Available online at www.sciencedirect.com

SCIENCE DIRECT®

Technological Forecasting & Social Change 72 (2005) 798–814

Technological Forecasting and Social Change

The structure and infrastructure of Mexico's science and technology☆

Ronald N. Kostoff[a,*], J. Antonio del Río[b], Héctor D. Cortés[b], Charles Smith[c], Andrew Smith[d], Caroline Wagner[e], Loet Leydesdorff[e], George Karypis[f], Guido Malpohl[g], Rene Tshiteya[h]

[a]*Office of Naval Research, 800 N. Quincy St., Arlington, VA 22217, USA*
[b]*Centro de Investigación en Energía, UNAM, Temixco, Mor. México*
[c]*Booz-Allen Hamilton, Bethesda, MD, USA*
[d]*University of Queensland, Brisbane, Australia*
[e]*University of Amsterdam, Amsterdam, the Netherlands*
[f]*University of Minnesota, Minneapolis, MN 55455, USA*
[g]*University of Karlsruhe, Postfach 6980, 76128 Karlsruhe, Germany*
[h]*DDL-OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600, Mclean, VA 22102, USA*

Received 11 January 2005; received in revised form 14 February 2005; accepted 17 February 2005

Abstract

The structure and infrastructure of the Mexican technical literature was determined. A representative database of technical articles was extracted from the Science Citation Index for the year 2002, with each article 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). One of the pervasive technical topics identified from the clustering, thin

☆ 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.

* Corresponding author. Tel.: +1 703 696 4198; fax: +1 703 696 3098.
E-mail addresses: kostofr@onr.navy.mil (R.N. Kostoff), malpohl@ipd.uka.de (G. Malpohl).

0040-1625/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.techfore.2005.02.001

films research, was analyzed further using bibliometrics, in order to identify the infrastructure of this technology.
Published by Elsevier Inc.

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

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:

a) priority technical areas for joint commercial or military ventures,
b) assessment of a country's military potential,
c) knowledge of emerging areas to avoid commercial or military surprise.

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

a) Much science and technology performed is not documented.
b) Much documented science and technology is not widely available.
c) Much available documented science and technology is expensive and difficult to acquire.
d) 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.

For the past half century, driven mainly by the Cold War, a large number of country technology assessments were performed [2–14]. The last decade has 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: World Technology Evaluation Center (WTEC—Loyola Univ) and 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. Mexican science and technology structure

Mexico is an important country with which the current USA President and Administration want to strengthen relationships and build a stronger partnership. In addition, there is a long common border, with common security concerns. To improve awareness of Mexico's S&T program, Mexico was selected as the prototype for a country core competency assessment. However, due to lack of space, only a brief overview of Mexican S&T is presented in this paper. A much more detailed description is contained in the unabridged DTIC report on this study [54].

The Federal Mexican S&T expenditures (FSTE) have been almost constant during the last decade, oscillating around 0.4% of the Gross Domestic Product (GDP). In terms of the Discretionary Federal Budget (DFB), the FSTE ratio has been of the order of 2.5%. This is the lowest FSTE in the thirty member nations of the Organization for Economic Cooperation and Development (OCDE).

The Government and the Universities are allocated most of the Gross Domestic Expenditures on Research and Development (GDERD), mainly in the natural sciences and engineering areas. The evolution of the GDERD invested in basic, applied and experimental development illustrates that government expenditures in basic and applied research are similar, while business expenditures are larger in applied research. Total expenditures are concentrated in government and education institutions.

1.4. Mexican researcher fellowships

About 20 years ago, the Mexican government created a researchers fellowship (Sistema Nacional de Investigadores—SNI). In this system, the government recognizes the research activity of people in higher education, government institutions, private sector and non-profit organizations. Selection of a fellow is made by a peer review commission. There are two main levels in this fellowship. The lower level is candidato, addressed to young people starting a researcher career. There are other levels for established researchers (investigador nacional).

In Mexico, there is less than one researcher per ten thousand habitants (one hundred million is the population in Mexico), with a low rate of young researchers. In 2002, there were 9200 SNI members, distributed as follows: Physics, Mathematics, and Earth Sciences (1771); Biology and Chemistry (1661); Medicine and Health Sciences (927); Humanity and Social Sciences (1552); Social Sciences (1096); Biotechnology and Agriculture (1011); and Engineering (1182).

2. Objectives

An objective is to identify the S&T core competencies of Mexico. Further, another objective is to generate a process that could be used efficiently and rapidly to assess the S&T core competencies in other countries of interest.

