SCIENCE AND TECHNOLOGY TEXT MINING: ORIGINS OF DATABASE TOMOGRAPHY AND MULTI-WORD PHRASE CLUSTERING

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

A. **ABSTRACT**

This report initially describes the motivations for co-word analysis in support of research policy formulation and research implementation evaluation. It places co-word analysis in perspective to other co-occurrence techniques such as co-citation and co-nomination analyses. It then traces the origins of co-word analysis in computational linguistics, describes in detail the development of co-word analysis for research evaluation, and concludes by presenting a new approach to co-word analysis for research evaluation (Database Tomography). The report shows that this new approach to co-word analysis, which requires no index or key words but deals with text directly, is a useful tool for scanning large bodies of text. It can identify pervasive thrust areas and their interrelationships, and serve as a starting point for further in-depth analysis of the text. Its value increases as: 1) the size of text increases and 2) the breadth of topical areas covered by the text increases beyond the expertise of a moderate number of panels of experts. A single link clustering example is shown that represents ***the first use of multi-word technical phrases in modern clustering***.

(The views in this report are solely those of the author and do not represent the views of the Department of the Navy or any of its components)

KEYWORDS: Text mining; Database Tomography; co-word; co-citation; co-nomination; research evaluation; clustering.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (FROM - TO)
15-08-2003	Technical	xx-xx-1995 to xx-xx-2003

4. TITLE AND SUBTITLE SCIENCE AND TECHNOLOGY TEXT MINING ORIGINS OF DATABASE TOMOGRAPHY AND MULTI-WORD PHRASE CLUSTERING 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

This report initially describes the motivations for co-word analysis in support of research policy formulation and research implementation evaluation. It places co-word analysis in perspective to other co-occurrence techniques such as co-citation and co-nomination analyses. It then traces the origins of co-word analysis in computational linguistics, describes in detail the development of co-word analysis for research evaluation, and concludes by presenting a new approach to co-word analysis for research evaluation (Database Tomography). The report shows that this new approach to co-word analysis, which requires no index or key words but deals with text directly, is a useful tool for scanning large bodies of text. It can identify pervasive thrust areas and their interrelationships, and serve as a starting point for further in-depth analysis of the text. Its value increases as: 1) the size of text increases and 2) the breadth of topical areas covered by the text increases beyond the expertise of a moderate number of panels of experts. A single link clustering example is shown that represents the first use of multi-word technical phrases in modern clustering.

15. SUBJECT TERMS
Text mining; Database Tomography; co-word; co-citation; co-nomination; research evaluation; clustering.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19. NAME OF RESPONSIBLE PERSON
			Same as Report (SAR)	85	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 -

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39.18

A-1. **TABLE OF CONTENTS**

B. **INTRODUCTION**

In formulating and executing broad spectrum research policy, it is important to understand how research thrusts have interrelated and evolved over time, how they are projected to evolve, and how different types of interventions from sponsors and policymakers can affect the evolution and impact of research. The problem is compounded because of the strong interconnectivity among the different areas of research and technology [Kostoff, 1991b, 1994]. While a panel of experts could provide an acceptable view of the trends and interrelationships within a narrowly-defined research area, identification of the connectivity of a broad range of areas is well beyond the expertise of any one panel of experts, and perhaps beyond a group of panels. An integration of topics and trends requires supplementation to the standard peer or analyst group evaluation. Much effort has been focused on development of more objective quantitative approaches for analyzing and integrating written and survey information to supplement analysts or groups of peers in understanding research trends.

Due to the rapid expansion of electronic media storage capabilities, research policy analysts now have available massive databases of research-relevant information that can be analyzed to supplement peer review processes. A major problem in practice is how to extract the essential information from these databases in a form readily amenable to analysis and interpretation. In other words, **how does the analyst extract the collective wisdom contained in these large databases in a concise, readily understandable form**.

Modern quantitative techniques utilize computer technology extensively, usually supplemented by network analytic approaches, and attempt to integrate disparate fields of research. One class of techniques exploits the use of co-occurrence phenomena, and it is this class, in particular co-word analysis, that will be addressed in this report. In co-occurrence analysis, **phenomena that occur together frequently in some domain are assumed to be related, and the strength of that relationship is assumed to be related to the co-occurrence frequency**. Networks of these co-occurring phenomena are constructed, and then maps of evolving scientific fields are generated using the link-node values of the networks. Using these maps of science structure and evolution, the research policy analyst can develop a deeper understanding of the interrelationships among the different research fields and the impacts of external intervention, and can recommend new directions for more desirable research portfolios.

The remainder of this report is structured as follows. Co-citation and co-nomination phenomena are described briefly, and some relevant references are provided for further reading (Section C). Then a short section follows on the origins of co-word analysis in linguistics, lexicography, and especially computational linguistics (Section D-1). This is provided mainly for the reader who wishes to have background context for the modern day applications of co-word analysis to research policy. After this brief section, a detailed description is provided of modern day development and applications of co-word analysis to research policy and issues (Section D-2). The positive features, as well as limitations, of present-day approaches are identified. This section is followed by the main focus of this report, a new approach to co-word analysis for research policy that uses textual database only, with no need for keywords or indexing (Section D-3). A detailed application of the method for scanning a promising research opportunities database, and preliminary results from the IR&D database, are presented. Section E, which follows, contains a summary, and Section 10 contains the Bibliography. The tables that follow contain the data from the co-word analyses of the research opportunities and IR&D databases.

The main body of this report was presented initially in part at the PICMET Meeting in 1991 (Portland, OR), and the remainder at the Third International Conference on Management of Technology in 1992 (Miami, FL). The present report aggregates these two components, and updates the applications of Database Tomography that have been performed since the initial technique was described at the above two meetings.

C. CO-CITATION AND CO-NOMINATION ANALYSIS

C-1. Co-citation Analysis

Three of the more applicable co-occurrence techniques to the science evolution problem, listed in order of level of development and frequency of utilization, are co-citation, co-word, and co-nomination. In co-citation analysis, the frequencies with which references in published documents are cited together are obtained, and are eventually used to generate maps of clusters of cohesive research themes.
Co-citation analysis was developed about three decades ago, when the Science Citation Index became more readily available for computer analysis, and it has spawned a number of studies and reviews, a few of which are listed here [Small,

1973; Small, 1977; Small, 1978; Garfield, 1978; Small, 1980; Small, 1985a; Small, 1985b; Small, 1986; Franklin, 1988; Oberski, 1988].

While the strengths and weaknesses of co-citation analysis are not the subject of this report, it should be noted that **co-citation is a rather indirect approach to obtaining connectivity among research areas, and it involves a number of abstract steps**. Querying the author(s) of a research paper about what other research areas are related to their work would be the most direct method of obtaining the desired data [Kostoff, 1991b]. Obtaining this information by analyzing the words in the paper and related papers would be the next most direct method. Obtaining this information by examining citations and co-citations restricts the types of documents that can be analyzed (essentially published papers) and requires the additional assumption that the themes of two articles co-cited many times by authors must be strongly related. While the co-citation proponents claim that "many potentially useful applications have been demonstrated" [Franklin, 1988], others conclude that "results of co-citation cluster analyses cannot be taken seriously as evidence relevant to the formulation of research policy" [Oberski, 1988].

C-2. Co-nomination

Co-nomination is a particular example of the more general social network analysis used to study communication among workers in the fields of science and technology. Generally, in co-nomination, experts in a given field are asked to identify other experts, and then a network is generated that shows the different linkages (and the strengths of these linkages) among all the experts (and possibly their organizations and technical disciplines) identified. A 1988 survey [Shrum, 1988] of the development of social network analysis traces studies in this area back at least three decades. Two of these studies are particularly relevant to the specific co-nomination approach that will be described in the following section, and these two studies are outlined briefly.

In a study of theoretical high energy physicists [Libbey, 1967], respondents were asked to name two persons outside their institution with whom they exchanged research information most frequently and no more than three who they believed to be doing the most important work in their area. A network analysis was done to identify communication linkages. In a later study of theoretical high energy physicists [Blau, 1978], respondents were asked to name two persons outside their institution with whom they exchanged information most frequently about their research. Again, communication networks were generated.

Co-nomination was developed to circumvent co-citation's dependence upon databases consisting of refereed scientific publications. **It is a more direct approach of obtaining links among researchers and, if combined with other network approaches that include both links between technical fields and the link strengths [Kostoff, 1991b, 1994], could potentially incorporate links among researchers and technical fields**. Since co-nomination is known less well than co-citation, its latest embodiment will be described briefly.

Researchers are sent a questionnaire inviting them to nominate other researchers whose work is most similar or relevant to their own. Based on the responses, networks are then constructed by assuming that links exist between co-nominated researchers and that the strength of each link is proportional to the frequency of co-nomination [Georghiou, 1988].

However, as is the case with co-citation, frequency of co-occurrence may not be a unique indicator of strength. One could postulate two cases: 1) researchers co-nominated were doing essentially identical work, and their linkages were very strong; and 2) researchers were doing vaguely similar work, and their linkages were very weak. In both cases, the frequency of co-occurrence would be the same, and the links on the network would have the same strength.

D. CO-WORD ANALYSIS

D-1. **Origins of co-word analysis in computational linguistics**

D-1-a. Introduction

The origins of co-word phenomena can be traced back at least six decades to the pioneering work in: 1) lexicography of Hornby [1942] to account for co-occurrence knowledge, and 2) linguistics of De Saussure [1949] to describe how affinity of two language units correlates with their appearance in the language.
For the reader interested in a detailed description of the evolution of co-word phenomena in linguistics, lexicography, and computer science over these past six decades, a 1991 dissertation on collocation phenomena (sequences of words

whose unambiguous meaning cannot be derived from that of their components, and which therefore require specific entries in the dictionary) is recommended highly [Smadja, 1991]. The remainder of this section will provide only summary information on the seminal publications that have advanced these fields, and will describe those works that are particularly relevant to the theme of the present report in more detail.

D-1-b. Linguistics and Lexicography

In early co-word studies, words were classified on the basis of their co-occurrence with other words as well as their meanings [Firth, 1957; Harris, 1968]. Chomsky [1965] added the observation that the reasons for two words co-occurring in the same context are not always relevant to a general linguistic description of a given language. Halliday [1966] related the well-formedness of sentences to their lexical levels; i.e., how sensitive the meaning of a sentence is to substitution for one member of a co-occurrence pair. A 1981 study included collocations as part of a linguistic model, whose goal was to relate any given meaning and all the texts that express it [Melcuk, 1981]. In 1986, a combinatory dictionary, which contained only general English collocations, was constructed [Benson, 1986].

D-1-c. Computational Linguistics

Computational linguistics interest in collocations has focused on information retrieval, computer assisted lexicography, stochastic language models, and natural language generation. Information retrieval research focused on designing more efficient indexing tools using pairwise lexical affinities instead of keywords [Sparck Jones, 1971; Van Rijsbergen; 1979; Salton, 1983; Maarek, 1989]. Computer assisted lexicography focused on making tools for assisting lexicographers to compile data. In Choueka's works, methods were developed for locating interesting collocational expressions in a large body of text. These methods were based principally on the distribution of types and tokens in the body of text and on the analysis of the statistical patterns of neighboring words [Choueka, 1983; Cheouka, 1988]. At the same time, other approaches used co-occurrence knowledge and statistical analysis of large bodies of text to help in language generation, as a basis for indexing, selection of lexical items, and generation of collocationally restricted sentences [Smadja, 1988; Smadja, 1989; Church, 1989; Maarek, 1989; Amsler, 1989]. Stochastic language models applications built on previous work in speech recognition and text compression, and treated collocations as statistical entities [Bahl, 1983; Mays, 1990; Church, 1990]. Efforts in the 1980s in natural language generation that account for collocational knowledge include McCardell [1988], Nirenburg [1988], Kittredge [1986], and Iordanskaja [1990].

D-1-d. Co-word Analysis for Thematic Relations

In the mid-1970s, a study was performed to examine relationships among themes in a Kierkegaard novel using co-occurrence phenomena [McKinnon, 1977]. An important term in the book, Systemet, was chosen, and a dictionary was constructed of all words in the book occurring in the same sentences as Systemet. A co-occurrence matrix that contained the co-occurrences among these related terms was constructed, and analyzed to eventually show the relations among all the associated terms in the mini-dictionary as they occured in the original text. While the dictionary was restricted to single words, and the co-occurrence domain was restricted to sentences, the methodology did represent a major step forward in extracting word relations from text by their co-occurrences.

An update of this method employed frequency of co-occurrence to extract relatedness information from text. The study looked at co-occurrence using the sense-definition as the textual unit (entire definition of a sense of a word). Database used was the Longman Dictionary of Contemporary English (LDOCE) rather than free text. The method used single word frequencies only, and resulted in construction of networks of related words. It was concluded that co-occurrences of words in the LDOCE-controlled vocabulary in the definitions in LDOCE appeared to provide some useful information about the meanings of those words. Co-occurrence frequency correlated significantly with human judgements of relatedness, and the relatedness functions on co-occurrences yielded even higher correlations. When the relatedness functions were used to derive Pathfinder networks on the LDOCE primitives, the intensional meaning of a word was represented by the collection of words that were nearby in the network. For correlation with human judges, conditional probability was better than just frequency of co-occurrence, and nothing was gained from the more complex measures [McDonald, 1990].

While the methods described above were useful for showing how relations among words and terms could be quantified and extracted from text, **none were applied**

to evaluating research trends or supporting research policy. The following sections describe the development and employment of co-word analysis techniques to extract relationships among research themes from large text databases.

D-2. **Development of co-word analysis for research evaluation**

D-2-a. Initial Motivations

Modern development of co-word analysis for purposes of evaluating research appears to have originated in the mid-1970s under Michel Callon at the Centre de Sociologie de l'Innovation at the Ecole des Mines de Paris [Callon, 1979; Callon, 1983; Callon, 1986]. The initial motivation was to develop a method to help evaluate the state of research that would have broader scope than, be more objective than, and provide a supplement to, panels of experts [Callon, 1979].

The method would also have to overcome the limitations that Callon viewed as inherent to co-citation analysis: The authors cite and co-cite what has already been sanctioned; citations permit an indirect access to a document's content and tend to respect tradition and reinforce existing hierarchies; technical and economic-industrial literature seldom use citations, but use other forms of expression.

D-2-b. Impact of French Government Intervention on Macromolecular Chemistry

The method developed initially by Callon focused on analyzing the content of articles and reports, rather than their citations. In one of the first descriptions and applications of the method [Callon, 1979], the impact of French government intervention in the field of macromolecular chemistry was examined. A database of over 4,000 articles covering the field of interest was generated. Key or index words were assigned to each article in the database. A basic assumption was that the key words describing an article had some linkages in the author's mind, and the different fields or functions represented by these words had some relation.

Each time a pair of words occurred together in the key word list of an article, it was counted as a co-occurrence of the pair. The number of co-occurrences for each pair was calculated for all the articles in the database. A co-occurrence matrix was constructed whose axes were the index words in the database and whose elements were the number of pair co-occurrences of the index words. A two-dimensional map was constructed that would display visually the positions of the key words relative to each other based on their co-occurrence values from the matrix.

There were at least two major problems with this approach: the text was not analyzed directly, and the analysis was performed on the key words. **The bias and error introduced from key word analysis was unknown, but use of key words continued to affect the credibility of the technique for years**. A 1986 study refers to this potential biasing phenomenon as the 'indexer effect' or, more descriptively: "In our study it often appeared to our Research Council experts that we were seeing science intellectually established in year X through the distorting class of conceptualization of indexers whose own intellectual formation was some years earlier" [Healey, 1986]. A 1987 study states: "It is well known that keywords selected by an indexer who is not a practicing scientist tend to be conservative: the keywords reflect the world of the scientists of two years ago. This problem has yet to be solved when working with keywords." [Leydesdorff, 1987]

Second, because Callon's matrices were limited to a size of 50x50, further aggregation was required to fit all the key words within a 50x50 matrix. The additional errors produced by this aggregation are unknown.

D-2-c. Tracking the Status of Biotechnology

A study in the early 1980s was aimed at tracking the status of biotechnology by performing a co-word analysis of articles in a biotechnology core journal (Biotechnology and Bioengineering) over a period of 10 years [Rip, 1984]. As in the Callon study described above, the authors constructed a co-occurrence matrix based on keywords of these biotechnology articles (or signal words, as the authors called them). To distinguish between co-occurrences that are interesting and co-occurrences that are uninteresting, the authors introduced indices that measured the strength of the co-occurrence linkage according to some formula and determined a threshold below that co-occurrence linkages were no longer considered to be interesting. These indices were then used to construct maps that portrayed the relationships among the 'signal words'.

D-2-c-i. Jaccard Index

The indices are essentially normalizing factors that relate the co-occurrences of word pairs to some function of the absolute frequency of occurrence of one or both members of the pair. One index the authors described is the Jaccard index, which is also used in co-citation analysis. The Jaccard index Jij is defined as: $Jij=(Cij/(Ci+Cj-Cij))$, where Cij is the co-occurrence frequency between words i and j, Ci is the absolute frequency of word i and Cj is the absolute frequency of word j. The authors presented Jaccard maps showing the links among the keywords, and they varied the threshold value of the index below that links did not appear on the map. As the threshold was lowered, the number of linkages portrayed increased and the fine structure became more evident, but the complexity and interpretability of the map increased as well.

One of the major problems in co-word mapping and analysis is to find normalization indices that minimize biasing of the results. For example, keywords will usually span a wide range of absolute frequencies of occurrence. In many cases, it would be useful to have indices that are relatively frequency independent. The Jaccard index cannot handle associations between low-frequency and high-frequency words very well because it will have low values even in cases where the low frequency word always appears together with the high frequency word [Courtial, 1986]. In the Jaccard index, if $Ci>>Cj$, large changes in Ci will produce large changes in Jij, but large changes in Cj will produce small changes in Jij.

D-2-c-ii. Inclusion Index

Another index discussed by the authors was the Inclusion index Iij. It is defined as $Iij=Cij/Ci$ (with $Ci<Cj$) and is a measure of the extent to which a less frequently occurring key word is joined to a more frequently occurring key word. Iij is not symmetrical, and inclusion maps that are oriented graphs can be generated that reflect the hierarchical relations between keywords. If a sufficiently high threshold is set for appearance of linkages in inclusion maps, then an overall picture is obtained of the 'master key words' dominating a tree of less frequently occurring key words.

D-2-c-iii. Proximity Index

A third index described was the Proximity index Pij. It is defined as $Pij=(Cij/\min(Ci,Cj))/(\max(Ci,Cj)/N)$ and is interpreted as the ratio of the conditional probability of finding word i, given word j, to to the unconditional probability of finding word i in any one of the N articles in the domain. It is symmetrical, and a proximity map consists of a collection of larger and smaller clusters of key words that, in their basic form, usually involve the co-occurrences of three or four words [Courtial, 1986].

D-2-c-iv. Statistical Index

A fourth index described was the Statistical index Sij, which is the normalized deviation from the expected value of the co-occurrence. It is defined as Sij=(1/SIGMA)*(Cij-(Ci*Cj)/N), where SIGMA is the standard deviation of the hypergeometric distribution function and Ci*Cj/N its mean, or expected, value. Comparisons of maps using the Statistical index and the Jaccard index led to the conclusion that the Jaccard maps provided a conservative picture of the linkages between key words, and were preferable because of ease of computation relative to Statistical index maps.

D-2-d. Impact of French Government Intervention on Aquaculture

Another study aimed at identifying the impact of the French government's efforts in creating a field of aquaculture [Bauin, 1986]. About 3,000 articles in this field were examined for the years 1979-1981. Inclusion and Proximity maps (two-dimensional portrayals of the main themes and subthemes of the body of articles and of the hierarchical relationships and interconnectivities among these themes and subthemes) using the Inclusion and Proximity indices were constructed to depict the field of aquaculture at the beginning and at the end of the time period of interest (1979-1981).

The Inclusion maps revealed the field of aquaculture to be very dispersed. Only in a few cases did meaningful structures and hierarchies exist. The Proximity maps revealed a very high degree of structuring in which no one issue was totally disconnected from the rest of the field.

The authors concluded that a unified field of aquaculture did not exist outside the political influence of decisionmakers. **Aquaculture appeared to be more of a bureaucratic category rather than a scientifically unified and integrated field**. Researchers appeared to have remained local and locally connected and maintained their respective approaches. No hierarchies appeared that would enable other scientists or companies to use the researchers' results when undertaking further research or product development. Frequent mention of geographical locations on the maps supported the interpretation of the research efforts remaining localized, and frequent occurrence of words like 'conference,' 'annual reports,' 'historical accounts,' supported the interpretation that decision-makers were attempting to improve communications to create a unified community.

D-2-e. Biotechnology Dynamics from Patent Analysis

Another study in the same time frame was targeted at improving evaluation of the contents of a large number of patents [Callon, 1986]. In the field of biotechnology, 268 patents were examined covering the years 1979-1981. Inclusion maps were generated to identify the main themes and the hierarchical relationships in the patent database, and then Proximity maps were drawn to identify small linked groups of related areas.

A new feature in this study was a 'zoom' about one of the themes (enzyme). In this technique, **attention is focused on a limited region of the Inclusion map and the threshold for inclusion of a word is lowered significantly. This allows more detail and consequently higher resolution in the region around the word being 'zoomed**.' In the specific case, all the words of the Inclusion maps that were linked directly or indirectly to 'enzyme' were included and the threshold for inclusion was lowered by more than a factor of three. A much more detailed analysis of the temporal evolution of 'enzyme' was made due to the significantly added information from the additional words and links in the 'zoom.'

Callon concluded that co-word analysis provided a unique capability of describing the mechanisms of innovations and the dynamics of technological development resulting from the patent literature. Use of the 'zoom' technique around enzymes showed the appearance of new centers of interest and reorganization of existing relationships over the short timeframe of database coverage. The patents responsible for these changes were identified easily. Callon also concluded that the patent indexation method used to generate index words for this study was too costly, and an improved method of indexing the database was required.

D-2-f. Key Words vs Titles

A 1989 study examined the difference in results when using key words vs. using the titles of articles [Whittaker, 1989]. Whittaker concluded that co-word analysis may be satisfactorily performed on a set of documents by using either title words or keywords and that (at least for the case in point) the main difference between the results obtained is that keywords provide a much more detailed account of the field of science studied.

Another 1989 study examined how words and co-words as scientometric indicators differ from citations in what they reveal about science [Leydesdorff, 1989]. The study included the use of title words, words from abstracts, and index words. Leydesdorff concluded that searching with index words generated more noise than searching with original title words. He also concluded that indexing subsumed different words under more general categories, and hence increased the number of co-word linkages substantially because the smaller set is more strongly tied together than the larger.

D-2-g. Industrial Ceramics Priorities for Ireland

The thinking and applications of the French group, which can be viewed as the state-of-the-art in co-word analysis as of 1990, is described in two major articles [Turner, 1988; Callon, 1991a]. The reader interested in applying co-word analysis should study these articles carefully, as well as the text that describes the state-of-the-art as of the mid-1980s [Callon, 1986]. Only the major advances reported in these two articles, relative to those described in previous sections, will be presented.

