results for automated classification with relative entropy defined by (A), (B) and (C) are given in Table 3.

Although there are some differences between these approaches and the manual characterization, all these results are statistically equivalent to the manual using the chi-squared statistical test.

5.3. Partitional clustering

5.3.1. Partitional clustering approach The approach presented in this section is based on a partitional clustering algorithm [41] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to

Table 3
Automated classification

A. Automated classification A formula	
Physics	23%
Biological and medical sciences	32%
Chemistry	8%
Agriculture	8%
Mathematical and computer sciences	9%
Earth sciences and oceanography	8%
Material sciences	12%
B. Automated classification B formula	
Physics	16%
Biological and medical sciences	37%
Chemistry	6%
Agriculture	7%
Mathematical and computer sciences	11%
Earth sciences and oceanography	4%
Material sciences	19%
C. Automated classification C formula	
Physics	16%
Biological and medical sciences	38%
Chemistry	6%
Agriculture	7%
Mathematical and computer sciences	11%
Earth sciences and oceanography	4%
Material sciences	18%

maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements.

5.3.2. Partitional clustering results In partitional clustering, the number of clusters desired is input, and all documents in the database are included in those clusters. The 64 clusters were aggregated into a hierarchical taxonomy using a hierarchical tree generated by the CLUTO software. The taxonomy is shown in Figure 1. The categories in the taxonomy levels, and the number of documents in each category, are described as follows.

In Figure 1, the columns represent the taxonomy levels. There are six levels depicted in this taxonomy. The highest level (two categories) is the first column, and the lowest level shown (approximately 64 levels) is the last column. The numbers in parentheses represent the number of records assigned to the category.

The first level has two categories: biomedical and ecological (2094) and engineering and physical science (2435). Percentage-wise, this is a split of 46/54%. In Table 2 (the manual assignment of GST clusters to categories defined by the word clustering approach) combining the biology, medicine, and environment categories is equivalent to the biomedical and ecological category in Figure 1, and combining the physics and chemistry categories is equivalent to the engineering and physical science category in Figure 1. In Table 2, the category split of 44/56% compares very favorably with the 46/54% split of Figure 1. In Table 1, the category split of 45/55% for the manual clustering of the full abstracts compares favorably as well.

In Figure 1, the second taxonomy level is generated by sub-dividing each first level category by two. Biomedical and ecological divides into biomedical (1267) and ecology (827), while engineering and physical science divides into materials and films (893) and mathematical, physics, and astrophysics modeling (1542).

Again, comparing Figure 1 with Table 2, biomedical (from Figure 1) is roughly equivalent to the combination of biology and medicine (from Table 2), and ecology (from Figure 1) is roughly equivalent to environment (from Table 2). The term 'roughly' is used because sometimes allocation to biology vs medicine is not overly clear, or assignment to biology vs environment is not overly clear. The biomedical/ecology ratio from Figure 1 (1.53) compares only modestly well with the (biology and medicine)/environment ratio from Table 2 (2.2). The definitional uncertainties are reflected in quantitative differences. Inspection of the GST clusters vs their partitional clustering counterparts shows that these quantitative differences represent manual assigment of clusters to categories vs computer assignment of clusters to categories, more than any intrinsic cluster differences.

Further, materials and films (from Figure 1) is roughly equal to chemistry (from Table 2), and mathematical, physics, and astrophysics (from Figure 1) is roughly equal to physics (from Table 2). The term 'roughly' is used here because sometimes the allocation to chemistry vs physics is not overly clear, especially for materials projects, where the physics of materials and the chemistry of materials are sometimes indistinguishable. The (materials and films)/(mathematical, physics, and astrophysics) ratio from Figure 1 (.58) compares reasonably well with the chemistry/physics ratio from Table 2 (.70). Also, the (materials and films)/(mathematical, physics and astrophysics) ratio from Figure 1 (.58) compares well with the (chemistry and materials sciences)/(physics and mathematical and computer science) ratio of full abstracts from Table 1 (.52).

One final comment about Figure 1. Using 64 clusters allows a reasonable picture to be drawn about broad areas of research. If detailed program thrusts were desired, however, many more clusters than 64 would be required. The specific number depends on the degree of focus desired.

From reference [27], the recent Mexico S&T expenditures are on the order of $2.5 b/year. If 64 clusters are used to categorize this S&T, then each cluster (on average) covers about $40m/year of S&T expenditure. This reflects rather broad categories. If, however, 512 clusters are used, then the resolution increases to about $5m/year for the category average. This level of resolution would cover small groups of projects.