3. Approach and results

3.1. 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, 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 both types of information, using multiple approaches for identifying the thrusts and their relationships.

Two types of results are presented, bibliometrics and taxonomies. Bibliometrics provide an indication of the technical infrastructure (prolific authors, journals, institutions, citations), while taxonomies provide an indication of major technology thrusts and their relationships.

Section 3.2 describes the database used for the bibliometrics and taxonomy analyses. Section 3.3 presents the bibliometrics approaches and results, where Section 3.3.1 presents the publication bibliometrics, and Section 3.3.2 presents the citation bibliometrics. Section 3.4 presents the taxonomy approaches and selected results, where Section 3.4.1 presents the manual taxonomy approaches and results, and Section 3.4.2 presents the statistical taxonomy approaches and results.

There are five manual taxonomy results presented, and two major classes of statistical taxonomy approaches presented (concept clustering, document clustering). Concept clustering includes factor matrix-based clustering and multi-link hierarchical aggregation clustering. Document clustering includes greedy string tiling, entropy-based data compression, partitional, journal, and latent semantic. Due to

space limitations, most of the clustering approaches are summarized very briefly. The reader interested in detailed descriptions of techniques and results should obtain Ref. [54].

3.2. Databases and information retrieval approach

For the present study, the science citation index database was used. The retrieved database used for analysis consists 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. 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, engineering, and life sciences basic research).

3.3. Bibliometrics

A total of 4529 records were retrieved, and the bibliometrics data were extracted.

3.3.1. Publication statistics on authors, journals, and organizations

The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality, although there is some threshold quality level inferred, since these papers are published in the (typically) high caliber journals accessed by the SCI.

In all previous text mining studies published by the first author's group, bibliometrics were performed on the overall database retrieved. Since all these previous studies focused on a specific technology, the resultant bibliometrics provided the technical infrastructure for that technology. In the present case, the focus is on the wide range of technologies being developed within a country. Applying the bibliometrics analysis to the total retrieved database for that country will not provide very useful results. Visitation strategies (one desired application) are typically developed for a specific technology using a group of experts for that technology.

The approach taken here is to identify the thematic thrust areas for the clustering performed in the latter part of this report, then retrieve documents that address each theme. The bibliometrics will then be performed on a theme by theme basis. For the present study, one theme is selected as an illustrative example for the bibliometrics.

Based on the computational linguistics (clustering) results, thin films is an important thrust area of Mexican research. A query for thin film research in Mexico was inserted into the science citation index, and 1727 records were recovered for the period 1991–2003, of which 1693 had abstracts. The bibliometrics analysis was performed on these 1727 records.

3.3.1.1. Prolific authors. Table 1 lists the twenty most prolific authors in Mexican thin film research, including their institutions. Two institutions predominate: UNAM and CINVESTAV, IPN. As Ref. [54] shows, the institution with the most scientists in Mexico is UNAM, followed by CINVESTAV. This institution Center was part of IPN some years ago. The area of thin films obeys a similar feature; these two institutions (UNAM and CINVESTAV) do most of the research in this area.

Table 1
Most prolific Mexican thin film authors

Author name	Institution	# Papers
Zelaya-Angel O	IPN	79
Nair PK	UNAM	78
Falcony C	CINVESTAV	71
Sebastian PJ	UNAM	70
Gonzalez-Hernandez J	CINVESTAV, IPN	66
Nair MTS	UNAM	55
Ramirez-Bon R	UNIV SONORA	47
Pena JL	CTR INVEST CIENTIFICA	43
Ortiz A	UNAM	42
Castro-Rodriguez R	CINVESTAV, IPN	40
Jergel M	CINVESTAV, IPN	37
Contreras-Puente G	CINVESTAV, IPN	37
Asomoza R	CINVESTAV, IPN	36
Jimenez-Sandoval S	CINVESTAV, IPN	35
Espinoza-Beltran FJ	CINVESTAV, IPN	35
Andrade E	UNAM	34
Alonso JC	UNAM	32
Haro-Poniatowski E	UAM-I	31

3.3.1.2. Prolific journals. Table 2 lists the fifteen most prolific thin film journals containing Mexican research papers. They appear to be top-quality journals, concentrated in physics and materials, with some emphasis on chemistry as well. All but one (Revista Mexicana de Fisica) are English language journals, and Revista Mexicana de Fisica is one of the most relevant peer reviewed physics journals in Latin America [27]. It publishes papers in both English and Spanish.

