SCIENCE AND TECHNOLOGY TEXT MINING: ELECTRIC POWER SOURCES

By

Dr. Ronald N. Kostoff*, Office of Naval Research, 800 North Quincy Street, Arlington, VA 22217

Mr. Rene Tshiteya, DDL OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600, Mclean, VA 22102

Ms. Kirstin M. Pfeil, Office of Naval Research, Arlington, VA 22217

Mr. James A. Humenik, Noesis, Inc, Manassas, VA 20109

Dr. George Karypis, University of Minnesota, Minneapolis, MN

ABSTRACT

Database Tomography (DT) is a textual database analysis system consisting of two major components: 1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment 2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a Power Sources database derived from the Science Citation Index (SCI). Phrase frequency analysis by the technical domain experts provided the pervasive technical themes of the Power Sources database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the Power Sources literature supplemented the DT results with author/ journal/ institution/ country publication and citation data.

KEYWORDS: Electrical Energy; Electrical Power; Energy Source; Energy Conversion; Energy Storage; Power Source; Power Conversion; Heat Engine; Direct Conversion; Renewable Source; Sustainable Energy; Power Generation; Fossil Fuel; Nuclear Power; Co-generation; Power Production; Energy Supply; Bio-mass Energy; Text Mining; Computational Linguistics; Bibliometrics; Scientometrics; Clustering; Taxonomy

(*The views expressed in this report are solely those of the authors, and do not represent the views of the Department of the Navy, DDL-OMNI, LLC, Noesis, Inc, or the University of Minnesota*)

*Corresponding Author :
PHONE: 703-696-4198, FAX: 703-696-4274, INTERNET: kostofr@onr.navy.mil

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
APR 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Science and Technology Text Mining: Electric Power Sources	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Dr. Ronald N. Kostoff; Mr. Rene Tshiteya; Ms. Kirstin M. Pfeil; Mr. James A. Humenik; Dr. George Karypis	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Office of Naval Research, 800 North Quincy Street, Arlington, VA 22217; DDL OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600,Mclean, VA 22102; Noesis, Inc,Manassas, VA 20109; University of Minnesota, Minneapolis, MN	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

The original document contains color images.

14. ABSTRACT

Database Tomography (DT) is a textual database analysis system consisting of two major components: 1) algorithms for extracting multi-word phrase frequencies and phrase proximities (physical closeness of the multi-word technical phrases) from any type of large textual database, to augment 2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a Power Sources database derived from the Science Citation Index (SCI). Phrase frequency analysis by the technical domain experts provided the pervasive technical themes of the Power Sources database, and the phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the Power Sources literature supplemented the DT results with author/ journal/ institution/ country publication and citation data.

15. SUBJECT TERMS

Electrical Energy; Electrical Power; Energy Source; Energy Conversion; Energy Storage; Power Source; Power Conversion; Heat Engine; Direct Conversion; Renewable Source; Sustainable Energy; Power Generation; Fossil Fuel; Nuclear Power; Co-generation; Power Production; Energy Supply; Bio-mass Energy; Text Mining; Computational Linguistics; Bibliometrics; Scientometrics; Clustering; Taxonomy

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**79**	

1. INTRODUCTION

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a commensurate increase in the need for scientific and technical intelligence to insure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While direct human intelligence gathering cannot be substituted, many techniques have become available that can support and complement it. In particular, techniques that identify, select, gather, cull, and interpret large amounts of technological information semi-automatically can expand greatly the capabilities of human beings in performing technical intelligence.

The combination of text mining and bibliometrics is being developed by different researchers for these, and many other, applications. Its component capabilities are as follows.

Science and technology (S&T) text mining [1-4] is a process for extracting useful information from large volumes of technical text, based mainly on the mechanics of computational linguistics. It identifies pervasive technical themes in large databases from frequently occurring technical phrases. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Text mining can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [5-7]
- Potential discovery and innovation based on merging common linkages between very disparate literatures [8-11]
- Uncovering unexpected asymmetries from the technical literature [12-13]
- Estimating global levels of effort in S+T sub-disciplines [14-16]
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors [15, 17]
- Tracking myriad research impacts across time and applications areas [18-19].

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the database using computational linguistics and bibliometrics, and integrates the processed information.

Evaluative bibliometrics [20-22] uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that 1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, 2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and 3) the citations from papers to papers, from patents to patents and from patents to papers provide indicators of intellectual linkages between the organizations which are producing the patents and papers, and knowledge linkage between their subject areas [23]. Evaluative bibliometrics can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain,
- Identify experts for innovation-enhancing technical workshops and review panels,
- Develop site visitation strategies for assessment of prolific organizations globally,
- Identify impacts (literature citations) of individuals, research units, organizations, and countries

One text mining approach developed by the first author's group is DT (Database Tomography) [24], a system for analyzing large amounts of textual computerized material. It includes algorithms for extracting multi-word phrase frequencies and phrase proximities from the textual databases, coupled with the topical expert human analyst to interpret the results and convert large volumes of disorganized data to ordered information. Phrase frequency analysis (occurrence frequency of multi-word technical phrases) provides the pervasive technical themes of a database, and the phrase proximity (physical closeness of the multi-word technical phrases) analysis provides the relationships among pervasive technical themes, as well as among technical themes and authors/journals/institutions/countries, etc. The present report describes use of the DT process, supplemented by literature bibliometric analyses, to derive technical intelligence from the published literature of Power Sources science and technology.

Power Sources, as defined by the authors for this study, consists of systems and processes for generating and converting power, and storing energy. It is defined operationally by a query with two components: 1) a phrase-based query, obtained by the iterative technique referenced in the next paragraph; and 2) a journal-title-based query, obtained by identifying non-technology-specific power source journals from the SCI journal listing under Energy and Fuels whose articles were deemed highly relevant to the Power Sources topic. Since one of the key outputs of the present study is a query that can be used by the community to access relevant Power Sources documents, a recommended query based on this study is presented in Appendix 1. This query serves as the operational definition of Power Sources, and its development is discussed in the database generation section.

To execute the study reported in this report, a database of relevant Power Sources articles is generated using the iterative search approach of Simulated Nucleation [25]. Then, the database is analyzed to produce the following characteristics and key features of the Power Sources field: recent prolific Power Sources authors; journals that contain numerous Power Sources papers; institutions that produce numerous Power Sources papers; keywords most frequently specified by the Power Sources authors; authors, papers and journals cited most frequently; pervasive technical themes of Power Sources; and relationships among the pervasive themes and sub-themes.

What is the importance of applying DT and bibliometrics to a topical field such as Power Sources? The roadmap, or guide, of this field produced by DT and bibliometrics provides the demographics and a macroscopic view of the total field in the global context of allied fields. This allows specific starting points to be chosen rationally for more detailed investigations into a specific topic of interest. DT and bibliometrics do not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but allow these detailed efforts to be executed more efficiently. DT and bibliometrics are quantity-based

measures (number of papers published, frequency of technical phrases, etc.), and correlations with intrinsic quality are less direct. The direct quality components of detailed literature investigation and interaction with performers, combined with the DT and bibliometrics analysis, can result in a product highly relevant to the user community.

2. BACKGROUND

2.1 Overview

The information sciences background for the approach used in this report is presented in [27]. This reference shows the unique features of the computer and co-word-based DT process relative to other roadmap techniques. It describes the two main roadmap categories (expert-based and computer-based), summarizes the different approaches to computer-based roadmaps (citation and co-occurrence techniques), presents the key features of classical co-word analysis, and shows the evolution of DT from its co-word roots to its present form.

The DT method in its entirety requires generically 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 are summarized after the present section. Time evolution of themes has not yet been studied.

At this point, a variety of different analyses can be performed. For databases of non-journal technical articles [26], the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting sub-thrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated bibliometric information (authors, journals, addresses, etc), the final results have also included relationships among the technical themes and authors, journals, institutions, etc [27-32].

