SCIENCE AND TECHNOLOGY TEXT MINING: PERVASIVE RESEARCH THRUSTS IN THE FORMER SOVIET UNION (FSU)

By

Dr. Ronald N. Kostoff
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217
Phone: 703-696-4198
Fax: 703-696-4274
Internet: kostofr@onr.navy.mil

(The views expressed in this paper are solely those of the author and do not necessarily represent the views of the Department of the Navy or the Office of Naval Research

ABSTRACT

A revolutionary approach to identifying pervasive thrust areas (and their relationships) in large textual databases was applied to a compendium of assessments of applied research areas in the Former Soviet Union (FSU) generated by the Foreign Applied Sciences Assessment Center (FASAC). Related thrust areas were combined to yield the following major thrust groupings:IONOSPHERIC HEATING/ MODIFICATION; IMAGE/ OPTICAL PROCESSING; AIR-SEA INTERFACE; LOW OBSERVABLE; EXPLOSIVE COMBUSTION; PARTICLE BEAMS; AUTOMATIC/ REMOTE CONTROL; FREQUENCY STANDARDS; RADAR CROSS SECTION. These represent a subset of FSU science of military and strategic interest to the United States.

KEYWORDS: textual database; Former Soviet Union; FASAC; taxonomy; research thrusts; text mining; co-word analysis; computational linguistics; database tomography; clustering; multi-word frequency analysis; equivalence index; inclusion index; ionospheric heating; ionospheric modification; image processing; optical processing; air-sea interface; low observable; explosive combustion; particle beams; automatic control; remote control; frequency standards; radar cross section.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY) 17-07-2003	2. REPORT TYPE Technical	3. DATES COVERED (FROM - TO) xx-xx-1995 to xx-xx-1995

4. TITLE AND SUBTITLE
SCIENCE AND TECHNOLOGY TEXT MINING
PERVASIVE RESEARCH THRUSTS IN THE FORMER SOVIET UNION (FSU)
Unclassified

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kostoff, Ronald N ;

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME AND ADDRESS
Office of Naval Research
800 N. Quincy St.
Arlington, VA22217

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES

14. ABSTRACT

A revolutionary approach to identifying pervasive thrust areas (and their relationships) in large textual databases was applied to a compendium of assessments of applied research areas in the Former Soviet Union (FSU) generated by the Foreign Applied Sciences Assessment Center (FASAC). Related thrust areas were combined to yield the following major thrust groupings: IONOSPHERIC HEATING/ MODIFICATION; IMAGE/ OPTICAL PROCESSING; AIR-SEA INTERFACE; LOW OBSERVABLE; EXPLOSIVE COMBUSTION; PARTICLE BEAMS; AUTOMATIC/ REMOTE CONTROL; FREQUENCY STANDARDS; RADAR CROSS SECTION. These represent a subset of FSU science of military and strategic interest to the United States.

15. SUBJECT TERMS

textual database; Former Soviet Union; FASAC; taxonomy; research thrusts; text mining; co-word analysis; computational linguistics; database tomography; clustering; multi-word frequency analysis; equivalence index; inclusion index; ionospheric heating; ionospheric modification; image processing; optical processing; air-sea interface; low observable; explosive combustion; particle beams; automatic control; remote control; frequency standards; radar cross section.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 25	19. NAME OF RESPONSIBLE PERSON Kostoff, Ronald kostofr@onr.navy.mil
a. REPORT Unclassified	b. ABSTRACT Unclassified	c. THIS PAGE Unclassified			19b. TELEPHONE NUMBER International Area Code Area Code Telephone Number 703696-4198 DSN -

INTRODUCTION

This report describes a revolutionary approach for identifying pervasive thrust areas of FSU applied science, the connectivity among these areas, and sub-thrust areas closely related to and supportive of the pervasive thrust areas. The approach utilizes a computer-based algorithm to extract and order data from a large body of textual material that describes a broad spectrum of FSU science (35 reports generated by the Foreign Applied Sciences Assessment Center). The algorithm extracts words and word phrases that are repeated throughout this large database, and allows the user to create a taxonomy of pervasive research thrusts from this extracted data. The algorithm then extracts words and phrases that occur physically close to the pervasive research thrusts throughout the text, and allows the user to determine interconnectivity among the research thrusts, as well as determine research sub-thrusts strongly related to the pervasive thrusts. While the focus of the present study was identifying technical thrusts and their interrelationships, the raw data obtained by the extraction algorithms allows the user to relate technical thrusts to institutions, journals, people, geographical locations, and other categories. The methodology can be applied to any text database, consisting of published papers, reports, memos, etc., that can be placed on computer storage media.

BACKGROUND

Foreign Applied Sciences Assessment Center

In the 1980s, the U. S. Federal Government established the Foreign Applied Sciences Assessment Center under the operation of Science Applications International Corporation (SAIC). The purpose of FASAC was to increase awareness of new foreign technologies with military, economic, or political importance. The emphasis was placed on 'exploratory research' (Department of Defense 6.1/ 6.2 equivalent) in the FSU, work that seeks to translate developments in fundamental research into new technology.