The goal of the Turner study was to perform a co-word analysis on industrial ceramics patents in order to determine priority areas for Ireland in this field. A computer-assisted indexing method, the LEXINET system, was employed to reduce the 'indexer effect.' An expert took over two months to index the full database of 16,000 patents, using the significant words extracted from the titles and summaries of each document.

The co-occurrence frequencies of the index words were obtained and a co-occurrence matrix was generated. The index of Mutual Inclusion $E_{ij}=(C_{ij}/C_i)*(C_{ij}/C_j)$ was used to normalize the co-occurrence frequencies. It measures the probability of word i being simultaneously present in a document set indexed by word j and, inversely, the probability of j if i, given the respective database frequencies of the two terms. The index avoids favoring any particular zone of the word frequency distribution curve.

The co-occurrence matrix, consisting of coefficients between all possible index word pairs, generated far too many links for graphical portrayal. An algorithm was used to generate subgroups of tightly linked words, called clusters. The clusters were kept relatively small, 10 words or fewer. Variable thresholds, characterized by the value of the first link refused, were used to form the cluster. The algorithm could also perform cluster nesting, or "clusters of clusters," in order to detect macro-subject areas. This allowed the co-word analysis user to choose the appropriate level of aggregation for his particular needs.

The subject areas identified by the clusters were described in terms of two policy relevant dimensions; internal coherence, and the strength of their specific relationships with other subject areas. The first dimension, a cluster's coherence or density, indicates the degree of overlap between the centers of interest shared by the particular group of authors working in the identified subject area. An indication of the coherence of the cluster is obtained by calculating the average value (of the index of Mutual Inclusion) of these internal links.

The second dimension, the centrality of a cluster within a research network, is obtained by using the sum (of the index of Mutual Inclusion) of a subject area's external links. The more the number of its connections with other subject areas, and the greater the strength of these connections, the more central a subject area will be in the research network.

The centrality and density measures were computed and used simultaneously to classify subject areas. A policy map was generated that situated each subject area within a two-dimensional space divided into four quadrants: the x-axis (centrality) served to locate subject-specific, as opposed to potential spill-over, areas; the y-axis served to locate subject areas internally well-structured as opposed to those that were weakly structured. The remainder of the document contained interpretation and discussion of the subject areas and their locations within the quadrants of the policy map.

D-2-h. Public Funding Impact on Polymer Science

In a 1991 study of the French group [Callon, 1991a], co-word analysis was used to describe the interactions that exist between different phases of the innovation process and to show if basic research or applied research is the moving force. The results presented came from a study concerning the impact of public funding on the development of polymer science.

Two databases were used for the co-word analysis: one consisted of all the international literature in the field of polymer science (basic and applied research), and the second (a subset of the first) consisted of the academic science (basic research) documents. Co-occurrence matrices were generated, and the co-occurrence frequencies were normalized using the same index as in the Turner study described above. Clusters were generated, and the four-quadrant policy maps of density and centrality were constructed.

An added feature of this study was plots of the evolution of cluster density and centrality with time. This was equivalent to following the motion of a cluster's position on the policy map as a function of time. These plots could show that a given cluster became more and more (or less and less) central over time, and that, at the same time, its density increased (or diminished).

D-2-i. Co-occurrence Research at Leiden

The Centre for Science and Technology Studies at the University of Leiden has been expanding its outputs of co-occurrence studies. The breadth and scope of its research efforts in co-word, co-citation, co-classification, and other bibliometric analysis can be seen in its list of projects and publications in its annual report.

One study used a unique two-step co-word analysis as the basis for bibliometric maps of neural network research. The maps portray neural networks embedded in the environment of related fields [Van Raan, 1991]. A second study applied co-word analysis to the field of chemical engineering. To improve the mapping, a combination of a clustering technique applied to the word co-occurrence data matrix and multidimensional scaling of the resulting word clusters was used [Peters, 1991]. A third group of studies combined word frequency analysis of citing articles with co-citation analysis [Braam, 1991a; Braam, 1991b]. This hybrid approach allows more accurate portrayal of cluster topics, and allows separate research specialties to be delineated more easily. However, key/ index words are used, not full text.

D-2-j. Summary

To summarize this section, the evolution of co-word analysis as a tool to support research policy was traced over the decade of the 1980s. Manual indexing is being reduced and gradually supplemented by computer-assisted indexing. Frequency insensitive normalization indices for co-occurrence frequencies are receiving wider use. Automatic clustering algorithms of variable threshold capabilities are becoming standard and nesting of clusters is available for aggregation studies. Maps that study the evolution of gross cluster measures, such as centrality and density, became available, and are much more comprehensible than the highly cluttered maps used previously.

D-3. **A new approach to co-word analysis for research evaluation**

The previous sections have described briefly different types of co-occurrence techniques used to evaluate research and have shown the background, evolution, and applications of co-word analysis in particular. While there has been some progress in overcoming the dependency of co-word analysis on key or index words, limitations remain. The remainder of this section describes a co-word method developed in the early 1990s that eliminates the requirement for any key or index words and deals with any form, or combination of forms, of text, be it published article, report, or memo.

The new method relies upon full-text analysis [Kostoff, 1991c]. In dealing with written material, it is the most direct method of extracting messages from large textual databases. It does not rely on interpreting intermediate abstractions of text such as citations, key or index words, or titles. The method displays the richness of the fine structure relations in the text, and provides orders of magnitude more detail and useful information than previous methods that relied upon aggregate measures such as key words, titles, citations, etc.

The method is more computer intensive than previous co-word approaches, since it requires examination of every word in the text database. However, compared with the cost of data analysis time in any of the bibliometric techniques, whether co-citation, co-word, or co-nomination, the cost of computer time in the new method is negligible. Given that computers are continually becoming more powerful, and computing costs per MB of text are decreasing, the method will be even more desirable in the future. **It is a method whose time has come**.

The method in its entirety requires three distinct steps. The first step is identification of the main themes of the text being analyzed. The second step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps, and associated results, will be described in the following sections. Time evolutions of themes have not yet been performed.

D-3-a. Theme Identification

D-3-a-i. Background

Careful reading of many technical papers, reports, and program descriptions has shown that technical writers tend to repeat continually the words describing a theme. Some experiments were performed at ONR using word frequency analysis on technical papers, reports, and technical program descriptions to develop key words for these documents. The experiments confirmed what the observations suggested: the frequency of appearance of words in a corpus was an excellent method for obtaining the text themes.

D-3-a-ii. Promising Research Opportunities Database

A method was developed [Kostoff, 1991a] to obtain frequencies of appearance of all single, adjacent double, and adjacent triple words in a body of text. The method was applied to a known and modest sized (~600KB) database, a compendium of promising research opportunities for the Navy developed by National Academy of Sciences panels and Navy internal experts. Using the single, double, and triple words from the frequency analysis, and ordering them by frequency as well as alphabetically (by first and second word in the case of word pairs, and by first, second, and third word in the case of word triplets), a clear picture of the pervasive themes (themes that in many cases cut across different disciplines) of the total text emerged. **This computerized scanning of the database provided a starting point for the development of technical guidance, which was eventually sent to members of the Navy research management community**.

D-3-a-iii. Industrial R&D (IR&D) Database

The method was then applied to a much larger database (~15MB), the list and description of IR&D projects supported by the Department of Defense. There were nearly 8,000 projects in this database circa 1990, and just reading the project descriptions, much less synthesizing and integrating their themes, would be a monumental task. The purpose of the application was twofold: First, using the overall technology taxonomy provided by the government to the principal investigators for categorization of their programs, determine whether the distributions within the taxonomy categories generated from the 'bottom-up' multiword frequency approach matched those obtained from principal investigator-supplied data; Second, determine whether a 'natural' taxonomy of the IR&D database, using no predetermined structure but only an orthogonalization of the themes resulting from the multiword frequency analysis, could be obtained.

For the first part of the problem, the main themes from the multi-word frequency analysis were placed in the appropriate taxonomy bins generated by the government. The distribution within the bins was obtained by summing the frequencies of the words in the bins. The conclusion drawn was that the **multiword frequency approach gave a taxonomy distribution similar to that using contractor-supplied data**, as well as could be expected given the subjectivity of characterizing both forms of data.

The second analysis, determination of a 'natural' taxonomy, showed that a 'bottom-up' taxonomy could be generated rapidly from the multiword frequencies. The conclusion drawn was that, **where a database exists and a new or modified taxonomy is desired, the multiword frequency approach provides an excellent starting point**.

D-3-a-iv. Practical Considerations

A number of practical lessons resulted from working with the above and other large databases. The raw data output is huge; a major problem is to reduce this mass of data to a manageable and interpretable form. One favorable characteristic of the raw data is that if frequency of occurrence is plotted on the abcissa, and number of words (or pairs or triplets) of a given frequency is plotted on the ordinate, then the plot of number of words vs. frequency is linear on semilog paper, especially at the lower frequencies. The bulk of the numbers of words are at the lowest frequencies. Eliminating the lowest frequency words (1, 2, and perhaps 3 occurrences) will result in the elimination of over 90% of the original words. Eliminating the nontechnical 'noise' words reduces the number of single words by a factor of two and the double words by a factor of four.

The multiword frequency approach is a powerful tool for making large databases transparent and for identifying pervasive themes within these databases. The quality of the results will be no better than the quality of the database. If the verbiage of the database has been skewed, the relative ordering of the main themes will be affected. Some of the skewing can be normalized from the results when the multiword frequency analysis technique is used with the co-word approach.

D-3-b. Theme Inter-relationships

The first part of this subsection will describe the theoretical background of the approach for determining theme inter-relationships and the methodology employed, and the second part will present some of the computational and analytic results.

D-3-b-i. Theory and Methodology

Once the main themes have been identified, the second step in using the new co-word approach is determining the quantitative and qualitative relationships among the main themes and their secondary themes. **The basic hypothesis underlying this step is that the closer words are to each other physically in the text, and the more times they co-occur, the stronger are their relationships**. Thus, unlike the other co-word (or co-citation or co-nomination) approaches that exist, in which co-occurrence frequency in a rather large domain (typically a published paper) is the sole indicator of the strength of the relationship between two words, **the present approach requires physical closeness of the words in addition to co-occurrence**. Two approaches that exploit the physical closeness of words in the text have been used to determine word relationships. These approaches have some degree of overlap.

Both of these approaches compute co-occurrences of words in relatively small physical domains. There were a number of choices for selecting the type and size of domain. The domain could have been restricted to the sentence in which the theme word occurs, as in McKinnon's case [McKinnon, 1977]. Since one object of the present approach is to look for relations between themes and concepts, limiting the domain to a sentence would be overly constrictive. **Related themes and concepts can certainly transcend sentences**. The domain could have been limited to a paragraph but again, depending on the author's style of writing, **related themes and concepts can transcend paragraphs**. It was decided to bound the domain by a specified number of words M away from the theme word, with the value of M to be determined by sensitivity studies. Some error will be introduced when the domain within M words from the theme word includes syntactic markers that separate unrelated concepts. It was felt that this error would have little impact on high frequency word associations; it could introduce spurious low frequency word associations that would fall below the filter thresholds and be eliminated; and having this small error was in any case far preferable to artificially limiting the domain to one with syntactic markers as boundaries.

D-3-b-i-A. Word Co-occurrence Frequencies - Nonzoom

In the first approach, hereafter called the nonzoom, the high frequency nontrivial words obtained by the multiword frequency analysis are assumed to be the major themes of the text database and are input into an algorithm. The algorithm computes co-occurrences of the high frequency words within a region of M words about the first high frequency word, and repeats this process around each high frequency word. An example will be instructive.

Assume one hundred high frequency words (single, double, triple) have been identified and their co-occurrence matrix is desired. Assume the first high frequency word whose co-occurrences will be computed is COMPOSITES, and assume the region M is 50 words. The algorithm goes to the first occurrence of COMPOSITES in the text, and creates a word frequency list of the subset of the hundred high frequency words that occur within 50 words of COMPOSITES. Then, the algorithm proceeds to the next occurrence of COMPOSITES in the text and adds the high frequency words within 50 words of COMPOSITES to the existing word frequency list. The process is repeated for each occurrence of COMPOSITES in the text, and the end result is a list of the co-occurrences (the word frequencies) of the high frequency words with COMPOSITES. The process used for COMPOSITES is then repeated for each of the other 99 high frequency words, and the final result is the complete co-occurrence matrix for the 100 high frequency words. Once the co-occurrence matrix has been obtained, then one of the standard mapping algorithms such as LEXIMAPPE (developed by the French research group under Michel Callon) [Callon, 1991b] can be used to generate clusters and track their evolution with time.

D-3-b-i-B. Word Co-occurrence Frequencies - Zoom

In the second approach, hereafter called the zoom, the high frequency words obtained from the multiword frequency approach are input to the algorithm. The algorithm computes frequencies of occurrence of all words (above a preselected frequency cutoff level) within a region of M words of each high frequency word. Thus, the results of the second approach include the results of the first approach plus lower co-occurrence frequency words in the region of extent M. An example will again be instructive.

Assume the same 100 high frequency words as in the first example have been identified by the multiword frequency approach. The algorithm goes again to the first word COMPOSITES. It then creates a list of all words that occur within 50 words of COMPOSITES. The algorithm proceeds to the next occurrence of COMPOSITES and repeats the process, adding all words within 50 words of COMPOSITES to the existing list. When the algorithm has repeated this process over all occurrences of COMPOSITES, the word list has become a word frequency dictionary of all words that occurred within 50 words of COMPOSITES. Obviously, a subset of this dictionary is the list of high co-occurrence frequency words of the first approach.

In practice, the second approach is much more useful than the first approach. It allows a detailed examination of specific research or technology areas and, especially, identification of those technologies that may not be major themes of the total database, but that are strongly supportive of the major themes. The second approach, unlike the orthodox co-word analysis approach, provides **research/ technology clusters directly without need for a mapping algorithm**. If the co-occurrence relationships among the words in a zoom are desired, the words in the zoom can be input to the non-zoom algorithm, and a co-occurrence matrix obtained. Mapping could be done using this matrix as input.

As in the orthodox co-word analysis, quantitative indicators may be assigned to the clusters to provide measures of cluster density and centrality. However, given the richness of detail and the fine structure resolution achieved with this form of zoom, **it has been found in practice that the quantitative measures (although very useful) provide a fraction of the information that can be obtained from examination of the word frequency dictionary for each cluster**. The quantitative measures are used as a starting point for the interpretation of the cluster data. More important are the patterns of relationships among words within and between the clusters and the messages provided by these patterns.

D-3-b-i-C. Application to Scanning Promising Research Opportunities Database

The new co-word analysis technique has been applied to a moderate sized database (~1.2MB). A description of the process and some results will now be presented.

The database was generated as follows. The Office of Naval Research (ONR) commissions the National Academy of Sciences (NAS) to convene 15 panels on a triennial basis to identify promising research opportunities in 15 areas of interest to ONR (Math, Physics, Chemistry, etc.). Each of these panels writes a short (~10-20 pages) report describing the promising opportunities. ONR also has 15 in-house experts who provide annual status reports (Research Planning Memoranda - RPMs) on these 15 research areas, and these short (~10 pages) reports include research forefronts, promising research opportunities, and research requirements based on naval needs. The NAS reports and the RPMs were combined into one database, and this database represents the major documented source of promising research opportunities provided to corporate ONR.

The purpose of selecting this database is to identify pervasive research thrusts, which in many cases transcend particular disciplines and can only be found within an integrated picture of research. In addition, identification of the relationships among these thrusts is desired to see what multidisciplinary thrusts are emerging. A serious question in this co-word approach, as in all co-word approaches, is how representative of each technical discipline is the volume of words describing each discipline. Also, how will a mismatch between the amount of verbiage actually describing a discipline, and the amount of verbiage that should describe the discipline, impact the final co-word results?

ONR administers its research program by the 15 technical disciplines, but the funding for each discipline varies from a few million dollars per year for Radiation Sciences to tens of millions for Ocean Sciences. Should the amount of verbiage in the database for the purpose of co-word analysis be about the same for each of the 15 disciplines, should it be proportional to funding of each discipline, or should some other measure be used? The approach taken in this particular instance was to use the database as written, but to remain aware of these potentially biasing issues when analyzing and interpreting the data. While most of the reports in the database tended to be similar in length, as mentioned above, there was one anomoly. The NAS Ocean Sciences and Ocean Geophysics report, whose development was managed by a separate group within the NAS, was about three times the length of the average NAS report. This extra verbiage resulted in more Ocean Sciences thrust areas identified above the frequency cutoff point.

A multiword frequency analysis was then performed on this database. The algorithm used was upgraded from the one described previously [Kostoff, 1991a] and employed a binary tree search approach. Both the zoom and nonzoom analyses were performed using the high frequency words. Since the nonzoom analysis could be recovered as a special case of the zoom analysis, most of the runs made were for the zoom analysis.

D-3-b-i-D. Double Counting

One problem was the existence of double counting. Suppose, for example, COMPOSITES was the high frequency word being analyzed. Suppose further, at some point in the text, there are two occurrences of COMPOSITES 50 words apart. If M is 50, then all the words between the two occurrences will get counted twice, once in the forward search around the earlier COMPOSITES, and once in the backward search around the later COMPOSITES. Since it is likely that the intermediate word related to only one of the COMPOSITES words, then the frequency of the intermediate word is being inflated artificially by this process.

To eliminate this effect, double counting was disallowed. In the COMPOSITES case with M of 50, after each word within 50 words of COMPOSITES was placed in the dictionary list, it was tagged. If a search around a later COMPOSITES within 50 words of the tagged word attempted to relist the tagged word, it was disallowed.

D-3-b-i-E. Input Words of Different Frequencies

The high frequency words input to the zoom algorithm were single, double, and triple words. On average, the high frequency single words occured about an order of magnitude more than the high frequency double words, and the high frequency double words occured about a factor of five more than the high frequency triple words. The size of the co-occurrence frequency dictionary is proportional to the frequency of the word being zoomed. Thus, for a given size of M, the co-occurrence frequency dictionary created around a high frequency single word would be about an order of magnitude larger [in terms of the product (number of words)x(word frequency)] than the co-occurrence frequency dictionary created around a high frequency double word (for a given cutoff threshold), and another factor of five larger than the co-occurrence frequency dictionary created around a high frequency triple word. At the same time, the information contained in a double word is substantially larger than that contained in a single word, there is added information in proceeding to a triple word, so that the smaller co-occurrence frequency dictionaries for double and triple high frequency words are clustered around more specific areas.

If the same value of cutoff frequency is used to limit the co-occurrence frequency dictionary size for each word being zoomed, then the double and especially the triple co-occurrence dictionaries run into limitations for the modest size database used here. The triple word co-occurrence dictionaries consist essentially of single words with low frequencies of occurrence. The double word co-occurrence dictionaries contain some double words with low to very low frequencies of occurrence. Questions of statistical meaning arise for double words in a co-occurrence dictionary with a frequency of 1 or 2. The single words have a sufficiently large co-occurrence dictionary to have viable representation from double and triple words. A database of at least 3 or 4 MB would yield a much richer co-occurrence dictionary for the high frequency double, and especially triple, words. Modification of the algorithm to allow for different cutoff frequencies for single, double, and triple words has been completed.

D-3-b-i-F. Cluster Formation

For each zoom around a high frequency word, the highest co-occurrence frequency single, double, and triple words in the dictionary were collected on a spreadsheet to form a cluster. Since the focus of this co-word technique is to form clusters whose members are **strongly related** to the high frequency word being zoomed, some filtering mechanism is required to remove words from the cluster that are weakly related to the central theme (high frequency word), independently of whether these cluster members are high frequency or not. Two normalization indices are used as a filter.

D-3-b-i-G. Normalization Indices

The Equivalence index [Callon, 1991], Eij, is written as: Eij=(Cij/Ci)*(Cij/Cj), where Ci is the frequency of word i, Cj the frequency of word j, and Cij the co-occurrence of the two. The individual frequencies of the terms in the two word pairs are used to normalize the co-occurrence count. More precisely, the coefficient will measure the probability of i being simultaneously present in a document set indexed by j and, inversely, the probability of j if i, given the respective database frequencies of the two terms. It is for this reason that the coefficient is sometimes called a coefficient of Mutual Inclusion [Turner, 1988]. It identifies the immediate proximity relationships of a word in a file and thereby avoids favoring any particular zone of the word frequency distribution curve.

When this index has a (relatively) high value, both components, Cij/Ci and Cij/Cj, tend to have relatively high values. However, there are groups of words in each cluster, typically double or triple words, whose absolute frequencies are well below those of the single words, but whose co-occurrence frequencies are a sizeable fraction of their frequency of occurrence. In other words, most of the time these words appear in the text, they co-occur with the theme word. However, because their co-occurrence frequencies are low in absolute value, and low in relation to the high frequency of occurrence of the theme word, their Mutual Inclusion coefficients tend to be low. Since these words are strongly tied with the theme (but the reverse is usually not true), it would be useful to keep them in the cluster and not have them filtered out by the Equivalence index Eij.

The second normalization index, sometimes called the Inclusion index, Iij (see section D-2-c-ii), is the conditional probability of i given j, and is expressed as: Iij=Cij/Ci, where Ci is a cluster member and not the theme word. Words that fail to pass through the first (high Eij value) filter, but have high values of Iij, are retained. Using Eij and Iij as numerical filters, the relatively uncoupled words in the initial cluster co-occurrence dictionaries can be discarded, the highly coupled double and triple words will be retained, and the remaining words will be closely integrated with the main theme. Following this procedure, a number of numerical measures can be computed to describe cluster properties.

D-3-b-i-H. Measures of Cluster Properties

The cluster density [Turner, 1988], a measure of the cohesiveness of the cluster, is computed as the average of the Equivalence indices, Eij, of selected members of the cluster. Only those cluster members whose Equivalence indices are larger than a pre-determined threshold, or whose Inclusion indices are larger than some threshold, are included for purposes of computing the density. This means that the cluster members whose value of Eij is small, but whose value of Iij is large, will have their Eij values included in the density computations. In a 1991 orthodox co-word analysis study [Callon, 1991a], the cluster density was calculated as the mean value of its internal links, in agreement with the present method.

The cluster centrality [Turner, 1988], a measure of the strength of the linkages between the cluster being analyzed and other clusters, is also computed using only those cluster members whose Equivalence indices are larger than a threshold, or whose Inclusion indices are larger than some preset threshold. The measure used to describe cluster centrality should have two main characteristics: It should increase as the cluster of interest has greater overlap (more cluster members in common) with a greater number of other clusters, and it should increase as the co-occurrence strength of the overlapping cluster members increases. However, in the simple case of two clusters that have one word in common, and this word has high co-occurrence strength in one cluster and low co-occurrence strength in the second cluster, the strength of the linkage between the clusters would be closer to the low co-occurrence strength than the high occurrence strength. This is analogous to a chain being no stronger than its weakest link.