Table 2
Most prolific journals—Mexican thin film research

Journal	# Papers
Thin Solid Films	147
Revista Mexicana de Fisica	80
Journal of Applied Physics	66
Solar Energy Materials and Solar Cells	62
Physical Review B	48
Applied Surface Science	46
Applied Physics Letters	36
Semiconductor Science and Technology	35
Journal of Vacuum Science and Technology A-Vacuum Surfaces and Films	34
Modern Physics Letters B	34
Solid State Communications	33
Journal of the Electrochemical Society	32
Materials Letters	27
Journal of Physics and Chemistry of Solids	25
Journal of Physics D-Applied Physics	23

3.3.1.3. Prolific institutions and countries. This section identifies the most prolific institutions producing Mexican-authored thin film papers, and the countries of the most prolific collaborators with Mexican authors of thin film papers.

Table 3A contains a list of the fifteen most prolific institutions for Mexican-authored thin film papers, and Table 3B contains a list of the eighteen most prolific countries associated with Mexican-authored thin film papers Two institutions seem to predominate (as found in the case of most prolific authors): UNAM and IPN, as do four countries (USA, Cuba, France, Spain).

As in the case of the affiliation of the most prolific authors, UNAM and CINVESTAV dominate as most prolific institutions, however, some other State Universities (Puebla, Sonora, San Luis Potosi) and some CONACyT Research Centers (CICESE, CIO) seem to have a role in this topic. One important feature of the institution analysis is that there is no industry involvement. On the other hand, it is noteworthy that non-Mexican institutions in this table are mainly from developing countries in collaboration with Mexican thin film groups. This confirms that most of the research on thin solid films in Mexico is dedicated to low cost technology, as it was found in Ref. [28].

The country collaborations were investigated further. To ascertain the impact resulting from these collaborations, the citations from different inter-country collaboration sub-sets were determined. The thin film papers that were published in 1998, and had the following country combinations in their address field (Mexico–USA; Mexico–Cuba; Mexico–France; Mexico–Spain), were examined for citations. The average and median citations are listed in Table 3B, in the two right-most columns, next to the respective collaborating countries.

The USA collaborations produced the most citations. While the median was similar to most of the other countries listed as collaborators, the average was substantially higher. Three of the twenty papers had over twenty cites, while France had only one paper over ten cites, Spain had one paper at ten cites, and Cuba's best paper had five cites. While there were modest differences in the citation distributions among the countries, the real difference was the number of highly cited papers.

Table 3A
Most prolific institutions—Mexican thin film research

Institution	# Papers
IPN, CINVESTAV	828
Univ NACL Autonoma Mexico	800
Univ Autonoma Metropolitana Iztapalapa	110
Univ Autonoma Puebla	102
Univ la Habana	94
Univ Sonora	68
Univ Autonoma San Luis Potosi	65
Inst Nacl Invest Nucl	53
Ctr Invest Opt	45
CNRS	42
Ctr Invest Cient and Educ Super Ensenada	40
CSIC	34
Inst Mexicano Petr	28
Ctr Invest Quim Aplicada	24
BUAP	23

3.3.2. Citation statistics on authors, papers, and journals

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics [29], much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers [30,31].

The citations in all the retrieved SCI papers were aggregated. The authors, specific papers, and journals cited most frequently were identified, and were presented in order of decreasing frequency. It should be emphasized that these citations are from papers in the retrieved database only. Total citations from all papers in the SCI could be substantially larger in some cases.

3.3.2.1. Most cited first authors. Table 4 lists the fifteen most cited first authors by Mexican thin film papers, unmodified for self-citations. In contrast to past text mining studies, where there was minimal overlap between most prolific authors and most cited first authors, in the present case, there are seven authors in common between the two lists (Nair, Nair, Sebastian, Falcony, Zelaya-Angel, Ortiz, Ramirez-Bon).

While there are a number of factors that could account for the disjointness between the two lists, in other text mining studies the main factor appeared to be that the most prolific authors tended to rarely be first authors (typically 10% of their SCI papers, or less). Thus, while papers to which they contributed may have received substantial citations, because of their rare appearance as first authors, they did not (on average) accumulate substantial citations as first authors.