One of the main products of FASAC has been reports on different areas of 'exploratory research'. FASAC would assemble panels of expert consultants from academia, industry, and government, typically six to eight members per panel. Each panel would assess the status and potential impacts of foreign applied science in selected areas, and write a report of its findings. Periodically, an Integration Report would be generated that described the trends in foreign research including pervasive issues which affect research capabilities. By early 1992, there were about 40 reports on different aspects of FSU applied science.

Co-word Analysis

Co-word analysis utilizes the proximity of words and their frequency of co-occurrence in some domain (sentence, paragraph, paper, etc.) to estimate the strength of their relationship. When applied to the literature in a technical field, co-word analysis allows a map of the relationship among technical themes to be constructed. A history of co-word analysis applied to research policy issues, its origins in computational linguistics, and its limitations due to previous dependence on the use of key words and index words, can be found in recent references (Kostoff, 1991;Kostoff, 1992a; Kostoff, 1992b).

In 1991-1993, a full text word association technique (database tomography, a variant of co-word analysis) was developed to allow rapid scanning of large text databases (Kostoff, 1991; Kostoff, 1992a; Kostoff, 1992b). The initial purpose of this development was to identify pervasive research thrusts (thrusts that transcend disciplines) from those large text databases that contain descriptions of many research programs or areas of research. Two applications were reported by 1992 (Kostoff, 1992a):

- Identification of pervasive research thrusts in a database describing promising research opportunities for the Navy (the database consisted of thirty reports produced by National Academy of Sciences panels and Office of Naval Research (ONR) internal experts on 15 technical disciplines);

- Identification of pervasive thrusts in the 7400 project Industrial R&D (IR&D) database.

Applications to other large databases of (mainly) research program descriptions were ongoing.

The reported studies, and the present study, have used the following procedure. First, the frequencies of appearance in the total text of all single words, adjacent double words, and adjacent triple words are computed (e.g., see Figure 1 for examples from the FASAC study). The highest frequency technical content words are selected as the pervasive themes of the full database (e.g., see Figure 2 for the 60 FASAC themes used in the analysis).

FIGURE 1

HIGH FREQUENCY SINGLE WORDS FROM FASAC

FREQ	**WORD**
4170	SYSTEMS
4139	INSTITUTE
2883	COMPUTER
2764	DATA
2611	PHYSICS
2587	WAVES
2564	MOSCOW
2466	CONTROL
2351	OPTICAL
2314	TIME
2268	MATERIALS
2248	SURFACE

HIGH FREQUENCY DOUBLE WORDS FROM FASAC

FREQ	**WORD**
2982	SOVIET UNION
0634	SHOCK WAVES
0503	INTERNAL WAVES
0461	QUANTUM ELECTRON
0425	PHASE CONJUGATION
0375	REMOTE SENSING
0374	IMAGE PROCESSING
0246	AKADEMII NAUK
243	MAGNETIC FIELD
216	CONTROL AVTOMATIKA
207	RADIO WAVES
193	FREQUENCY STANDARDS
176	CATALYSIS KINETIKA
0171	PARTICLE ACCELERATORS

HIGH FREQUENCY TRIPLE WORDS FROM FASAC

FREQ	**WORD**
0371	EXPLOS SHOCK WAVES
0159	CHARGED PARTICLE ACCELERATORS
0134	VYCHISLITEL NAYA TEKHNIKA

0127 SPATIAL LIGHT MODULATORS
0127 IZVESTIYA AKADEMII NAUK
0115 IMAGE PATTERN RECOGNITION
0112 STIMULATED BRILLOUIN SCATT ERING
0095 ATOMIC ENERGY INSTITUTE
0073 OPTICAL PHASE CONJUGATION
0071 FUELS AND OILS
0068 SPACE RESEARCH INSTITUTE
0064 SYNTHETIC APERTURE RADAR
0060 VYCHISLITEL NOY MATEMATIKI
0054 SOVIET ASTRONOMY LETTERS

FIGURE 2

FASAC TECHNICAL THEMES

INTERNAL WAVE
SEA SURFACE
SHOCK WAVE
COMPOSITE MATERIAL
QUANTUM ELECTRON
CROSS SECTION
PHASE CONJUGATION
INTEGRAL EQUATION
REMOTE SENSING
SOLID FUEL
IMAGE PROCESSING
BOUNDARY LAYER
PATTERN RECOGNITION
PLASMA PHYSICS
OCEANIC PHYSICS
COMPUTER SOFTWARE
RADIO WAVE
LIQUID CRYSTAL
MAGNETIC FIELD
DATA PROCESSING
COMPUTER SCIENCE
NEUTRAL BEAM
HYDROGEN MASER
DIGITAL COMPUTER
REMOTE CONTROL
ELECTRIC FIELD
FREQUENCY STANDARD
ELECTROMAGNETIC WAVE