In the Callon study cited above, centrality of a cluster was calculated as the mean of the Equivalence indices of the first six links with other clusters. This means that if two clusters were compared, and one had 10 links with other clusters, while the second had six links with other clusters, and all links had equal Equivalence index values, then the two clusters would have the same value of centrality. Since one would expect that the cluster with a greater number of external linkages would be more central to the network, and give a greater value for centrality, another measure is required that would satisfy this limit.

A different method of computing centrality was devised. An Overlap index for each externally-linked word in the cluster was computed by summing the value of Eij for each of these words over every cluster in which they appeared. For example, if COMPOSITES appeared in three different clusters, and had Eij values of .1, .2, .3 in these clusters, then the Overlap index for COMPOSITES would be (.1+.2+.3), or .6. Centrality of the cluster (CEN1) was then computed as the sum of the Overlap indices for each externally-linked member of the cluster. This measure satisfied the initial requirements of increasing as numbers of external links increase and increasing as strength of the links increase, but the measure also took into account the overlapping link values in the adjacent clusters. However, this measure of centrality favored large clusters over small clusters. Two other measures of centrality, which normalized on cluster size, were also examined. The first was the ratio of CEN1 to the total number of links in the cluster, and the second was the ratio of CEN1 to the number of links in the cluster that had overlap with other clusters.

D-3-b-i-I. Qualitative Cluster Studies

In addition to the quantitative studies of the clusters and thrust areas, qualitative studies were also performed. The thrust subareas in each cluster were examined, and patterns that emerged were analyzed and interpreted. Another significant advantage of the direct text analysis of the present method became apparent during the qualitative cluster analysis. If a word, or words, appeared in the cluster, and the relationship of this word to the cluster theme was not obvious, the context of this word in the body of the text was examined. A text retrieval software package, ZyINDEX, was used to see how the word in question related to the theme word within the text by examining every occurrence of the word in the text. Then a decision was made as to whether the relation of the word to the theme was spurious, or whether a unique tie to the theme actually existed. **Other co-occurrence methods that use key words or citations do not have the capability for this type of validation analysis.**

D-3-b-i-J. Indirect Impacts of Cluster Members

A further analysis was identification of indirect impacts by members of one cluster on another cluster. Assume that cluster A has a number of high frequency subareas (A1, A2,..), and the impact of these subareas on the advancement of the technical discipline of cluster A can be ascertained. Suppose subarea A1 is also a cluster, and it has a subarea A11 identified in its cluster listing upon which it depends heavily for advancement. Then the advancement of cluster A has a strong indirect dependence on subarea A11 through the direct dependence of cluster A on subarea A1. The construction of the spreadsheet that contains the clusters and subsequent analysis allowed these indirect, but important, impacts of one subarea on another cluster(s) to be determined.

D-3-b-ii. Analysis and Results

D-3-b-ii-A. Multiword Frequency Analysis of Promising Research Opportunities Database

A single, double, and triple word frequency analysis was performed on the ONR promising research opportunities database. The 20 highest frequency technical single words, and 30 highest frequency double words, were defined as themes and extracted, and zooms were performed about each one of these themes. These 50 themes are listed in Table 1. Triple words were not chosen because of the relatively small size of the text database and the consequent relatively small number of high frequency words in the dictionary.

Since the 50 theme words selected drove the remainder of the analysis and results in this section, their selection and impact on the zoom process will be discussed further. The multiword frequency analysis provided thousands of single, double, and triple words, and hundreds of these could have been classified as high frequency and used for the zoom.

One criterion was that the words selected as themes would have technical content. Thus, a word such as ACOUSTIC would be a candidate, whereas BASIC would not. Identification of technical content themes was not a problem for double words and, in most cases, was not a problem for single words.

Another set of criteria related to the total number of words chosen as themes, the number of single and double words in that total, and the cutoff frequency of words chosen. One goal of the selection process was to insure that the very highest frequency words were chosen as themes, at a minimum. The frequency of absolute occurrence for the single words chosen ranged from 592 for MATERIALS to 180 for MODELING. Obviously, while the highest frequency single words were selected, many other high frequency words were available as themes, and other considerations dominated the single word cutoff. The frequency of absolute occurrence for the double words chosen ranged from 88 for REMOTE SENSING to 18 for a number of different words. Thus, while more double words could have been chosen as zoom themes, far less high frequency double words remained than single words.

Based on the analysis reported in this section, the major impact on this study was the choice of single word themes vis a vis double word themes. A mix of single and double word themes was desired, to ascertain how the contents of the word frequency dictionaries resulting from single word themes differed from those of the double word themes. It was obvious that the double word themes were far more focused than the single word themes, but it was felt that a broader range of supporting fields and potentially interesting, but less obvious, thematic relations might emerge from selection of single word themes. The double word themes were chosen initially, and the 30 highest frequency word themes appeared to constitute a fairly diverse first layer of interest. The remaining 20 themes were then chosen from the high frequency single words.

D-3-b-ii-B. Selection of Window Size Around Theme Words

The size and content of the word frequency dictionaries resulting from the zoom around each theme word were determined by the size of M chosen (M is the extent of the region around each theme word in the text from which the dictionary's single, double, and triple words were chosen). When M is very small, two effects occur. Many words in the dictionary will be related syntactically to the theme word, and the total number of words in the dictionary will be small. When M is very large, two opposite effects occur. Many words in the dictionary will be weakly related to the theme word, and the total number of words in the dictionary will be large.

Some initial experiments were performed where M was varied, and it was found, in order to get reasonable word count statistics for the lower frequency double word themes, M values on the order of 40 or 50 words were necessary. While M values in this range include words syntactically related to the theme word, they also include words representing concepts related to the theme, and these different conceptual relations are the desired targets of co-word analysis. Thus, as M increases in size, a broader range of concepts related to the theme word is included, but the bonds between these concepts and the theme word are weaker. The numerical filters employed set a threshold for bond strength.

In further experiments on the size of M, the ratio of co-occurrence frequency of selected high frequency words (in the zoom dictionaries for REMOTE SENSING and SIGNAL PROCESSING) to M was plotted as a function of M. Results showed large gradients between M of 0 to 20 for most of the words, and small gradients for larger M. The interpretation of these results is that the highly bonded words would be very evident with Ms of about 20 to 30. When the results of the above experiments, and the other considerations mentioned above, were taken into account, it appeared that an M of 50 for the database chosen would be a reasonable compromise to satisfy the multiple constraints.

D-3-b-ii-C. Description of Cluster Members

The groups of high frequency words in the dictionary resulting from the zooms around each of the themes are defined as clusters, and the total words (before the numerical filters were used) in the first cluster only are listed in Table 2. A cutoff co-occurrence frequency of three was used to limit words in the clusters.

The columns in Table 2 are defined as the following, going from left to right:

- CL. # is the number of the cluster;
- the next column is headed by the theme of each cluster (enclosed by asterisks) and contains the members of each cluster;
- the third column Cij is the co-occurrence frequency of each cluster member with the cluster theme in the region within 50 words of the theme word;
- Cj is the absolute occurrence frequency of the theme word in the entire database;
- Ci is the absolute occurrence frequency of the cluster member in the entire database; Cij^2/CiCj is the Equivalence index;
- Dpr ij is identical to the value of the Equivalence index for those cluster members that survived the numerical filters, and
- the final column on the right, Cij/Ci, is the Inclusion index.

D-3-b-ii-D. Filter Conditions for Cluster Members

The numerical filter conditions for a cluster member being included in the computations for density and centrality are as follows: If a cluster member's value of Equivalence index (Cij^2/CiCj) is greater than or equal to 67% of the largest value of Equivalence index of a member in the cluster, or if a cluster member's value of Inclusion index (Cij/Ci) is greater than or equal to 50% of the largest value of Inclusion index of a member in the cluster, then that cluster member is included in the computations. Those cluster members in Table 2 with entries in the Dpr ij column survived these filter conditions. Since the filter conditions are arbitrary, all the words chosen for the clusters will be used as data for the **qualitative** analysis of the clusters.

The single word theme clusters contain more of the higher co-occurrence frequency words than the double word theme clusters. This is a direct result of the higher absolute occurrence frequencies of the single word themes, and the consequent larger size of the dictionaries. The single word members of the clusters tend to have two main characteristics: higher values of Equivalence index than the double words, and usually one or more single words are at the top of each cluster. They also are broader and contain less information than the double words and provide a useful first step in identifying subcluster categorizations.

For example, in cluster 1, ACOUSTIC, the first four words in the cluster (highest equivalence index) are the single words PROPAGATION, SCATTERING, OCEAN, and BOTTOM. While they contain far less information than, say, 'WAVEGUIDE INVERSE SCATTERING,' 'OCEANOGRAPHIC SAMPLING NETWORK,' or 'COASTAL TRANSITION ZONE,' they do provide some broad structuring and categorization for the cluster, as well as the potential for broader overlap with other clusters. In fact, the set of single words included in the ACOUSTIC cluster (PROPAGATION, SCATTERING, OCEAN, BOTTOM, SENSORS, ARCTIC, WATER, WAVE, MODELING, ENERGY, DATA) provides a reasonable taxonomy for categorizing the double and triple words contained in the ACOUSTIC cluster. There are probably too many terms in this taxonomy for practical purposes, and the taxonomy is probably not complete (for example, **acoustic sources** are not part of the taxonomy, nor are they mentioned in the double or triple words).

Interestingly enough, the first four single words in the ACOUSTIC cluster appear to cover the main two subthemes within the cluster, namely, **acoustic propagation within the ocean environment**, and **acoustic wave interactions with the boundaries** (mainly ocean bottom).

D-3-b-II-E. Themes Within Clusters

A number of additional analyses were performed on the clusters. The purpose of these studies was to:

- define the next level down (from the theme) of the structure within each cluster,
- define cohesiveness of each cluster,
- identify the relation of each cluster with neighboring clusters, and
- determine the existence and extent of mega-clusters.

In terms of cluster categorization, as a compromise between detail and conciseness, each cluster could be subdivided into from 2 to 4 categories. For those themes that were fairly specific, such as INTEGER PROGRAMMING, subcategorization was straightforward. For those themes that were fairly general, and perhaps ambiguous in meaning, such as a homonym like CURRENT, subcategorization was much more difficult, and an integrated set of categories was in some cases impossible. Usually, though not always, the single word themes were harder to categorize because of the broader implications of the themes. The conclusion to be drawn is that **cluster subcategorization is useful for integrating the disparate members into related topical groups when a focused theme exists, but subcategorization serves less of a purpose when the theme is diffuse.**

D-3-b-ii-F. Cluster Figures of Merit

For further analysis, it is useful to understand how closely integrated are the members of each cluster, and what is the nature of the ties between one cluster and neighboring clusters. A number of figures of merit were defined to describe the cluster characteristics quantitatively. These figures of merit for the top 10 clusters and the bottom 10 clusters are presented in Table 3. The entries in the table are the cluster numbers, the upper half of the table represents the clusters that had the top 10 values of figure of merit, the lower half of the table represents the clusters that had the bottom 10 values of figure of merit, and the columns represent different figures of merit. These figures of merit are defined as follows. For any cluster i:

- TOTLINK is the total number of links (members) in cluster i;
- EXTLINK is the number of links in cluster i with at least one overlap with another cluster;
- EXT/TOT is the ratio of EXTLINK to TOTLINK;

- OVERLAP# is the number of words in other clusters overlapped by words in cluster i;
- OVER/TOT is the ratio of OVERLAP# to TOTLINK;
- OVER/LINK is the ratio of OVERLAP# to EXTLINK and can be viewed as an overlap efficiency per overlapping link;
- CLOVER represents the number of clusters overlapped by cluster i;
- DENS represents the strength of the bonds between the members of cluster i and the theme of cluster i, and is calculated as the average of the Equivalence index for each member of cluster i that passed the numerical filters;
- DENSOVER represents the strength of the bonds between the theme of cluster i and those members of cluster i that have at least one overlap with another cluster; and
- CEN1 is a measure of the centrality of cluster i.

If SUMOVER(j) represents the sum of the Equivalence indices for the members of the clusters overlapped by member j of cluster i (including in the sum the equivalence index for member j), then:

- CEN1 is the sum of SUMOVER(j) over all members j of cluster i that have overlaps with other clusters;
- CEN2 is another measure of cluster centrality and is the ratio of CEN1 to TOTLINK; and
- CEN3 is a measure of centrality efficiency per overlapping link and is the ratio of CEN1 to EXTLINK.

The main figure of merit for cluster density is represented by DENS, column 8, with DENSOVER, column 9, as a supplementary figure.

The first observation is that the top 10 clusters in terms of density, column 8 in the upper half of Table 3, all have double word themes, and the bottom 10 clusters, column 8 in the lower half of Table 3, all have single word themes. This is because the double word themes tend to be more focused and specific, and the members in the cluster that survive the numerical filters tend to be closely related to the theme. The single word themes may have different meanings in different contexts (e.g., CURRENT), and while there may be a broad grouping of terms contained within the cluster, there is no strong bond that ties these members to the cluster theme (on average).

The top 3 clusters in terms of density, (INVERSE SCATTERING (22), INTEGER PROGRAMMING (19), SEA ICE (40)), all have relatively modest absolute frequencies of occurrence of the theme words (about 20), very focused topical areas, relatively few cluster members on average, and tend to have a few members whose very high Inclusion indices and relative occurrence frequencies give them very high Equivalence indices. The bottom 3 clusters in terms of density, (SURFACE (47), MODELS (25), CONTROL (6)), all have very high absolute frequencies of the theme words (~300, or more), extremely broad areas with multiple contexts, and consequently, don't contain members that tend to co-occur often with the theme word. It may be concluded that **many single word themes may be generic thrust areas, but focused thrust areas should have reasonable density values and may require double (or higher) word themes or unique single word themes**.

The main figure of merit for centrality is CEN1, column 10, with CEN2 and CEN3, columns 11 and 12, as supplementary figures of merit. The centrality data is based on the list of words common to more than one cluster, and those words with the highest overlap are shown in Table 4. The first observation on centrality from Table 3 is that the top 10 clusters are split evenly between single and double word themes, while the bottom 10 clusters are almost exclusively double word themes. In general, clusters with broader themes will tend to have more overlaps with other clusters, but the strengths of the overlaps will be modest. The focused double word clusters that have high centrality tend to have fewer overlaps, but some of the overlaps are strongly bonded members with high Equivalence indices. Since the centrality measure used, CEN1, increases with the number of overlaps and the strength of the overlap, a few very strong overlaps (high Equivalence indices of the overlapping members) can outweigh the lack of a large number of overlaps.

D-3-b-ii-G. Generation of Megaclusters

Another approach used to explore centrality was the generation and analysis of megaclusters, or cluster strings. The objective of effectively regrouping themes, or even theme members, to form new types of clusters is to maximize word bonding for all the words extracted from the database; i.e., a global optimization as opposed to a local bonding optimization as applied at the high frequency theme level only. In its purest form, this global optimization is a combinatorial optimization on the total extracted database level. There are many practical ways to generate megaclusters. Two will be discussed: a 'top-down' approach, and a 'bottom-up' approach.

In a 1988 study of the French group [Turner, 1988], the co-occurrence matrix of high frequency words is used as a starting point. Based on the word pair association values, groupings of words are formed whose main links are statistically stronger with one another than with other words in the data file. The size of these clusters is allowed to vary. This process can be viewed as a 'top-down' approach, since the clusters are formed based on high frequency word information.

For the present study, a 'bottom-up' approach was used. This approach started with examination of the overlaps of the members of the high frequency theme clusters; thus, the more numerous mid-frequency words could be included in determination of overlap along with the higher frequency words. A 50 x 50 matrix was generated, with the axes of the matrix being the cluster numbers, and the elements of the matrix being the number of overlaps between any two clusters. As an example, the entry of 3 in the matrix element [2,19] means that clusters 2 (ARTIFICIAL INTELLIGENCE) and 19 (INTEGER PROGRAMMING) have three members in common.

Two levels of strings were identified in the following manner. For the highest level of string creation, a cluster that had at least three overlapping members with any other individual cluster was selected arbitrarily as the starting point. By tracing through the rows and columns of the matrix, all clusters that had three or more overlapping links with the initial cluster were listed. After further tracing, all clusters that had three or more links with any of the listed clusters were added to the list. The process was continued until there were no other clusters linked by three or more overlaps to any of the listed clusters. This completed list constituted the first string. Then the process was repeated by again arbitrarily selecting a cluster that had at least three overlapping members with any other individual cluster, but that was not on the initial list. The process was completed when every cluster that had three or more overlaps with at least one other cluster was listed on one of the strings. The listed strings of clusters constitute the highest level strings.

The second level strings were obtained by extending the above process to a threshold of two overlaps, but with one restriction. Existing strings from the highest level string creation were expanded by adding clusters that had two overlaps with those clusters already in the string. These newly added members of the cluster strings were not used to generate further members through secondary overlaps. New strings were generated without restriction at a level of two overlaps for those clusters not already in a string. The clusters added to each string, and the new strings created, by lowering the overlap threshold, constitute the second level.

At the highest string level, five cluster strings (I-V, below) were produced. At the second string level, strings I-V roughly doubled in size (size being defined as the number of clusters in the string), and two additional strings appeared (VI and VII, below).

The clusters that constitute strings I-VII follow:

I. Primary - ACOUSTIC, DYNAMICS, PHYSICAL OCEANOGRAPHY, REMOTE SENSING, SPACE;
Secondary - INTERNAL WAVES, OCEAN, SHALLOW WATER, DATA, MODELS

II. Primary - ARTIFICIAL INTELLIGENCE, INTEGER PROGRAMMING;
Secondary - COMPUTER SCIENCE

III. Primary - CONTROL, DEVICES, ENERGY, INFORMATION, MATERIALS, PROCESSING, RADIATION, STRUCTURES;
Secondary - CHEMICAL, ENERGETIC MATERIALS, PHYSICAL, ENERGY CONVERSION, INFORMATION PROCESSING, MODELS, NEURAL NETWORKS, OPTICAL, SIGNAL PROCESSING, SURFACE

IV. Primary - CURRENT, DATA, MODELING, MODELS, OCEAN, SURFACE;
Secondary - DYNAMICS, INFORMATION, INTERNAL WAVES, NEURAL NETWORKS, PHYSICAL, PROCESSING, REMOTE SENSING

V. Primary - NEURAL NETWORKS, PATTERN RECOGNITION;
Secondary - INFORMATION, MODELS

VI. Primary - ELECTRONIC DEVICES, THIN FILMS

VII Primary - INTEGRATED CIRCUITS, SOLID STATE

The strings that do not have single words in the primary segment (II, V, VI, VII) have very strongly focused themes and relatively few clusters. They seem to revolve around the themes of electronics, information, and computers. The strings with single words in the primary segment are relatively large, relatively broad, and two of these three strings (I, IV) are related to ocean issues. The third string (III) revolves around materials, but because of the multiple meanings of some of the single word cluster themes, mainly PROCESSING, INFORMATION, and CONTROL, clusters related to information and signal processing are integrated into the string. It appears that if all cluster themes were double words, then smaller but more tightly focused strings would be formed.

There is an interesting pattern to observe in the highest level strings (the primary segments of I-V). Strings I and IV could be broadly categorized as Ocean Sciences, strings II and V as Information Sciences, and string III as Materials. ONR identifies three areas of emphasis in its investment strategy, namely, Ocean Sciences, Materials, and Information Sciences. Thus, the **three broad string areas identified by an analysis of experts' recommendations to ONR of promising research opportunities coincide with ONR's stated areas of emphasis.**

The single link clustering approach described above was the first instance of multi-word technical phrases reported as used for technical text clustering. Prior to the time of this study, only single technical words had been used for technical text clustering. See the dissertation of Oren Zamir (University of Washington) for further discussion on this point [Zamir, 1999].

D-3-c. Applications to IR&D Database

In section D-3-a-iii, results of a multiword frequency analysis applied to the IR&D database were described briefly. The present section describes further studies on the IR&D database, both multiword and co-word zoom analyses. Because of time constraints, development of cluster strings, as was done for the promising opportunities database, remains to be completed for the IR&D database.

For these studies, the FY90 incarnation of the IR&D database, representing about

7400 FY90 programs, was used. Single, double, and triple word frequency analyses of the database were performed. Table 5 contains the highest frequency single words ordered by decreasing frequency; Table 6 contains the highest frequency adjacent double words ordered by decreasing frequency; and Table 7 contains the highest frequency adjacent triple words ordered by decreasing frequency. These tables contain the raw data; thus, there are many non-technical content words in the tables. Table 8 contains the high frequency adjacent double words with the non-technical content words removed.

A careful study of Table 8, supported by studies of Tables 5, 6, and 7, as well as the zoom analyses to be described later, shows at least three major pervasive thrust areas. From Table 8, these areas may be categorized as:

- Information Technology (SIGNAL PROCESSING, CONTROL SYSTEM, DATA BASE, SOFTWARE DEVELOPMENT, NEURAL NETWORK, SYSTEM ARCHITECTURE, ARTIFICIAL INTELLIGENCE, IMAGE PROCESSING, TARGET RECOGNITION, DATA PROCESSING, DATA FUSION, COMPUTER SIMULATION, DATA LINK, SOFTWARE ENGINEERING, PARALLEL PROCESSING,.....),

- Materials Technology (FIBER OPTIC, COMPOSITE MATERIALS, ADVANCED MATERIALS, COMPOSITE STRUCTURES, MECHANICAL PROPERTIES, MATRIX COMPOSITES, THIN FILM, ADVANCED COMPOSITES, METAL MATRIX, ...), and

- Aerospace Technology (FLIGHT CONTROL, PROPULSION SYSTEM, HEAT TRANSFER, WIND TUNNEL, LAUNCH VEHICLE, ENGINE CONTROL, ENGINE TESTING, GAS TURBINE ENGINES,.....).

These pervasive thrust areas are not necessarily independent. As the zoom analyses will show, much of the Materials and some of the Information work is directed towards Aerospace applications.

As Tables 5, 6, and 7 show, the high frequency single words are about an order of magnitude higher in frequency than the high frequency double words, and the high frequency double words are about a factor of three higher in frequency than the high frequency triple words. To perform the zooms about themes of similar levels of description and moderately similar frequencies, only the double (and two triple) words of Table 8 were chosen as zoom theme words. This is in contrast to the selection of single and double zoom theme words that was done for the promising research opportunities database study. Selection of only double words as themes should reduce some of the ambiguities that resulted from the single word themes in the promising opportunities study (especially from the homonyms), and is viewed as an experiment in the data analysis of this technique.

Zooms around each of the sixty words in Table 8 were performed, and the single, double, and triple words in each resulting zoom dictionary were retained. In the cluster overlap analysis to be performed, it is expected that the single, double, and triple words will be retained. While there is insufficient space in this document to present zoom results for all sixty themes, summary results from four zooms will be presented. Table 9 contains zooms about IMAGE PROCESSING, Table 10 contains zooms about ADVANCED MATERIALS, Table 11 contains zooms about PROPULSION SYSTEM, and Table 12 contains zooms about SIGNAL PROCESSING. These zoom tables will now be discussed.