THE STRUCTURE OF MEXICO RESEARCH - 64 CLUSTERS

Biomedical and ecological (2094)	Biomedical (1267)	Microbiology laboratory studies (699)	Proteins, genetics (307)	Protein activity (207)	Calcium channel currents, sperm modulation (45)
					Large protein activity (162)
				Gene transcripts, sequencing, and expression (100)	Gene transcripts, sequencing, and expression (100)
			Laboratory cell experiments, receptors (392)	Cell infections, immunology, mice (193)	DNA analysis of cell cultures (132)
					Infection immunology, mice (61)
				Receptors, rats (199)	Neuron receptors, rats, sleep induction (114)
					Rats, liver, dialysis (85)
		Clinical studies (568)	Clinical studies, diseases (269)	Patient congenital syndromes (93)	Patient congenital syndromes (93)
				Patient infectious diseases (176)	Patient infectious diseases (176)
			Clinical studies, women and children (299)	Insulin and diabetes, women, men (90)	Women, HPV, cervical (41)
					Women, insulin, diabetes, obesity, BMI (49)
				Children's health, Mexico City (209)	Children, blood tests, lead, infections (119)
					Health, Mexico City, water, radon (90)
	Ecology (827)	New species (267)	New species (86)	New species (86)	New species (86)
			Mexican ecology species (181)	Species, forest habitation (104)	Species, forest habitation (104)
				Species, Mexican fish (77)	Species, Mexican fish (77)
		Food population and environment (560)	Sediment, fish abundance, Gulf of California (145)	Sediments, Gulf of California, river water (70)	Sediments, Gulf of California, river water (70)
				Seasonal fish abundance (75)	Seasonal fish abundance (75)
			Plant and fruit populations (415)	Plant and fruit populations, soils, seeds (227)	Population genetics, wheat genotypes (104)
					Plants and fruits, soils, seeds (123)
				Growth, diet, food (188)	Food, diet, growth (62)
					Grain processing (126)

Fig. 1. Partitional document clustering taxonomy.

5.4. Journal clustering

In the information provided by ISI there is a register indicating category or categories of the journal. This section utilizes this classification of journals by categories, and papers are associated in accordance with the category in the ISI.

5.4.1. Journal clustering approach The simplest form of clustering the journals is to use the register provided by ISI. However, the criteria used by ISI in the classification are not in agreement with the DTIC taxonomy and there are several hundred categories. For this reason, we group the categories provided by ISI manually with the goal of obtaining a classification as close as possible to that of DTIC, and then we count the number of papers with the register in the ISI. Thus, the use of ISI classification provides useful information, as can be seen in the results.

5.4.2. Journal clustering results Table 4 presents the results of the automated classification. These results seem to be in agreement with the manual classification according with DTIC, at least in names. Please note that some papers appear in two or more categories, because ISI gives this possibility. However, these cases are less than 5% of the total sample.

5.5. Self-organizing named concept extraction and clustering (latent semantic)

5.5.1. Concept extraction and clustering approach This approach to concept extraction and clustering employs a Bayesian analysis of word co-occurrences, but one that includes nonlinear machine learning algorithms. The method passes through four stages of processing. The first stage involves the seeding of named concepts via extraction from the text of seed terms which possess particular statistical characteristics. The second stage learns a family of related terms around each seeded concept by means of an iterative optimizer with feedback. The result of the first two stages is referred to as a thesaurus, since it bears some resemblance to the thesauri used in information science applications. At this stage, the thesaurus has no hierarchy – it is flat. In the third stage, the thesaurus is used to classify the text at a two-sentence resolution. The tagging of each two-sentence segment with multiple concepts generates a directed network of concept co-occurrences. The final stage treats the network of concept co-occurrences as a complex system in order to extract emergent thematic groupings of concepts. This stage results in an interactive visualization of the concept network. For non-interactive publication, the spatial proximity of clustered concepts and the connectedness of each concept are

Table 4
Automated classification according to ISI

Category	Number	Fraction
Astronomy	217	0.046229
Atmosphere	45	0.009587
Behavior	96	0.020452
Biology	1825	0.388794
Computer	63	0.013421
Chemistry	464	0.09885
Electronics	117	0.024925
Energy	63	0.013421
Engineering	101	0.021517
Environmental	170	0.036216
Geosciences	105	0.022369
Materials	276	0.058798
Mathematics	157	0.033447
Mechanics	30	0.006391
Multidisciplinary	32	0.006817
Ocean	92	0.019599
Physics	819	0.174478
Radiation	22	0.004687

used to generate a ranked recursive schedule of concept groups. At the lowest level, each concept is described by the lexical term list from the thesaurus.

More details of the method are given in Reference [42].

5.5.2. Concept extraction and clustering results Table 5 contains some examples of thesaurus entries (not in strict rank order), which form the lowest level of the hierarchy. After classification of the data using the thesaurus, and subsequent emergent clustering, a hierarchical concept net was obtained. An annotated screen shot of this, taken from the interactive browser, is shown in Figure 2.

For the purposes of non-interactive publication, this 2D clustering of the hierarchical network is then serialized into a ranked recursive list of thematic concept groups. Some of these are listed in Table 6 (not in strict rank order). The interactive version of the full network is currently available from www.leximancer.com/documents/mexico_report/report.html.

Finally, it should be noted that this approach naturally results in automatic classification of the text. This classification system can be used to explore the collection.

6. Network mapping of word co-occurrence

This section discusses the data sources and methods, the use of network analysis, and the results of the analysis.

6.1. Approach

6.1.1. Data sources The materials consist of the titles and abstracts of 4529 documents collected from various sources on the selection criterion of an institutional address in Mexico. Abstracts and titles are studied separately. The titles contain 10,956 words that occur in total 40,852 times. The abstracts contain 31,724 unique words that occur in total 482,922 times.