SIGNAL PROCESSING
LOW OBSERVABLE
ARTIFICIAL INTELLIGENCE
PARALLEL PROCESSING
LIGHT MODULATOR
AUTOMATIC CONTROL
SURFACE WAVE
ATOMIC ENERGY
RADIO ENGINEERING
WAVE PROPAGATION
CONTROL SYSTEM
IONOSPHERIC MODIFICATION
CHARGED PARTICLE
FRACTURE MECHANICS
DIFFERENTIAL EQUATION
CHEMICAL PHYSICS
OPTICAL PROCESSING
HIGH-POWER MICROWAVE
PARTICLE ACCELERATOR
EXPLOS SHOCK WAVE
THIN FILM
KINETICS AND CATALYSIS
PROGRAMMING LANGUAGE
EXPLOSION AND SHOCK
STRENGTH MATER
SPATIAL LIGHT MODULATOR
COMPUTER VISION
CHARGED PARTICLE ACCELERATORS
ELECTRON BEAM
ATMOS OCEANIC PHYS
DATA BASE
MOLECULAR ELECTRONIC

Second, for each theme word, the frequencies of words within +/- 50 words of the theme word for every occurrence in the full text are computed, and a word frequency dictionary is constructed. This dictionary shows the words closely related to the theme word. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed for each dictionary (hereafter called cluster) yielding, among many results, those sub-themes closely related to and supportive of the main cluster theme. Third, threshold values are assigned to the numerical indices, and these indices are used to filter out the most closely related words to the cluster theme (e.g., see Figure 3 for part of a typical filtered cluster from the FASAC study).

FIGURE 3

REMOTE SENSING CLUSTER - CLOSELY RELATED WORDS

Cij	Ci	Ii (Cij/Ci)	Eij (Cij^2/CiCj)	CLUSTER MEMBER
022	0036	0.611	0.0359	THERMAL INFRARED
056	0323	0.173	0.0259	ICE
070	0522	0.134	0.0250	SATELLITE
041	0228	0.180	0.0197	OCEANOGRAPHIC
012	0020	0.600	0.0192	ATMOSPHERIC CORRECTIONS
109	1707	0.064	0.0186	SPACE
012	0024	0.500	0.0160	AEROSOL OPTICAL
012	0025	0.480	0.0154	IMAGING SYSTEMS
006	0007	0.857	0.0137	MICROWAVE SENSORS
074	1072	0.069	0.0136	RADAR
012	0037	0.324	0.0104	VEGETATION

CODE:
Cij IS CO-OCCURENCE FREQUENCY, OR NUMBER OF TIMES CLUSTER MEMBER APPEARS WITHIN +/- 50 WORDS OF CLUSTER THEME IN TOTAL TEXT; Ci IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER MEMBER; Cj IS ABSOLUTE OCCURRENCE FREQUENCY OF CLUSTER THEME; Ii, THE INCLUSION INDEX BASED ON CLUSTER MEMBER, IS RATIO OF Cij TO Ci; AND Eij, THE EQUIVALENCE INDEX, IS PRODUCT OF INCLUSION INDEX BASED ON CLUSTER MEMBER Ii (Cij/Ci) AND INCLUSION INDEX BASED ON CLUSTER THEME Ij (Cij/Cj).

These subsets of closely related words are combined into one file, and words that are common to more than one subset (cluster overlaps) are identified. Mega-clusters, or strings of overlapping clusters (based on a threshold of numbers of common words or overlaps), are constructed, showing umbrella areas of related research.

The final results have been identification of the pervasive themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. Because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allows an understanding of the theme interrelationships not heretofore possible with previous text abstraction techniques (using index words, key words, etc.).

Application of Co-word Analysis to FASAC Database

The FSU was a major contributor to many areas of science and technology, and the FASAC reports help document and provide insight to these contributions. There is present interest in preserving the basic science capability of the FSU, as evidenced by a 1992 workshop of leaders of the U.S. science and engineering community (NAS, 1992). This task would benefit by improved understanding of the FSU science and technology capability.

Application of full text co-word analysis to the FSU component of the FASAC database has the potential of providing a unique perspective on the FSU science and technology capability. Even though this database has a different structure from the databases analyzed previously (FASAC contains topical area assessments vs. program, project, or promising opportunity descriptions that characterize the other databases analyzed), it was felt that full text co-word analysis is sufficiently powerful and flexible to be applicable to FASAC as well.

Only unclassified FASAC reports were used. The topics of the FASAC reports used in this study are contained in Figure 4.

FIGURE 4

FASAC UNCLASSIFIED REPORT TOPICS USED IN PRESENT STUDY

*APPLIED INFORMATION SCIENCES
*OCEANOGRAPHIC SYNTHETIC APERTURE RADAR
*CHEMICAL PROPELLANT
*COMBUSTION
*OPTICAL PROCESSING
*PRECISION TIMEKEEPING
*SATELLITE COMMUNICATIONS
*ATMOSPHERIC ACOUSTICS
*IMAGE PATTERN RECOGNITION
*LOW OBSERVABLE MATERIALS
*MOLECULAR ELECTRONICS
*PHASE CONJUGATION
*RADAR CROSS SECTIONS
*APPLIED DISCRETE MATHEMATICS
*HIGH-STRENGTH STRUCTURAL MATERIALS