These more recent DT/ bibliometrics studies were conducted of the technical fields of: 1) Near-earth space (NES) [28]; 2) Hypersonic and supersonic flow over aerodynamic bodies (HSF) [27]; 3) Chemistry (JACS) [29] as represented by the Journal of the American Chemical Society; 4) Fullerenes (FUL) [30]; 5) Aircraft (AIR) [31]; 6) Hydrodynamic flow over surfaces (HYD); 7) Electrochemical Power Sources (ECHEM) [32]; and 8) the non-technical field of research impact assessment (RIA) [29]. Overall parameters of these studies from the SCI database results and the current EPS study are shown in Table 1.

First Step

The frequencies of appearance in the total text of all single word phrases (e.g., Matrix), adjacent double word phrases (e.g., Metal Matrix), and adjacent triple word phrases (e.g., Metal Matrix Composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database.

Second Step

Numerical Boundaries

For each theme phrase, the frequencies of phrases within +/-M (nominally 50) words of the theme phrase are computed for every occurrence of the theme phrase in the full text, and a phrase frequency dictionary is constructed. This dictionary contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster) yielding, among many results, those sub-themes closely related to and supportive of the main cluster theme.

Threshold values are assigned to the numerical indices, and these indices are used to filter out the phrases most closely related to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and sub-themes, the qualitative analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allow an understanding of the theme inter-relationships not heretofore possible with previous text abstraction techniques (using index words, key words, etc.).

Semantic Boundaries

The approach is conceptually similar to that of the previous section, with the difference being that semantic boundaries are used to define the co-occurrence domain rather than numerical boundaries. The only semantic boundaries used for the present studies were paper Abstract boundaries. Software is being developed that will allow paragraphs or sentences to be used as semantic boundaries.

It is an open question as to whether semantic boundaries or numerical boundaries provide more accurate results. The elemental messages of text are contained in concepts or thoughts. Sentences or paragraphs are the vehicles by which the concepts or thoughts are expressed. The goal of text mining is to usually quantify relationships occurring in the concepts or thoughts, not in the fragments of their vehicles of expression. In particular, while intra-sentence relationships will be very strong, they may be overly restrictive for text mining purposes, and many cross-discipline relationships can be lost by adhering to intra-sentence relationships only. Intra-paragraph relationships are more inclusive and reasonable. For journal paper Abstracts of the type found in SCI, many Abstracts constitute a single paragraph.

More generally, there is a tradeoff between recall (completeness of information retrieval) and precision (correctness of information retrieval) as the domain in which co-occurrence is measured changes. Co-occurrence within the smallest units (e.g., phrases) provides high precision, while co-occurrence within the largest units (e.g., full article or full report) provides high recall. In the latter case, if the two co-occurring words/ phrases are physically positioned very far apart, co-occurrence may have no meaning. A recent study quantifies some of the precision-recall tradeoffs for different units, ranging from phrases to Abstracts [33].

2.2 Unique Study Features

The study reported in the present report is in the latter (journal article abstract) category. It differs from the previous published papers in this category [27-32] in four respects. First, the topical domain (power sources) is completely different. Second, a more rigorous technical theme clustering approach is used. Third, the phrase-based query approach has been supplemented by the journal-title-based query approach. Fourth, since estimation of relative global levels of emphasis in power sources was desired, a generic power sources query was used in both the phrase-based and journal-title-based queries (e.g., ELECTRICITY PRODUCTION), rather than using power source-specific terms (e.g., FUEL CELL). A companion study [32] examines the more specific sub-area of ELECTROCHEMICAL POWER SOURCES using specific terms rather than the generic terms.

3. DATABASE GENERATION

The key step in the power source literature analysis is the generation of the database. There are three key elements to database generation: the overall objectives, the approach selected, and the database used. Each of these elements is described.

3.1 Overall Study Objectives

The main objective was to identify global S&T that had both direct and indirect relations to Power Sources. One sub-objective was to estimate the overall level of global effort in Power Sources S&T, as reflected by the emphases in the published literature. Another sub-objective was to determine whether any radically new power sources were under development.

It was believed that if known specific technical terms were used for the query, there would be three negative impacts relative to the objectives above. First, the query would be biased toward the specific technologies reflected in the query, and the records retrieved would reflect this bias. The relative global efforts devoted toward each technology would have little credibility. Second, use of specific technical terms in the query would identify advances made in existing technologies, but might not access radically new technologies. Third, the query size would have been unmanageable, and unusable in present search engines. An unpublished study of controlled fusion energy resulted in a query of hundreds of terms after only the first iteration. The companion study to the present study, on the topic of electrochemical power sources, generated a query with hundreds of terms. Summing this experience over all the source, converter, and storage technologies contained within the umbrella of power sources S&T would have generated many hundreds or thousands of query terms.

Thus, it was decided to use generic energy or power-related terms for the query, relatively independent of any specific power supply, conversion, or storage system (e.g., ELECTRICITY PRODUCTION vs LIGHT-WATER REACTOR). This approach would retrieve documents that described technologies specifically related to power production, conversion, and storage. To retrieve documents related to power production, but where the author may not have used specific terminology relating the technology to power production in the write-up, the journal-based

approach was added. The concept was to identify power source journals that were generic, not source specific, and add their articles to the phrase-based query database.

However, even with the use of both approaches, one class of articles will not be retrieved. These are power source-related articles that do not contain the generic terms relating them to power sources, nor are published in a journal with a dedicated power source emphasis. Thus, an article on a new scientific phenomenon potentially related to power sources that was published in, for example, Science or Nature would not appear in this retrieval. To retrieve such articles, a detailed technology-specific query, such as the type developed in past DT studies, is required. A companion study on Electrochemical Power Sources developed such a query [32].

3.2 Databases and Approach

The Science Citation Index [34] was the database used for the present study. The approach used for query development was the DT-based iterative relevance feedback concept [25].

The database consists of selected journal records (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for power source articles. At the time the present report was written, the Web version of the SCI accessed about 5600 journals (mainly in physical, engineering, and life sciences basic research).

The SCI database selected represents a fraction of the available Power Source (mainly research) literature, that in turn represents a fraction of the Power Source S&T actually performed globally [35]. It does not include the large body of classified literature, or company proprietary technology literature. It does not include technical reports or books or patents on Power Sources. It covers a finite slice of time (1991 to late 2000). The database used represents the bulk of the peer-reviewed high quality Power Source science and technology documented.

To extract the relevant articles from the SCI, the phrase-based query and the journal-title-based query were used, and the results combined with duplications eliminated. For application of the phrase-based query, the Title, Keyword, and Abstract fields were searched using phrases relevant to power sources. The resultant Abstracts were culled to those relevant to power sources. The search was performed with the aid of two powerful DT tools (multi-word phrase frequency analysis and phrase proximity analysis) using the process of Simulated Nucleation [25].

An initial query of generic power source-related terms produced two groups of papers: one group was judged by domain experts to be relevant to the subject matter, the other was judged to be non-relevant. Gradations of relevancy or non-relevancy were not considered. An initial database of Titles, Keywords, and Abstracts was created for each of the two groups of papers. Phrase frequency and proximity analyses were performed on this textual database for each group. The high frequency single, double, and triple word phrases characteristic of the relevant group, and their boolean combinations, were then added to the query to expand the papers retrieved. Similar phrases characteristic of the non-relevant group were effectively subtracted from the query to contract the papers retrieved. The process was repeated on the new database of Titles, Keywords, and Abstracts obtained from the search. A few more iterations were performed until the number

of records retrieved stabilized (convergence). The final approximately 400 term phrase-based query used for the Power Source study is shown in Appendix 1.

The query consists of two components. The first component consists of phrases and phrase combinations designed to access mainly relevant records (e.g., bio-mass energy, power conversion, energy storage). The second component consists of phrases and phrase combinations designed to remove non-relevant records (e.g., leptin, lunch, spawning, muscle, women). Thus, the first component increases the comprehensiveness of the retrieval (recall), while the second component increases the signal-to-noise ratio (precision) by removing the noise.