The title words are packed more densely than the abstract words. Note that the ratio is 40,852/10,956 = 3.73 for title words and 481,922/31,724 = 15.18 for abstract words. This accords with previous research in which it was shown that abstract words are less codified than title words [43]. Sentences indicating copyright issues were removed from the abstracts. The stop word list available at www.uspto.gov/patft/help/stopword.htm was used as a corrective to the inclusion of common words. Otherwise, the words were corrected only for the plural 's'.

Table 5
Concepts and their related lexical terms

Concept	Lexical terms
Cells	cells Trh internalization C × 43 cell Sertoli transfected macrophage Sf9 lymphocyte germinal dendritic proliferate cancers monocytic
Species	species helminths Monstrilla subgenus Atlantic_ocean monstrilloid Coreidae Hemiptera tribe synonym Cercidium digenean Qpf niche greggii
Surface	surface plasmon adsorbed passivation broadening Bet pacificus higher-mode probing Fvc radiometry wafer 4 × 2 acetylene scribeline
Films	films thin Cds spray sputtering ellipsometry foils Cdo Cbd Films as-deposited co-sputtering F-7 filamentous Sb2s3-cus
Acid	acid acetic lactic bell linoleic nucleic uric arachidonic lysophosphatidic demineralization niflumic glutamic aminolevulinic Taurine retinoic
Gene	gene encodes encoded Streptomyces reporter undetectable exons di-rhamnolipid Drd4 Recr Rhlc St ichthyosis Ais rhamnosyltransferase
Quantum	quantum dots dilatonic Thomas-Fermi exciton undetected excitons reflectometry spins mechanics worlds billiard inter-band polarization-modulation rigorously

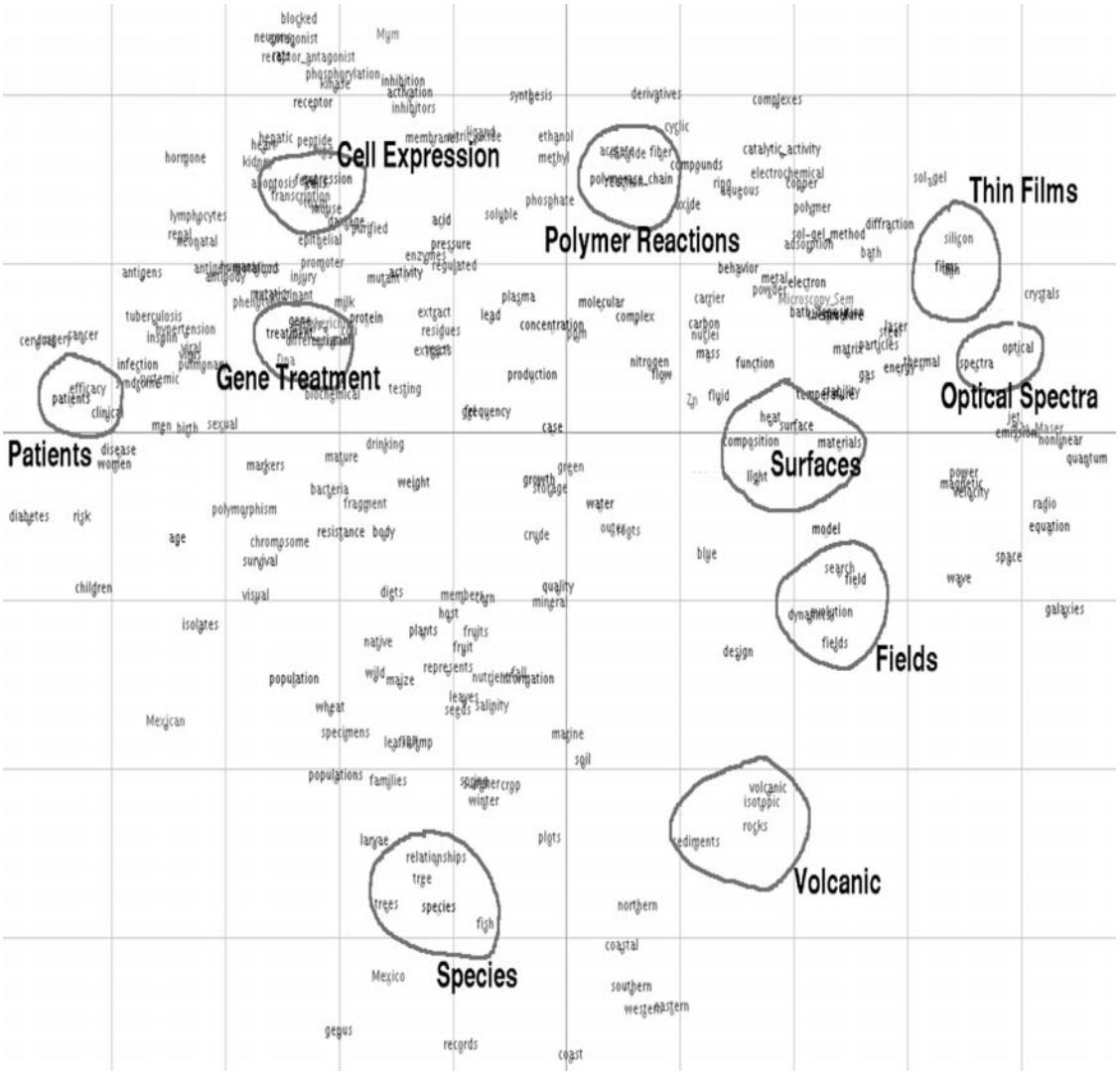

Fig. 2. Hierarchical concept net.