*PHYSICAL OCEANOGRAPHY
*MICROELECTRONICS
*COMPUTER SCIENCE
*APPLIED MATHEMATICS-MATHEMATICAL THEORY OF SYSTEMS, CONTROL, AND STATISTICAL SIGNAL PROCESSING
*ROBOTICS AND ARTIFICIAL INTELLIGENCE
*EFFECTS OF EDUCATION REFORM ON THE MILITARY
*SCIENCE AND TECHNOLOGY EDUCATION
*HETEROGENEOUS CATALYSIS
*SPACESCIENCE
*SPACECRAFT ENGINEERING
*TRIBOLOGY
*MAGNETIC CONFINEMENT FUSION
*HIGH-POWER RADIO FREQUENCY
*APPLIED SCIENCE
*REMOTE SENSING
*DYNAMIC FRACTURE MECHANICS
*IONOSPHERIC MODIFICATION
*EXOATMOSPHERIC NEUTRAL PARTICLE BEAM
*SOVIET SCIENCE AS VIEWED BY WESTERN SCIENTISTS
*SYSTEM SOFTWARE FOR COMPUTERS.

The FASAC database has a moderate density of technical terms, mostly scientific, but in addition has many institute names, journal names, publishers, and people names. This makes the determination of the relationship among technical areas more difficult than in some purely technically focused databases that were analyzed previously, but does allow for analyses beyond purely technical relationships not possible with the other databases.

RESULTS

Multiword Frequency Analysis

One type of benefit from the output of the multiword frequency analysis is the ability to construct a multi-level taxonomy of the full database. There is a major difference between the taxonomy obtained by this approach and other taxonomies. The present taxonomy derives from the language and natural divisions of the database (analogous to a natural coordinate system of the database), and therefore database entries are easily categorized.
Other taxonomies are usually generated top-down and usually attempt to force-fit database subjects into pre-determined categories.

One of the advantages of the present full text approach, relative to the index or key word approach, is that many types of taxonomies can be generated: i.e., science, technology, institution, journal, person name, etc. Even within one of these categories, such as science, many types of taxonomies can be developed, depending on the interests of the analyst and the reason for the taxonomy. An example of one science taxonomy of the FASAC
database will be shown.

The highest level taxonomy of science from the FASAC database can be seen from the high frequency single words (the following capitalized words are high frequency words from the multiword frequency analyses):

COMPUTER,
DATA,
PHYSICS,
WAVES,
CONTROL,
OPTICAL,
MATERIALS,
COMBUSTION,
SPACE,
INFORMATION,
ENERGY,
SOFTWARE,
PLASMA,
IMAGE,
LASER,
OCEAN.

Broadly speaking, these areas could be subsumed into an Information category

COMPUTER,
DATA,
INFORMATION,
IMAGE,
SOFTWARE,

a Physics category

PHYSICS,
WAVES,
OPTICAL,
PLASMA,

LASER, and, to a lesser extent, an Environment category

OCEAN, SPACE, and a Materials category

MATERIALS.

The high frequency double words reinforce this categorization:

Information

REMOTE SENSING,
IMAGE PROCESSING,
PATTERN RECOGNITION,
COMPUTER SCIENCE,
SIGNAL PROCESSING,
ARTIFICIAL INTELLIGENCE,
OPTICAL PROCESSING,

Physics

SHOCK WAVE,
QUANTUM ELECTRON,
PHASE CONJUGATION,
RADIO WAVE,
MAGNETIC FIELD,
HYDROGEN MASER,

Environment

INTERNAL WAVE,
OCEANIC PHYSICS,
SEA SURFACE,
IONOSPHERIC MODIFICATION,

Materials

THIN FILM,
STRENGTH MATER,
COMPOSITE MATERIAL,
FRACTURE MECHANICS.

The high frequency triple words further amplify the categorization:

Information

IMAGE PATTERN RECOGNITION,
DIGITAL IMAGE PROCESSING,
STATISTICAL PATTERN RECOGNITION,
INTELLIGENCE ANDINFORMATION-CONTROL,

Physics

CHARGED PARTICLE ACCELERATORS,
RADIOPHYS QUANTUM ELECTRON,
STIMULATED BRILLOUIN SCATTERING,
OPTICAL PHASE CONJUGATION,
RADAR CROSS SECTION,

Environment

ATMOSPHERIC AND OCEANIC,
SPACE RESEARCH INSTITUTE,
SYNTHETIC APERTURE RADAR,
RADIO WAVE PROPAGATION,
INTERNAL GRAVITY WAVES,

Materials

COMBUST EXPLOS SHOCK,
SOLID FUEL CHEMISTRY,
METAL MATRIX COMPOSITES,
MAGNETIC THIN FILMS.

Caution should be exercised in relating the above taxonomy based on FASAC to the actual taxonomy of all of FSU science. The FASAC reports represent selected areas of FSU science, and how representative all the FASAC reports are of total FSU science is unknown. The FASAC reports tend to reflect the open FSU literature; how well this open literature represents all of FSU science, including classified work and other work unreported, is unknown. The above taxonomy reflects frequency of word usage, and therefore represents the numbers of words written about technical areas in the FASAC reports. However, dollars spent on these areas, or other measures of FSU priorities, were not taken into account, and thus the taxonomy could be skewed relative to FSU importance attached to these areas. Nevertheless, the above taxonomy does offer insight into areas of FSU science of interest to the U. S.