An interesting finding occurred in the present case. Consider a highly prolific and most cited author, PK Nair. This author has 94 entries in the SCI. If these entries are listed in order of citations, in descending frequency, then the fractions of papers first authored are as follows (arranged by first ten

Table 3B
Most prolific countries—Mexican thin film research

Country	# Papers	Average Cites	Median Cites
Mexico	1727		
USA	246	8.3	3
Cuba	103	3.2	3
France	81	3.3	1
Spain	72	5	3
England	39		
Ukraine	32		
Germany	31		
Russia	30		
Japan	29		
Slovakia	26		
India	20		
Brazil	18		
Poland	13		
Canada	13		
Italy	12		
Venezuela	12		
Colombia	11		

Table 4
Most cited first authors—Mexican thin film research

Author	Times cited
Nair PK	231
Nair MTS	149
Chopra KL	105
Sebastian PJ	96
Ortiz A	67
Zelayaangel O	60
Aspnes DE	59
Brinker CJ	58
Rakhshani AE	57
Falcony C	55
Lucovsky G	52
Bhattacharya RN	52
Ramirezbon R	51
Sanchezgil JA	50
Major S	47

most cited, second ten most cited, etc): 7/10, 4/10, 3/10, 3/10, 1/10, 2/10, 1/10, 2/10, 1/10, 0/4. Overall, 26% of Nair's papers are first authored, somewhat higher than most prolific authors in other text mining studies. However, a substantial fraction of Nair's most cited papers are first authored, and a much smaller fraction of Nair's least cited papers are first authored. This selectivity is the reason for the high citations.

3.3.2.2. Most cited papers. Table 5A lists the fifteen most cited papers by Mexican thin film authors (cited by the retrieved papers in the 1727 paper database only). This is a reasonable mix of papers from the past two decades, and reflects a dynamic field of research.

Table 5A
Most cited papers—Mexican thin film research

Paper	# Cites
Britt J, 1993, Appl Phys Lett, v62, p2851	44
Kaur I, 1980, J Electrochem Soc, v127, p943	33
Zelayaangel O, 1994, Appl Phys Lett, v64, p291	31
Hodes G, 1987, Phys Rev B, v36, p4215	27
Chopra KL, 1983, Thin Solid Films, v102, p1	26
Chopra KL, 1982, Phys Thin Films, v12, p201	24
Orton JW, 1982, J Appl Phys, v53, p1602	21
Nair PK, 1991, J Phys D Appl Phys, v24, p441	20
Nair MTS, 1994, J Appl Phys, v75, p1557	19
Aranovich J, 1979, J Vac Sci Technol, v16, p994	19
Nair PK, 1998, Sol Energ Mat Sol C, v52, p313	18
Nair MTS, 1989, Semicond Sci Tech, v4, p191	18
Doolittle LR, 1986, Nucl Instrum Meth B, v15, p227	18
Manifacier JC, 1976, J Phys E Sci Instrum, v9, p1002	18
Nair PK, 1988, Semicond Sci Tech, v3, p134	17

In this case, two of the most cited papers have been published in one of the best ranked journals in applied physics (APL). Also, it seems to be that Chopra's and Nair's works play a seminal role in the research of thin films in Mexico.

Table 5B compares the most and least cited Mexican papers published in 1998, retrieved with the same thin film query. There were 157 papers retrieved. Eight papers had twenty or more citations (most cited), 27 papers had zero citations (least cited), and these two groups were compared. The most cited have over 50% more authors and references compared to the least cited. The most cited were all laboratory demonstrations, typically of novel material formation, or material growth and deposition processes. The least cited were mainly laboratory demonstrations, typically of property measurements, but 20% were theoretical studies, and 10% were long-term demonstrations, such as corrosion development. The most cited researchers were all from universities, while the least cited were mainly from universities, but 25% were from research institutes. Interestingly, none of the articles had industry representation. Of the most cited papers' first authors, five were from Mexico, two were from England, and one from the USA. Other non-first author countries represented were USA (three articles). Of the least cited papers' first authors, 23 were from Mexico, two from France, one from Wales, and one from Slovakia. Other non-first author countries represented were France (two papers), Slovakia [1], Ukraine [1], Canada [1], Columbia [1], Venezuela [1].

3.3.2.3. Most cited journals. Table 6 lists the fifteen most cited journals by Mexican thin film authors. The highest ranked journals are in applied physics, and others listed focus on materials, with some chemistry. In common with past text mining studies, there is substantial overlap (nine journals) of the most prolific journals with the most cited journals. The list of most cited journals indicates the predominance of four mainly physics journals as the source literature for Mexican research on thin films.

3.4. Taxonomies

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.