In each of the zooms, a dictionary of all words and their frequencies of occurrence within 50 words of the theme word was constructed. Cutoff frequencies of five for single words, four for double words, and three for triple words were used. Sortings of the lists by different parameters were performed. Three sortings for each of the four zooms are presented.

In Table 9A, the results are contained in six columns. Starting from the lefthand side, the columns are:

- CL# represents the theme number and ranges from one to sixty;
- Cij is the co-occurrence frequency;
- Ci is the absolute occurrence frequency of the theme **member** in the text;
- Cij/Ci is the Inclusion index based on the theme member;
- Cij^2/CiCj is the Equivalence index, where Cj is the absolute frequency of the theme **word** in the text; and the righthand column is the theme **member**.

Sortings were done using three of the column headings as sort parameter: Cij by descending order, Cij/Ci by descending order, and Cij^2/CiCj by descending order. Since Cj is constant for each theme sort, then sorting by Cij for a theme is equivalent to sorting by Cij/Cj. Thus, the sorts were performed by the two Inclusion indices and by the Equivalence index.

The sort in Table 9A is by Cij. As in the previous zooms reported, when sorting is done by absolute co-occurrence frequency, the single words predominate at the high frequency end, since their absolute frequency occurrence in the text is an order of magnitude greater than that of the double words. Thus, Table 9A can be considered as a compendium of the broad sub-thrust areas that constitute IMAGE PROCESSING. The sort in Table 9B is by Cij/Ci. A value of unity, which characterizes the highest ranking theme members, means that whenever that member appears in the IR&D database, it appears within fifty words of the theme, IMAGE PROCESSING. Thus, whenever CORRELATION TRACKER appears in the IR&D database, it appears within fifty words of IMAGE PROCESSING. The words in the high end of this sort tend to be double and triple words. For the most part, they appear rather infrequently in the text, but are very specific terms tied closely to the theme. The sort in Table 9C is by Cij^2/CiCj. Since this Equivalence index is the product of the two Inclusion indices, the ordering that it generates in the sorting process represents a compromise between the orderings of Tables 9A and 9B. Thus, while there are some high frequency single words near the top of the sort, there are still many low frequency high Inclusion index words that predominate at the top of the sort. This parameter appears to combine the most important broad sub-thrust area descriptors with the most important specific supporting areas tied closely to the theme.

Tables 10A, 10B, and 10C contain the same parameter sorts for ADVANCED MATERIALS. As Table 10A shows (AIRCRAFT, ENGINE, TURBINE, LANDING GEAR), and as Table 10B shows more graphically (ROUGH FIELD LANDING GEAR, THRUST BEARING SYSTEM, SPACE LAUNCH VEHICLE, CREW ESCAPE, SPACE TRANSPORTATION VEHICLES, GENERATION AIRCRAFT), there is a strong Aerospace Technology flavor in the members that constitute ADVANCED MATERIALS. This result, which pervades many of the themes examined, supports the earlier statement that much of the Materials Technology themes tends to support Aerospace Technology.

Tables 11A, 11B, and 11C contain the same parameter sorts for PROPULSION SYSTEM. Again, the thrusts seem to be propulsion that supports Aerospace Technology.

Finally, Tables 12A, 12B, and 12C contain the same parameter sorts for SIGNAL PROCESSING. There are some ocean applications mentioned (SURFACE SHIP

SONAR, SURFACE SHIP APPLICATIONS, NON-TRADITIONAL ACOUSTIC PROCESSING, ACOUSTIC INTERCEPT RECEIVER), as was the case with IMAGE PROCESSING. Generally, the Materials Technology themes seem to be much more closely related to the Aerospace Technology application than do the Information Technology themes. This may reflect the more generic nature and applicability of the Information Technology at the development stage. Another interpretation is that the Materials Technology development is defense-requirements driven, while the Information Technology development is market driven primarily and defense-requirements driven secondarily. More analysis is required before this interpretation can be strongly substantiated.

Compared to the promising opportunities database, which is research oriented and whose themes and sub-themes are expressed in the research language, the IR&D database is technology development oriented, and its themes and sub-themes are expressed in the technology development language. The next step in the analysis is to develop cluster strings as was done in the promising opportunities database analysis.

E. **SUMMARY AND CONCLUSIONS**

Analysis of co-occurrence phenomena is useful for identifying research thrusts, connectivities between these thrusts, the evolution of these thrusts and connectivities and, potentially, the determination of research and sponsor impact on these evolutionary trends. In particular, **co-word analysis offers the potential of rapidly identifying research trends from large bodies of textual information.**

The main strengths and weaknesses of co-word analysis relative to co-citation and co-nomination analysis were discussed. Co-word was shown to be a more direct way of identifying research trends than co-citation and a more automated and less labor intensive way of identifying research trends than co-nomination in its present incarnation.

The origins of co-word analysis in computational linguistics were traced, and the development of co-word analysis, as applied especially to research policy, was described in detail. Limitations of both approaches in performing direct text co-word analysis were presented. In a review of an early draft of this report, Dr.

Yaacov Choueka, one of the world's experts in collocations, made the following comments on the different evolutionary paths of co-word analysis as described above: "***It is obvious now that research on co-phenomena in textual databases was pursued in the last decade or so by two different groups of researchers, the first mainly interested in computational linguistics and corpus processing, and the second in evaluating research trends and supporting research policy. Each one of these groups was almost totally unaware of the work of the other one (and this becomes obvious when reference lists in papers originating from the two groups are compared), and therefore methods developed and results obtained in one of these areas were not applied to the other one....you would be doing a good service to both groups by acting as 'common denominator'....***"

A new co-word approach that deals directly with text and requires no indexing or key words was described in detail. The first phase of this approach was identification of thrust areas by multiword frequency analysis of large text databases. Applications of the multiword frequency analysis to support identification of promising research opportunities for ONR and to identification of R&D taxonomies for the Industrial R&D database were presented.

The second phase of this approach was identification of the interrelationships among the thrust areas defined in the first approach and identification of supporting research areas for each thrust area. The main technique described was creation of a dictionary of high frequency words occurring in a physical region of limited extent around each major thrust area identified by the multiword frequency analysis. Filter conditions based on closeness of bonding between the words in the dictionary and the thrust area were used to select important words from the dictionaries and form finite-sized clusters of words around each thrust area. The cohesiveness of each cluster and its central (or isolated) position, relative to that of the other clusters, were calculated with the use of density and centrality measures. Megaclusters, or strings of clusters whose members have significant overlap, were constructed for the promising research opportunities database, and these megaclusters showed concisely the major themes of the total database. These megaclusters were shown to have a direct mapping to ONR emphasis areas selected by an entirely different technique. ***The single link clustering approach used to generate these megaclusters was the first reported use of multi-word technical phrases for technical text clustering.***

While this new co-word approach was shown to work, and to supply a richness of detailed information about research themes and sub-themes unavailable by any other approach, a number of areas in which improvements could be made were identified. These areas are:

***How should synonyms be treated in theme identification;**
***How should synonyms be treated when applying filter conditions;**
***What types of single words should be considered for themes;**
***What types of single words should be included as cluster members;**
***When should clusters be fusioned or fissioned to form aggregated superclusters or fragmented subclusters.**

F. **BIBLIOGRAPHY**

Amsler, B., "Research Towards the Development of a Lexical Knowledge Base for Natural Language Processing," Proceedings of the 1989 SIGIR Conference, Association for Computing Machinery, Cambridge, MA 1989.

Bahl, L., Jelinek, F., and Mercer, R., "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 5, No. 2, March 1983.

Bauin, S., "Aquaculture: A Field by Bureaucratic Fiat,"in: Callon, M., Law, J., and Rip, A., (eds.) "Mapping the Dynamics of Science and Technology", Macmillan Press Ltd., London, 1986.

Bauin, S., Courtial, J.P., Law, J., and Whittaker, J., "Policy and the Mapping of Scientific Change: A Co-word Analysis of Research into Environmental Acidification," Scientometrics, Vol. 14, No. 3, 1988.

Benson, M., Benson, E., and Ilson, R., "The BBI Combinatory Dictionary of English: A Guide to Word Combinations," John Benjamins, Amsterdam and Philadelphia, 1986.

Blau, J.R., "Sociometric Structure of a Scientific Discipline," in: Jones, R.A. (ed.) Research in the Sociology of Knowledge, Sciences, and Art: An Annual Compilation of Research I, JAI Press, Greenwich, CT, 1978.

Braam, R., Moed, H., and Van Raan, A., "Mapping of Science by Combined Co-Citation and Word Analysis. 1. Structural Aspects," Journal of the American Society for Information Science, 42 (4), 1991a; "Mapping of Science by Combined Co-Citation and Word Analysis. 2. Dynamical Aspects," Journal of the American Society for Information Science, 42 (4), 1991b.

Callon M., Courtial, J. P., and Turner, W. A., "PROXAN: A Visual Display Technique for Scientific and Technical Problem Networks," Second Workshop on the Measurement of R&D Output, Paris, France, December 5-6, 1979.

Callon, M., Courtial, J.P., Turner, W.A., and Bauin, S., "From Translations to Problematic Networks: An Introduction to Co-word Analysis," Social Science Information, 22, 1983.

Callon, M., "Pinpointing Industrial Invention: An Exploration of Quantitative Methods for the Analysis of Patents," in: Callon, M., Law, J., and Rip, A., (eds.) "Mapping the Dynamics of Science and Technology", Macmillan Press Ltd., London, 1986.

Callon, M., Courtial, J.P., and Laville, F., "Co-word Analysis as a Tool for Describing the Network of Interactions between Basic and Technological Research: The Case of Polymer Chemistry," Scientometrics, Vol. 22, No. 1, 1991a.

Callon, M., Courtial, J.P., Crance, P., Laredo, P., Mauguin, P., Rabeharisoa, V., Rocher, Y.A., and Vinck, D., "Tools for the Evaluation of Technological Programmes: an Account of Work Done at the Centre for the Sociology of Innovation," Technology Analysis and Strategic Management, Vol. 3, No. 1, 1991b.

Chomsky, N., "Aspect of the Theory of Syntax," MIT Press, New York, NY, 1965.

Choueka, Y., Klein, T., and Neuwitz, E., "Automation Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus," ALLC Journal, Vol. 4, 1983.

Choueka, Y., "Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases," Proceedings of the RIAO

Conference on User-Oriented Content-Based Text and Image Handling, Cambridge, MA, March 21-24, 1988.

Church, K., and Hanks, K., "Word Association Norms, Mutual Information, and Lexicography," Proceedings of the 27th Meeting of the Association of Computational Linguistics, Vancouver, BC, 1989.

Church, K., and Gale, W., "Poor Estimates of Context are Worse Than None," DARPA Speech and Natural Language Workshop, Hidden Valley, PA, June 1990.

Courtial, J. P., "Technical Issues and Developments in Methodology," in: Callon, M., Law, J., and Rip, A., (eds), Mapping the Dynamics of Science and Technology: Sociology of Science in the Real World, the MacMillan Press Ltd, London, England, 1986.

Courtial, J.P., and Law, J., "A Co-word Study of Artificial Intelligence," Social Studies of Science, Vol. 19, 1989.

Courtial, J.P., Turner, W.A., and Michelet, B., "Scientific and Technological Information Banks for Research Networks Management Purposes," Research Policy, Vol.19, 1990a.

Courtial, J.P., "A Mathematical Model of Development in a Research Field," Scientometrics, Vol.19, No.1, 1990b.

De Saussure, F., "Cours de Linguistique Generale," 4eme Edition, Librairie Payot, Paris, 1949.

Firth, J., "A Synopsis of Linguistic Theory 1930-1955", in Studies in Linguistic Analysis, Philological Society, Oxford; Reprinted in Palmer, F. (ed), Selected Papers of J. R. Firth, Longman, Harlow, 1968.

Franklin, J. J. and Johnston, R., "Co-citation Bibliometric Modeling as a Tool for S&T Policy and R&D Management: Issues, Applications, and Developments," in Van Raan, A.F.J. (ed.), Handbook of Quantitative Studies of Science and Technology, North Holland, 1988.

Garfield, E., Malin, M.V., and Small, H., "Citation Data as Science Indicators," in:

Elkana, Y., Lederberg, J., Merton, R.K., Thackray, A., and Zuckerman, H., (eds), Toward a Metric of Science: The Advent of Science Indicators, John Wiley and Sons, New York, 1978.

Georghiou, L., Giusti, W.L., Cameron, H.M. and Gibbons, M., "The Use of Co-nomination Analysis in the Evaluation of Collaborative Research," in Van Raan, A.F.J. (ed.), Handbook of Quantitative Studies of Science and Technology, North Holland, 1988.

Halliday, M. A. K., "Lexis as a Linguistic Level," in: Bazell et al (eds), In Memory of J. R. Firth," Longmans Linguistic Library, London, England, 1966.

Harris, Z., "Mathematical Structures of Language," New York, Wiley, 1968.

Healey, P., Rothman, H., and Hoch, P., "An Experiment in Science Mapping for Research Planning," Research Policy, Vol. 15, 1986.

Hornby, A.S., Gatenby, E. V., and Wakefield, H., "Idiomatic and Syntactic English Dictionary," Kaitakusha, Tokyo, Japan, 1942.

Iordanskaja, L., Kittredge, R., and Polguere, A., "Lexical Selection and Paraphrase in a Meaning-text Generation Model," in: Paris, C. et al (eds), "Natural Language Generation in Artificial Intelligence and Computational Linguistics," Kluwer Academic Publishers, 1990.

Kittredge, R., Polguere, A., and Goldberg, E., "Synthesizing Weather Forecasts from Formatted Data," Proceedings of the 11th COLING, Int'l Conference on Computational Linguistics, 1986.

Kostoff, R. N., "Word Frequency Analysis of Text Databases", ONR Memorandum 5000 Ser 10P4/ 1443, April 12, 1991a.

Kostoff, R. N., "A Quantitative Approach to Determining the Impact of Research", Presented at: Twenty-Second Annual Pittsburgh Conference on Modeling and Simulation, May 2-3, 1991b.

Kostoff, R. N., "Database Tomography: Multidisciplinary Research Thrusts from Co-Word Analysis," Presented at Portland International Conference on

Management of Engineering and Technology, October 27-31, 1991c.

Kostoff, R. N., "Research Impact Quantification," R&D Management, 24:3, July 1994.

Leydesdorff, L., "Various Methods for the Mapping of Science," Scientometrics, Vol. 11, 1987.

Leydesdorff, L., "Words and Co-words as Indicators of the Intellectual Organization of the Sciences," Presented to the EASST workshop in Amsterdam, December 1987.

Leydesdorff, L., "Words and Co-words as Indicators of Intellectual Organization," Research Policy, Vol. 18, 1989.

Libbey, M. and Zaltman, G. "The Role and Distribution of Written Informal Communication in Theoretical High Energy Physics," American Institute of Physics, New York, 1967.

Maarek, Y. S., and Smadja, F. A., "Full Text Indexing Based on Lexical Relations, An Application: Software Libraries," Proceedings of the 12th International SIGIR, Conference on Research and Development in Information Retrieval, Cambridge, MA, 1989.

Mays, E., Damerau, F., and Mercer, R., "Context Based Spelling Correction," IBM Nat'l Language ITL, IBM, Paris, France, Mar 1990.

McCardell, R., "Lexical Selection for Natural Language Generation," Technical Report, Computer Science Department, Univ of Md, 1988.

Melcuk, I. A., "Meaning-Text Models: A Recent Trend in Soviet Linguistics," The Annual Review of Anthropology, 1981.

McDonald James E., Plate, Tony A., and Schvaneveldt, Roger W., "Using Pathfinder to Extract Semantic Information From Text," in: Schvaneveldt, Roger W.,ed., Pathfinder Associative Networks: Studies in Knowledge Organization, Ablex Publishing Corp., 1990.

McKinnon, A., "From Co-occurrences to Concepts," Computers and the Humanities, Vol. 11, Pergamon Press, 1977.

Moed, H.F. and Van Raan, A.F.J., "Indicators of Research Performance: Applications in University Research Policy," in: Van Raan, A.F.J., ed., Handbook of Quantitative Studies of Science and Technology, North Holland, 1988

Mullins, N., Snizek, W., and Oehler, K., "The Structural Analysis of a Scientific Paper," in: Van Raan, A.F.J., ed., Handbook of Quantitative Studies of Science and Technology, North Holland, 1988

Nirenburg, S., "Lexicon Building in Natural Language Processing," Program and Abstracts of the 15th International ALLC, Conference of the Association for Literary and Linguistic Computing, Jerusalem, Israel, June 1988.

Oberski, J. E. J., "Some Statistical Aspects of Co-citation Cluster Analysis and a Judgement by Physicists," in Van Raan, A.F.J. (ed.), Handbook of Quantitative Studies of Science and Technology, North Holland, 1988.

Peters, H. and Van Raan, A., "Co-Word Based Science Maps of Chemical Engineering," Research Report to the Netherlands Foundation for Technological Research (CWTS-91-03), April 1991.

Rip, A., and Courtial, J.P., "Co-word Maps of Biotechnology: An Example of Cognitive Scientometrics," Scientometrics, Vol. 6, No. 6, 1984.

Rip, A., "Mapping of Science: Possibilities and Limitations," in:
Van Raan, A.F.J., ed., Handbook of Quantitative Studies of Science and Technology, North Holland, 1988

Salton, G., and McGill, M. J., "Introduction to Modern Information Retrieval," Computer Series, McGraw Hill, New York, NY, 1983.

Salton, G., "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer," Addison-Wesley Publishing Company, New York, NY, 1989.

Shrum, W. and Mullins, N., "Network Analysis in the Study of Science and

Technology," in: Van Raan, A.F.J., ed., Handbook of Quantitative Studies of Science and Technology, North Holland, 1988

Smadja, F., "Lexical Cooccurrence: The Missing Link in Language Acquisition," 15th International ALLC, Conference of the Association for Literary and Linguistic Computing, Jerusalem, Israel, 1988.

Smadja, F., "Macrocoding the Lexicon with Cooccurrence Knowledge for Language Generation," Columbia University, Computer Science Department, Technical Report CUCS-630-89, 1989.

Smadja, F., "Extracting Collocations from Text. An Application: Language Generation," PhD Thesis, Columbia University, 1991.

Small, H., "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents," Journal of the American Society for Information Science, 24, July/August, 1973.

Small, H., "A Co-citation Model of a Scientific Speciality: A Logitudinal Study of Collagen Research," Social Studies of Science, 7, 1977.

Small, H., "Cited Documents as Concept Symbols," Social Studies of Science, 8, 1978.

Small, H. and Greenlee, E., "Citation Context Analysis of a Co-citation Cluster: Recombinant DNA," Scientometrics, 2, 1980.

Small, H. and Sweeney, E., "Clustering the Science Citation Index using Co-citations," Scientometrics, 7, 1985a.

Small, H., Sweeney, E., and Greenlee, E., "Clustering the Science Citation Index using Co-citations. II. Mapping Science," Scientometrics, 8, 1985b.

Small, H. and Greenlee, E., "Collagen Research in the 1970s," Scientometrics, 10, 1986.

Sparck Jones, K., "Automatic Keyword Classification for Information Retrieval," Butterworths, London, 1971.

Sparck Jones, K. and Tait, J. I., "Automatic Search Variant Generation," Journal of Documentation, Vol. 40, No. 1, March 1984.

Turner, W.A., Chartron, G., Laville, F., and Michelet, B., "Packaging Information for Peer Review: New Co-Word Analysis Techniques," in: Van Raan, A.F.J., ed., Handbook of Quantitative Studies of Science and Technology, North Holland, 1988

Van Raan, A. and Tijssen, R., "The Neural Net of Neural Network Research: An Exercise in Bibliometric Mapping," Centre for Science and Technology Studies, University of Leiden, To Be Published.

Van Rijsbergen, K., "Information Retrieval," 2nd Edition, Butterworths, London, 1979.

Whittaker, J., "Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis," Social Studies of Science, Vol. 19, 1989.

Zamir, O., "Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results", Ph. D. Dissertation, University of Washington, 1999.

G. **SUPPLEMENT TO BIBLIOGRAPHY**

This Technical Report was written in the early 1990s. Since that time, the new co-word method described in the report has been used extensively for a number of text mining studies. For the reader interested in pursuing the Database Tomography methodology and its text mining applications further, the following text mining-related references are suggested.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography". International Journal of Bifurcation and Chaos. In Press.

Kostoff, R. N. "Text Mining for Global Technology Watch". In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. 2003. Vol. 4. 2789-2799.

Kostoff, R. N. "Stimulating Innovation". International Handbook of Innovation. In Press.

Kostoff, R. N. "Bilateral Asymmetry Prediction". Medical Hypotheses. August 2003.

Kostoff, R.N. "Role of Technical Literature in Science and Technology Development." Journal of Information Science. In Press.

Hartley, J. and Kostoff, R. N. "How Useful are 'Key Words' in Scientific Journals?" Journal of Information Science. October 2003.

Kostoff, R. N. "The Practice and Malpractice of Stemming". JASIST. 54: 10. June 2003.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". Fractals. December 2003.

Kostoff, R. N., Shlesinger, M., and Malpohl, G. "Fractals Roadmaps using Bibliometrics and Database Tomography". SSC San Diego SDONR 477, Space and Naval Warfare Systems Center. San Diego, CA. June 2003.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power: Military Requirements and Literature Structure." Academic and Applied Research in Military Science. In Press.

Kostoff, R. N. "Data – A Strategic Resource for National Security". Academic and Applied Research in Military Science. In Press.

Kostoff, R. N. "Disruptive Technology Roadmaps". Technology Forecasting and Social Change. In Press.

Kostoff, R. N., and DeMarco, R. A. "Science and Technology Text Mining: Analytical Chemistry". DTIC Technical Report Number ADA?????? In Press.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Science and Technology Text Mining: Citation Mining of Dynamic Granular Systems." DTIC Technical Report Number ADA?????? In Press.

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Science and Technology Text Mining: Electrochemical Power." DTIC Technical Report Number ADA?????? In Press.

Kostoff, R. N. "Science and Technology Text Mining: Cross-Disciplinary Innovation". DTIC Technical Report Number ADA?????? In Press.

Losiewicz, P., Oard, D., and Kostoff, R. N. "Science and Technology Text Mining: Basic Concepts". DTIC Technical Report Number ADA?????? In Press.

Kostoff, R. N. "Science and Technology Text Mining: Global Technology Watch". DTIC Technical Report Number ADA?????? In Press.

Kostoff, R. N., Andrews, J., Buchtel, H., Pfeil, K., Tshiteya, R., and Humenik, J. A. "Text Mining and Bibliometrics of the Journal Cortex". Cortex. Invited for Publication.

Kostoff, R.N., Del Rio, J. A., Bedford, C.W., Garcia, E.O., and Ramirez, A.M. "Macromolecule Mass Spectrometry-Citation Mining of User Documents". Submitted for Publication.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". Submitted for Publication.