For application of the journal-title-based query to the SCI database, articles contained in the 68 journals classified by the SCI under the category Energy and Fuels were sampled. Those journals that were not power-source specific, and that contained a very high fraction of articles deemed relevant to the Power Source topic, were identified, and all their articles were included in the retrieved database. The final journal title-based query used for the Power Source study identified the eleven journals shown in the Introduction.

The authors believe that queries of these magnitudes and complexities are required when necessary to provide a tailored database of relevant records that encompasses the broader aspects of target disciplines. In particular, if it is desired to enhance the transfer of ideas across disparate disciplines, and thereby stimulate the potential for innovation and discovery from complementary literatures [36-38], then even more complex queries using Simulated Nucleation may be required.

4. RESULTS

The results from the publications bibliometric analyses are presented in section 4.1, followed by the results from the citations bibliometrics analysis in section 4.2. Results from the DT analyses are shown in section 4.3. The SCI bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, and Keywords. In addition, the SCI included references for each paper.

4.1 Publication Statistics on Authors, Journals, Organizations, Countries

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

Author Frequency Results

There were 20825 papers retrieved, 34808 different authors, and 60493 author listings. The occurrence of each author's name on a paper is defined as an author listing. While the average number of listings per author is about 1.7, the ten most prolific authors (see Table 2) have listings more than an order of magnitude greater than the average. The number of papers listed

for each author are those in the database of records extracted from the SCI using the query, not the total number of author papers listed in the source SCI database.

Of the ten most prolific authors listed in Table 2, four are from India, three are from the UK, and one each from the USA, Japan, and Saudi Arabia. All are from universities. This prolific author country distribution differs radically from any in previous studies [27-32], with the high concentration from India. These prolific author countries in previous text mining studies tended to be dominated by Northern America countries (United States and Canada), the most developed Western European nations (UK, Germany, France, Italy), and the major oriental Asian countries (Japan, China, South korea). In these previous text mining studies, the prolific author country distributions tended to align with the prolific country distributions. In the present report, the prolific country distributions follow the conventional pattern above (shown later), contrary to the prolific author country distributions. The electrochemical power sources study [32] showed 65% of the prolific authors from the Far East, mainly Japan and China.

Because of the nature of the query used in the present study, many traditional energy production and conversion technologies were included (solar cooking, solar drying, solar distillation, biomass, coal combustion, etc). Reading of thousands of Abstracts confirmed that much of the Power Sources S&T focused on relatively low technology traditional approaches, especially research from the developing countries. The most prolific Indian authors addressed the solar and biomass topics. Interestingly, the most prolific British authors all concentrated on coal, including combustion, properties, and gasification.

Journals Containing Most Power Sources Papers

There were 1422 different journals represented. This is twice the number of journals from any of the previous studies [27-32], and again reflects the multi-disciplined nature of EPS. There was an average of 14.64 papers per journal. This number is somewhat inflated compared to the journal averages from other text mining studies [27-32]. In the journal-derived component of the present study, all the papers in eleven journals were used. Nevertheless, even for those journals identified by the query-derived component of the database, the journals containing the most Power Source papers had in some cases an order of magnitude more papers than the average (See Table 3).

The journals cover a wide range of energy themes. These include Combustion/ Propulsion (Journal of Propulsion and Power, Combustion Science and Technology, Combustion and Flame, Combustion and Explosion), Converters (Journal of Engineering for Gas Turbines and Power-Transactions of the ASME, Brennstoff-Warme-Kraft , IEEE Transactions of Energy Conversion, IEEE Transactions of Power Systems), Thermal Engineering (Applied Thermal Engineering, JSME International Journal Series B – Fluids Thermal Engineering), Renewables (International Journal of Hydrogen Energy, Biomass and Bioenergy, Solar Energy), Electrochemistry (Solid State Ionics, Journal of the Electrochemical Society), Physics/ Magnetics (IEEE Transactions on Magnetics, Journal of Applied Physics, Fusion Technology), and General/ Policy (Energy Policy, Applied Energy). They do not cover the more fundamental science journals (e.g., Science, Nature, Physics of Fluids, Journal of Chemical Physics), since the query had a power/ energy sources focus.

Institutions Producing Most Power Sources Papers

A similar process was used to develop a frequency count of institutional address appearances. It should be noted that many different organizational components may be included under the single organizational heading (e.g., Harvard Univ could include the Chemistry Department, Biology Department, Physics Department, etc.). Identifying the higher level institutions is instrumental for these DT studies. Once they have been identified through bibliometric analysis, subsequent measures may be taken (if desired) to identify particular departments within an institution.

Of the ten most prolific institutions listed in Table 4, four are from the Far East, two are from Western Europe, two from the USA, one from Eastern Europe, and one from the Middle East. Five are universities, and the remaining five institutions are research institutes. Compared to previous studies [27-32], the ratios of research institutes to universities is relatively high in this study.

Typically, the ratio of research institutes to universities has been in the vicinity of 10-20%. The higher ratio in the present study is indicative of the applied focus of the query and retrievals, where it would be expected that more of the effort would be conducted in research institutes or industry.

Countries Producing Most Power Sources Papers

There are 78 different countries listed in the results. The country bibliometric results are summarized in Table 5. The dominance of a handful of countries is clearly evident.

There appear to be three dominant groups in the twenty most prolific countries. The US and Japan constitute the most dominant group. England, India, Germany, Canada, and France constitute the next group, and the remaining countries constitute the third group. This is the prolific country distribution pattern typical of past text mining studies [27-32].

Of these top twenty countries, two are from North America, five are from the Far East, nine are from Western Europe, two are from Eastern Europe, and two are from the Middle East. South America and Africa are not represented.

Weighting these regions by number of papers, the ranking is North America (6282), Western Europe (5803), Far East (4970), Eastern Europe (720), and Middle East (542). When total population and GDP are taken into account, some dramatic changes occur. For papers per unit of population in the top twenty, the top five are mainly Western European and English-speaking nations (SWEDEN, CANADA, AUSTRALIA, UK, NETHERLANDS), and the bottom five are dominated by Asia and Eastern Europe (CHINA, INDIA, RUSSIA, EGYPT, POLAND). For papers per unit of GDP in the top twenty, the top five are mainly developed nations (SWEDEN, AUSTRALIA, CANADA, GREECE, EGYPT), and the bottom five are a more amorphous mix (CHINA, SOUTH KOREA, RUSSIA, ITALY, USA). Interestingly, for all three productivity measures, Canada, Australia, and Sweden rank high.

Figure 1 contains a co-occurrence matrix of the top 15 countries. In terms of absolute numbers of co-authored papers, the USA major partners are Canada, Japan, Germany, England, China, and France. Overall, countries in similar geographical regions tend to co-publish substantially, although the larger producers (e.g., USA, Japan) are universal in their co-publishing.

Figure 2 contains a Country-Time matrix, where the matrix elements are numbers of papers produced. The year 2000 results are only partially complete. Country productivity varied considerably as a function of time. For example, over the decade the USA increased number of papers by only a few percent. Japan doubled, England, India, Germany increased by about 50%, and China, South Korea, and Turkey approximately quintupled.

Figure 3 contains a Country-Journal matrix, for the top fifteen countries and top seventeen journals. The matrix entries are expressed in decimal fraction of each country's total papers in the seventeen journals. For each country, the bulk of its papers are contained in about four of the seventeen journals (i.e., journals containing about ten percent or more of a country's total papers).

In decreasing order, the four main journals for USA papers are: ENERGY & FUELS, FUEL, J POWER SOURCES, ENERGY. The papers in Energy & Fuels focus mainly (not exclusively) on fossil fuel properties, combustion efficiencies and pollution. The papers in Fuel focus mainly (with some biomass exceptions) on fossil fuel properties, additives, and reactant product properties and utilization. The papers in Journal of Power Sources focus on electrochemical power supply, with main emphasis on batteries and fuel cells. The papers in Energy focus on energy utilization, with emphasis on increasing efficiency and alternatives to reduce pollution.