6.1.2. Analysis An analysis of the data shows that 100 abstract words occur more then 500 times, and that 108 title words occur more then 40 times. In both cases, an asymmetrical matrix was constructed containing the 4592 documents as the cases and the respective word set as the variables. From this matrix a symmetrical matrix of co-occurrences among the words was generated and a second symmetrical matrix was constructed based on the cosine as a similarity criterion between the words as variables [44–48].

The symmetrical matrices are analyzed using Pajek [49]. The asymmetrical ones are factor analyzed using SPSS (Varimax rotation and Kaiser normalization). Figure 3 provides an example of a co-occurrence map (of abstract words) and Figure 4 an example of a vector-space model based on the cosine matrix using title words.

6.2. Results

6.2.1. Abstracts Of the 100 abstract words used, 63 co-occur more than 500 times. These are depicted in Figure 3. They form a star-shaped network with some interconnecting hubs. The words 'effect' and 'result' function as hubs and represent the methodologies and their outputs; thus, these

Table 6
Thematic concept groups

Group name	Child groups and leaf concepts
Cells	cells protein expression treatment gene human blood receptor damage DNA coli Escherichia antibody apoptosis heart recombinant fetal mouse resistant epithelial mutations hepatic mutant milk purified toxin antigen injury promoter biochemical peptide lung assays differentiation phenotype mutation transcription kidney expressing inhibit gland peripheral mitochondrial epithelial_cells regulatory mild actions disorder apoptotic potent saline participation protection organs subunit peripheral_blood initiation pathogenic cells_expressing Western_Blotting
Surface	surface electron materials chemical bath_deposition composition behavior sol-gel_method gas particles metal matrix laser stability heat Microscopy_Sem adsorption powder polymer bath steel alloy aluminum coatings electrode oxides Sem eta reactor silica reversible Pb Ti ionization chains tau UV loop microscopic Ftir Cr decomposition surface_tension crude_oil
Patient	patients disease infection women clinical risk insulin cancer virus men syndrome tuberculosis hypertension cervical antigens birth pulmonary viral surgery efficacy systemic surgical parasite men_women oral care diabetes_mellitus cervical_cancer hospital cardiac birth_weight mycobacterium_tuberculosis systemic_lupus divided_groups multivariate_analysis intestinal_metaplasia pulmonary_tuberculosis patients_underwent
Optical	optical emission spectra thermal magnetic H2o_Maser velocity nonlinear power jet radio transverse excited disk Gaas transitions charged tension photon detector formula oscillations mechanics neutron transverse_momentum quantum_wells excited_states phase_transitions porous_media
Plants	plants body host fruit leaves wild diets corn shrimp maize native spp members salinity seeds fruits leaf represents germination nutrient comparative recovered juvenile nutritional winter_spring white difficult spring_summer segment requirements eggs head crude_protein similarity movement majority superior date white_shrimp
Species	species Mexico larvae genus fish tree relationships records trees habitat vegetation seasons genera larval forests
Space	space galaxies wave radio scalar disk gravity compact dual algebra metric formula black_holes matrices expressions scalar_field quantum_wells

results are not highly indicative of capacity. Other words that act as hubs may be more indicative of capacity, including 'cell,' 'patient,' and 'model'. In particular, cell and patient may be aligned with biomedical or biotechnology research.

Normalization of the word occurrences using the cosine as a similarity criterion does not change this picture qualitatively, although some of the stronger relations are highlighted because the star shape is less pronounced in the vector-space model.

6.2.2. Title words Among the 108 title words that occur more than 40 times in the set, 53 words co-occur more than 10 times. If the threshold for the cosine is set at 0.1, 75 words are included in the vector-space model This results in an informative picture (Figure 4).

The map shows that several groupings in the data can be distinguished. The clusters appear to bolster the suggestion drawn from mapping the co-occurrences among title words (not shown here) that there are capacities in biomedicine, biotechnology, materials science, and possibly chemistry. These can be further refined to show the possibility of a specialty in materials related to semiconductors (E1 and E2), biotechnology related to genetic expression within human cells (F), and chemical synthesis at the molecular level – nanotechnology? – (G1 and possibly G2).

In addition, this level of analysis suggests several capacities that are not revealed in any other figure. These include a cluster (H) which may suggest capacity in physics and/or astronomy. The cluster revealed in (J) suggests capacities related to semiconductors, polymers and/or geophysics. The cluster (K) also shows a co-occurrence among the words related to optical research, possibly indicating capacities in lasers or other optical research.