Kostoff, R. N., Karpouzian, G., and Malpohl, G. "Abrupt Wing Stall Roadmaps Using Database Tomography and Bibliometrics". TR NAWCAD PAX/RTR-???? Naval Air Warfare Center, Aircraft Division, Patuxent River, MD. In Press.

Kostoff, R. N., and Block, J. A. "Factor Matrix Text Filtering and Clustering." Submitted for Publication.

Kostoff, R. N., Tshiteya, R., Humenik, J. A., and Pfeil, K M. "Power Source Roadmaps Using Database Tomography and Bibliometrics". Submitted for Publication.

Kostoff, R. N., Del Rio, J. A., Bloomfield, L.A., Shlesinger, M. F., Malpohl, G., and Smith, A. "Dual-Use Publishing." To be Submitted for Publication.

Kostoff, R. N., Del Rio, J. A., Smith, C., and Malpohl, G. "Mexico Technology Assessment using Text Mining." To be Submitted for Publication.

Kostoff, R. N., Del Rio, J. A., Briggs, M., and Malpohl, G. "China Technology Assessment using Text Mining." To be Submitted for Publication.

Kostoff, R. N., Tshiteya, R., and Stump, J. "Wireless LAN Roadmaps using Bibliometrics and Database Tomography". To be Submitted for Publication.

Kostoff, R. N., Braun, T., Schubert, A., Pfeil, K. M., and Malpohl, G. "Fullerene Applications from Text Mining". To be Submitted for Publication.

Kostoff, R. N., Zablotska, L., and Neugut, A. "Factor Matrix Filtering and Clustering for Bilateral Asymmetry Prediction." To be Submitted for Publication.

Kostoff, R. N., and Block, J. A. "Literature-based Discovery and Innovation". To be Submitted for Publication.

Kostoff, R. N., Block, J. A., and Pfeil, K. M. "Information Content in Medline Record Fields". To be Submitted for Publication.

Kostoff, R. N., Hartley, J., and Smith, C. "Abstract and Keyword Field Quantitative Characteristics for Different Technical Disciplines." To be Submitted for Publication.

Kostoff, R. N., Del Rio, J. A., and Malpohl, G. "SBIR Technology Thrusts using Text Mining". To be Submitted for Publication.

Kostoff, R. N., Culpepper, R., Del Rio, J. A., and Malpohl, G. "ILIR Technology Thrusts using Text Mining". To be Submitted for Publication.

Kostoff, R. N., Coder, D., Wells, S., Toothman, D. R., and Humenik, J. "Surface Hydrodynamics Roadmaps Using Bibliometrics and Database Tomography". To be Submitted for Publication.

Kostoff, R. N., and Humenik, J. A. "Text Mining for Technical Intelligence". To be Submitted for Publication.

Kostoff, R. N., Pfeil, K. M., and Tshiteya, R. "Text Clustering and Taxonomies". To be Submitted for Publication.

Kostoff, R. N., Toothman, D. R., Humenik, J. A., and Pfeil, K. M. "Textual Data Mining Study of Textual Data Mining". To be Submitted for Publication.

Kostoff, R. N., and Toothman, D. R. "Simulated Nucleation for Information Retrieval". To be Submitted for Publication.

Kostoff, R. N., and Hartley, J. "Science and Technology Text Mining: Structured Papers". DTIC Technical Report Number ADA??????

Kostoff, R. N., and Geisler, E. "Science and Technology Text Mining : Strategic Management and Implementation in Government Organizations." DTIC Technical Report Number ADA??????

Kostoff, R. N., Bedford, C., Del Rio, J. A., García, E. O., and Ramírez, A. M. "Science and Technology Text Mining : Citation Mining of Macromolecular Mass Spectrometry." DTIC Technical Report Number ADA??????

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Information Retrieval from the Technical Literature". DTIC Technical Report Number ADA??????

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Science and Technology Text Mining: Nonlinear Dynamics". DTIC Technical Report Number ADA??????

Kostoff, R. N., and Tshiteya, R. "Science and Technology Text Mining: Wireless LANs". DTIC Technical Report Number ADA??????

Kostoff, R. N., Braun, T., Schubert, A., Pfeil, K. M., and Malpohl, G. "Science and Technology Text Mining: Fullerene Research and Applications" DTIC Technical Report Number ADA??????

Kostoff, R. N., Del Rio, J. A., and Malpohl, G. "Science and Technology Text Mining: SBIR". DTIC Technical Report Number ADA??????

Kostoff, R. N., Culpepper, R., Del Rio, J. A., and Malpohl, G. "Science and Technology Text Mining: ILIR ". DTIC Technical Report Number ADA??????

Kostoff, R. N., Tshiteya, R., Humenik, J. A., and Pfeil, K M. "Science and Technology Text Mining: Electric Power Sources". DTIC Technical Report Number ADA??????

Kostoff, R. N., Coder, D., Wells, S., Toothman, D. R., and Humenik, J. "Science and Technology Text Mining: Surface Hydrodynamics". DTIC Technical Report Number ADA??????

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Science and Technology Text Mining: Hypersonic and Supersonic Flow". DTIC Technical Report Number ADA??????

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Science and Technology Text Mining: Near-Earth Space". DTIC Technical Report Number ADA??????

Kostoff, R. N., "Science and Technology Text Mining: Comparative Analysis of the Research Impact Assessment Literature and the Journal of the American Chemical Society." DTIC Technical Report Number ADA??????

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., and Humenik, J. A. " Science and Technology Text Mining: A Review". DTIC Technical Report Number ADA??????

Kostoff, R. N., Andrews, J., Buchtel, H., Pfeil, K., Tshiteya, R., and Humenik, J. A. "Science and Technology Text Mining: Cortex". DTIC Technical Report Number ADA??????

Kostoff, R. N., Del Rio, J. A., Smith, C., and Malpohl, G. "Science and Technology Text Mining: Mexico Core Competencies" DTIC Technical Report Number ADA??????

Kostoff, R. N., Del Rio, J. A., Smith, C., and Malpohl, G. "Science and Technology Text Mining: China Core Competencies." DTIC Technical Report Number ADA??????

Kostoff, R. N. and Block, J. A. "Factor Matrix TextFiltering and Clustering". DTIC Technical Report Number ADA??????

Kostoff, R. N., Del Rio, J. A., Bloomfield, L.A., Shlesinger, M. F., and Malpohl, G. "Dual-Use Publishing." DTIC Technical Report Number ADA??????

Kostoff, R. N. and Block, J. A. "Literature-based Discovery and Innovation". DTIC Technical Report Number ADA??????

Kostoff, R. N., Block, J. A., and Pfeil, K. M. "Information Content in Medline Record Fields". DTIC Technical Report Number ADA??????

Kostoff, R. N. "Bilateral Asymmetry Prediction". DTIC Technical Report Number ADA??????

Kostoff, R. N. "Science and Technology Metrics". DTIC Technical Report Number ADA??????

Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. "Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography". Journal of Power Sources. 110:1. 163-176. 2002.

Kostoff, R. N., and Hartley J. "Structured Abstracts for Technical Journals". Journal of Information Science. 28:3. 257-261. 2002.

Del Rio, J. A., Kostoff, R. N., Garcia, E. O., Ramirez, A. M., and Humenik, J. A. "Phenomenological Approach to Profile Impact of Scientific Research: Citation Mining." Advances in Complex Systems. 5:1. 19-42. 2002.

Braun, T., Schubert, A., and Kostoff, R. N. "A Chemistry Field in Search of Applications: Statistical Analysis of U. S. Fullerene Patents". Journal of Chemical Information and Computer Science. 42:5. 1011-1015. 2002.

Kostoff, R. N. "Biowarfare Agent Prediction". Homeland Defense Journal. 1:4. 1-1. 2002.

Kostoff, R. N. "Overcoming Specialization." BioScience. 52:10. 937-941. 2002.

Kostoff, R. N. "TexTosterone-A Full-Spectrum Text Mining System". Provisional Patent Application. Filed 30 September 2002.

Kostoff, R. N. "The Extraction of Useful Information from the BioMedical Literature". Academic Medicine. 76:12. December 2001.

Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". JASIST. 52:13. 1148-1156. 52:13. November 2001.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., and Humenik, J. A. "Text Mining Using Database Tomography and Bibliometrics: A Review". Technology Forecasting and Social Change. 68:3. November 2001.

Kostoff, R. N. "Predicting Biowarfare Agents Takes on Priority". The Scientist. 26 November 2001.

Kostoff, R. N. "Stimulating Discovery". Proceedings: Discovery Science Workshop. November 2001.

Kostoff, R. N., and DeMarco, R. A. "Science and Technology Text Mining". Analytical Chemistry. 73:13. 370-378A. 1 July 2001.

Kostoff, R. N. "Intel Gold". Military Information Technology. 5:6. July 2001.

Kostoff, R. N. "Extracting Intel Ore". Military Information Technology. 5:5. 24-26. June 2001.

Kostoff, R. N., and Del Rio, J. A. "Physics Research Impact Assessment". Physics World. 14:6. 47-52. June 2001.

Kostoff, R. N., and Hartley, J. "Structured Abstracts for Technical Journals". Science. 11 May. p.292 (5519):1067a. 2001.

Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. 40:1. 19-39. Jan-Feb 2000.

Braun, T., Schubert, A. P., and Kostoff, R. N. "Growth and Trends of Fullerene Research as Reflected in its Journal Literature." Chemical Reviews. 100:1. 23-27. January 2000.

Losiewicz, P., Oard, D., and Kostoff, R. N. "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15. 99-119. 2000.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". Journal of Aircraft, 37:4. 727-730. July-August 2000.

Kostoff, R. N. "High Quality Information Retrieval for Improving the Conduct and Management of Research and Development". Proceedings: Twelfth International Symposium on Methodologies for Intelligent Systems. 11-14 October 2000.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations". Proceedings: Federal Data Mining Symposium and Exposition. 28-29 March 2000.

Kostoff, R. N. "The Underpublishing of Science and Technology Results". The Scientist. 14:9. 6-6. 1 May 2000.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. A. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". TR NAWCAD PAX/RTR-2000/84. Naval Air Warfare Center, Aircraft Division, Patuxent River, MD.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining Citing Population Profiling using Bibliometrics and Text Mining". Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico. http://www.cie.unam.mx/W_Reportes.

Kostoff, R. N. "Science and Technology Text Mining". Keynote presentation/ Proceedings. TTCP/ ITWP Workshop. Farnborough, UK. 12 October 2000.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations". Proceedings: Federal Data Mining Symposium and Exposition, 28-29 March 2000.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography". Journal of the American Society for Information Science. 50:5. 427-447. 15 April 1999.

Kostoff, R. N. "Science and Technology Innovation". Technovation. 19:10. 593-604. October 1999.

Kostoff, R. N., and Geisler, E. "Strategic Management and Implementation of Textual Data Mining in Government Organizations". Technology Analysis and Strategic Management. 11:4. 1999.

Kostoff, R. N. "Implementation of Textual Data Mining in Government Organizations", Presented at American Society for Information Science Annual Conference. Special Interest Group on Automated Language Processing. 3 November 1999.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature". Information Processing and Management. 34:1. 1998.

Kostoff, R. N. "Science and Technology Innovation". http://www.dtic.mil/dtic/kostoff/index.html. 1998.

Kostoff, R. N. "Science and Technology Innovation". http://www.scicentral.com. 1998.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997.

Kostoff, R. N., "Database Tomography for Technical Intelligence: Analysis of the Research Impact Assessment Literature", Competitive Intelligence Review, 8:2, Summer 1997.

Kostoff, R. N., "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.

Kostoff, R. N., Eberhart, H. J., and Miles, D., "System and Method for Database Tomography", U. S. Patent Number 5440481, August 8, 1995.

Kostoff, R.N., "Database Tomography: Origins and Applications," Competitive Intelligence Review, Special Issue on Technology, 5:1, Spring 1994. 48-55.

Kostoff, R. N. and Eberhart, H. J., "Database Tomography: Applications to Information, Logistics, and Personnel Management", Proceedings: Advanced Information Systems & Technology for Acquisition, Logistics, & Personnel Applications, Williamsburg, VA, March 28-30, 1994.

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.

Kostoff, R. N., "Database Tomography for Technical Intelligence", Proceedings: Eighth Annual Conference of the Society for Competitive Intelligence Professionals, Los Angeles, CA 1993.

Kostoff, R. N., "Database Tomography for Technical Intelligence," Competitive Intelligence Review, 4:1, Spring 1993. 38-43.

Kostoff, R.N. and Eberhart, H.J., "Database Tomography: Applications to Technical Intelligence," Proceedings: Technology 2003, Vol. 2, Anaheim, CA, Dec. 7-9, 1993.

TABLE 1 - CLUSTER THEMES

CL#	CLUSTER THEME	Ci	CL#	CLUSTER THEME	Ci
1	ACOUSTIC	190	26	NEURAL NETWORKS	21
2	ARTIFICIAL INTELLIGENCE	23	27	NUMERICAL MODELS	24
3	BOUNDARY LAYER	34	28	OCEAN	478
4	CHEMICAL	230	29	OCEAN BASINS	20
5	COMPUTER SCIENCE	47	30	OCEAN FLOOR	20
6	CONTROL	242	31	OPTICAL	185
7	CURRENT	191	32	PATTERN RECOGNITION	22
8	DATA	313	33	PHYSICAL	234
9	DECISION MAKING	25	34	PHYSICAL OCEANOGRAPHY	30
10	DEVICES	182	35	PROCESSING	244
11	DYNAMICS	189	36	PROPULSION SYSTEMS	22
12	ELECTRONIC DEVICES	20	37	RADIATION	279
13	ELECTRONIC MATERIALS	18	38	RADIATION SOURCES	18
14	ENERGETIC MATERIALS	30	39	REMOTE SENSING	88
15	ENERGY	278	40	SEA ICE	18
16	ENERGY CONVERSION	47	41	SEA SURFACE	21
17	INFORMATION	193	42	SHALLOW WATER	23
18	INFORMATION PROCESSING	20	43	SIGNAL PROCESSING	68
19	INTEGER PROGRAMMING	22	44	SOLID STATE	41
20	INTEGRATED CIRCUITS	20	45	SPACE	192
21	INTERNAL WAVES	29	46	STRUCTURES	196
22	INVERSE SCATTERING	18	47	SURFACE	299
23	MATERIALS	592	48	THIN FILMS	28
24	MODELING	180	49	UPPER OCEAN	27
25	MODELS	296	50	WATER COLUMN	26

TABLE 2 - ACOUSTIC CLUSTER AND MEMBERS

CL. #	******ACOUSTIC CLUSTER******	Cij	Cj	Ci	Cij^2/CiCj	Dpr ij	Cij/Ci
1.000	PROPAGATION	35.000	190.000	98.000	0.066	0.066	0.357
1.000	SCATTERING	34.000	190.000	95.000	0.064	0.064	0.358
1.000	OCEAN	76.000	190.000	478.000	0.064	0.064	0.159
1.000	BOTTOM	31.000	190.000	91.000	0.056	0.056	0.341
1.000	SHALLOW WATER	12.000	190.000	23.000	0.033	0.033	0.522
1.000	ACOUSTIC WAVE	6.000	190.000	6.000	0.032	0.032	1.000
1.000	THICK SECTION COMPOSITES	6.000	190.000	6.000	0.032	0.032	1.000
1.000	INVERSE SCATTERING	10.000	190.000	18.000	0.029	0.029	0.556
1.000	SCATTERING PROBLEMS	5.000	190.000	6.000	0.022	0.022	0.833
1.000	SAMPLING NETWORK	4.000	190.000	4.000	0.021	0.021	1.000
1.000	ACOUSTIC MODELS	4.000	190.000	4.000	0.021	0.021	1.000
1.000	GEOLOGIC PROCESSES	4.000	190.000	4.000	0.021	0.021	1.000
1.000	WAVEGUIDE INVERSE SCATTERING	6.000	190.000	9.000	0.021	0.021	0.667
1.000	UNDERWATER ACOUSTIC	6.000	190.000	10.000	0.019	0.019	0.600
1.000	OCEAN BOTTOM	6.000	190.000	11.000	0.017	0.017	0.545
1.000	SEAFLOOR MORPHOLOGY	4.000	190.000	5.000	0.017	0.017	0.800
1.000	BOTTOM INTERACTION	4.000	190.000	5.000	0.017	0.017	0.800
1.000	TRANSIENT SIGNALS	4.000	190.000	5.000	0.017	0.017	0.800
1.000	OCEANOGRAPHIC SAMPLING NETWORK	3.000	190.000	3.000	0.016	0.016	1.000
1.000	INVERSE SCATTERING THEORY	3.000	190.000	3.000	0.016	0.016	1.000
1.000	PHYSICAL FORCING	4.000	190.000	7.000	0.012	0.012	0.571
1.000	SHALLOW WATER ACOUSTICS	3.000	190.000	4.000	0.012	0.012	0.750
1.000	FRONTS AND EDDIES	3.000	190.000	4.000	0.012	0.012	0.750
1.000	AUTONOMOUS OCEANOGRAPHIC SAMPLING	3.000	190.000	4.000	0.012	0.012	0.750
1.000	SUVEILLANCE SYSTEMS	4.000	190.000	8.000	0.011	0.011	0.500
1.000	AIR-SEA FLUXES	4.000	190.000	8.000	0.011	0.011	0.500

TABLE 2 - ACOUSTIC CLUSTER AND MEMBERS (CONT'D)

CL. #	******ACOUSTIC CLUSTER******	Cij	Cj	Ci	Cij^2/CiCj	Dpr ij	Cij/Ci
1.000	ACTIVE AND PASSIVE	7.000	190.000	15.000	0.017		0.467
1.000	NUMERICAL MODELING	5.000	190.000	11.000	0.012		0.455
1.000	ACOUSTIC ENERGY	4.000	190.000	9.000	0.009		0.444
1.000	WAVE PROPAGATION	7.000	190.000	16.000	0.016		0.438
1.000	COASTAL TRANSITION ZONE	3.000	190.000	7.000	0.007		0.429
1.000	ACTIVE CONTROL	4.000	190.000	10.000	0.008		0.400
1.000	INTERNAL WAVE	8.000	190.000	24.000	0.014		0.333
1.000	AIR-SEA INTERACTION	4.000	190.000	12.000	0.007		0.333
1.000	SENSORS	32.000	190.000	126.000	0.043		0.254
1.000	PHYSICS OF ACOUSTICS	4.000	190.000	17.000	0.005		0.235
1.000	ARCTIC	32.000	190.000	158.000	0.034		0.203
1.000	PHYSICAL OCEANOGRAPHY	6.000	190.000	30.000	0.006		0.200
1.000	WATER	35.000	190.000	177.000	0.036		0.198
1.000	WAVE	29.000	190.000	149.000	0.030		0.195
1.000	INTERNAL WAVES	5.000	190.000	29.000	0.005		0.172
1.000	WATER COLUMN	4.000	190.000	26.000	0.003		0.154
1.000	MODELING	25.000	190.000	180.000	0.018		0.139
1.000	REMOTE SENSING	11.000	190.000	88.000	0.007		0.125
1.000	ENERGY	33.000	190.000	278.000	0.021		0.119
1.000	DATA	30.000	190.000	313.000	0.015		0.096
1.000	MODELS	25.000	190.000	296.000	0.011		0.084

TABLE 3 - FIGURES OF MERIT

TOTLINK	EXTLINK	EXT/TOT	OVERLAP#	OVER/TOT	OVER/LINK	CLOVER	DENS	DENSOVER	CEN1	CEN2	CEN3
1	2	3	4	5	6	7	8	9	10	11	12
28	25	10	25	46	35	28	22	19	35	2	19
25	28	46	35	24	21	25	19	22	28	35	2
23	23	25	1	34	10	35	40	48	2	40	12
1	39	24	39	7	1	1	30	2	25	34	14
19	46	17	28	42	41	33	9	40	19	12	32
39	1	33	23	10	18	39	20	34	1	17	18
15	35	39	46	8	11	46	2	43	39	18	48
45	33	7	10	47	42	24	36	26	26	4	40
37	8	4	37	25	34	7	13	14	17	32	26
35	47	35	33	17	39	6	41	12	34	26	22
22	5	22	19	22	38	22	1	24	45	21	7
13	22	13	5	13	27	13	7	33	20	37	33
18	20	9	41	9	20	5	24	5	5	13	24
12	13	14	20	14	13	50	23	37	13	20	5
50	50	20	13	20	9	38	8	25	9	5	10
40	41	50	50	19	5	20	45	7	41	50	23
27	38	19	38	50	2	19	28	45	38	45	37
9	9	38	9	38	19	9	6	23	50	38	45
49	49	49	49	49	49	49	25	49	49	49	49
41	36	36	36	36	36	36	47	36	36	36	36

TE: TABLE ELEMENTS ARE CLUSTER NUMBERS
UPPER TABLE HALF CONTAINS TOP 10 CLUSTERS
LOWER TABLE HALF CONTAINS BOTTOM 10 CLUSTERS

TABLE 4 - HIGH OVERLAPPING CLUSTER MEMBERS

CL. #	PERVASIVE THRUST AREA	Cij	Cj	Ci	Dij	Dpr ij	Eij	SUMOVER
4.000	MATERIALS	149.000	230.000	592.000	0.163	0.163	0.252	0.721
6.000	MATERIALS	121.000	242.000	592.000	0.102	0.102	0.204	0.721
31.000	MATERIALS	120.000	185.000	592.000	0.131	0.131	0.203	0.721
35.000	MATERIALS	151.000	244.000	592.000	0.158	0.158	0.255	0.721
46.000	MATERIALS	139.000	196.000	592.000	0.167	0.167	0.235	0.721
17.000	NEURAL	34.000	193.000	91.000	0.066	0.066	0.374	0.615
18.000	NEURAL	17.000	20.000	91.000	0.159	0.159	0.187	0.615
26.000	NEURAL	17.000	21.000	91.000	0.151	0.151	0.187	0.615
32.000	NEURAL	15.000	22.000	91.000	0.112	0.112	0.165	0.615
35.000	NEURAL	53.000	244.000	91.000	0.127	0.127	0.582	0.615
1.000	OCEAN	76.000	190.000	478.000	0.064	0.064	0.159	0.339
8.000	OCEAN	92.000	313.000	478.000	0.057	0.057	0.192	0.339
11.000	OCEAN	86.000	189.000	478.000	0.082	0.082	0.180	0.339
25.000	OCEAN	96.000	296.000	478.000	0.065	0.065	0.201	0.339
39.000	OCEAN	55.000	88.000	478.000	0.072	0.072	0.115	0.339
1.000	SAMPLING NETWORK	4.000	190.000	4.000	0.021	0.021	1.000	0.201
21.000	SAMPLING NETWORK	3.000	29.000	4.000	0.078	0.078	0.750	0.201
34.000	SAMPLING NETWORK	2.000	30.000	4.000	0.033	0.033	0.500	0.201
39.000	SAMPLING NETWORK	3.000	88.000	4.000	0.026	0.026	0.750	0.201
42.000	SAMPLING NETWORK	2.000	23.000	4.000	0.043	0.043	0.500	0.201