Megaclusters

The cluster overlaps were determined, and those clusters that had three or more overlaps (three or more common members) were combined to form strings of related clusters, or megaclusters. The following megaclusters (1-9), and their component clusters (*), were obtained:

1. IONOSPHERIC HEATING/ MODIFICATION:

*RADIO WAVE;
*WAVE PROPAGATION;
*QUANTUM ELECTRON;
*IONOSPHERIC MODIFICATION;
*PHASE CONJUGATION.

2. IMAGE/ OPTICAL PROCESSING:

*PARALLEL PROCESSING;
*PATTERN RECOGNITION;
*IMAGE PROCESSING;
*COMPUTER VISION;
*DIGITAL COMPUTER;
*ARTIFICIAL INTELLIGENCE;
*DATA PROCESSING
*COMPUTER SCIENCE
*OPTICAL PROCESSING
*SPATIAL LIGHT MODULATOR
*SIGNAL PROCESSING
*LIQUID CRYSTAL
*LIGHT MODULATOR
*PROGRAMMING LANGUAGES
*INTEGRAL EQUATIONS.

3. AIR-SEA INTERFACE:
*SURFACE WAVE
*OCEANIC PHYSICS
*INTERNAL WAVE
*SEA SURFACE
*BOUNDARY LAYER
*ATMOS OCEANIC PHYS
*REMOTE SENSING.

4. LOW OBSERVABLE:

*LOW OBSERVABLE
*THIN FILM.

5. EXPLOSIVE COMBUSTION:

*KINETICS AND CATALYSIS
*SOLID FUEL;
*EXPLOSION AND SHOCK
*SHOCK WAVE
*CHEMICAL PHYSICS
*EXPLOS SHOCK WAVE
*STRENGTH MATER
*FRACTURE MECHANICS
*COMPOSITE MATERIALS.

6. PARTICLE BEAMS:

*NEUTRAL BEAM
*PARTICLE ACCELERATOR
*ATOMIC ENERGY
*PLASMA PHYSICS
*ELECTRON BEAM
*CHARGED PARTICLE ACCELERATOR
*CHARGED PARTICLE.

7. AUTOMATIC/ REMOTE CONTROL:

*AUTOMATIC CONTROL
*REMOTE CONTROL.

8. FREQUENCY STANDARDS:

*FREQUENCY STANDARD
*HYDROGEN MASER.

9. RADAR CROSS SECTION:

*CROSS SECTION
*ELECTROMAGNETIC WAVE;
*RADIO ENGINEERING.

Of the 60 cluster themes, 52 were in one of the nine Mega-clusters above. The remaining eight cluster themes are:

ELECTRIC FIELD,
MAGNETIC FIELD,

HIGH POWER MICROWAVE,
MOLECULAR ELECTRONICS,
CONTROL SYSTEM,
DIFFERENTIAL EQUATION,
DATA BASE,
COMPUTER SOFTWARE.

Most of these eight remaining themes could be subsumed under the nine megaclusters. ELECTRIC FIELD and MAGNETIC FIELD could be placed in megacluster 6 (PARTICLE BEAMS) or megacluster 1 (IONOSPHERIC HEATING/ MODIFICATION); HIGH POWER MICROWAVE could be placed in megacluster 1 (IONOSPHERIC HEATING/ MODIFICATION); MOLECULAR ELECTRONICS could be placed in megacluster 2 (IMAGE/ OPTICAL PROCESSING); CONTROL SYSTEM could be placed in megacluster 7 (AUTOMATIC/ REMOTE CONTROL); DATA BASE and COMPUTER SOFTWARE could be placed in mega-cluster 2 (IMAGE/ OPTICAL PROCESSING).

From the multiword frequency analysis, the science discipline taxonomy for the FASAC database was defined as Information, Physics, Environment, and Materials. In terms of the megaclusters, Information would encompass
IMAGE/ OPTICAL PROCESSING and
AUTOMATIC/ REMOTE CONTROL;

Physics would encompass
IONOSPHERIC HEATING/ MODIFICATION,
PARTICLE BEAMS,
FREQUENCY STANDARDS, and
RADAR CROSS SECTION;

Environment would encompass
AIR-SEA INTERFACE;

and Materials would encompass
EXPLOSIVE COMBUSTION and
LOW OBSERVABLE.

Categorizing the database with the mega-cluster subcategories allows a re-interpretation of the FASAC database. FASAC can be viewed as a compendium of those aspects of FSU science of interest to the U. S. for strategic and military purposes rather than viewed as a microcasm of all of FSU science

For example, many classes of materials were researched and developed in the FSU. Yet, the materials subcategory in the FASAC analysis focuses on FSU capabilities in energetic materials and coatings to reduce radar cross sections, both important classes from a military viewpoint. The main environmental focus is air-sea interface, with little mention of the terrestrial environment. Coupled with the information category focus on image and optical processing, and the secondary information category focus on remote control, it could be concluded that the FASAC concern was FSU capability in sensing the ocean for ship and submarine activity, and remotely processing and interpreting this information. The secondary environmental focus of FASAC was on the ionosphere, specifically on FSU capabilities for modifying the ionosphere through high power radio wave heating and exploiting its use as a communication medium. One focus of the physics category is particle beams, which could have dual applications of high energy directed weapons and heaters for magnetically confined plasmas and inertial fusion targets.