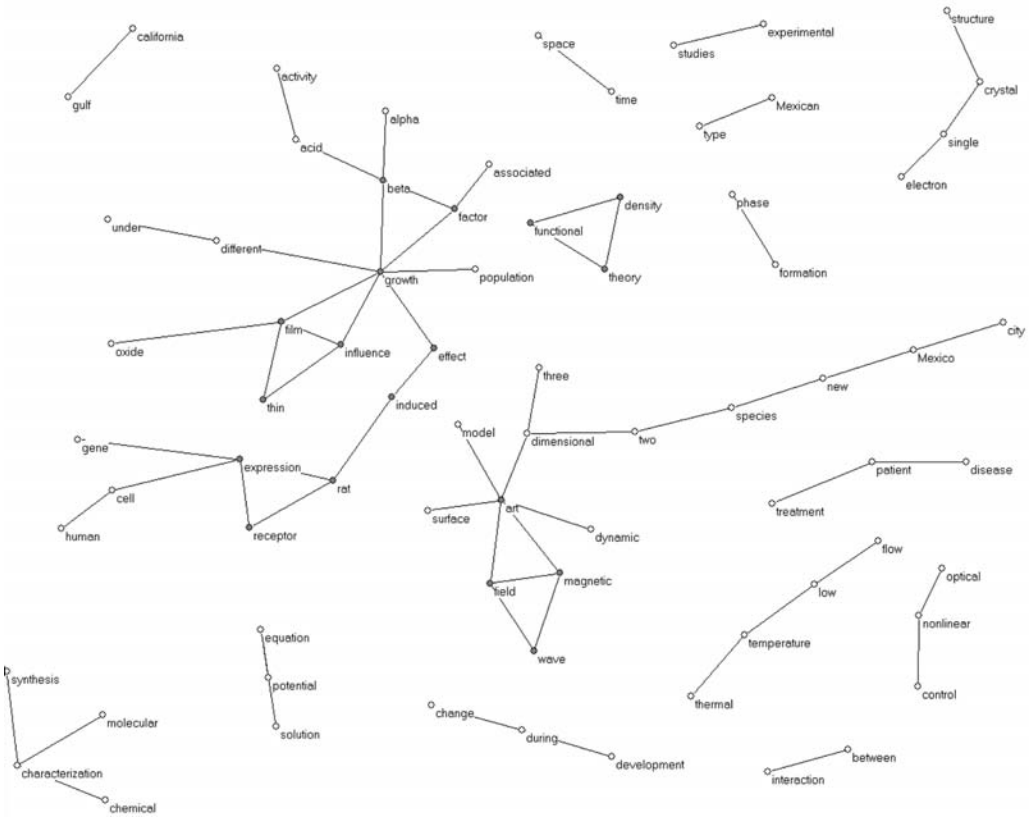

Fig. 3. Co-occurrence map of 63 abstract words co-occurring more than 500 times.

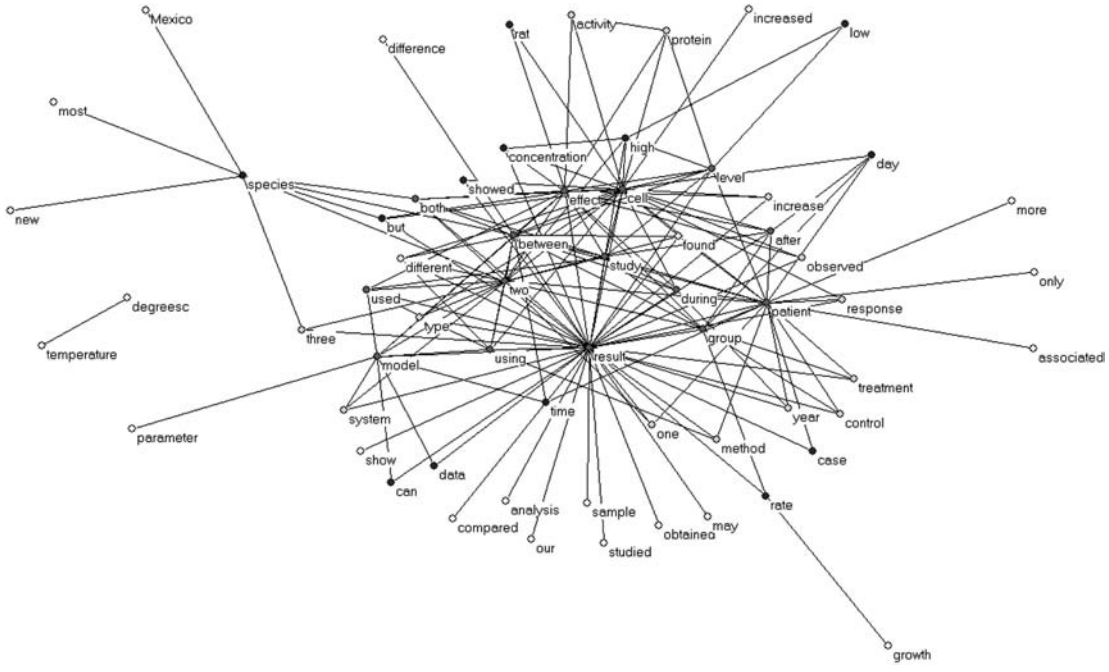

Fig. 4. Cosine-based map: 75 title words are included at the threshold level of cosine ≥ 0.1.

6.2.3. Observations on network mapping results The data is weakly codified. This is a consequence of the selection criterion of the retrieval (i.e. an address in Mexico). Different lines of research are drawn into the set and the set is therefore very heterogeneous. Small groups of co-occurring words can be distinguished in the set of title words, but the abstract words are mainly tied together because of the words related to the word 'results'.