TABLE 4 - HIGH OVERLAPPING CLUSTER MEMBERS (CONT'D)

CL. #	PERVASIVE THRUST AREA	Cij	Cj	Ci	Dij	Dpr ij	Eij	SUMOVER
10.000	ARTIFICIALLY STRUCTURED MATERIALS	5.000	182.000	8.000	0.017	0.017	0.625	0.046
23.000	ARTIFICIALLY STRUCTURED MATERIALS	7.000	592.000	8.000	0.010	0.010	0.875	0.046
35.000	ARTIFICIALLY STRUCTURED MATERIALS	4.000	244.000	8.000	0.008	0.008	0.500	0.046
46.000	ARTIFICIALLY STRUCTURED MATERIALS	4.000	196.000	8.000	0.010	0.010	0.500	0.046
1.000	AUTONOMOUS OCEANOGRAPHIC SAMPLING	3.000	190.000	4.000	0.012	0.012	0.750	0.062
21.000	AUTONOMOUS OCEANOGRAPHIC SAMPLING	2.000	29.000	4.000	0.034	0.034	0.500	0.062
28.000	AUTONOMOUS OCEANOGRAPHIC SAMPLING	3.000	478.000	4.000	0.005	0.005	0.750	0.062
39.000	AUTONOMOUS OCEANOGRAPHIC SAMPLING	2.000	88.000	4.000	0.011	0.011	0.500	0.062
10.000	COHERENT X-RAY SOURCES	6.000	182.000	9.000	0.022	0.022	0.667	0.065
15.000	COHERENT X-RAY SOURCES	6.000	278.000	9.000	0.014	0.014	0.667	0.065
23.000	COHERENT X-RAY SOURCES	7.000	592.000	9.000	0.009	0.009	0.778	0.065
37.000	COHERENT X-RAY SOURCES	7.000	279.000	9.000	0.020	0.020	0.778	0.065
10.000	FERROELECTRIC THIN FILMS	6.000	182.000	7.000	0.028	0.028	0.857	0.169
12.000	FERROELECTRIC THIN FILMS	4.000	20.000	7.000	0.114	0.114	0.571	0.169
23.000	FERROELECTRIC THIN FILMS	7.000	592.000	7.000	0.012	0.012	1.000	0.169
35.000	FERROELECTRIC THIN FILMS	5.000	244.000	7.000	0.015	0.015	0.714	0.169
10.000	RADIATION DETECTORS	8.000	182.000	13.000	0.027	0.027	0.615	0.091
15.000	RADIATION DETECTORS	7.000	278.000	13.000	0.014	0.014	0.538	0.091
23.000	RADIATION DETECTORS	9.000	592.000	13.000	0.011	0.011	0.692	0.091
37.000	RADIATION DETECTORS	12.000	279.000	13.000	0.040	0.040	0.923	0.091
10.000	RADIATION-INDUCED DEFECT	5.000	182.000	5.000	0.027	0.027	1.000	0.078
24.000	RADIATION-INDUCED DEFECT	4.000	180.000	5.000	0.018	0.018	0.800	0.078
46.000	RADIATION-INDUCED DEFECT	4.000	196.000	5.000	0.016	0.016	0.800	0.078
35.000	RADIATION-INDUCED DEFECTS	4.000	244.000	4.000	0.016	0.016	1.000	0.078
1.000	SHALLOW WATER ACOUSTICS	3.000	190.000	4.000	0.012	0.012	0.750	0.124
28.000	SHALLOW WATER ACOUSTICS	3.000	478.000	4.000	0.005	0.005	0.750	0.124
33.000	SHALLOW WATER ACOUSTICS	3.000	234.000	4.000	0.010	0.010	0.750	0.124
42.000	SHALLOW WATER ACOUSTICS	3.000	23.000	4.000	0.098	0.098	0.750	0.124
1.000	THICK SECTION COMPOSITES	6.000	190.000	6.000	0.032	0.032	1.000	0.082
6.000	THICK SECTION COMPOSITES	3.000	242.000	6.000	0.006	0.006	0.500	0.082
7.000	THICK SECTION COMPOSITES	4.000	191.000	6.000	0.014	0.014	0.667	0.082
46.000	THICK SECTION COMPOSITES	6.000	196.000	6.000	0.031	0.031	1.000	0.082
10.000	ULTRAFAST MATERIALS RESPONSES	4.000	182.000	5.000	0.018	0.018	0.800	0.057
23.000	ULTRAFAST MATERIALS RESPONSES	5.000	592.000	5.000	0.008	0.008	1.000	0.057
35.000	ULTRAFAST MATERIALS RESPONSES	4.000	244.000	5.000	0.013	0.013	0.800	0.057
37.000	ULTRAFAST MATERIALS RESPONSES	5.000	279.000	5.000	0.018	0.018	1.000	0.057

TABLE 5

IRAD DATABASE
SINGLE WORDS
ORDERED BY FREQUENCY

10393	SYSTEM
9996	DESIGN
8212	SYSTEMS
6407	PROJECT
6075	DEVELOPMENT
5774	PERFORMANCE
5299	OBJECTIVE
5258	DATA
4184	ADVANCED
4121	TEST
4069	HIGH
3811	CONTROL
3519	OVERALL
3307	COST
3234	SOFTWARE
3119	AIRCRAFT
3064	ANALYSIS
2958	PROCESSING
2908	OBJECTIVES
2771	POWER
2511	CURRENT
2490	MATERIALS
2447	APPLICATIONS
2257	ENGINE
2212	CAPABILITY
2190	IMPROVED
2183	HARDWARE
2176	SUPPORT
2146	TESTING
2072	LOW
2030	COMPONENTS
1964	INTEGRATED
1941	PROCESS
1914	METHODS
1913	SPECIFIC
1888	COMPLETED
1885	MODEL
1878	FUTURE
1855	CONCEPTS
1851	MILITARY
1840	IR
1787	COMPUTER
1737	BASED
1731	TWO
1728	SIGNAL
1688	RELIABILITY
1627	CAPABILITIES
1624	DESIGNS
1609	ENVIRONMENT
1598	WEIGHT
1576	TASK
1529	EVALUATE
1516	DEMONSTRATE
1505	RANGE
1499	SIMULATION
1495	SENSOR
1464	DESIGNED
1460	TIME
1453	MEET
1441	CLASSIFICATION
1435	SPACE
1390	APPLICATION
1373	FLIGHT
1371	ARCHITECTURE
1370	TOOLS
1346	EFFORT
1336	PROTOTYPE
1329	AIR
1329	OPTICAL
1328	AREAS
1328	MISSION
1292	EVALUATION
1290	LARGE
1288	REDUCE
1287	INFORMATION
1266	RADAR
1236	MATERIAL
1234	MODELS
1233	CONCEPT
1233	INTERFACE
1229	SUBMITTED
1213	DIGITAL
1203	PROBLEMS
1200	DETERMINE
1199	TEMPERATURE
1197	ALGORITHMS
1197	TARGET
1192	LASER
1170	MISSILE
1166	COMPOSITE
1160	TACTICAL
1148	INCREASED
1140	APPLICABLE
1136	FABRICATION
1119	INTEGRATION
1104	COMPLETE
1104	DEVICES
1103	CONTINUE
1085	ELECTRONIC
1073	SIZE
1072	TESTED
1068	EQUIPMENT
1065	STUDY
1058	FREQUENCY
1052	TESTS
1047	TURBINE
1046	DEMONSTRATED
1043	VEHICLE
1041	POTENTIAL
1031	STUDIES
1024	HIGHER
1024	TASKS
1004	PROCESSOR
997	OPERATING
993	IMPROVEMENTS
991	IDENTIFY
990	SENSORS
985	MAJOR
985	THERMAL
966	CODE
958	COMMUNICATIONS
956	INCREASE
955	OPERATIONAL
955	STRUCTURES
944	NETWORK
942	WEAPON
932	CRITICAL
924	GENERATION
923	COMPONENT
910	PROGRAMS
909	IDENTIFIED
909	PERFORM
897	MANAGEMENT
895	STRUCTURAL
882	ARRAY
882	TECHNICAL
876	CIRCUIT
868	COMPLEX
858	DETECTION
857	ENGINEERING
854	CAPABLE
851	LIFE
850	PHASE
844	EVALUATED
842	BASE
838	SMALL
837	PROCESSES
836	EFFECTIVE
833	REDUCED
832	OPERATION
830	DEFENSE
827	EFFICIENCY
824	FUNCTIONS
823	DISPLAY
823	SET
822	LEVEL
815	ACHIEVE
814	CHARACTERISTICS

TABLE 5 (CONT'D)

806 VEHICLES	649 MODULES	537 GENERAL
803 MODELING	649 PROPULSION	536 ACTIVE
801 SELECTED	648 EFFICIENT	536 METHODOLOGY
798 FLOW	647 REAL-TIME	535 ENERGY
794 MANUFACTURING	642 INCREASING	535 MISSILES
790 ADDITION	639 ELEMENTS	534 PRODUCTS
790 DEVICE	639 METHOD	534 STATE-OF-THE-ART
787 PROPERTIES	628 MULTIYEAR	533 CONFIGURATIONS
786 SPEED	625 PRESSURE	532 MICROWAVE
785 PERFORMED	618 MAINTENANCE	531 PRODUCT
780 ENGINES	617 IMAGE	531 SUCCESSFULLY
777 MODULE	617 LIMITED	529 FOUR
775 SINGLE	616 PACKAGING	527 OUTPUT
772 NOISE	613 CONDITIONS	525 ACQUISITION
771 EFFECTS	611 MISSIONS	523 HIGHLY
771 FOLLOWING	610 LAUNCH	523 PLAN
768 FUEL	606 OPERATIONS	520 AUTOMATED
767 CURRENTLY	604 ASSOCIATED	520 STATE
767 SURFACE	604 LONG	518 FORCE
757 ANALYTICAL	602 CONVENTIONAL	517 BUILT
744 STRUCTURE	600 PARAMETERS	516 ACCURATE
741 DEMONSTRATION	599 COMMUNICATION	515 ELECTRICAL
734 PRODUCTION	596 DETAILED	515 FABRICATE
733 ANTENNA	596 RELIABLE	510 KNOWLEDGE
732 NUMBER	595 ESTABLISH	509 SUBSYSTEM
731 MECHANICAL	595 INITIAL	509 VARIETY
731 PRELIMINARY	595 TRAINING	508 TRACKING
726 AVIONICS	594 DEFINED	503 IMPLEMENT
724 CONFIGURATION	593 ACHIEVED	502 FULL
724 GOAL	590 EFFECTIVENESS	501 CYCLE
717 PRESENT	588 LEVELS	501 IMPLEMENTED
716 MULTIPLE	586 CONDUCTED	501 LIGHTWEIGHT
714 INVESTIGATE	585 IMPROVEMENT	500 AGAINST
713 GAS	584 ALGORITHM	500 GAAS
704 THREAT	583 QUALITY	499 CONTINUED
702 GOALS	578 COMMON	497 ADVANCES
702 IMPLEMENTATION	574 ACCURACY	497 DATABASE
696 CIRCUITS	572 COMMERCIAL	496 ELEMENT
696 TARGETS	571 FEATURES	493 IMPACT
694 FABRICATED	565 PARALLEL	490 INTEGRATE
690 FIELD	563 DYNAMIC	488 TRANSFER
689 BUILD	563 GROUND	487 ENHANCE
687 PRODUCE	559 TOOL	486 ADDRESS
685 RECEIVER	558 BASIC	486 PLANNING
685 REDUCTION	556 ROTOR	484 EXPERIMENTAL
682 ENVIRONMENTS	553 ABILITY	483 ACCOMPLISHED
681 RF	553 ARCHITECTURES	483 APPLIED
681 STANDARD	553 ASSEMBLY	483 DENSITY
674 RATE	553 PRIMARY	483 TECHNIQUE
671 LOWER	552 WEAPONS	483 WIDE
666 ADA	546 FEASIBILITY	482 AERODYNAMIC
665 ORDER	543 ESTABLISHED	480 INTELLIGENCE
656 FIBER	541 USER	479 COMPOSITES
654 UNIT	537 ADDITIONAL	478 HEAT

TABLE 6

IRAD DATABASE
DOUBLE WORDS
ORDERED BY FREQUENCY

1632	OVERALL OBJECTIVE
575	SIGNAL PROCESSING
550	LOW COST
475	HIGH PERFORMANCE
448	CONTROL SYSTEM
447	SPECIFIC OBJECTIVES
378	DATA BASE
360	CONTROL SYSTEMS
325	SYSTEM DESIGN
321	WEAPON SYSTEMS
317	OVERALL MULTIYEAR
292	GAS TURBINE
288	HIGH SPEED
281	WEAPON SYSTEM
273	SOFTWARE DEVELOPMENT
264	SYSTEM PERFORMANCE
250	AIR FORCE
236	HIGH TEMPERATURE
231	COST EFFECTIVE
231	FIBER OPTIC
219	NEURAL NETWORK
219	PRELIMINARY DESIGN
217	EXPERT SYSTEM
216	COMPOSITE MATERIALS
208	INTEGRATED CIRCUITS
206	SIGNAL PROCESSOR
204	HIGH POWER
200	AIR DEFENSE
197	LONG TERM
194	FLIGHT CONTROL
191	TURBINE ENGINE
185	SIZE WEIGHT
184	TEST DATA
183	OVERALL OBJECTIVES
183	SYSTEM ARCHITECTURE
178	MILITARY AIRCRAFT
177	INTEGRATED CIRCUIT
177	LIFE CYCLE
175	LONG-TERM OBJECTIVE
174	DESIGN CONCEPTS
174	LOWER COST
168	MULTIYEAR OBJECTIVE
164	PROPULSION SYSTEM
160	WIDE RANGE
159	REAL TIME
154	PROPULSION SYSTEMS
154	TEST BED
153	ADVANCED MATERIALS
150	TURBINE ENGINES
149	FLIGHT TEST
149	TRADE STUDIES
146	HEAT TRANSFER
145	ARTIFICIAL INTELLIGENCE
144	IMAGE PROCESSING
143	FIRE CONTROL
143	MILITARY SYSTEMS
141	DIGITAL SIGNAL
141	WIND TUNNEL
140	MISSION PLANNING
137	TARGET RECOGNITION
135	CONTINUING PROJECT
135	DATA PROCESSING
134	FINITE ELEMENT
134	IMPROVED PERFORMANCE
133	POWER SUPPLY
133	USER INTERFACE
132	NEURAL NETWORKS
130	SYSTEM CONCEPTS
127	MANAGEMENT SYSTEM
126	CONCEPTUAL DESIGN
125	OPERATING SYSTEM
125	POWER CONSUMPTION
122	DESIGN PROCESS
122	SYSTEM DEVELOPMENT
121	DETAILED DESIGN
120	ANALYSIS TOOLS
120	SPACE STATION
119	FIGHTER AIRCRAFT
116	CONTINUE DEVELOPMENT
113	MILITARY APPLICATIONS
112	DATA BASES
112	DATA FUSION
112	DESIGN FABRICATE
112	PHASED ARRAY
111	WIDE VARIETY
110	LAUNCH VEHICLE
110	PRIMARY OBJECTIVE
109	COMPUTER SIMULATION
109	LONG RANGE
108	DESIGN TOOLS
106	MISSILE SYSTEMS
105	COMPOSITE STRUCTURES
105	EXPERT SYSTEMS
104	FUTURE MILITARY
104	HIGH RELIABILITY
102	LAUNCH VEHICLES
102	TACTICAL AIRCRAFT
101	HIGH RESOLUTION
100	COMMUNICATIONS SYSTEMS
100	PROCESSING ALGORITHMS
99	ADVANCED AIRCRAFT
99	DATA LINK

TABLE 6 (CONT'D)

99 ELECTRONIC SYSTEMS	84 ADVANCED MILITARY
99 FUTURE AIRCRAFT	84 DESIGN METHODS
99 MECHANICAL PROPERTIES	84 MILLIMETER WAVE
99 SOLID STATE	84 OVERALL PROJECT
98 DEFENSE SYSTEMS	84 RADIO FREQUENCY
98 HIGHER PERFORMANCE	84 SYSTEM COMPONENTS
98 TERM OBJECTIVE	84 TRANSPORT AIRCRAFT
97 MULTI-YEAR PROJECT	83 COMPUTATIONAL FLUID
97 SOFTWARE ENGINEERING	83 MANUFACTURING PROCESSES
96 ENGINE CONTROL	82 MULTIYEAR PROJECT
96 LOW POWER	82 RADAR SYSTEMS
96 MATRIX COMPOSITES	82 RELIABILITY MAINTAINABILITY
96 SPECIFIC OBJECTIVE	81 COMMAND CONTROL
96 WEAPONS SYSTEMS	81 DESIGN SYSTEM
95 COST REDUCTION	81 LASER RADAR
95 DYNAMIC RANGE	80 KNOWLEDGE BASE
95 FUEL CONSUMPTION	80 STRATEGIC DEFENSE
94 ADVANCED TACTICAL	80 TEST EQUIPMENT
94 DESIGN METHODOLOGY	79 ADVANCED COMPOSITE
94 HIGH DENSITY	79 COMPLETED PROJECT
94 PARALLEL PROCESSING	79 FLUID DYNAMICS
94 PROCESSING SYSTEMS	79 MECHANICAL DESIGN
93 AVIONICS SYSTEMS	79 SENSOR DATA
93 FULL SCALE	78 HIGH QUALITY
93 LOW OBSERVABLE	78 INTEGRATED HIGH
91 ANALYTICAL TOOLS	78 METAL MATRIX
91 DATA ACQUISITION	78 OUTPUT POWER
91 ENGINE TESTING	77 FOLLOWING TASKS
90 COMPUTER CODE	77 PERFORMANCE TURBINE
90 DESIGN FABRICATION	77 PROJECT ADDRESSES
90 HARDWARE SOFTWARE	77 SOFTWARE TOOLS
90 POWER SYSTEM	76 COMPUTER SYSTEM
90 SENSOR SYSTEMS	76 SPACE COMPANY
90 TARGET DETECTION	75 DATA BUS
89 GATE ARRAY	75 DATA RATE
89 HIGH- PERFORMANCE	75 DESIGN CONCEPT
89 LANDING GEAR	75 LOCKHEED MISSILES
89 MANAGEMENT SYSTEMS	75 MATERIAL PROPERTIES
89 UNITED STATES	75 PROCESSING SYSTEM
88 SYSTEM CONCEPT	75 PROJECT OBJECTIVES
87 ANALYSIS METHODS	74 ENGINE TEST
87 DAMAGE TOLERANCE	74 FOUR TASKS
87 DESIGN ANALYSIS	74 GAS GENERATOR
87 DEVELOPMENT ENVIRONMENT	74 MATERIAL SYSTEMS
87 DEVELOPMENT IR	73 BUILDING BLOCKS
87 FOCAL PLANE	73 FREQUENCY RANGE
87 LOW NOISE	73 RAPID PROTOTYPING
86 COMMUNICATION SYSTEMS	73 SPACE SYSTEMS
86 TECHNICAL DATA	73 SYSTEM INTEGRATION
86 THIN FILM	73 THERMAL MANAGEMENT
85 DATA RATES	72 AUTOMATIC TARGET
85 PROJECT OBJECTIVE	72 DESIGN BUILD
85 PROTOTYPE SYSTEM	72 SIGNAL PROCESSORS
85 SOFTWARE DESIGN	72 SYSTEM CAPABLE

TABLE 7

IRAD DATABASE
TRIPLE WORDS
ORDERED BY FREQUENCY

861	CLASSIFICATION NOT SUBMITTED
321	HARDWARE AND SOFTWARE
286	APPLICABLE NOT APPLICABLE
154	COMMAND AND CONTROL
133	DESIGN AND ANALYSIS
129	FABRICATE AND TEST
123	OVERALL MULTIYEAR OBJECTIVE
115	RELIABILITY AND MAINTAINABILITY
114	GAS TURBINE ENGINES
111	FABRICATED AND TESTED
110	WEIGHT AND COST
109	DESIGN AND DEVELOPMENT
106	CURRENT AND FUTURE
99	DESIGN AND FABRICATION
97	SIZE AND WEIGHT
96	BUILT AND TESTED
92	BUILD AND TEST
86	WEIGHT AND POWER
85	REDUCE THE COST
84	DEVELOPMENT OF ADVANCED
83	MATERIALS AND PROCESSES
75	HIGH PERFORMANCE TURBINE
73	DESIGNED AND FABRICATED
73	GAS TURBINE ENGINE
72	ADVANCE THE STATE
72	INTEGRATED HIGH PERFORMANCE
71	SPACE COMPANY INC
70	COMPUTATIONAL FLUID DYNAMICS
70	DIGITAL SIGNAL PROCESSING
70	MILITARY AND COMMERCIAL
70	PERFORMANCE TURBINE ENGINE
69	DESIGN AND FABRICATE
68	PERFORMANCE AND RELIABILITY
67	LOCKHEED MISSILES SPACE
67	LONG TERM OBJECTIVE
67	MISSILES SPACE COMPANY
64	CONTINUE THE DEVELOPMENT
64	DESIGN AND BUILD
62	DEMONSTRATE THE FEASIBILITY
62	DESIGN AND TEST
61	PROJECT WAS TERMINATED
60	SYSTEMS THE OVERALL
58	COMPANY INC LMSC
56	RADIO FREQUENCY RF
55	DESIGNED AND BUILT
55	PROOF OF CONCEPT
54	TEST AND EVALUATION
53	AUTOMATIC TARGET RECOGNITION
53	LIFE CYCLE COST
52	ANALYSIS AND DESIGN
50	COMPLETE THE DEVELOPMENT
50	GUIDANCE AND CONTROL
49	COMPLETE THE DESIGN
49	TACTICAL AND STRATEGIC
48	ARTIFICIAL INTELLIGENCE AI
48	DETAILED TECHNICAL DATA
48	FLIGHT CONTROL SYSTEM
48	ORDER TO MEET
48	SYSTEMS THE OBJECTIVE
47	OVERALL PROJECT OBJECTIVE
47	SOFTWARE AND HARDWARE
46	ACTIVE AND PASSIVE
46	FABRICATION AND TESTING
46	SIGNAL PROCESSING ALGORITHMS
46	TESTS WERE CONDUCTED
45	OBJECTIVES WERE MET
44	COMPLETED THE DESIGN
44	COST AND SCHEDULE
44	COVERING DETAILED TECHNICAL
44	DEVELOPMENT AND DEMONSTRATIO
44	MONOLITHIC MICROWAVE INTEGRA
44	RADAR CROSS SECTION
44	REPORT COVERING DETAILED
44	SUPPORT AND READINESS
43	CAPABLE OF OPERATING
43	COST AND WEIGHT
43	DEVELOPMENT AND TESTING
43	FABRICATION AND TEST
43	SIZE AND COST
43	SOFTWARE DEVELOPMENT ENVIRON
42	CAPABLE OF MEETING
42	DESIGN WAS COMPLETED
42	METAL MATRIX COMPOSITES
42	SPECIFIC FUEL CONSUMPTION
41	APPLICATION SPECIFIC INTEGRAT
41	CHEMICAL VAPOR DEPOSITION
41	DETERMINE THE FEASIBILITY
41	DEVELOPMENT OF IMPROVED
40	FLIGHT CONTROL SYSTEMS
40	OVERALL MULTIYEAR OBJECTIVES
40	PRESENT AND FUTURE
39	ENGINE TECHNOLOGY IHPTET
39	FLUID DYNAMICS CFD
39	TEST AND EVALUATE
38	DETECTION AND CLASSIFICATION
37	ORDER OF MAGNITUDE
37	PERFORMANCE AND COST
37	STRATEGIC AND TACTICAL
36	ADVANCED COMPOSITE MATERIALS
36	APPLICATION OF ADVANCED
36	DESIGN AND MANUFACTURING
35	DIGITAL SIGNAL PROCESSOR
35	STRATEGIC DEFENSE INITIATIVE
35	WEIGHT AND VOLUME