Cluster Theme/ Member Relationships

The final display, Figures 5 and 6, shows high technical content words from two of the 60 clusters. The selection cutoff criterion was an Equivalence Index (see Figure 3 for definition) greater than or equal to .001. Many different methods of displaying the relationships of the cluster members to the cluster theme and to each other were examined, including advanced graphical packages. It was concluded that a simple division of word categories into quadrants based on Inclusion Index values was most appropriate for the present analysis.

In Figure 5, the underlined line under the figure number, ATMOS OCEANIC PHYS, is the cluster theme. The cluster members are segregated into quadrants. The quadrants are headed by their values of Inclusion Indices. Ij is the ratio of Cij to Cj, and is the Inclusion index based on the theme word. Ii is the ratio ofCij to Ci, and is the Inclusion Index based on the cluster member.

The dividing points between high and low Ij and Ii were selected after examining the distribution functions of numbers of cluster members vs. values ofIj and Ii, and choosing the middle of the 'knee' of the distribution functions as the dividing points. All cluster members with Ij greater than or equal to .1 were defined as having high Ij, and all cluster members with Ii greater than or equal to .5 were defined as having high Ii.

A high value of Ij means that, whenever the theme word appears in the text, there is a high probability that the cluster member will appear within +/- 50 words of the theme word. A high value of Ii means that, whenever the cluster member appears in the text, there is a high probability that the theme word will appear within +/- 50 words of the cluster member.

Thus, words located in the upper quadrant (high Ij high Ii) are coupled very strongly to the theme word. Whenever the theme word appears, there is a high probability that the cluster member will be physically close, and whenever the cluster member appears, there is a high probability that the theme word will be physically close. Essentially, whenever either word appears in the text, the other will be physically close.

For words located in the left quadrant (high Ij low Ii), whenever the cluster member appears in the text, there is a low probability that it will be physically close to the theme word, but whenever the theme word appears in the text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster memberCi is substantially larger than the frequency of occurrence of the theme wordCj, but the cluster member and the theme word have some related meaning. As shown previously (Kostoff, 1991; Kostoff, 1992a), single words have absolute frequencies an order of magnitude higher than double words. Thus, the words in the left quadrant are typically high frequency single words (but not always, as Figure 6 shows), related to the theme word but much broader in meaning than the theme word. A small fraction of the time these broad single words appear, the more narrowly defined double word theme will appear physically close. However, whenever the narrowly defined double word theme appears, the broader related single word cluster member will appear. The words in the left quadrant can also be viewed as a higher level taxonomy of technical disciplines related to the theme ATMOS OCEANIC PHYS.

For words located in the right quadrant (low Ij high Ii), whenever the cluster member appears in the text, there is a high probability that it will be physically close to the theme word, but whenever the theme word appears in the text, there is a low probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster memberCi is substantially smaller than the frequency of occurrence of the theme wordCj, but the cluster member and the theme word have some related meaning. Thus, the words in the right quadrant tend to be low frequency double and triple words, related to the theme word but very narrowly defined. A large fraction of the time these very narrow double and triple words appear, the relatively broader double word theme will appear physically close. However, a small fraction of the time that the relatively broad double word theme appears, the more narrow double and triple word cluster member will appear. This quadrant grouping has the potential for identifying 'needle-in-a-haystack' type thrusts that occur infrequently but strongly support the theme when they do occur. One of many advantages of full text over key or index words is this illustrated ability to retain low frequency but highly important words, since the key word approach must of necessity ignore the low frequency words.

The words in the bottom quadrant (low Ij low Ii) are the remainder of the culled words. They are related to and supportive of the theme, but do not have the strong inclusions based on theme or cluster member occurrence of the members of the other quadrants. Since the upper quadrant typically contains very few or no words, the left quadrant contains very broad words related to the theme, the right quadrant contains extremely narrow words related to the theme, the bottom quadrant contains words related to the theme of the same level of specificity as the theme (on average).

Figure 5, ATMOS OCEANIC PHYS, has a null upper quadrant (typical of the majority of clusters for the threshold values of Equivalence Index chosen). The left quadrant, the broad taxonomy of related areas, appears to describe two major thrusts. One is underwater related (SEA, INTERNAL WAVE, ACOUSTIC, SCATTERING), and focuses on sound propagation through the sea. The other is atmosphere related (ATMOSPHERE, RADAR, SEA SURFACE, SCATTERING), and focuses on radar propagation through the atmosphere. The thrusts have a common juncture at the sea surface, where both acoustic and radar scattering occur on different sides. The right quadrant focuses on very specific sub-areas related to acoustics (mainly), including acoustics applied to the atmosphere (RADIOACOUSTIC SOUNDING), and other aspects of atmospheric science (THEORY OF WIND). The bottom quadrant provides the most balanced view of the two thrusts, expanding on the underwater propagation medium (STRATIFIED FLUID, SHEAR FLOW, INTERNAL GRAVITY WAVES), the radar platform and issues (SATELLITE, PROCESSING OF RADAR), and further amplifying the ocean surface issues (WIND WAVES, TURBULENT, OCEAN SURFACE). The integrated picture presented by the three quadrants is the use of radar from a space platform to view the ocean surface, and the research problems arising from the wind and undersea flows governing the conditions and structure of the ocean surface and impacting the interpretation of the radar images.