3.4.1. Manual

Five manual categorization techniques were compared: article titles, journal titles, keywords, full abstracts, journals. Table 7 compares the different manual categorization results. If manual categorization

Table 5B
Most and least cited papers—1998

	Ave # Auth	Med # Auth	Ave # Ref	Med # Ref	Ave # Cites	Med # Cites
Most cited	7	6	32	27.5	31	27
Least cited	4	4	19	17	0	0

Table 6
Most cited journals—Mexican thin film research

Journal	# Cites
Phys Rev B	2072
J Appl Phys	1723
Appl Phys Lett	1482
Thin Solid Films	1324
Phys Rev Lett	952
J Electrochem Soc	924
J Vac Sci Technol A	427
J Chem Phys	376
Sol Energ Mat Sol C	368
J Cryst Growth	353
Solid State Commun	338
Appl Optics	324
Nucl Instrum Meth B	251
Semicond Sci Tech	248
Nature	242

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.

The manual characterization contrasts with the distribution of fellows in SNI by areas, where physics and earth sciences represents less than 20% and chemistry–biology and health sciences are around 30%. This means that the other areas are under-represented in terms of Mexican papers appearing in 2002 in the international scientific literature.

3.4.2. Statistical clustering

Two generic types of statistical clustering were used, concept clustering and document clustering. In concept clustering, words or phrases are clustered based on their co-occurrence in the same text unit. In document clustering, documents are clustered based on their overall text similarity.

3.4.2.1. 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

Table 7
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

taxonomy category structure from the document clustering approach described later. A synergistic combination of factor matrix and multi-link clustering was used that offers substantial improvement in the quality of the resultant clusters by eliminating those words that are trivial operationally in the application context [32,33].

The remainder of this section summarizes briefly the multi-link clustering only. See Ref. [54] for factor matrix and multi-link clustering technique details and results.

Multi-link clustering results. A normalized symmetrical co-occurrence matrix of the highest frequency high technical content words/phrases was generated. At the top level in the cluster hierarchy, five broad topics can be discerned. These include biology, medicine, physics, chemistry, and environment. Each of these highest level clusters will be divided into smaller clusters, as follows.

1) Biology—There are four main groupings: membrane biology/cell–cell recognition; microbial molecular biology/gene expression; recombinant DNA biology; plant population genetics.
2) Medicine—There are five main groupings: cardiopulmonary; reproductive; liver damage; immunology; chronic disease treatment.
3) Physics—There are four main groupings: quantum and dynamical systems; accelerator physics; solid-state; astrophysics.
4) Chemistry—There are three main groupings: polymers; molecular characterization; thin films.
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 7, especially those determined by manual categorization of the full abstracts. In Table 7, Agriculture and Earth Sciences and Oceanography were listed as separate themes, whereas the present taxonomy lists them under Environment.

3.4.2.2. Document clustering. Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [34–41]. Five approaches were examined in this paper: Greedy String Tiling, Entropy-based Data Compression, Partitional Clustering, Automatic Journal Categorization, and Latent Semantic Clustering.

Greedy string tiling approach. The approach presented in this section is based on a Greedy String Tiling (GST) text matching algorithm [42,43]. 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.

Data compression clustering approach. The compression algorithm approach [44] 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.

Partitional clustering approach. The approach presented in this section is based on a partitional clustering algorithm [53] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to 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.

Partitional clustering results. In partitional clustering, the number of clusters desired is input, and all documents in the database are included in those clusters. Sixty-four clusters were selected, and aggregated into a hierarchical taxonomy using a hierarchical tree generated by the CLUTO software. The numbers of papers in each of the 64 elemental clusters, as well as in hierarchical aggregates of these clusters, was obtained. Thin films was a separate third-level category, and was chosen for the bibliometrics example in the previous sections of this paper.

Three final comments about these results. First, 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 Ref. [54], the recent Mexico S&T expenditures are on the order of Mexican Pesos 2.5 billion/y. If 64 clusters are used to categorize this S & T, then each cluster (on average) would cover about US$40 million/y of S&T expenditure. This reflects rather broad categories. If, however, 512 clusters were used, then the resolution increases to about US$5 million/y for the category average. This level of resolution would cover small groups of projects.

Second, the Physical, Chemical, and Material Sciences topics identified appear to address forefront areas of research, areas also addressed by other technologically sophisticated countries. Research conducted by Mexican scientists somewhat distinctive from that of other countries is concentrated in the Ecology and Biomedicine. In Ecology, fish, animal, and plant species (and other foods) indigenous to Mexico are focal points, as well as geographical and climatic phenomena. In Biomedicine, the distinctive aspects focus on health problems indigenous to Mexico, related to geography, environment, and diet.