For India, the five journals are: ENERGY CONV MANAG, INT J ENERGY RES, J POWER SOURCES, RENEW ENERGY, FUEL. The papers in Energy Conversion & Management focus on energy utilization, aimed at improving energy efficiency and reducing pollutants, with balanced emphasis given to solar and biomass systems. The papers in International Journal of Energy Research focus on performance of total energy systems and components, with reasonable emphasis provided to solar energy systems. The papers in Journal of Power Sources focus on rechargeable batteries and fuel cells. The papers in Renewable Energy focus on alternative energy sources and utilization, with focus on solar, but inclusion of biomass and other renewables like wind as well. The papers in Fuel focus on properties and combustion products of (mainly) fossil fuels. While there is overlap with the USA in technical areas studies, there appears to be much more relative emphasis in solar-based systems and alternative power supplies in India relative to the USA.

For China, the four journals are: J POWER SOURCES, FUEL, ENERGY CONV MANAG, ENERGY. The papers in Journal of Power Sources focus on batteries (mainly rechargeable lithium) and fuel cells. The papers in Fuel focus on properties, combustion, and products of (mainly) fossil fuels, and, of those, almost exclusively on coals. The papers in Energy Conversion and Management focus on analysis of energy conversion and utilization across a wide variety of systems and applications. The papers in Energy focus on analysis and modeling of energy utilization in a wide variety of systems and applications. Relative to India, China has less focus on the solar and other alternative supplies, and more on fossil fuel combustion. All the

above conclusions are based on these four or five major publishing journals' contents only, for each country.

4.2 Citation Statistics on Authors, Papers, and Journals

The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics [39], much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers [40-41].

The citations in all the retrieved SCI papers were aggregated, the authors, specific papers, years, journals, and countries cited most frequently were identified, and were presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations. From the citation year results, the most recent papers tended to be the most highly cited. This reflected rapidly evolving fields of research.

4.2.1 Most Cited Authors

The most highly cited authors are listed in Table 6.

Of the twenty most cited authors, eight are from the USA, four are from Japan, five are from Western Europe, one from Israel, one from Bulgaria, and one from China. This is a far different distribution from the most prolific authors, where half were from Asia, and ten percent from the USA. There are a number of potential reasons for this difference, including difference in quality and late entry into the research discipline. In another three or four years, when the papers from present-day authors have accumulated sufficient citations, firmer conclusions about quality can be drawn.

Ten of the most cited authors worked on fossil fuels (mainly coal, mainly combustion), five worked in thermodynamics, three worked on batteries (mainly lithium), one worked on solar, and one worked on polymers.

The lists of most prolific authors and most highly cited authors only had one name in common (WU, C). This phenomenon of minimal intersection has been observed in all other text mining studies performed by the first author. The time frame of interest for most prolific authors is present time, whereas the time frame of interest for most cited authors can span many decades. Researchers who may very well have been prolific when their most citable work was done may no longer be prolific. They may have left the discipline, may have assumed non-research duties, or may have slowed down. As the gap between their most citable work and the present widens, the validity of this statement increases.

Sixteen of the authors' institutions are universities, two are government-sponsored research laboratories, and two are private companies. The appearance of the companies on this list is another differentiator from the list of most prolific authors.

The citation data for authors and journals represents citations generated only by the specific records extracted from the SCI database for this study. It does not represent all the citations received by the references in those records; these references in the database records could have been cited additionally by papers in other technical disciplines.

Most Cited Papers

The most highly cited papers are listed in Table 7.

The theme of each paper is shown in italics on the line after the paper listing. The order of paper listings is inverse number of citations by other papers in the extracted database analyzed. The total number of citations from the SCI paper listing, a more accurate measure of total impact, is shown in the last column on the right. Papers more closely linked to energy applications, such as those on coal, capture many of the total citations (about half) within the present database. The more fundamental science-oriented papers tend to be referenced by myriad disciplines, and the papers within the present database capture a much smaller fraction of the total citations (in some cases, near ten percent of the total).

Energy and Fuels contains the most papers, four out of the ten listed. Most of the journals are fundamental science journals, and most of the topics have a fundamental science theme. Most of the papers are from the 1989-1990 time frame. This reflects a dynamic research field, with seminal works being performed in the recent past.

Six papers focus on coal issues, one on combustion, one on thermodynamics, and two on secondary lithium battery issues. Thus, the intellectual heritage focus is on conversion to electricity with a thermal step, as opposed to direct conversion to electricity. Even though the text analysis will show later a significant effort on renewables, this level of effort is not reflected in the intellectual heritage.

Most Cited Journals

Fuel received almost as many citations as the next three journals combined. Most of the highly cited journals are fossil fuel/ combustion oriented or electrochemical power source oriented. These are followed by some fundamental Chemistry and Physics journals. The only renewables journal interspersed is Solar Energy. These results are fully in line with those of the most cited authors and papers, and suggest that consensus seminal works have yet to be established for many of the renewables areas.

The authors end this bibliometrics section by recommending that the reader interested in researching the topical field of interest would be well-advised to, first, obtain the highly-cited papers listed and, second, peruse those sources that are highly cited and/or contain large numbers of recently published papers.

4.3 Database Tomography Results

There are two major analytic methods used in this section to generate taxonomies of the SCI databases: non-statistical clustering, based on phrase frequency analysis, and statistical clustering, based on phrase proximity analysis. Non-statistical clustering is performed on the Keywords and Abstracts fields. Statistical clustering is performed on the Abstracts field only.

Non-Statistical Clustering Taxonomies

Keyword Taxonomy

All the Keywords from the extracted SCI records, and their associated frequencies of occurrence, were tabulated, and then grouped into categories by visual inspection. The phrases were of two types: system-related and tech base-related. While the system sub-categories were relatively independent, there was substantial overlap between some of the tech base categories. These results are summarized now.

There are three Source categories: fossil, renewables, nuclear. Fossil focuses on COAL and its components, OIL, and GAS; Renewables focuses on BIOMASS, SOLAR, HYDROGEN, WIND, and GEOTHERMAL; Nuclear focuses on FISSION and FUSION.

Fossil and renewables dominate in terms of phrase frequencies, with much less emphasis on nuclear. This is due to the following. There are three major journal types in the SCI that serve as sources of papers. First, there are the fundamental multi-discipline journals, such as Science and Nature. These journals would contain papers focused on the fundamental energy conversion phenomena. Because of the high tech nature of these journals, they would have a higher fraction of nuclear-related articles than are reflected in the Keyword analysis of the present study. These papers would have a higher probability of being accessed through phenomena-related terms, rather than the specific energy production and conversion terms in the query used to generate part of the overall database in this study.

The second journal type is generic power-oriented. These journals constituted the journal-derived component of the total database used in this study, and are listed in the Introduction. The journals in this category contain basic and applied research papers, but on average, as will be shown later, tend to emphasize fossil, electrochemical, and traditional renewables, with very modest representation of fusion, fission, MHD, and more exotic renewables.

The third journal type is specific power-oriented, and the thirty journals in this category are listed in Table 9. These journals were not added to the total database in full, as were the generic power-oriented, for the reasons provided in the database generation section. Their representation in the total database derived from their papers that were accessed by the query. Half of these journals were devoted to nuclear energy and power. It appears that the nuclear S&T community publishes mainly in the first and third types of journals, especially in their dedicated literatures for the more applied S&T.

Thus, the observation that nuclear Keywords/ frequencies are a small fraction of the fossil and renewables Keywords/ frequencies should not be interpreted that nuclear source S&T is not being performed or is not important. The proper interpretation is that when power source-related

nuclear S&T is examined within the overall power source-related S&T, the high and low tech non-nuclear S&T performed globally dominate the higher tech nuclear S&T performed in a smaller number of the more developed countries. To obtain a more detailed picture of the advances in nuclear power S&T, a standard DT focused analysis of the literature would need to be performed. Detailed technical terms would be used in the query, and the fifteen nuclear-specific journals listed in Table 9 could be added to form the total database.