The structure in the title words can be appreciated as intellectually meaningful despite the weak structure in the network among the words. Analysis of the title words is in some ways more suggestive than the abstract words. The vector-space model of the title words suggests certain capacities within Mexican technology relating to biotechnology, biomedicine, materials research, chemistry, and physics. This can be checked against overall publications records and citations, which suggest Mexican strength in physics and chemistry [50].

7. Taxonomy comparisons

Three generic approaches to taxonomy construction were presented: manual clustering, statistical concept clustering, and statistical document clustering. The manual clustering of abstracts was used as the benchmark, and was approximated most closely in the manual group by manual clustering of titles.

The concept clustering approaches (factor matrix, multi-link word/phrase, self-organizing concept extraction, network analysis) provided complementary perspectives, and all identified the major thrust areas. The document clustering approaches (greedy string tiling, partitional clustering, data compression, journal clustering) showed reasonable agreement among each other, and with the manual abstract clustering (see Table 7 below). The main differences appear to be among biomedicine, chemistry/materials, and environment. Chemical reactions and biological organisms play a role in all three literatures, and slight differences in similarity determination could result in transference of documents among these three clusters.

8. Summary and conclusions

The main objective of this study was to identify and assess the technical core competencies of Mexico. This was accomplished using a variety of manual and statistical clustering approaches. There appear to be four major technical core competencies: biomedical sciences includes about 35% of Mexican research; physics/mathematics includes about 30%; chemistry/material sciences covers about 15%; and environmental sciences includes about 10%. The remaining 10% of Mexican research is allocated to myriad other research topics.

If manual clustering is to be used for taxonomy development, the full abstract is preferable. If the full abstract is not available, manual clustering of titles is an acceptable alternative.

The different concept clustering approaches provided complementary perspectives. The factor matrix approach provided good intra-theme word/phrase quantification linkages, while the network-based approaches provided excellent maps of related concepts.

Table 7
Technical category vs document clustering technique (matrix elements in percentages)

Taxonomy	Biomedicine	Physics/mathematics	Chemistry/material sciences	Environmental sciences
GST	30.4	32.8	23	13.8
CLUTO	28	34	19.8	18.3
DATACOMP A	32	32	20	18
DATACOMP B	37	27	25	11
DATACOMP C	38	27	24	11
Journals	41	34	16	9
Manual	38.6	32.7	17	11.1

The document clustering approaches provided reasonable agreement among each other and the benchmark manual abstract clustering. All the document clustering approaches need improvement in handling multi-theme documents and eliminating low technical content words/phrases.

For multi-theme documents some types of fuzzy clustering [51] will be required, where a document can be allocated fractionally to different clusters. The CLUTO partitional clustering algorithm is presently being upgraded to incorporate fuzzy clustering. Elimination of low technical content words/phrases can be done manually and/or statistically. The manual approach involves creation of larger stop word lists. This is a laborious process, and has an intrinsic deficiency. The judgment of whether a word/phrase has high or low technical content is context-dependent, and accurate word/phrase characterizations require context-dependency as part of the selection algorithm. Various statistical approaches have been proposed for context-dependent stop word selection [52, 53]. In the present study, none of the document clustering techniques used a statistical approach for stop word removal, but the multi-link word/phrase clustering approach used a unique quasi-statistical approach [54]. Improved elimination of low technical content words/phrases is mandatory for clustering accuracy gains.

Finally, another clustering accuracy limitation which all the concept clustering and most of the document clustering approaches did not address was the treatment of related concepts that used different terminology. Most of the clustering approaches examined here used text matching for generating cluster similarity. To overcome this limitation, some types of thesaurus need to be employed to standardize terminology and/or some form of latent semantic approach is required.

Greedy string tiling was developed, and is an excellent tool, for detecting plagiarism based on similarity of long text sections. Much of its powerful capability goes unused in the present document clustering application, since it would be rare for non-plagiarized text to contain identical long text strings, and the algorithm operationally ends up comparing word or short phrase similarities. Running times are very long for the clustering application.

The network mapping approaches appear to have strength in determining technical thrust relationships, and offer a complementary perspective to the phrase/document clustering approaches.

The clustering appears useful for generating the structure of a country's S&T. Continual upgrades in the clustering algorithms insure that the accuracy of the clusters and categories will continue to improve.

Acknowledgements

The component of work on this paper conducted in Mexico was partially supported by CONACyT-FOMIX 9250. (The views in this paper are solely those of the authors, and do not necessarily represent the views of the Department of the Navy or any of its components, the UNAM, Booz-Allen Hamilton, DDL-OMNI, the University of Queensland, the University of Amsterdam, the University of Karlsruhe, or the University of Minnesota.)

References

[1] R.N. Kostoff, Text mining for Global Technology Watch. In: M. Drake (ed.), *Encyclopedia of Library and Information Science, Vol. 4*, Second Edition (Marcel Dekker, New York, 2003), 2789–2799.

[2] C.W. Bostian, W.T. Brandon, A.U. MacRae, C.E. Mahle and S.A. Townes, Key technology trends – satellite systems, *Space Communications* 16(2–3) (2000) 97–124.