Third, based on the partitional document clustering taxonomy, are there any large research gaps evident? As in the multi-link clustering taxonomy shown previously, most of the major research areas appear to be represented, but engineering science (other than materials engineering) does not play a prominent role at the upper taxonomy levels. As a test, a brief comparison of Mexican and USA papers in a couple of engineering topics was made. The fraction of Mexican papers that contained the word ‘aircraft’ was .00025, while the fraction of USA papers that contained the word ‘aircraft’ was .027, or two orders of magnitude difference. For the term ‘aerodynamic’, the respective fractions were .00037 and .0137, a factor of 37 difference.

However, here it is important to note that in Mexican science the area of materials science is treated as physics or chemistry. Incorporating the information (from Ref. [54]) of number of members in SNI by scientific area, it is clear that engineering science is under-represented in the Mexican scientific community, at least as represented in the SCI open research literature.

Journal clustering approach. This section utilizes the ISI classification of journals by categories, and papers are associated in accordance with the category in the ISI. This classification is not in agreement with DTIC, and does not obey criteria as DTIC.

Self-organising named concept extraction and clustering. This approach to concept extraction and clustering employs a Bayesian analysis of word co-occurrences, but one which includes nonlinear machine learning algorithms. The method passes through four stages of processing. The first stage involves the seeding of named concepts via extraction of seed terms from the text which possess particular statistical characteristics. The second stage learns a family of related terms around each seeded concept by means of an iterative optimiser 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 2-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 visualisation of the concept network. For non-interactive publication, the spatial proximity of clustered concepts and the connectedness of each concept are 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.

Network analysis of word co-occurrence. This section presents analysis of Mexico's technology capabilities using network analysis of word occurrence [45–52] to reveal patterns within the data. These patterns can provide information that would not be evident from a visual examination of the data. Ref. [54] discusses the data sources and methods, the use of network analysis and the results of the analysis.

3.4.3. Taxonomy comparisons

Three generic approaches to taxonomy construction were presented: manual clustering, statistical concept clustering, 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 Compressin, Journal Clustering) showed reasonable agreement among each other, and with the manual abstract clustering (see table below). The main differences appear to be among Biomed, 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.

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

Taxonomy	Biomed	Physmath	Chemat'ls	Environment
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 first author's very recent unpublished studies on clustering show three main sources of error for all present clustering techniques: 1) excessive trivial words that influence the clustering process; 2) use of different terminology to describe the same concept; and 3) assignment of records to one cluster only. Improved techniques for eliminating trivial words, use of thesauri to normalize terminology, and use of fuzzy clustering to assign individual records to multiple categories would increase the quality of taxonomies substantially. These clustering improvements would also reduce the spread in results of the high quality clustering and network approaches presented in this paper.

4. Summary and conclusions

The main objective of this study was to assess the technical core competencies of Mexico. This was accomplished using a variety of 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.

The Physical, Chemical, and Material Sciences topics identified appear to address forefront areas of research, areas also addressed by other technologically sophisticated countries. Research conducted by Mexican scientists somewhat distinctive from that of other countries is concentrated in the Ecology and Biomedical. In Ecology, fish, animal, and plant species (and other foods) indigenous to Mexico are focal points, as well as geographical and climatic phenomena. In Biomedical, the distinctive aspects focus on health problems indigenous to Mexico, related to geography, environment, and diet.

The Engineering Sciences appear to be under-represented, based on the open source SCI research literature. The Engineering Sciences were not visible at the higher taxonomy levels, and only started to emerge at a few of the lowest level document clusters.

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.

The document clustering approaches provided good agreement among each other and the benchmark manual abstract clustering. For more detailed technical analyses, hundreds of clusters would be required. All the document clustering approaches need improvement in handling multi-theme documents and eliminating low technical content words/phrases. These required improvements are being implemented presently.

The clustering appears useful for generating the structure of a country's S&T, while the bibliometrics appears useful for identifying centers of excellence and prolific performers for specific technology areas. 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 has been partially supported by CONACyT-FOMIX 9250.

References

[1] R.N. Kostoff, Text mining for global technology watch, in: M. Drake (Ed.), Second Edition. Encyclopedia of Library and Information Science, vol. 4, Marcel Dekker, Inc., New York, NY, 2003, pp. 2789–2799.