Now the description of the specific Keyword results of this study continues. Following the Fuel Sources category, there is a Fuel Processing category that includes fossil, renewables, and nuclear. The capitalized phrases within a category are listed in approximate declining occurrence frequency order, and therefore provide some indication of relative emphasis.

The main fossil component includes GASIFICATION, LIQUEFACTION, ALKYLATION, DESULFERIZATION, and ELECTROCATALYSIS.

The secondary renewables component includes SUPERCRITICAL FLUID EXTRACTION, FERMENTATION, BIOMASS GASIFICATION, WATER VAPOR GASIFICATION, BIOMASS LIQUEFACTION, MICROBIAL DESULFURIZATION, BIODESULFURIZATION, THERMAL-DECOMPOSITION, and BIODEGREDATION. At the higher Keyword frequencies, nothing was evident for nuclear.

There are two major categories of Converters: Thermal and Direct. The Thermal Converters involve a high temperature heat engine cycle step in the conversion to electricity, while the Direct Converters bypass the thermal step.

Thermal Converter categories include conversion Processes, Products, Processed Products, Product Impacts, Components, and Systems.

Processes include COMBUSTION, PYROLYSIS, CATALYSIS, and INCINERATION.

Products generated include EMISSIONS, CHAR, POWER, HEAT, and ASH. These Products may be Processed (CO2 REMOVAL, DC-DC POWER CONVERSION, EMISSION CONTROL), and their major side impacts are global warming and climate.

Major Components used include CATALYSTS (See CATALYSIS above), FLUIDIZED BEDS, and SOLAR COLLECTORS. Major Converter Systems examined include HEAT PUMP, HEAT ENGINES, TURBINES, and SOLAR.

Direct Converter categories include Reactants, Processes, Products, Components, and Systems.

Direct Converters emphasize Lithium Reactants, the three Processes of ELECTROCHEMISTRY, MHD, and PHOTOSYNTHESIS, and yield Products of essentially POWER, with no negative impacts emphasized. Major Components used include ELECTRODES, ELECTROLYTES, MEMBRANES, and SOLAR COLLECTORS. Major Direct Converter Systems include FUEL CELLS and PHOTOVOLTAICS.

Storage has two major sub-divisions, Electrochemical and Mechanical. Electrochemical Storage may be divided further into Reactants, Process, Products, Components, and Systems.

Electrochemical Reactants emphasize Lithium, and Processes include DISCHARGE, ELECTROCHEMISTRY, OXYGEN REDUCTION, CYCLIC VOLTAMMETRY, and PREMATURE CAPACITY LOSS.

Components include ELECTRODES, ELECTROLYTES, MEMBRANES, POLYANILINE, and POLYPYRROLE, and Systems emphasize BATTERIES.

Mechanical Storage focuses almost exclusively on flywheels, and is sub-divided into Components (SUPERCONDUCTING MAGNETIC BEARINGS, COMPOSITE FLYWHEEL ROTOR, CONTROL SYSTEM), Operating Characteristics (HIGH CURRENT DENSITY, HIGH PEAK POWER OUTPUT, HIGH MAGNETIC FIELD, HIGH SPEED, HIGH ENERGY DENSITY), and Phenomena (TORQUE FLUCTUATION, MAGNETIC LEVITATION, FRICTION/ ROTATIONAL LOSS, ENERGY LOSS).

The above categorizations have been based on phrases that could be associated with specific Source, Converter, or Storage concepts. However, there were many generic Keywords that could not be associated with specific concepts, especially since co-occurrence matrices were not generated to identify such associations. These generic Keywords represent technology base efforts that underlay a number of the specific concepts. They are classified in the categories of Theory, Experiment/ Diagnosis, Parameters, Properties, Phenomena, Materials, and Geometries.

Theory includes MODELS and SIMULATION, while Experiment/ Diagnosis includes SPECTROSCOPY, SPECTROMETRY, CHROMATOGRAPHY, CALORIMETRY, DIFFRACTION, XPS, THERMOGRAVIMETRY, LASER, and APPARATUS.

Parameters/ Variables include TEMPERATURE, PRESSURE, ENVIRONMENT, ECONOMICS, DENSITY, TIME, CYCLE LIFE, ENTHALPY, COST, DEMAND, and THERMAL EFFICIENCY.

Properties include CONDUCTIVITY, SOLUBILITY, ELECTRICAL, THERMODYNAMIC, ELECTROCHEMICAL, OPTICAL, MAGNETIC, THERMOPLASTIC, MECHANICAL, PHYSICAL, FUEL, STRUCTURAL, CAKING, TRANSPORT, SURFACE, LOW-TEMPERATURE, THERMAL, COKING, PHOTOELECTROCHEMICAL, PHYSIOCHEMICAL, RHEOLOGICAL, SPECTROSCOPIC, THERMOCHEMICAL, THERMOPHYSICAL, TENSILE, COAL, COMBUSTION, COLD FLOW, RESISTIVITY, PASSIVATION, MOLECULAR WEIGHT, REFLECTANCE, HEAT CAPACITY, ACTIVATION ENERGY, and ENTROPY.

Phenomena include KINETICS, OXIDATION, REDUCTION, DECOMPOSITION, INTERCALATION, DEVOLATILIZATION, TRANSPORT, EXTRACTION, HYDROGENATION, DIFFUSION, EVOLUTION, CORROSION, STABILITY, INSERTION, ABSORPTION, SEPARATION, DEPOSITION, DYNAMICS, and a range of other broad phenomena including EXERGY, RECYCLING, RADIATION, REFRIGERATION,

DISSOLUTION, DRYING, FLUORESCENCE, RECOVERY, PROPAGATION, RELAXATION, COOLING, HEATING, CONVECTION, CLEAVAGE, DEACTIVATION, ACTIVATION, SCATTERING, DISPERSION, HYDRODENITROGENATION, RHEOLOGY, BOND SCISSION, HEAT TREATMENT, SORPTION, AGGREGATION, COMPRESSION, DIFFRACTION, DISTILLATION, DEMINERALIZATION, DESORPTION, INHIBITION, LATENT HEAT STORAGE, PRECIPITATION, CHEMISORPTION, FRACTIONATION, HYDROLYSIS, INSOLATION, INSTABILITY, IRRADIANCE, SOLIDIFICATION, INJECTION, IRREVERSIBILITY, MOISTURE CONTENT, POLARIZATION, SUBLIMATION, and SULFATION.

Materials include CARBON, LITHIUM, WATER,, SULFUR, OXYGEN, GRAPHITE, IRON, NITROGEN, NICKEL, AIR, and many others including ALLOYS, LEAD, POLYMERS, METALS, SOLVENTS, CALCIUM, PLATINUM, ALUMINUM, SILICON, MANGANESE DIOXIDE, PYRIDINE, STEAM, LIMESTONE, COBALT, TETRALIN, SEDIMENTS, TIN, AMMONIA, PITCH, COPPER, MINERALS, MANGANESE, MOLYBDENUM, CERAMICS, PEROVSKITE, ZIRCONIA, ZEOLITE, ZINC, ANTIMONY, POLYETHELYNE, CERIA, RESINS, COMPOSITES, POWDERS, SODIUM, CHLORINE, GAAS, PHASE CHANGE MATERIAL and POROUS MEDIA.

Geometries include FILMS, SURFACE, SIZE, PLATE, and LAYERS.

Abstract Taxonomy

A taxonomy of all energy-related technologies was developed through visual inspection of the Abstract phrase frequencies, and manual assignment of the phrases to categories. In this section, a four level taxonomy was necessary to provide sufficient detail on the various energy-related technologies. The first three levels of the taxonomy were developed using a phrase frequency-only analysis. Phrases generated with the phrase frequency analysis could be classified into two types of categories: system specific (e.g., COAL STRUCTURE, TOKAMAK, LITHIUM SECONDARY BATTERIES) and generic (CARBON, THERMAL DIFFUSION, REACTION RATES). Since one feature of the manually generated taxonomy was allocation of Abstract phrases and associated frequencies to specific categories, a method was required to relate the generic phrases to their associated specific systems (e.g., what fraction of the THERMAL DIFFUSION frequencies should be allocated to the Geothermal Sources category?). The method selected was to perform a proximity analysis using the third level taxonomy categories as themes. The third level of the taxonomy consisted exclusively of high technical content phrases that actually appeared in the phrase frequency analysis data, and were deemed as specific or systems technologies.