[3] B. Leneman, Automation in Soviet industry, 1970–83 – an assessment of the present state of robot-technology, *Revue d'Etudes Comparatives Est-Ouest* 15(1) (1984) 75–112.

[4] P. Stares, United States and Soviet military space programs – a comparative-assessment, *Daedalus* 114(2) (1985) 127–145.

[5] R.C.W. Hutubessy, P. Hanvoravongchai and T.T.T. Edejer, Diffusion and utilization of magnetic resonance imaging in Asia, *International Journal of Technology Assessment in Health Care* 18(3) (2002) 690–704.

[6] B. Mooney and R. Seymour, WTEC panels survey Russian maritime technologies, *Marine Technology Society Journal* 30(1) (1996) 71–2.
[7] L.V. McIntire, WTEC panel report on tissue engineering (reprinted), *Tissue Engineering* 9(1) (2003) 3–7.
[8] R. Campbell, H.D. Balzer, J. Berliner, R. Dobson and P. Gregory, *Soviet Science and Technology* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1985).
[9] A. Klinger (ed.), *Soviet Image Pattern Recognition Research* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1990).
[10] R.M. Gray (ed.), M. Cohn, L.W. Craver, A. Gersho, T. Lookabaugh, F. Pollara and M. Vetterli, *Non-US Data Compression and Coding Research; a Foreign Applied Sciences Assessment Center (FASAC) report prepared for Science Applications International Corporation (SAIC) under U.S. Government sponsorship* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1993).
[11] L.J. Lanzerotti, R.C. Henry, H.P. Klein, H. Masursky, G.A. Paulikas, F.L. Scarf, G.A. Soffen and Y. Terzian, *Soviet Space Science Research, FASAC Technical Assessment Report FASAC-TAR-3060* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1986).
[12] L.M. Duncan, F.T. Djuth, J.A. Fejer, N.C. Gerson, T. Hagfors, D.B. Newman Jr, R.L. Showen, *Soviet Ionospheric Modification Research*, *Foreign Applied Sciences Assessment Center Technical Assessment Report 4040* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1988).
[13] W.J. Spencer, J.Y. Chen, A. Chiang, W. Frieman, E.S. Kuh, J.L. Moll, R.F. Pease and K.C. Saraswat, *Chinese microelectronics, Foreign Applied Sciences Assessment Center Technical Assessment Report* (SAIC/Foreign Applied Sciences Assessment Center, San Diego, 1989).
[14] R.C. Davidson, M.A. Abdou, L.A. Berry, C.W. Horton, J.F. Lyon, and P.H. Rutherford, *Japanese Magnetic Confinement Fusion Research*, *Foreign Applied Sciences Assessment Center Technical Assessment Report* (Science Applications International, 1990).
[15] R.N. Kostoff, H.J. Eberhart, D.R. Toothman and R. Pellenbarg, Database tomography for technical intelligence: comparative analysis of the research impact assessment literature and the Journal of the American Chemical Society, *Scientometrics* 40(1) (1997) 103–138.
[16] R.N. Kostoff, H.J. Eberhart and D.R. Toothman, Database tomography for technical intelligence: a roadmap of the near-earth space science and technology literature, *Information Processing and Management* 34(1) (1998) 69–85.
[17] R.N. Kostoff, H.J. Eberhart and D.R. Toothman, Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography, *Journal of the American Society for Information Science* 50(5) (1999) 427–447.
[18] R.N. Kostoff, T. Braun, A. Schubert, D.R. Toothman and J.A. Humenik, Fullerene roadmaps using bibliometrics and database tomography, *Journal of Chemical Information and Computer Science* 40(1) (2000) 19–39.
[19] R.N. Kostoff, K.A. Green, D.R. Toothman and J.A. Humenik, Database tomography applied to an aircraft science and technology investment strategy, *Journal of Aircraft* 37(4) (2000) 727–730.
[20] R.N. Kostoff and R.A. DeMarco, Science and technology text mining, *Analytical Chemistry* 73(13) (2001) 370–378A.
[21] R.N. Kostoff, J.A. Del Rio, E.O. García, A.M. Ramírez and J.A. Humenik, Citation mining: integrating text mining and bibliometrics for research user profiling, *Journal of the American Society for Information Science and Technology* 52(13) (2001) 1148–1156.
[22] R.N. Kostoff, R. Tshiteya, K.M. Pfeil and J.A. Humenik, Electrochemical power source roadmaps using bibliometrics and database tomography, *Journal of Power Sources* 110(1) (2002) 163–176.
[23] R.N. Kostoff, M.F. Shlesinger and G. Malpohl, Fractals roadmaps using bibliometrics and database tomography, *Fractals* 12(1) (2004) 1–16.
[24] R.N. Kostoff, M.F. Shlesinger and R. Tshiteya, Nonlinear dynamics roadmaps using bibliometrics and database tomography, *International Journal of Bifurcation and Chaos* 14(1) (2004) 61–92.
[25] R.N. Kostoff, C.W. Bedford, J.A. Del Rio, H. Cortes and G. Karypis, Macromolecule mass spectrometry: citation mining of user documents, *Journal of the American Society for Mass Spectrometry* 15(3) (2004) 281–287.
[26] R.N. Kostoff, G. Karpouzian and G. Malpohl, Text mining the global abrupt wing stall literature, *Journal of Aircraft* 42(3) (2005) 661–4.
[27] R.N. Kostoff, J.A. Del Rio, H.D. Cortes, C. Smith, A. Smith, C.S. Wagner, L. Leydesdorff, G. Karypis, G. Malpohl and R. Tshiteya, *Science and Technology Text Mining: Mexico Core Competencies. DTIC ADA Number 430724* (Defense Technical Information Center, Fort Belvoir, 2005).
[28] R.N. Kostoff, The practice and malpractice of stemming, *Journal of the American Society for Information Science and Technology* 54(10) (2003) 984–985.
[29] R.N. Kostoff and J.A. Block, Factor matrix text filtering and clustering, *Journal of the American Society for Information Science and Technology* 56(9) (2005) 946–968.