FIGURE 5

ATMOS OCEANIC PHYS CLUSTER - HIGH TECHNICAL CONTENT WORDS

HIGH Ij HIGH Ii

NULL

HIGH Ij LOW Ii

SEA
INTERNAL WAVE
ACOUSTIC
SCATTERING
RADAR
SEA SURFACE
ATMOSPHERE

LOW Ij HIGH Ii

RADIOACOUSTIC SOUNDING
ACOUSTIC SOUNDING
THEORY OF WIND
MODELING OF SURFACE
WIND WAVES ATMOS
INFRASOUND AND INTERNAL
ATTENUATION OF SOUND
THEORY OF WAVE

LOW Ij LOW Ii

WIND WAVES
SHEAR FLOW
PROCESSING OF RADAR
SOUND PROPAGATION
TURBULENT
WAVE PROPAGATION
OCEAN SURFACE
SATELLITE
WIND VELOCITY
GRAVITY WAVES
INTERNAL GRAVITY WAVES
POINT SOURCE
STRATIFIED FLUID
SOUND WAVES

Figure 6, PATTERN RECOGNITION, has the quadrants shown sequentially because of space limitations. It has a null upper quadrant. The other three quadrants provide an example of the potential of each cluster to relate the research, technology, and applications for a selected theme. This type of information helps to overcome the problem of language changes as a research area progresses through different phases of development. It aids the construction of forward-looking technology roadmaps (that attempt to predict and display the metamorphosis of research to technology to systems), and retrospective maps (that show how research evolved into technology and existing systems).

The figure contains:

potential applications
IMAGE PROCESSING,
COMPUTER VISION,
MACHINE VISION,
SMART SENSORS,
CLASSIFICATION OF BLOOD,
MAGNETIC RESONANCE IMAGING,
MEDICAL IMAGING SYSTEMS,
BIOMEDICAL IMAGE ANALYSIS,
BLOOD CELL COUNTING,
ANALYSIS OF INTERFEROGRAMS

supporting research areas
ARTIFICIAL INTELLIGENCE,
FUZZY SETS,
DECISION RULES,
FINITE ABELIAN GROUPS,
RANDOM PROCESSES,
STOCHASTIC SYSTEMS, and

technological issues
INFORMATION TRANSMISSION PROBLEMS,
ERROR RATES,
CONVERTING SEQUENTIAL ALGORITHMS.

If it is desired to widen the vocabulary of terms relating research/ technology/systems for a given theme, then the threshold could be lowered for the numerical filter that converts the raw data cluster into a cluster of the type shown in Figure 6, and more applications- oriented words could be included as well.

FIGURE 6

PATTERN RECOGNITION CLUSTER - HIGH TECHNICAL CONTENT WORDS

HIGH Ij HIGH Ii
(UPPER QUADRANT)

NULL

HIGH Ij LOW Ii
(LEFT QUADRANT)

IMAGE PATTERN RECOGNITION
COMPUTER VISION
IMAGE PROCESSING
ARTIFICIAL INTELLIGENCE

LOW Ij HIGH Ii
(RIGHT QUADRANT)

MACHINE VISION
PICTORIAL INFORMATION
FUZZY SETS
MAGNETIC RESONANCE IMAGING
THREE-DIMENSION SCENES
MASSIVELY PARALLEL PROCESSOR
IMAGE CODING
MISSILE MACHINE VISION
OPTICAL TRANSFORMS
MEDICAL IMAGING SYSTEMS
PROBABILITY DISTRIBUTION ESTIMATION
OPTICAL REPROCESSING
IMAGE INPUT DEVICES
HOLOGRAPHIC FILTERS
IMAGE CODING PATTERN
IMAGE PROCESSING HARDWARE
IMAGING AND ULTRASOUND
LARGE PROTEIN MOLECULES
INTEGRATED CIRCUIT ASSEMBLY
STATISTICAL CLASSIFICATION PROBLEMS
SMART SENSORS
BIOMEDICAL IMAGE ANALYSIS
TOMOGRAPHY MAGNETIC RESONANCE
DIGITAL HOLOGRAPHIC FILTERS
VIDEO INFORMATION INPUT