All the high to mid-frequency system specific phrases and system-related generic phrases could be rationally allocated to the categories in this taxonomy. The absence of any categories/ sub-categories in this taxonomy (e.g., Thermionics in the Direct Electrical Conversion) should not be interpreted that S&T efforts are not being pursued in these areas. The correct interpretation is that within the constraints of the EPS database, mid-high frequency phrases related to these categories do not appear.

Table 11 presents the taxonomy. The phrase frequency summations are shown in parentheses after each taxonomy category for the first three levels. Sample categories are shown for the fourth level. The categories will now be described.

Abstract Taxonomy Level 1

The highest taxonomy level consists of three categories: Primary Energy Sources (23422), Energy Converters (17481), and Energy Storage Devices (2901). The numbers in parenthesis after each category reflect the sum of the phrase frequencies in each category. While the sum of phrase frequencies in a category may give some indication of activity in that category, this approach intrinsically provides only a very approximate estimate of activity. A more accurate approach for estimating activity is presented later under document clustering, where the number of documents in each category is counted, and used to estimate activity.

These results suggest that Primary Energy Sources have more research activity than Energy Converters, and substantially more research than Energy Storage Devices. In an environment of increasingly scarce energy resources, developing new and affordable sources is of primary concern. Once the sources are defined, then focus on conversion and storage is appropriate. Additionally, energy needs to be converted to more usable forms before it can be stored in such forms. Therefore, substantially more research is performed on converters relative to storage.

Abstract Taxonomy Level 2

Primary Energy Sources – Level 2

Each of the categories in taxonomy level 1 can be subdivided into level 2 categories. Primary Energy Sources can be subdivided into Fossil Fuels (9509), Renewable Energy/ Alternative Fuels (12874), and Nuclear Fuels (1039). Renewable Energy/ Alternative Fuels has a modestly higher level of activity than Fossil Fuels. In the past, substantial R&D was performed on Fossil Fuels, with relatively smaller amounts of research on renewable sources. Because of the foreseeable future decline in Fossil Fuel resources, and the perceived reduced environmental impacts of renewable sources, there are a wealth of opportunities for advancement in renewable sources research, and this is reflected in the relative levels of effort. The reasons for low frequencies related to Nuclear are stated at the beginning of the Keyword taxonomy section (4.3.1.2).

The technical emphases of Fossil Fuel research are primarily increasing efficiency (THERMAL EFFICIENCY, CONVERSION EFFICIENCY, COMBUSTION EFFICIENCY, ENERGY CONSUMPTION) and reducing emissions (NITROGEN, SULFUR, ASH, CO2, SO2), with some emphasis on widening usage (GASIFICATION, LIQUEFACTION). The technical emphases of Renewable Energy/ Alternative Fuels are increased efficiency, reduced production and maintenance costs, increased commercial interest, and reduce environmental impact. The technical emphases of Nuclear Fuels research are safety, waste disposal, increased efficiency, and reduced life cycle costs.

The above technical emphases strictly apply to the full conversion cycle, not to the source fuels alone. It is very difficult to separate the conversion from the fuels for specific systems in research articles, since a research article on fuel sources (other than exploration or perhaps some stages of pre-processing) tends to incorporate some aspect of conversion.

Each of the categories in level two can be sub-divided into level 3 categories. Fossil Fuels was subdivided into Coal (4753), Oil (3148), and Natural Gas (1608). The major sub-categories of Coal were constituents/ characteristics/ properties and pre-processing/cleansing/ combustion. The major sub-categories of Oil were constituents/ types, conversion processes, and by-products. The major sub-categories of Natural Gas were types, cleansing, and by-products. The relative magnitudes of research reflect the relative usage diversity of each type, the magnitude of perceived resources available, the energy potentially extractable per resource unit, and the perceived marginal utility of additional research for increased energy extraction. It should be re-emphasized at this point that these conclusions are based on the published literature. If there is substantial proprietary research being done in one of these technology sub-areas relative to another sub-area (e.g., if the oil companies were doing substantially more proprietary research than the coal companies), then the total relative efforts among Coal, Oil, and Natural Gas would not be reflected by the numbers above.

Renewable Energy/ Alternative Fuels was subdivided into Solar Energy (4285), Hydrogen (3917), Biomass (2701), Wind Energy (1063), Geothermal Energy (844), and Hydropower (64). These five level 3 categories can be stratified into three groups. The largest group (Solar Energy, Hydrogen, and Biomass) has the common characteristics of non-site specificity and effective transportability. The next largest group (Wind Energy, Geothermal Energy) is constrained to geographical regions with favorable operating environments, but additional research is perceived as having the potential to produce substantial benefits at those sites. The smallest group (Hydropower), is also site constrained, but in addition is a mature technology. Hydropower articles address environmental issues (flood control, ecological damage) as much as technology improvement issues.

The major sub-categories of Solar Energy were conversion system characteristics, conversion system components, conversion system processes, and applications. Photovoltaics is classified under Converters. The major sub-categories of Hydrogen were materials/ compounds and conversion processes. The major sub-categories of Biomass were sources, types, and conversion processes. The major sub-categories of Wind Energy were converter systems and applications. The major sub-categories of Geothermal Energy were sources and applications, and the major sub-categories of Hydropower were environmental protection and applications.

Nuclear Fuels was subdivided into Fission (712) and Fusion (327). The Fission component is a mature technology (proof-of-principle was demonstrated sixty years ago), and the research focuses on cost, safety, environmental, and health issues resulting from operational experiences. The Fusion component is in the proof-of-principle stage, and the research focuses on predicting/ demonstrating ignition and burn, as well as cost and size reduction, and maintenance and cleanup issues. Because of the nature of the query used (linked to power plant production issues), the Fusion papers are further under-represented relative to Fission papers due to the different levels of maturity and linkage to power production terminology.

Energy Converters – Level 2

Energy Converters can be divided into Thermal Converters (12514), Direct Electric Converters (4441), and Nuclear Converters (526). The research effort in Thermal Converters is significantly larger than in Direct Electric Converters because of the larger embedded operational base in Thermal Converters (and therefore larger payoffs for small improvements), and the higher technology threshold required to perform research in Direct Electric Converters. Nuclear Converters is substantially smaller than either because of the reasons described in the Keywords section.

Thermal Converters can be subdivided into Engines (7543) and Turbines (4971). There is more research effort on Engines because of the diversity of types and applications of Engines, as well as the pollution control issues unique to automotive engines, where a main target of pollution reduction research is improvement of the combustion process. The major Engine sub-categories include engine types, engine components, engine characteristics, conversion processes, conversion by-products, and engine fuels. The latter sub-category contained a number of examples of mixed fossil-alternative fuel combinations. The major Turbine sub-categories include fuels, turbine and conversion cycle types, and conversion processes. Acoustics, mixing, and combustion chemistry are focal research areas in the combustion chamber. Heat transfer at the blade, and the underlying flowfield and turbulence transition phenomena, tend to dominate the conversion section research.

Direct Electric Converters can be subdivided into Fuel Cells (3154), Photovoltaics (1096), Thermoelectric (106), and MHD (85). Fuel Cells are researched most heavily because of wider diversity applications, higher efficiency potential, and higher power density. Photovoltaics is researched more than Thermoelectrics because the light sources (sun, room lighting) required for input are readily available, compared to the requirement for high temperature heat sources for Thermoelectrics. In addition, the light sources are lower entropy than the heat sources, offering the potential for higher conversion efficiency, and the potential improvement in conversion efficiency for Photovoltaics has been, and promises to be, substantially higher than for Thermoelectrics. MHD research is minimal due to technical difficulties caused by very high temperature gases operating in close proximity to super-cooled magnets.