[30] D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections. In: N.J. Belkin et al. (eds), *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)* (ACM, New York, 1992) 318–29.
[31] S. Guha, R. Rastogi and K. Shim, CURE: an efficient clustering algorithm for large databases. In: A. Tiwary and M. Franklin (eds), *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data (SIGMOD'98)* (ACM, Seattle, 1998) 73–84.
[32] M.A. Hearst, The use of categories and clusters in information access interfaces. In: T. Strzalkowski (ed.), *Natural Language Information Retrieval* (Kluwer, Dordrecht, 2000).
[33] G. Karypis, E.H. Han and V. Kumar, Chameleon: a hierarchical clustering algorithm using dynamic modeling, *IEEE Computer* 32(8) (1999) 68–75. [Special Issue on Data Analysis and Mining]
[34] E. Rasmussen, Clustering algorithms. In: W. B. Frakes and R. Baeza-Yates (eds), *Information Retrieval Data Structures and Algorithms* (Prentice Hall, Englewood Cliffs, 1992).
[35] M. Steinbach, G. Karypis and V. Kumar, *A Comparison of Document Clustering Techniques. Technical Report #00–034* (Department of Computer Science and Engineering, University of Minnesota, 2000).
[36] P. Willet, Recent trends in hierarchical document clustering: a critical review, *Information Processing and Management* 24 (1988) 577–597.
[37] O. Zamir and O. Etzioni, Web document clustering: a feasibility demonstration. In: H.P. Frei et al. (eds), *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)* (ACM, Zurich, 1998) 46–54.
[38] L. Prechelt, G. Malpohl and M. Philippsen, Finding plagiarisms among a set of programs with JPlag, *Journal of Universal Computer Science* 8(11) (2002) 1016–1038.
[39] M.J. Wise, Neweyes: a system for comparing biological sequences using the running Karp-Rabin Greedy String-Tiling algorithm. In: C. Rawlings et al. (eds), *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, Cambridge, UK 1995* (AAAI, Menlo Park, 1995) 93–401.
[40] D. Benedetto, E. Caglioti and V. Loreto, Language trees and zipping, *Physical Review Letters* 88(4) (2002) Art. No. 048702.
[41] G. Karypis, *CLUTO – a clustering toolkit* (2005). Available at: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download (accessed 5 September 2006).
[42] A.E. Smith and M.S. Humphreys, Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping, *Behavior Research Methods* (in press).
[43] L. Leydesdorff, Words and co-words as indicators of intellectual organization, *Research Policy* 18(4) (1989) 209–223.
[44] P. Ahlgren, B. Jarneving and R. Rousseau, Requirement for a cocitation similarity measure, with special reference to Pearson's Correlation Coefficient, *Journal of the American Society for Information Science and Technology* 54(6) (2003) 550–560.
[45] C.S. Wagner and L. Leydesdorff, Mapping global science using international co-authorships: a comparison of 1990 and 2000, *International Journal of Technology and Globalization* (in press).
[46] J.L. Ortega Priego, A vector space model as a methodological approach to the triple helix dimensionality: a comparative study of biology and biomedicine centres of two European National Councils from a Webometric view, *Scientometrics,* 58(2) (2003) 429–443.
[47] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, Auckland, 1983).
[48] H.D. White, Author cocitation analysis and Pearson's *r, Journal of the American Society for Information Science and Technology* 54(13) (2003) 1250–1259.
[49] V. Batagelj and A. Mrvar, *Pajek – Program for Large Network Analysis* (). Available at: http://vlado.fmf.uni-lj.si/pub/networks/pajek/ (accessed 14 December 2005).
[50] C.S. Wagner and S. Popper, *Technology Use and Productivity in Mexico, Final Report, RAND Europe, 2002* (Unpublished monograph).
[51] M.P. Windham, cluster validity for fuzzy clustering algorithms, *Fuzzy Sets and Systems* 5(2) (1981) 177–85.
[52] W.J. Wilbur and K. Sirotkin, The automatic identification of stop words, *Journal of Information Science* 18(1) (1992) 45–55.
[53] A. Bookstein, S.T. Klein and T. Raita, Clumping properties of content-bearing words, *Journal of the American Society for Information Science* 49(2) (1998) 102–114.
[54] R.N. Kostoff and J.A. Block, Factor matrix text filtering and clustering, *Journal of the American Society for Information Science and Technology* 56(9) (2005) 946–968.