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

[3] B. Leneman, Automation in Soviet industry, 1970–1983—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, T.T.T. Edejer, Diffusion and utilization of magnetic resonance imaging in Asia, International Journal of Technology Assessment in Health Care 18 (3) (2002 (SUM)) 690–704.
[6] B. Mooney, R. Seymour, WTEC panels survey Russian maritime technologies, Marine Technology Society Journal 30 (1) (1996 (SPR)) 71–72.
[7] L.V. McIntire, WTEC panel report on tissue engineering, Tissue Engineering 9 (1) (2003 (FEB)) 3–7 Reprinted.
[8] Robert Campbell, H.D. Balzer, J. Berliner, R. Dobson, P. Gregory, "Soviet science and technology," Foreign Applied Sciences Assessment Center, 1985 (October 15).
[9] A. Klinger, editor, A. Klinger, et al., "Soviet Image Pattern Recognition Research," Jan. 1990, Foreign Applied Sciences Assessment Center, Science Applications International Corp., 10260 Campus Point Drive, San Diego, CA 92121, and 1710 Goodridge Drive, McLean VA 22102.
[10] Non-US Data Compression and Coding Research, R.M. Gray (Ed.), M. Cohn, L.W. Craver, A. Gersho, T. Lookabaugh, F. Pollara, M. Vetterli, A Foreign Applied Sciences Assessment Center (FASAC) report prepared for Science Applications International Corporation (SAIC) under U.S. Government sponsorship, 1993 November.
[11] L.J. Lanzerotti, R.C. Henry, H.P. Klein, H. Masursky, G.A. Paulikas, F.L. Scarf, G.A. Soffen, Y. Terzian, "Soviet space science research," FASAC Technical Assessment Report FASAC-TAR-3060, Foreign Applied Sciences Assessment Center, 1986
[12] "Soviet Ionospheric Modification Research," with L.M. Duncan, F.T. Djuth, J.A. Fejer, N.C. Gerson, t. Hagfors, D.B. Newman, Jr., R.L. Showen, Foreign Applied Sciences Assessment Center, Technical Assessment Report 4040, 1988.
[13] W.J. Spencer, J.Y. Chen, A. Chiang, W. Frieman, E.S. Kuh, J.L. Moll, R.F. Pease, K.C. Saraswat, "Chinese microelectronics," Foreign Applied Sciences Assessment Center Technical Assessment Report, Science Applications International Corporation, April 1989.
[14] R.C. Davidson, M.A. Abdou, L.A. Berry, C.W. Horton, J.F. Lyon, P.H. Rutherford, Japanese magnetic confinement fusion research, Foreign Applied Sciences Assessment Center Technical Assessment Report, Science Applications International Corporation, 1990.
[15] R.N. Kostoff, 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, D.R. Toothman, Database tomography for technical intelligence: a roadmap of the near-earth space science and technology literature, Information Processing & Management 34 (1) (1998) 69–85.
[17] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography, Journal of the American Society for Information Science 50 (5) (1999 (April 15)) 427–447.
[18] R.N. Kostoff, T. Braun, A. Schubert, D.R. Toothman, J. Humenik, Fullerene roadmaps using bibliometrics and database tomography, Journal of Chemical Information and Computer Science 40 (1) (2000 (Jan–Feb)) 19–39.
[19] R.N. Kostoff, K.A. Green, D.R. Toothman, J. Humenik, Database tomography applied to an aircraft science and technology investment strategy, Journal of Aircraft 37 (4) (2000 (July–August)) 727–730.
[20] R.N. Kostoff, R.A. DeMarco, Science and technology text mining, Analytical Chemistry 73 (13) (2001 (July 1)) 370A–378A.
[21] R.N. Kostoff, J.A. Del Rio, E.O. García, A.M. Ramírez, J.A. Humenik, Citation mining: integrating text mining and bibliometrics for research user profiling, JASIST 52 (13) (2001 (November)) 1148–1156.
[22] R.N. Kostoff, R. Tshiteya, K.M. Pfeil, 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. Shlesinger, G. Malpohl, Fractals roadmaps using bibliometrics and database tomography, Fractals 12 (1) (2004 (March)) 1–16.
[24] R.N. Kostoff, M. Shlesinger, R. Tshiteya, Nonlinear dynamics roadmaps using bibliometrics and database tomography, International Journal of Bifurcation and Chaos in Applied Sciences and Engineering 14 (1) (2004 (January)) 61–92.
[25] R.N. Kostoff, C.W. Bedford, J.A. Del Rio, H. Cortes, G. Karypis, Macromolecule mass spectrometry: citation mining of user documents, Journal of the American Society for Mass Spectrometry 15 (3) (2004 (March)) 281–287.
[26] R.N. Kostoff, Bilateral asymmetry prediction, Medical Hypotheses 61 (2) (2003 (August)) 265–266.
[27] E.O. García, J.A. del Río, A.M. Ramírez, Analisis de la evaluacion de las revistas latinoamericanas a traves del factor de impacto renormalizado, Revista Española de Documentación Científica 25 (2002) 467–476.