Fuel Cell sub-categories include: higher longevity and efficiency component technologies (ELECTROLYTES, ANODES, CATHODES); diverse fuel cell types (SOLID OXIDE, MOLTEN CARBONATE, POLYMER ELECTROLYTE, DIRECT METHANOL, PHOSPHORIC ACID, PROTON EXCHANGE MEMBRANE); candidate fuels (HYDROGEN, METHANOL, NATURAL GAS), and component materials (NAFION, YSZ, POLYMERS, CERAMICS, LANTHANUM, PLATINUM, NICKEL, CARBON). Photovoltaic sub-categories include conversion/ quantum efficiency improvement and cost reduction, with emphasis on: component materials (AMORPHOUS SILICON, CRYSTALLINE SILICON, TIN OXIDE, GA SE-2, LITHIUM NIOBATE, INDIUM TIN, CADMIUM TELLURIDE, GAAS, RU NCS); electrical properties (ELECTRON TRANSFER, BAND GAP, OPEN-CIRCUIT VOLTAGE, CHARGE TRANSFER, SHORT-CIRCUIT CURRENT, CHARGE SEPARATION, DIFFUSION LENGTH, CONDUCTION BAND, CHARGE CARRIERS, CURRENT

DENSITY); optical properties (SPECTRAL RESPONSE, PHOTOVOLTAIC RESPONSE, LIGHT ABSORPTION, OPTICAL ABSORPTION, ABSORPTION COEFFICIENT, ABSORPTION SPECTRA); fabrication techniques (CHEMICAL VAPOR DEPOSITION, GLASS SUBSTRATES, COMPOSITE FILMS, CHEMICAL BATH DEPOSITION, MOLECULAR BEAM EPITAXY) and applications (SOLAR CELLS, PHOTOVOLTAIC DEVICES/ MODULES/ SYSTEMS, ELECTRIC ENERGY, RURAL ELECTRIFICATION, POWER PLANTS).

Energy Storage Devices – Level 2

Energy Storage Devices can be divided into Electric (2774) and Mechanical (127). With no rotating parts and high energy density per unit weight, Electric storage is the preferred approach. Electric can be sub-divided into Battery (2400), Capacitor (334), and Superconducting Magnetic Energy Storage (SMES) (40). Relative to batteries, capacitors have a virtually unlimited cycle life and rapid charging, but low energy density and high self discharge. Even the most promising capacitors, electrochemical super-capacitors, have an energy density an order of magnitude or more less than batteries. Further, their thin insulators limit voltages because of breakdown, and slow ionic liquid conduction limits discharge rate. For these reasons, battery research substantially outpaces capacitor research for energy storage. SMES differs from the other storage approaches in its ability to charge and discharge energy rapidly. The SMES technology is therefore suitable in applications that require repeated pulses of large amounts of active power for a short duration of time. Because it is viewed presently as a niche technology, research level is limited.

Major battery sub-categories include: Types (LITHIUM RECHARGEABLE, LITHIUM ION, LITHIUM POLYMER, LEAD-ACID, NICKEL-METAL HYDRIDE, ALKALINE, SILVER-ZINC, NICKEL-ZINC); Components (ELECTRODES [COMPOSITE CATHODE, CARBON ANODE], ELECTROLYTES [POLYMER, LIQUID, GEL, FLOODED], SEPARATORS, PLATES, STRAPS, COPPER CURRENT COLLECTOR); Materials (LITHIUM [LI, LIMN2O4, LICOO2, LITHIUM METAL, LINIO2], POLYMER, ALLOYS, CARBON [GRAPHITE], METAL, ACID, NICKEL [NI, NICKEL HYDROXIDE, NICKEL-CADMIUM, HYDROGEN [HYDRIDE]); Processes/ Phenomena (DISCHARGE, CYCLING/ CYCLES, INTERCALATION, CORROSION, CHARGING, CHARGE-DISCHARGE, OXIDATION, RECOVERY, REDOX FLOW, CAPACITY LOSS, SELF-DISCHARGE, OVERCHARGE, GRID CORROSION); Properties (CAPACITY [DISCHARGE CAPACITY, SPECIFIC CAPACITY], ENERGY DENSITY, VOLTAGE, INTERNAL IMPEDANCE, ONDUCTIVITY, COULOMBIC EFFICIENCY; RESISTANCE); and Characteristics (RECHARGEABILITY, CYCLE PERFORMANCE, ELECTROCHEMICAL STABILITY, SEALED, HIGH ENERGY, AMORPHOUS, PORTABLE, AQUEOUS, LITHIATED, HIGH CAPACITY, CONDUCTIVE, IMPLANTABLE, LAMINATED, LIGHTWEIGHT).

Major capacitor sub-categories include: Structure (THIN FILMS, OXIDE FILMS, DOUBLE LAYER, SI SUBSTRATES, BUFFER LAYER, BOTTOM ELECTRODE, TOP ELECTRODES), Fabrication (CHEMICAL VAPOR DEPOSITION, DEPOSITION TEMPERATURE, MAGNETRON SPUTTERING, SINTERING TEMPERATURE, FILMS ANNEALED, PULSED LASER DEPOSITION), Materials (SIO2, PZT, BA SR, ZR TI, PT,

SRBI2TA2O9 SBT, ACTIVATED CARBON, SR TIO3, LEAD ZIRCONATE, TI O-3, RUTHENIUM OXIDE, PB ZR, LEAD ZIRCONATE TITANATE),
Properties/ Characteristics/ Environment (DIELECTRIC CONSTANT, ELECTRICAL PROPERTIES, DIELECTRIC PROPERTIES, SPECIFIC CAPACITANCE, FERROELECTRIC PROPERTIES, BARRIER HEIGHT, ACTIVATION ENERGY, REMANENT POLARIZATION, GRAIN SIZE, OXIDE THICKNESS, SURFACE AREA, GRAIN BOUNDARIES, POWER CONSUMPTION, DYNAMIC RANGE, TEMPERATURE RANGE, THERMAL STABILITY, SURFACE ROUGHNESS, ELECTRIC FIELD, CURRENT DENSITY, ROOM TEMPERATURE, COERCIVE FIELD, LOW TEMPERATURE, HIGH TEMPERATURE, BIAS VOLTAGE, POWER DENSITY, ENERGY DENSITY), Phenomena (LEAKAGE CURRENT DENSITY, PHASE TRANSITION, OXYGEN VACANCIES, HYSTERESIS LOOPS, DISSIPATION FACTOR, DIELECTRIC LOSS, RUTHERFORD BACKSCATTERING), Experiment (TRANSMISSION ELECTRON MICROSCOPY, SCANNING ELECTRON MICROSCOPY, X-RAY DIFFRACTION, ATOMIC FORCE MICROSCOPY, PHOTOELECTRON SPECTROSCOPY, CYCLIC VOLTAMMETRY, AUGER ELECTRON SPECTROSCOPY), System (MOS CAPACITORS, POWER SUPPLY, CAPACITOR BANK, FERROELECTRIC CAPACITORS, THIN FILM CAPACITORS, ENERGY STORAGE, MEMORY CELL, RANDOM ACCESS MEMORY, TRANSMISSION LINE, CERAMIC CAPACITORS, SUPERCAPACITORS).

The SMES study emphasis appears focused on cost reduction through use of high temperature superconductors and optimized coil configurations. Systems studies and testing appear to receive more emphasis than research.

Abstract Journal and Query-based Taxonomies

Traditionally, for DT studies, only the phrase-based query method has been used for database generation. In the EPS study, the hybrid information retrieval approach (phrase-based and journal-based queries) was utilized to ensure that the final, combined database of energy literature was comprehensive. As previously mentioned, the EPS database was constructed with two queries:

1. A Journal Title query where all SCI articles (1991 – 2000 inclusive) from 11 identified relevant energy journals were retrieved (JOURNAL QUERY)
2. A Phrase query, where SCI articles were retrieved by searching Title/ Keywords/ Abstract fields with a query of phrases and phrase combinations (PHRASE QUERY).