DIFFRACTION PATTERN
WHITE BLOOD CELL
DEDICATED SYSTEMS
RECURSIVE SCENE MATCHING
DECISION RULES
SOLID-STATE TELEVISION CAMERAS
ERROR RATES
CIRCUITS OPTOELECTRON COMPONENTS
TRAINING SETS
SAMPLE SIZE
BLOOD CELL COUNTING
PHASES OF OPTICS
CONVERTING SEQUENTIAL ALGORITHMS
OBJECT ROTATION
COMPUTER-GENERATED DATA SETS
ONE-DIMENSIONAL SIGNALS
ANALYSIS OF INTERFEROGRAMS
CLASSIFICATION OF BLOOD
ALL-OPTICAL DIGITAL COMPUTERS
FINITE ABELIAN GROUPS
DIGITAL DATA ACQUISITION
DECISION RULE GENERATION
INCOHERENT-TO-COHERENT CONVERTER
REPRESENTATION OF KNOWLEDGE

LOW Ij LOW Ii
(BOTTOM QUADRANT)

AUTOM CONTROL
INCOHERENT ILLUMINATION
AUTOM REMOTE CONTROL
CHARACTER RECOGNITION
STATISTICAL PATTERN RECOGNITION
RANDOM PROCESSES
SYSTEMS OF ROBOTS
OPTICAL COMPUTING
INTELLIGENCE AND INFORMATION-CONTROL
IMAGE ANALYSIS
COMPUTER TOMOGRAPHY
EXPERT SYSTEMS
LINEAR RECOGNITION MACHINES
CYBERNETICS
DIGITAL DATA PROCESSING
HOLOGRAPHIC MEMORIES
STOCHASTIC SYSTEMS
DIGITAL IMAGE PROCESSING

DIGITAL SIGNAL PROCESSING
INFORMATION TRANSMISSION PROBLEMS
SPATIAL LIGHT MODULATOR
OPTICAL PROCESSING

DISCUSSION AND CONCLUSIONS

The results of a multiword frequency analysis performed on the total FASAC database allowed a high level science taxonomy of four broad categories to be generated: Information, Physics, Environment, and Materials. A co-word analysis on the 60 highest frequency pervasive themes identified by the multiword frequency analysis, and a subsequent (effective) renormalization of the pervasive themes due to linkages among sub-themes allowed nine 'umbrella' themes to be generated:

- Ionospheric Heating/ Modification;
- Image/ Optical Processing;
- Air-Sea Interface;
- Low Observable;
- Explosive Combustion;
- Particle Beams;
- Automatic/ Remote Control;
- Frequency Standards;
- Radar Cross Section.

Based on the results and interpretation of the multiword frequency analysis and the co-word analysis, it could be concluded that the FASAC database used in this study is a compendium of those aspects of FSU science of interest to the U. S. for strategic and military purposes. The microlevel analysis of selected theme clusters, showing how the cluster members related to each theme, reinforced this conclusion and provided more detail about those aspects of each theme on which FASAC concentrated.

A wealth of information resulted from the FASAC output, and only a small fraction of that information was presented and analyzed in this report. The analysis was restricted to technical themes and their relationships, but raw data was available for relating technical themes to non-technical themes such as institutions, scientists, journals, geographical regions, etc. Experts in the technical themes were not utilized in the data analysis, which limited the level of detail of the analysis.

In the future, full text co-word analysis could be used to obtain even a more representative structure of FSU (or any other country's) science. If a large number of randomly selected published FSU scientific papers were entered into a database (optically scanned if not already on computer storage media), then a multiword frequency analysis and co-word analysis could be performed on this text database. The algorithm used presently (1992) requires about twice as much RAM as the size of the database. If a paper is assumed to occupy 20KB of storage, then 1000 papers would require 40MB of RAM (well within today's microcomputer capability), and 10000 papers would require 400MB of RAM (bordering on today's microcomputer capability). Assuming that a paper represents about $100K worth of effort, then a 10000 paper database would represent $1B worth of effort, and would offer a very representative sample of FSU science output. The critical path would be assembling this database, not analyzing it.

The major purposes of this report were: 1) to demonstrate that full text co-word analysis could allow useful information to be extracted from a large text database consisting of seemingly heterogeneous reports, and 2) to apply the technique to a timely, important body of information. Both of these targets were achieved. Full text co-word analysis is in its formative stages, and much development remains to be done to utilize its full potential. This potential includes understanding the breadth of analyses that can be performed and the breadth of applications that can be covered. It is hoped that the initial techniques and results reported in this study will motivate and stimulate other organizations and researchers to develop and apply the general technique of full text co-word analysis on a much broader scale.

BIBLIOGRAPHY

Kostoff, R. N., "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Proceedings: Portland International Conference on Management of Engineering and Technology, October 27-31, 1991. More detailed paper available from author.

Kostoff, R. N., "Research Impact Assessment," Presented at Third International Conference on Management of Technology, Miami, FL, February 17-21, 1992a. Larger text available from author.

Kostoff, R. N., "Co-Word Analysis," in Assessing R&D Impacts:Method and Practice, Bozeman, B. and Melkers, J., Eds. (Kluwer Academic Publishers, Norwell, MA) 1993.

NAS, "Reorientation of the Research Capability of the Former Soviet Union," Results of a Workshop on March 3, 1992, NAS, NAE, IOM, National Academy Press, Wash., DC 1992.