[28] J.A del Río, R.N. Kostoff, E.O. García, A.M. y Ramírez, J.A. Humenik, Phenomenological approach to profile impact of scientific research: citation mining, Advances in Complex Systems 5 (2002) 19–42.
[29] E. Garfield, History of citation indexes for chemistry—a brief review, JCICS 25 (3) (1985) 170–174.
[30] R.N. Kostoff, The use and misuse of citation analysis in research evaluation, Scientometrics 43 (1) (1998) 27–43.
[31] M. MacRoberts, B. MacRoberts, Problems of citation analysis, Scientometrics 36 (3) (1996) 435–444.
[32] R.N. Kostoff, The practice and malpractice of stemming, JASIST 54 (10) (2003 (June)).
[33] R.N. Kostoff, J.A. Block, "Factor matrix text filtering and clustering" JASIST, 56 (in Press).
[34] D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/Gather: a cluster-based approach to browsing large document collections, Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92), 1992, pp. 318–329.
[35] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clustering algorithm for large databases, Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data (SIGMOD '98), 1998, pp. 73–84.
[36] M.A. Hearst, The use of categories and clusters in information access interfaces, in: T. Strzalkowski (Ed.), Natural Language Information Retrieval, Kluwer Academic Publishers, 2000.
[37] G. Karypis, E.H. Han, V. Kumar, Chameleon: a hierarchical clustering algorithm using dynamic modeling, IEEE Computer: Special Issue on Data Analysis and Mining 32 (8) (1999) 68–75.
[38] E. Rasmussen, Clustering algorithms, in: W.B. Frakes, R. Baeza-Yates (Eds.), Information Retrieval Data Structures and Algorithms, Prentice Hall, N.J., 1992.
[39] M. Steinbach, G. Karypis, V. Kumar. A comparison of document clustering techniques. Technical Report #00–034. 2000. Department of Computer Science and Engineering. University of Minnesota.
[40] P. Willet, Recent trends in hierarchical document clustering: a critical review, Information Processing & Management 24 (1988) 577–597.
[41] O. Zamir, O. Etzioni, Web document clustering: a feasibility demonstration, Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98), 1998, pp. 46–54.
[42] L. Prechelt, G. Malpohl, M. Philippsen, Finding plagiarisms among a set of programs with JPlag, Journal of Universal Computer Science 8 (11) (2002) 1016–1038.
[43] M.J. Wise. String similarity via greedy string tiling and running Karb–Rabin matching. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, 1992. Dept. of CS, University of Sidney.
[44] D. Benedetto, E. Caglioti, V. Loreto, Language trees and zipping, Physical Review Letters 88 (4) (2002 (JAN 28)) Art. No. 048702.
[45] L. Leydesdorff, Words and co-words as indicators of intellectual organization, Research Policy 18 (1989) 209–223.
[46] P. Ahlgren, B. Jarneving, 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.
[47] C.S. Wagner, L. Leydesdorff, Mapping global science using international co-authorships: a comparison of 1990 and 2000, International Journal of Technology and Globalization (2004) in press, in print.
[48] 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.
[49] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Auckland, 1983.
[50] 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.
[51] L. Leydesdorff, Meaning and translation at the interfaces of science: mapping the case of 'stem-cell research', Paper Presented at the Annual Meeting of the Society for the Social Studies of Science 4S, Atlanta, October 20032003, http://www.leydesdorff.net/stemcell.
[52] C.S. Wagner, and S. Popper, (2002). Technology use and productivity in Mexico, RAND Europe, Final Report.
[53] G. Karypis, CLUTO—a clustering toolkit. http://www.cs.umn.edu/~cluto.
[54] R.N. Kostoff, J.A. Del Rio, H.D. Cortes, C. Smith, A. Smith, C. Wagner, L. Leydesdorff, G. Karypis, G. Malpohl, R. Tshiteya. Science and technology text mining: Mexico core competencies. DTIC Technical Report number ADA 430724 (http://www.dtic.mil/).