TABLE 7 (CONT'D)

34 ADVANCED TACTICAL FIGHTER	26 LONG RANGE OBJECTIVE
34 WEAPON SYSTEM SUPPORT	26 ORDER TO REDUCE
33 AMOUNTS OF DATA	26 PERFORMED TO DETERMINE
33 COMMAND CONTROL COMMUNICATIONS	25 ASSEMBLY AND TEST
33 DETECTION AND TRACKING	25 CONTROL SYSTEM DESIGN
33 DEVELOPMENT AND EVALUATION	25 ELECTRONIC SUPPORT MEASURES
33 FIELD OF VIEW	25 FLAT PANEL DISPLAYS
33 GALLIUM ARSENIDE GAAS	25 FUTURE WEAPON SYSTEMS
33 LOCAL AREA NETWORK	25 HIGH DATA RATE
33 MICROWAVE INTEGRATED CIRCUITS	25 ORDERS OF MAGNITUDE
33 SPACE STATION FREEDOM	25 SUPPORT MEASURES ESM
33 SPECIFIC INTEGRATED CIRCUITS	24 ADVANCE THE STATE-OF-THE-ART
32 AIR DEFENSE SYSTEMS	24 DESIGN AND PERFORMANCE
32 ANALYTICAL AND EXPERIMENTAL	24 INTEGRATED CIRCUITS ASIC
32 COMMERCIAL AND MILITARY	24 MATERIALS AND MANUFACTURING
32 FINITE ELEMENT ANALYSIS	24 MILLIMETER WAVE MMW
31 ADVANCED THE STATE	24 PROJECT IS CONCERNED
31 COST OF OWNERSHIP	24 REDUCED INSTRUCTION SET
31 DESIGN AND IMPLEMENTATION	24 RELIABILITY AND PERFORMANCE
31 DEVELOPMENT AND TEST	23 ADVANCED GAS TURBINE
31 DEVELOPMENT OF HIGH	23 ASSEMBLED AND TESTED
31 ELECTRONIC WARFARE EW	23 DESIGN AND DEMONSTRATE
31 EVALUATE THE PERFORMANCE	23 DEVELOPMENT AND VALIDATION
31 ORDER TO ACHIEVE	23 ENGINE CONTROL SYSTEMS
31 SUPPORT THE DEVELOPMENT	23 ENVIRONMENTAL CONTROL SYSTEM
30 VOLUME AND WEIGHT	23 EVALUATE THE FEASIBILITY
30 WIND TUNNEL TEST	23 FIRE CONTROL SYSTEMS
29 COST THE OVERALL	23 LOW COST HIGH
29 DESIGNED AND TESTED	23 LOW OBSERVABLE COMPONENTS
29 DETERMINE THE BEST	23 MICROWAVE INTEGRATED CIRCUIT
29 DIVIDED INTO TWO	23 POSITIONING SYSTEM GPS
29 FOCAL PLANE ARRAYS	23 PROCESSING AND DISPLAY
29 LARGE SCALE INTEGRATION	23 REDUCE THE SIZE
29 OVERALL TECHNICAL OBJECTIVE	23 RING LASER GYRO
29 PRINTED WIRING BOARD	23 SIZE AND POWER
29 REDUCE THE NUMBER	23 STATIC AND DYNAMIC
29 RELIABILITY AND COST	23 SYSTEMS TO MEET
29 TARGET RECOGNITION ATR	23 TERMINATED TO ACCOMMODATE
29 VOICE AND DATA	22 ANGLE OF ATTACK
28 ANALOG AND DIGITAL	22 CAPABILITY TO PERFORM
28 COMMUNICATIONS AND INTELLIGENCE	22 CROSS SECTION RCS
28 DESIGN AND MANUFACTURE	22 DEMONSTRATED THE FEASIBILITY
28 PROBABILITY OF INTERCEPT	22 EW RF MANAGER
28 SIZE WEIGHT POWER	22 FABRICATION AND ASSEMBLY
28 SURVEILLANCE AND TRACKING	22 HIGH SPEED INTEGRATED
27 APPLICATIONS THE OVERALL	22 INTEGRATED CIRCUIT VHSIC
27 CAPABLE OF PERFORMING	22 PERFORMANCE AND DURABILITY
27 DESIGN AND EVALUATION	22 SCALE INTEGRATION VLSI
27 DESIGN AND IMPLEMENT	22 STUDIES WERE CONDUCTED
27 GLOBAL POSITIONING SYSTEM	22 TACTICAL FIGHTER ATF
27 REDUCING THE COST	22 TEST AND DEMONSTRATE
27 TIME AND COST	22 TESTING AND EVALUATION
26 CONDUCTED TO DETERMINE	22 WIND TUNNEL TESTING
26 INTEGRATED CIRCUIT ASIC	21 ANTISUBMARINE WARFARE ASW

TABLE 8 - PERVASIVE THRUST AREAS

575	SIGNAL PROCESSING
448	CONTROL SYSTEM
378	DATA BASE
288	HIGH SPEED
273	SOFTWARE DEVELOPMENT
236	HIGH TEMPERATURE
231	FIBER OPTIC
219	NEURAL NETWORK
217	EXPERT SYSTEM
216	COMPOSITE MATERIALS
208	INTEGRATED CIRCUITS
204	HIGH POWER
194	FLIGHT CONTROL
183	SYSTEM ARCHITECTURE
177	LIFE CYCLE
164	PROPULSION SYSTEM
159	REAL TIME
153	ADVANCED MATERIALS
146	HEAT TRANSFER
145	ARTIFICIAL INTELLIGENCE
144	IMAGE PROCESSING
143	FIRE CONTROL
141	WIND TUNNEL
137	TARGET RECOGNITION
135	DATA PROCESSING
134	FINITE ELEMENT
133	POWER SUPPLY
112	DATA FUSION
112	PHASED ARRAY
110	LAUNCH VEHICLE
109	COMPUTER SIMULATION
105	COMPOSITE STRUCTURES
101	HIGH RESOLUTION
99	DATA LINK
99	MECHANICAL PROPERTIES
99	SOLID STATE
97	SOFTWARE ENGINEERING
96	ENGINE CONTROL
96	MATRIX COMPOSITES
94	PARALLEL PROCESSING
93	LOW OBSERVABLE
91	DATA ACQUISITION
91	ENGINE TESTING
87	FOCAL PLANE
87	LOW NOISE
86	THIN FILM
85	DATA RATES
85	SOFTWARE DESIGN
84	MILLIMETER WAVE
84	RADIO FREQUENCY
82	RADAR SYSTEMS
81	LASAR RADAR
79	ADVANCED COMPOSITES
79	FLUID DYNAMICS
79	SENSOR DATA
78	METAL MATRIX
75	DATA BUS
74	GAS GENERATOR
114	GAS TURBINE ENGINES
70	DIGITAL SIGNAL PROCESSING

TABLE 9A - IMAGE PROCESSING - SORTED BY Cij

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
40	100	10393	0.01	0.0067	SYSTEM
40	86	5258	0.016	0.0098	DATA
40	85	2958	0.029	0.017	PROCESSING
40	84	617	0.136	0.0794	IMAGE
40	81	8212	0.01	0.0055	SYSTEMS
40	53	3234	0.016	0.006	SOFTWARE
40	51	6075	0.008	0.003	DEVELOPMENT
40	45	6407	0.007	0.0022	PROJECT
40	42	9996	0.004	0.0012	DESIGN
40	39	2183	0.018	0.0048	HARDWARE
40	39	3064	0.013	0.0034	ANALYSIS
40	38	1197	0.032	0.0084	ALGORITHMS
40	37	1197	0.031	0.0079	TARGET
40	37	5774	0.006	0.0016	PERFORMANCE
40	36	1627	0.022	0.0055	CAPABILITIES
40	36	2447	0.015	0.0037	APPLICATIONS
40	36	5299	0.007	0.0017	OBJECTIVE
40	32	188	0.17	0.0378	IMAGERY
40	31	4069	0.008	0.0016	HIGH
40	29	4184	0.007	0.0014	ADVANCED
40	27	422	0.064	0.012	IMAGING
40	25	144	0.174	0.0301	IMAGE PROCESSING
40	25	415	0.06	0.0105	RECOGNITION
40	25	565	0.044	0.0077	PARALLEL
40	25	1004	0.025	0.0043	PROCESSOR
40	25	2908	0.009	0.0015	OBJECTIVES
40	24	824	0.029	0.0049	FUNCTIONS
40	24	1609	0.015	0.0025	ENVIRONMENT
40	23	343	0.067	0.0107	WORKSTATION
40	23	1728	0.013	0.0021	SIGNAL
40	23	3519	0.007	0.001	OVERALL
40	22	1329	0.017	0.0025	OPTICAL
40	21	990	0.021	0.0031	SENSORS
40	20	647	0.031	0.0043	REAL-TIME
40	20	1441	0.014	0.0019	CLASSIFICATION
40	20	1737	0.012	0.0016	BASED
40	19	414	0.046	0.0061	AUTOMATIC
40	19	584	0.033	0.0043	ALGORITHM
40	19	2190	0.009	0.0011	IMPROVED
40	18	1213	0.015	0.0019	DIGITAL
40	18	1840	0.01	0.0012	IR
40	18	1964	0.009	0.0011	INTEGRATED
40	18	3811	0.005	0.0006	CONTROL
40	17	439	0.039	0.0046	NEURAL
40	17	1287	0.013	0.0016	INFORMATION
40	17	2176	0.008	0.0009	SUPPORT
40	16	508	0.031	0.0035	TRACKING
40	16	1328	0.012	0.0013	MISSION
40	16	1328	0.012	0.0013	AREAS
40	16	1495	0.011	0.0012	SENSOR

TABLE 9B - IMAGE PROCESSING - SORTED BY Cij/Ci

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
40	6	6	1	0.0417	PRO SCANNER
40	4	4	1	0.0278	MACHINE VISION PROGRAMMING
40	4	4	1	0.0278	PRO SCANNER SUBSYSTEM
40	4	4	1	0.0278	CORRELATION TRACKER
40	4	4	1	0.0278	VISION PROGRAMMING
40	4	4	1	0.0278	SCANNER SUBSYSTEM
40	4	4	1	0.0278	ILLUMINATOR AND RECEIVER
40	3	3	1	0.0208	SONAR AND VIDEO
40	3	3	1	0.0208	GEOGRAPHIC INFORMATION SYSTEM
40	4	5	0.8	0.0222	IMAGING RADIOMETER
40	3	4	0.75	0.0156	HIGH RESOLUTION SONAR
40	3	4	0.75	0.0156	ALGORITHMS TO DETECT
40	4	6	0.667	0.0185	PROTOTYPING CAPABILITIES
40	3	5	0.6	0.0125	AUTOMATIC PATTERN RECOGNITION
40	4	7	0.571	0.0159	GEOMETRIC ARITHMETIC PARALLEL
40	4	7	0.571	0.0159	RECEIVER AIRCRAFT
40	4	7	0.571	0.0159	ARITHMETIC PARALLEL
40	4	7	0.571	0.0159	PROCESSOR GAPP
40	4	7	0.571	0.0159	ARITHMETIC PARALLEL PROCESSOR
40	4	7	0.571	0.0159	PARALLEL PROCESSOR GAPP
40	4	7	0.571	0.0159	GEOMETRIC ARITHMETIC
40	4	7	0.571	0.0159	BASE GENERATION
40	3	6	0.5	0.0104	DATA BASE GENERATION
40	3	6	0.5	0.0104	TARGETS IN CLUTTER
40	8	17	0.471	0.0261	TOOLKIT
40	3	7	0.429	0.0089	IMAGE PROCESSING SYSTEM
40	4	10	0.4	0.0111	REAL-TIME IMAGE
40	6	16	0.375	0.0156	POSIX
40	4	11	0.364	0.0101	IMAGE ENHANCEMENT
40	4	12	0.333	0.0093	PROTOTYPING ENVIRONMENT
40	4	13	0.308	0.0085	IMAGE PROCESSING ALGORITHMS
40	4	14	0.286	0.0079	IMAGE EXPLOITATION
40	7	25	0.28	0.0136	IMAGE ANALYSIS
40	9	37	0.243	0.0152	MACHINE VISION
40	6	25	0.24	0.01	GAPP
40	8	36	0.222	0.0123	SCANNER
40	6	28	0.214	0.0089	PARALLEL PROCESSOR
40	4	21	0.19	0.0053	NDE METHODS
40	5	27	0.185	0.0064	DATA COMPRESSION
40	5	28	0.179	0.0062	DNA
40	25	144	0.174	0.0301	IMAGE PROCESSING
40	4	23	0.174	0.0048	IMAGING SENSORS
40	32	188	0.17	0.0378	IMAGERY
40	4	24	0.167	0.0046	OPERATOR WORKLOAD
40	5	32	0.156	0.0054	CT
40	5	33	0.152	0.0053	GENERAL-PURPOSE
40	5	36	0.139	0.0048	IMAGE DATA
40	84	617	0.136	0.0794	IMAGE
40	5	37	0.135	0.0047	DEFECT
40	4	30	0.133	0.0037	PROCESSING HARDWARE

TABLE 9C - IMAGE PROCESSING - SORTED BY Cij^2/CiCj

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
40	84	617	0.136	0.0794	IMAGE
40	6	6	1	0.0417	PRO SCANNER
40	32	188	0.17	0.0378	IMAGERY
40	25	144	0.174	0.0301	IMAGE PROCESSING
40	4	4	1	0.0278	VISION PROGRAMMING
40	4	4	1	0.0278	SCANNER SUBSYSTEM
40	4	4	1	0.0278	PRO SCANNER SUBSYSTEM
40	4	4	1	0.0278	MACHINE VISION PROGRAMMING
40	4	4	1	0.0278	ILLUMINATOR AND RECEIVER
40	4	4	1	0.0278	CORRELATION TRACKER
40	8	17	0.471	0.0261	TOOLKIT
40	4	5	0.8	0.0222	IMAGING RADIOMETER
40	3	3	1	0.0208	GEOGRAPHIC INFORMATION SYSTEM
40	3	3	1	0.0208	SONAR AND VIDEO
40	4	6	0.667	0.0185	PROTOTYPING CAPABILITIES
40	85	2958	0.029	0.017	PROCESSING
40	4	7	0.571	0.0159	ARITHMETIC PARALLEL
40	4	7	0.571	0.0159	PROCESSOR GAPP
40	4	7	0.571	0.0159	ARITHMETIC PARALLEL PROCESSOR
40	4	7	0.571	0.0159	RECEIVER AIRCRAFT
40	4	7	0.571	0.0159	GEOMETRIC ARITHMETIC
40	4	7	0.571	0.0159	PARALLEL PROCESSOR GAPP
40	4	7	0.571	0.0159	BASE GENERATION
40	4	7	0.571	0.0159	GEOMETRIC ARITHMETIC PARALLEL
40	3	4	0.75	0.0156	HIGH RESOLUTION SONAR
40	6	16	0.375	0.0156	POSIX
40	3	4	0.75	0.0156	ALGORITHMS TO DETECT
40	9	37	0.243	0.0152	MACHINE VISION
40	7	25	0.28	0.0136	IMAGE ANALYSIS
40	3	5	0.6	0.0125	AUTOMATIC PATTERN RECOGNITION
40	8	36	0.222	0.0123	SCANNER
40	27	422	0.064	0.012	IMAGING
40	4	10	0.4	0.0111	REAL-TIME IMAGE
40	23	343	0.067	0.0107	WORKSTATION
40	25	415	0.06	0.0105	RECOGNITION
40	3	6	0.5	0.0104	DATA BASE GENERATION
40	3	6	0.5	0.0104	TARGETS IN CLUTTER
40	4	11	0.364	0.0101	IMAGE ENHANCEMENT
40	6	25	0.24	0.01	GAPP
40	86	5258	0.016	0.0098	DATA
40	4	12	0.333	0.0093	PROTOTYPING ENVIRONMENT
40	3	7	0.429	0.0089	IMAGE PROCESSING SYSTEM
40	6	28	0.214	0.0089	PARALLEL PROCESSOR
40	4	13	0.308	0.0085	IMAGE PROCESSING ALGORITHMS
40	38	1197	0.032	0.0084	ALGORITHMS
40	37	1197	0.031	0.0079	TARGET
40	4	14	0.286	0.0079	IMAGE EXPLOITATION
40	25	565	0.044	0.0077	PARALLEL
40	12	137	0.088	0.0073	TARGET RECOGNITION
40	11	116	0.095	0.0072	EXPLOITATION

TABLE 10A - ADVANCED MATERIALS - SORTED BY Cij

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
1	134	2490	0.0540	0.0471	MATERIALS
1	125	4184	0.0300	0.0244	ADVANCED
1	97	9996	0.0100	0.0062	DESIGN
1	66	8212	0.0080	0.0035	SYSTEMS
1	65	5774	0.0110	0.0048	PERFORMANCE
1	58	6075	0.0100	0.0036	DEVELOPMENT
1	49	1914	0.0260	0.0082	METHODS
1	49	3119	0.0160	0.0050	AIRCRAFT
1	48	1236	0.0390	0.0122	MATERIAL
1	48	2190	0.0220	0.0069	IMPROVED
1	47	3519	0.0130	0.0041	OVERALL
1	47	4069	0.0120	0.0035	HIGH
1	47	6407	0.0070	0.0023	PROJECT
1	46	5299	0.0090	0.0026	OBJECTIVE
1	45	895	0.0500	0.0148	STRUCTURAL
1	45	10393	0.0040	0.0013	SYSTEM
1	40	2908	0.0140	0.0036	OBJECTIVES
1	40	3064	0.0130	0.0034	ANALYSIS
1	39	3307	0.0120	0.0030	COST
1	36	5258	0.0070	0.0016	DATA
1	35	2257	0.0160	0.0035	ENGINE
1	34	1855	0.0180	0.0041	CONCEPTS
1	32	479	0.0670	0.0140	COMPOSITES
1	32	787	0.0410	0.0085	PROPERTIES
1	32	1598	0.0200	0.0042	WEIGHT
1	32	4121	0.0080	0.0016	TEST
1	31	1047	0.0300	0.0060	TURBINE
1	31	1913	0.0160	0.0033	SPECIFIC
1	30	955	0.0310	0.0062	STRUCTURES
1	30	1166	0.0260	0.0050	COMPOSITE
1	30	1624	0.0180	0.0036	DESIGNS
1	29	1199	0.0240	0.0046	TEMPERATURE
1	27	153	0.1760	0.0311	ADVANCED MATERIALS
1	26	794	0.0330	0.0056	MANUFACTURING
1	26	837	0.0310	0.0053	PROCESSES
1	26	1529	0.0170	0.0029	EVALUATE
1	25	188	0.1330	0.0217	GEAR
1	24	1878	0.0130	0.0020	FUTURE
1	23	2030	0.0110	0.0017	COMPONENTS
1	23	2447	0.0090	0.0014	APPLICATIONS
1	22	392	0.0560	0.0081	DAMAGE
1	21	199	0.1060	0.0145	LANDING
1	21	382	0.0550	0.0075	MATRIX
1	21	2958	0.0070	0.0010	PROCESSING
1	21	3811	0.0060	0.0008	CONTROL
1	20	89	0.2250	0.0294	LANDING GEAR
1	20	610	0.0330	0.0043	LAUNCH
1	20	1136	0.0180	0.0023	FABRICATION
1	20	1328	0.0150	0.0020	AREAS
1	20	1390	0.0140	0.0019	APPLICATION
1	20	1632	0.0120	0.0016	OVERALL OBJECTIVE

TABLE 10B - ADVANCED MATERIALS - SORTED BY Cij/Ci

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
1	6	6	1.0000	0.0392	ROUGH FIELD LANDING
1	6	6	1.0000	0.0392	FIELD LANDING
1	6	6	1.0000	0.0392	FIELD LANDING GEAR
1	4	4	1.0000	0.0261	FIELD GEAR DESIGN
1	4	4	1.0000	0.0261	ROUGH FIELD GEAR
1	4	4	1.0000	0.0261	FIELD GEAR
1	3	3	1.0000	0.0196	THRUST BEARING SYSTEM
1	4	5	0.8000	0.0209	GEAR DESIGN CONCEPTS
1	3	4	0.7500	0.0147	HEAT TREATMENT STUDIES
1	12	18	0.6670	0.0523	ROUGH FIELD
1	3	5	0.6000	0.0118	IMPROVED ANALYSIS METHODS
1	3	5	0.6000	0.0118	SPACE LAUNCH VEHICLE
1	3	5	0.6000	0.0118	MANUFACTURING AND PROCESSING
1	4	7	0.5710	0.0149	STRUCTURAL CONFIGURATIONS
1	5	10	0.5000	0.0163	BEARING SYSTEM
1	4	8	0.5000	0.0131	CREW ESCAPE
1	3	6	0.5000	0.0098	SPACE TRANSPORTATION VEHICLES
1	4	9	0.4440	0.0116	OPERATIONS SIMULATION
1	4	9	0.4440	0.0116	HIGH TEMPERATURE MATERIALS
1	6	14	0.4290	0.0168	GEAR DESIGN
1	14	39	0.3590	0.0328	ROUGH
1	8	27	0.2960	0.0155	ADVANCED MATERIAL
1	6	21	0.2860	0.0112	NDE METHODS
1	4	15	0.2670	0.0070	SIMULATION AND MODELING
1	4	17	0.2350	0.0062	TEMPERATURE MATERIALS
1	4	17	0.2350	0.0062	SYSTEM TASK
1	20	89	0.2250	0.0294	LANDING GEAR
1	4	19	0.2110	0.0055	LIGHTWEIGHT MATERIALS
1	6	29	0.2070	0.0081	SPACE LAUNCH
1	3	15	0.2000	0.0039	SPACE LAUNCH VEHICLES
1	5	26	0.1920	0.0063	STRUCTURAL MATERIALS
1	3	16	0.1880	0.0037	NONDESTRUCTIVE EVALUATION NDE
1	5	27	0.1850	0.0061	ESCAPE
1	4	22	0.1820	0.0048	GENERATION AIRCRAFT
1	27	153	0.1760	0.0311	ADVANCED MATERIALS
1	17	97	0.1750	0.0195	NDE
1	5	30	0.1670	0.0054	TOUGHENED
1	4	25	0.1600	0.0042	HEAT TREATMENT
1	5	32	0.1560	0.0051	INTERMETALLIC
1	5	33	0.1520	0.0050	ALUMINUM-LITHIUM
1	4	29	0.1380	0.0036	ADVANCED COMPOSITES
1	25	188	0.1330	0.0217	GEAR
1	6	45	0.1330	0.0052	CERAMICS
1	5	38	0.1320	0.0043	LAMINATE
1	6	47	0.1280	0.0050	TEMPERATURE CAPABILITY
1	6	47	0.1280	0.0050	ELECTRONIC PACKAGING
1	15	118	0.1270	0.0125	METALLIC
1	5	42	0.1190	0.0039	METAL MATRIX COMPOSITES
1	9	78	0.1150	0.0068	METAL MATRIX
1	21	199	0.1060	0.0145	LANDING
1	5	47	0.1060	0.0035	PROCESSING METHODS

TABLE 10C - ADVANCED MATERIALS - SORTED BY Cij^2/CiCj

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
1	12	18	0.6670	0.0523	ROUGH FIELD
1	134	2490	0.0540	0.0471	MATERIALS
1	6	6	1.0000	0.0392	FIELD LANDING GEAR
1	6	6	1.0000	0.0392	FIELD LANDING
1	6	6	1.0000	0.0392	ROUGH FIELD LANDING
1	14	39	0.3590	0.0328	ROUGH
1	27	153	0.1760	0.0311	ADVANCED MATERIALS
1	20	89	0.2250	0.0294	LANDING GEAR
1	4	4	1.0000	0.0261	FIELD GEAR
1	4	4	1.0000	0.0261	ROUGH FIELD GEAR
1	4	4	1.0000	0.0261	FIELD GEAR DESIGN
1	125	4184	0.0300	0.0244	ADVANCED
1	25	188	0.1330	0.0217	GEAR
1	4	5	0.8000	0.0209	GEAR DESIGN CONCEPTS
1	3	3	1.0000	0.0196	THRUST BEARING SYSTEM
1	17	97	0.1750	0.0195	NDE
1	6	14	0.4290	0.0168	GEAR DESIGN
1	5	10	0.5000	0.0163	BEARING SYSTEM
1	8	27	0.2960	0.0155	ADVANCED MATERIAL
1	4	7	0.5710	0.0149	STRUCTURAL CONFIGURATIONS
1	45	895	0.0500	0.0148	STRUCTURAL
1	3	4	0.7500	0.0147	HEAT TREATMENT STUDIES
1	21	199	0.1060	0.0145	LANDING
1	32	479	0.0670	0.0140	COMPOSITES
1	4	8	0.5000	0.0131	CREW ESCAPE
1	15	118	0.1270	0.0125	METALLIC
1	48	1236	0.0390	0.0122	MATERIAL
1	3	5	0.6000	0.0118	MANUFACTURING AND PROCESSING
1	3	5	0.6000	0.0118	IMPROVED ANALYSIS METHODS
1	3	5	0.6000	0.0118	SPACE LAUNCH VEHICLE
1	4	9	0.4440	0.0116	OPERATIONS SIMULATION
1	4	9	0.4440	0.0116	HIGH TEMPERATURE MATERIALS
1	6	21	0.2860	0.0112	NDE METHODS
1	3	6	0.5000	0.0098	SPACE TRANSPORTATION VEHICLES
1	32	787	0.0410	0.0085	PROPERTIES
1	49	1914	0.0260	0.0082	METHODS
1	6	29	0.2070	0.0081	SPACE LAUNCH
1	22	392	0.0560	0.0081	DAMAGE
1	21	382	0.0550	0.0075	MATRIX
1	14	174	0.0800	0.0074	DESIGN CONCEPTS
1	4	15	0.2670	0.0070	SIMULATION AND MODELING
1	48	2190	0.0220	0.0069	IMPROVED
1	9	78	0.1150	0.0068	METAL MATRIX
1	5	26	0.1920	0.0063	STRUCTURAL MATERIALS
1	4	17	0.2350	0.0062	TEMPERATURE MATERIALS
1	4	17	0.2350	0.0062	SYSTEM TASK
1	30	955	0.0310	0.0062	STRUCTURES
1	97	9996	0.0100	0.0062	DESIGN
1	5	27	0.1850	0.0061	ESCAPE
1	31	1047	0.0300	0.0060	TURBINE
1	26	794	0.0330	0.0056	MANUFACTURING

TABLE 11A - PROPULSION SYSTEM - SORTED BY Cij

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
16	144	10393	0.014	0.0122	SYSTEM
16	115	649	0.177	0.1243	PROPULSION
16	110	9996	0.011	0.0074	DESIGN
16	106	2257	0.047	0.0304	ENGINE
16	82	8212	0.01	0.005	SYSTEMS
16	74	5774	0.013	0.0058	PERFORMANCE
16	69	4184	0.016	0.0069	ADVANCED
16	67	3119	0.021	0.0088	AIRCRAFT
16	59	3519	0.017	0.006	OVERALL
16	53	3811	0.014	0.0045	CONTROL
16	53	5299	0.01	0.0032	OBJECTIVE
16	52	6407	0.008	0.0026	PROJECT
16	49	6075	0.008	0.0024	DEVELOPMENT
16	44	3307	0.013	0.0036	COST
16	44	4069	0.011	0.0029	HIGH
16	41	2030	0.02	0.005	COMPONENTS
16	37	164	0.226	0.0509	PROPULSION SYSTEM
16	35	1878	0.019	0.004	FUTURE
16	35	4121	0.008	0.0018	TEST
16	34	2146	0.016	0.0033	TESTING
16	33	1043	0.032	0.0064	VEHICLE
16	31	1885	0.016	0.0031	MODEL
16	30	2190	0.014	0.0025	IMPROVED
16	29	780	0.037	0.0066	ENGINES
16	29	1964	0.015	0.0026	INTEGRATED
16	28	1855	0.015	0.0026	CONCEPTS
16	28	2212	0.013	0.0022	CAPABILITY
16	26	154	0.169	0.0268	PROPULSION SYSTEMS
16	25	1632	0.015	0.0023	OVERALL OBJECTIVE
16	24	1688	0.014	0.0021	RELIABILITY
16	24	1851	0.013	0.0019	MILITARY
16	24	1888	0.013	0.0019	COMPLETED
16	24	2072	0.012	0.0017	LOW
16	24	5258	0.005	0.0007	DATA
16	23	1047	0.022	0.0031	TURBINE
16	23	1373	0.017	0.0023	FLIGHT
16	23	1453	0.016	0.0022	MEET
16	23	1913	0.012	0.0017	SPECIFIC
16	22	814	0.027	0.0036	CHARACTERISTICS
16	22	923	0.024	0.0032	COMPONENT
16	22	997	0.022	0.003	OPERATING
16	22	2447	0.009	0.0012	APPLICATIONS
16	21	424	0.05	0.0063	PROPELLANT
16	21	2511	0.008	0.0011	CURRENT
16	21	2908	0.007	0.0009	OBJECTIVES
16	20	724	0.028	0.0034	CONFIGURATION
16	20	806	0.025	0.003	VEHICLES
16	20	1041	0.019	0.0023	POTENTIAL
16	20	1598	0.013	0.0015	WEIGHT
16	19	355	0.054	0.0062	THRUST

TABLE 11B - PROPULSION SYSTEM - SORTED BY Cij/Ci

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
16	4	4	1	0.0244	ENGINE AND SYSTEM
16	3	3	1	0.0183	COMMONALITY OF COMPONENTS
16	3	3	1	0.0183	UNIT OR DESIGN
16	3	3	1	0.0183	MODES AND FAULT
16	3	3	1	0.0183	ADVANCED MISSILE ENGINE
16	4	5	0.8	0.0195	SONIC ENVIRONMENT
16	4	6	0.667	0.0163	CONFIGURATION HARDWARE
16	4	6	0.667	0.0163	BOOSTER SYSTEM
16	4	6	0.667	0.0163	TOTAL IMPULSE
16	5	8	0.625	0.0191	MAGNETOHYDRODYNAMIC
16	3	5	0.6	0.011	SMALL ORBIT TRANSFER
16	3	5	0.6	0.011	CAPABILITIES IMPROVED WEAPON
16	3	5	0.6	0.011	INCREASED MISSION CAPABILITIES
16	3	5	0.6	0.011	TRENDING TOWARD INCREASED
16	3	5	0.6	0.011	MISSION CAPABILITIES IMPROVED
16	4	7	0.571	0.0139	PROPULSION MODULE
16	6	11	0.545	0.02	MONOPROPELLANT
16	5	10	0.5	0.0152	CYANIDE
16	4	8	0.5	0.0122	PROPULSION SYSTEM COMPONENTS
16	3	6	0.5	0.0091	ACQUISITION AND LIFE
16	3	7	0.429	0.0078	IMPROVED WEAPON SYSTEM
16	3	7	0.429	0.0078	CONTROL AND MONITORING
16	3	7	0.429	0.0078	FLUID AND PROPULSION
16	3	8	0.375	0.0069	SYSTEM SUPPORT READINESS
16	4	11	0.364	0.0089	MISSION CAPABILITIES
16	5	14	0.357	0.0109	ORBIT TRANSFER
16	5	14	0.357	0.0109	PROPULSION CONCEPTS
16	3	10	0.3	0.0055	THRUST TO WEIGHT
16	7	24	0.292	0.0124	THRUSTERS
16	10	36	0.278	0.0169	IHPTET INITIATIVE
16	4	15	0.267	0.0065	ELECTRIC PROPULSION
16	5	20	0.25	0.0076	SONIC
16	5	20	0.25	0.0076	UNDUCTED FAN ENGINE
16	6	25	0.24	0.0088	GEOSYNCHRONOUS
16	5	21	0.238	0.0073	ACCELERATOR
16	5	21	0.238	0.0073	ENGINE SYSTEM
16	9	38	0.237	0.013	STOVL
16	5	22	0.227	0.0069	FAN ENGINE
16	37	164	0.226	0.0509	PROPULSION SYSTEM
16	5	25	0.2	0.0061	EJECTOR
16	3	15	0.2	0.0037	GROUND AND FLIGHT
16	5	26	0.192	0.0059	FUEL BURN
16	8	42	0.19	0.0093	PROPFAN
16	7	37	0.189	0.0081	HYDRAZINE
16	115	649	0.177	0.1243	PROPULSION
16	4	23	0.174	0.0042	PROPULSION CONTROL
16	26	154	0.169	0.0268	PROPULSION SYSTEMS
16	6	36	0.167	0.0061	UNDUCTED FAN
16	6	37	0.162	0.0059	ADVANCED PROPULSION
16	5	34	0.147	0.0045	HIGH MACH

TABLE 11C - PROPULSION SYSTEM - SORTED BY Cij^2/CiCj

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
16	115	649	0.177	0.1243	PROPULSION
16	37	164	0.226	0.0509	PROPULSION SYSTEM
16	106	2257	0.047	0.0304	ENGINE
16	26	154	0.169	0.0268	PROPULSION SYSTEMS
16	4	4	1	0.0244	ENGINE AND SYSTEM
16	6	11	0.545	0.02	MONOPROPELLANT
16	4	5	0.8	0.0195	SONIC ENVIRONMENT
16	5	8	0.625	0.0191	MAGNETOHYDRODYNAMIC
16	3	3	1	0.0183	ADVANCED MISSILE ENGINE
16	3	3	1	0.0183	MODES AND FAULT
16	3	3	1	0.0183	UNIT OR DESIGN
16	3	3	1	0.0183	COMMONALITY OF COMPONENTS
16	10	36	0.278	0.0169	IHPTET INITIATIVE
16	4	6	0.667	0.0163	TOTAL IMPULSE
16	4	6	0.667	0.0163	BOOSTER SYSTEM
16	4	6	0.667	0.0163	CONFIGURATION HARDWARE
16	5	10	0.5	0.0152	CYANIDE
16	4	7	0.571	0.0139	PROPULSION MODULE
16	9	38	0.237	0.013	STOVL
16	7	24	0.292	0.0124	THRUSTERS
16	4	8	0.5	0.0122	PROPULSION SYSTEM COMPONENTS
16	144	10393	0.014	0.0122	SYSTEM
16	3	5	0.6	0.011	SMALL ORBIT TRANSFER
16	3	5	0.6	0.011	CAPABILITIES IMPROVED WEAPON
16	3	5	0.6	0.011	TRENDING TOWARD INCREASED
16	3	5	0.6	0.011	INCREASED MISSION CAPABILITIES
16	3	5	0.6	0.011	MISSION CAPABILITIES IMPROVED
16	5	14	0.357	0.0109	PROPULSION CONCEPTS
16	5	14	0.357	0.0109	ORBIT TRANSFER
16	8	42	0.19	0.0093	PROPFAN
16	3	6	0.5	0.0091	ACQUISITION AND LIFE
16	4	11	0.364	0.0089	MISSION CAPABILITIES
16	6	25	0.24	0.0088	GEOSYNCHRONOUS
16	67	3119	0.021	0.0088	AIRCRAFT
16	7	37	0.189	0.0081	HYDRAZINE
16	3	7	0.429	0.0078	FLUID AND PROPULSION
16	3	7	0.429	0.0078	IMPROVED WEAPON SYSTEM
16	3	7	0.429	0.0078	CONTROL AND MONITORING
16	5	20	0.25	0.0076	SONIC
16	5	20	0.25	0.0076	UNDUCTED FAN ENGINE
16	11	98	0.112	0.0075	BOOSTER
16	110	9996	0.011	0.0074	DESIGN
16	5	21	0.238	0.0073	ENGINE SYSTEM
16	5	21	0.238	0.0073	ACCELERATOR
16	13	144	0.09	0.0072	ELECTRIC
16	3	8	0.375	0.0069	SYSTEM SUPPORT READINESS
16	5	22	0.227	0.0069	FAN ENGINE
16	69	4184	0.016	0.0069	ADVANCED
16	14	181	0.077	0.0066	IHPTET
16	29	780	0.037	0.0066	ENGINES

TABLE 12A - SIGNAL PROCESSING - SORTED BY Cij

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
18	344	1728	0.199	0.1191	SIGNAL
18	337	2958	0.114	0.0668	PROCESSING
18	305	8212	0.037	0.0197	SYSTEMS
18	300	10393	0.029	0.0151	SYSTEM
18	239	9996	0.024	0.0099	DESIGN
18	215	6407	0.034	0.0125	PROJECT
18	189	5774	0.033	0.0108	PERFORMANCE
18	179	5258	0.034	0.0106	DATA
18	173	1197	0.145	0.0435	ALGORITHMS
18	160	6075	0.026	0.0073	DEVELOPMENT
18	157	5299	0.03	0.0081	OBJECTIVE
18	149	1213	0.123	0.0318	DIGITAL
18	148	4184	0.035	0.0091	ADVANCED
18	127	4069	0.031	0.0069	HIGH
18	121	1004	0.121	0.0254	PROCESSOR
18	113	575	0.197	0.0386	SIGNAL PROCESSING
18	112	2447	0.046	0.0089	APPLICATIONS
18	104	2771	0.038	0.0068	POWER
18	102	3519	0.029	0.0051	OVERALL
18	101	1266	0.08	0.014	RADAR
18	101	4121	0.025	0.0043	TEST
18	100	1371	0.073	0.0127	ARCHITECTURE
18	97	1329	0.073	0.0123	OPTICAL
18	97	1495	0.065	0.0109	SENSOR
18	95	3234	0.029	0.0049	SOFTWARE
18	93	1964	0.047	0.0077	INTEGRATED
18	93	2183	0.043	0.0069	HARDWARE
18	88	3811	0.023	0.0035	CONTROL
18	84	685	0.123	0.0179	RECEIVER
18	84	3064	0.027	0.004	ANALYSIS
18	83	858	0.097	0.014	DETECTION
18	82	2511	0.033	0.0047	CURRENT
18	79	2908	0.027	0.0037	OBJECTIVES
18	77	1197	0.064	0.0086	TARGET
18	77	3307	0.023	0.0031	COST
18	76	392	0.194	0.0256	SIGNALS
18	75	1627	0.046	0.006	CAPABILITIES
18	74	1840	0.04	0.0052	IR
18	71	1888	0.038	0.0046	COMPLETED
18	71	2212	0.032	0.004	CAPABILITY
18	69	824	0.084	0.01	FUNCTIONS
18	68	1104	0.062	0.0073	DEVICES
18	67	958	0.07	0.0081	COMMUNICATIONS
18	66	1851	0.036	0.0041	MILITARY
18	66	1913	0.035	0.004	SPECIFIC
18	65	1505	0.043	0.0049	RANGE
18	63	476	0.132	0.0145	ACOUSTIC
18	63	1058	0.06	0.0065	FREQUENCY
18	61	437	0.14	0.0148	PROCESSORS
18	59	156	0.378	0.0388	DSP

TABLE 12A - SIGNAL PROCESSING - SORTED BY Cij

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
18	344	1728	0.199	0.1191	SIGNAL
18	337	2958	0.114	0.0668	PROCESSING
18	305	8212	0.037	0.0197	SYSTEMS
18	300	10393	0.029	0.0151	SYSTEM
18	239	9996	0.024	0.0099	DESIGN
18	215	6407	0.034	0.0125	PROJECT
18	189	5774	0.033	0.0108	PERFORMANCE
18	179	5258	0.034	0.0106	DATA
18	173	1197	0.145	0.0435	ALGORITHMS
18	160	6075	0.026	0.0073	DEVELOPMENT
18	157	5299	0.03	0.0081	OBJECTIVE
18	149	1213	0.123	0.0318	DIGITAL
18	148	4184	0.035	0.0091	ADVANCED
18	127	4069	0.031	0.0069	HIGH
18	121	1004	0.121	0.0254	PROCESSOR
18	113	575	0.197	0.0386	SIGNAL PROCESSING
18	112	2447	0.046	0.0089	APPLICATIONS
18	104	2771	0.038	0.0068	POWER
18	102	3519	0.029	0.0051	OVERALL
18	101	1266	0.08	0.014	RADAR
18	101	4121	0.025	0.0043	TEST
18	100	1371	0.073	0.0127	ARCHITECTURE
18	97	1329	0.073	0.0123	OPTICAL
18	97	1495	0.065	0.0109	SENSOR
18	95	3234	0.029	0.0049	SOFTWARE
18	93	1964	0.047	0.0077	INTEGRATED
18	93	2183	0.043	0.0069	HARDWARE
18	88	3811	0.023	0.0035	CONTROL
18	84	685	0.123	0.0179	RECEIVER
18	84	3064	0.027	0.004	ANALYSIS
18	83	858	0.097	0.014	DETECTION
18	82	2511	0.033	0.0047	CURRENT
18	79	2908	0.027	0.0037	OBJECTIVES
18	77	1197	0.064	0.0086	TARGET
18	77	3307	0.023	0.0031	COST
18	76	392	0.194	0.0256	SIGNALS
18	75	1627	0.046	0.006	CAPABILITIES
18	74	1840	0.04	0.0052	IR
18	71	1888	0.038	0.0046	COMPLETED
18	71	2212	0.032	0.004	CAPABILITY
18	69	824	0.084	0.01	FUNCTIONS
18	68	1104	0.062	0.0073	DEVICES
18	67	958	0.07	0.0081	COMMUNICATIONS
18	66	1851	0.036	0.0041	MILITARY
18	66	1913	0.035	0.004	SPECIFIC
18	65	1505	0.043	0.0049	RANGE
18	63	476	0.132	0.0145	ACOUSTIC
18	63	1058	0.06	0.0065	FREQUENCY
18	61	437	0.14	0.0148	PROCESSORS
18	59	156	0.378	0.0388	DSP

TABLE 12C - SIGNAL PROCESSING - SORTED BY Cij^2/CiCj

CL#	Cij	Ci	Cij/Ci	Cij^2/CiCj	
18	344	1728	0.199	0.1191	SIGNAL
18	337	2958	0.114	0.0668	PROCESSING
18	173	1197	0.145	0.0435	ALGORITHMS
18	59	156	0.378	0.0388	DSP
18	113	575	0.197	0.0386	SIGNAL PROCESSING
18	149	1213	0.123	0.0318	DIGITAL
18	76	392	0.194	0.0256	SIGNALS
18	121	1004	0.121	0.0254	PROCESSOR
18	305	8212	0.037	0.0197	SYSTEMS
18	84	685	0.123	0.0179	RECEIVER
18	44	206	0.214	0.0163	SIGNAL PROCESSOR
18	41	186	0.22	0.0157	MODEM
18	40	181	0.221	0.0154	SONAR
18	300	10393	0.029	0.0151	SYSTEM
18	61	437	0.14	0.0148	PROCESSORS
18	63	476	0.132	0.0145	ACOUSTIC
18	101	1266	0.08	0.014	RADAR
18	83	858	0.097	0.014	DETECTION
18	10	13	0.769	0.0134	MODEM SIGNAL
18	100	1371	0.073	0.0127	ARCHITECTURE
18	215	6407	0.034	0.0125	PROJECT
18	11	17	0.647	0.0124	SIGNAL RECOGNITION
18	97	1329	0.073	0.0123	OPTICAL
18	26	100	0.26	0.0118	PROCESSING ALGORITHMS
18	52	415	0.125	0.0113	RECOGNITION
18	30	141	0.213	0.0111	DIGITAL SIGNAL
18	97	1495	0.065	0.0109	SENSOR
18	17	46	0.37	0.0109	SIGNAL PROCESSING ALGORITHMS
18	189	5774	0.033	0.0108	PERFORMANCE
18	179	5258	0.034	0.0106	DATA
18	6	6	1	0.0104	BQQ-5
18	69	824	0.084	0.01	FUNCTIONS
18	239	9996	0.024	0.0099	DESIGN
18	20	72	0.278	0.0097	SIGNAL PROCESSORS
18	10	18	0.556	0.0097	ECHO
18	12	26	0.462	0.0096	ANS
18	56	584	0.096	0.0093	ALGORITHM
18	54	553	0.098	0.0092	ARCHITECTURES
18	10	19	0.526	0.0092	COCHANNEL
18	148	4184	0.035	0.0091	ADVANCED
18	17	56	0.304	0.009	TRANSPONDER
18	112	2447	0.046	0.0089	APPLICATIONS
18	5	5	1	0.0087	NON-TRADITIONAL ACOUSTIC
18	77	1197	0.064	0.0086	TARGET
18	10	21	0.476	0.0083	SIGNAL PROCESSING SYSTEMS
18	67	958	0.07	0.0081	COMMUNICATIONS
18	157	5299	0.03	0.0081	OBJECTIVE
18	39	326	0.12	0.0081	HIGH-SPEED
18	18	70	0.257	0.008	DIGITAL SIGNAL PROCESSING
18	10	22	0.455	0.0079	ROTHR