Subsequently, taxonomies were developed for each database (JOURNAL QUERY and PHRASE QUERY). The results were then merged to provide the overall EPS taxonomy structure in the previous section.

In this section, the two component taxonomy results are presented to elucidate the differences between the JOURNAL QUERY and PHRASE QUERY databases approaches.

In each case, the taxonomies were developed through visual inspection of the Abstract phrase frequencies, and manual assignment of the phrases and their frequencies to categories. This

resulted in system specific phrases and generic phrases. The third level phrases (system specific) were then used as themes in a proximity analysis. The generic phrases closely related to system specific phrases were identified through the proximity analysis, and grouped into categories (taxonomy level four).

A comparison of phrases selected to illuminate the differences between the two databases from the results of the JOURNAL QUERY and PHRASE QUERY database taxonomy development is presented in Table 10.

The journal database has a higher fossil emphasis compared to the query database, with additional concentration on the traditional combustion vessels (FURNACES, BOILERS). While the query database had more generic representation in biomass, the journal database had noticeably higher representation in the traditional types of biomass (FIREWOOD, RICE HUSK). The journal database had noticeably higher representation in the other types of renewables (WIND, GEOTHERMAL, HYDROPOWER, SOLAR). Not only are the numbers higher in the renewables for the journal database, but the emphases are different for the query and journal databases. For example, the PHOTOVOLTAICS component of solar, targeted at higher direct electricity conversion efficiencies, is substantially higher in the query database than the journal database. On the other hand, the non-direct electricity conversion component of solar (heat engine boiler, desalinization, hot water heater, solar refrigerator, distillation, water sterilization), as reflected in SOLAR COLLECTOR, is substantially larger in the journal database.

The nuclear energy technologies, high temperature plasma-based technologies, and mechanical energy storage had modest representation in the query database (for database selection reasons explained previously), and essentially no representation in the journal database.

Thermal Conversion methods were accessed equally by the journal and phrase queries. Direct Electric Conversion methods were also accessed equally by the journal and phrase queries. This is only because the Journal of Power Sources, which tends to have a heavy focus on "direct" electric converters and electric storage, especially electrochemical, was selected as one of the eleven journal query journals. The other direct electric converters, such as thermoelectric or MHD, were not well represented by the journal query.

The journal query retrieved most of the battery articles because of the Journal of Power Sources. Relatively few capacitor articles were retrieved. Mechanical Energy Storage articles were retrieved almost exclusively by the phrase query.

With the exception of the Journal of Power Sources, the journal query approach accessed generic energy related journals that, for the most part, focused on applied energy research. These journals reported on the numerous processes that utilize energy, and the potential that developed / developing energy sources / conversion methods could provide. Many of the contributors were from the developing countries, where those types of technologies could be readily produced and implemented.

This is substantially different from the articles retrieved from the specific phrase query, where the focus was well distributed among existing and developing primary sources of energy and the

fundamental technology issues with converting these sources in various energy-requiring applications. The contributors reflected, on average, the more developed countries, that have the resources to both develop and implement these technologies.

The absence of any categories/ sub-categories in this taxonomy should not be interpreted that S&T efforts are not being pursued in those areas. The correct interpretation is that within the frequency threshold constraints of the Power Sources database, mid-high frequency phrases related to these categories do not appear.

Statistical Clustering

Two generic types of statistical clustering were performed, concept clustering and document clustering. In concept clustering, words/ phrases are combined into groups based on their co-occurrence in documents. In document clustering, documents are combined into groups based on their text similarity. Document clustering yields number of documents in each cluster directly, a proxy metric for level of emphasis in each taxonomy category.

Statistical Concept Clustering

The purpose of the analysis was to identify relationships among the major technical themes, and among the major and minor themes, in the Abstract databases. The generic approach used was to identify the themes by extracting the high technical content phrases and their frequencies of occurrence, and then use statistical methods to relate the themes by combining similar phrases into thematically-related groups. While this approach has the similar overall objective of generating an EPS taxonomy as the manual approach described in the phrase frequency section, it has one critical difference. The manual approach defines phrase similarity by visual inspection based on analyst experience. The statistical approach defines the similarity of two phrases by the similarity of their co-occurrence profiles with other phrases. Neither approach is inherently superior. Each offers a unique perspective on the database structure.

To obtain the theme and sub-theme relationships, a phrase proximity, or clustering, analysis is performed about each selected theme phrase. Two clustering variants are used, and are eventually combined to exploit the strengths of each variant synergistically. The first variant uses the TextSlicer software from DT. All technical phrases are retrieved, but extensive manual cleanup is required. The second variant uses the TechOasis software from Search Technology. It is more automated than TextSlicer presently, and provides co-occurrence matrices (required as a quantitative basis for the statistical phrase clustering algorithms). It uses Natural Language Processing (NLP) to generate the technical phrases, and is subject to the limitations of any NLP package (not all technical phrases recovered, extensive manual cleanup still required for high quality results). Combining the two variants allows the co-occurrence matrix of technical phrases to be used as the basis for statistical clustering algorithms, with any missing phrases supplied by the TextSlicer results.

In the first variant, multi-word phrase themes are selected from a multi-word phrase analysis of the type shown above. For each theme phrase, the frequencies of phrases within +-50 words of the theme phrase are computed for every occurrence of the theme phrase in the full text, using

the TextSlicer software from DT. A phrase frequency dictionary is constructed that contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those sub-themes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the cluster member phrases most closely related to the cluster theme.

In the second variant, all the phrases generated by NLP analysis of the Abstracts' text are examined, and the low or non-technical content phrases removed. Lists of authors, institutions, journals, etc. are also generated, with relatively little cleanup required. These various lists are matrixed against each other, to ascertain co-occurrence frequencies. Standard clustering packages (e.g., WINSTAT, an Excel add-in) group these list elements into thematic areas.

Thus, the matrixing of an Abstract phrase list against itself will generate purely technical theme relationships. Matrixing of an author list against and Abstract phrase list will relate specific authors to specific technical themes.

The specific clustering approach consists of the following steps:

1) Import the Abstract database into TechOasis, a text mining software package produced by Search Technology.
2) Generate lists of high technical content phrases. This involves manual examination of all phrases output by TechOasis, and selection of only the high technical content phrases.
3) Generate co-occurrence phrase-phrase matrices, where each matrix element represents the frequency of co-occurrence of the ordinate and abcissa phrases.
4) Import the matrices into Excel spreadsheets.
5) Normalize the matrix elements, typically non-dimensionalizing on combinations of the ordinate and abcissa values.
6) Use Excel add-in clustering software (WINSTAT) to relate phrases quantitatively.
7) Manually generate groups of thematically-similar phrases, based on quantitative phrase relationships, initial clustering software groupings, and criteria for taxonomy categories (e.g., groups of similar extent, groups of same type, groups of equal strength of relationship, etc)
8) Select high frequency phrases. For each high frequency phrase, identify the low frequency phrases (located in the same matrix column) that are strongly related to the high frequency phrase. Use threshold values of the Inclusion Index to filter out those strongly related low frequency phrases. Supplement this list with phrases from a proximity analysis of each selected high frequency phrase using the TextSlicer software from DT, to insure all phrases within the cluster are retrieved. Categorize the low frequency phrases, and identify any low frequency phrases that appear anomolous.
9) Select low frequency phrases. For each low frequency phrase, identify the high frequency phrases (located in the same matrix row) to which the low frequency phrase is strongly related. Examine the high frequency phrase categories; identify any high frequency phrase combinations that appear unusual.

Three types of raw data output result from each clustering run:

1) A dendogram that shows the quantitative linkages among closely-related phrases. Figure 4, for example, is a dendogram that portrays linkages among the twenty highest frequency technical content phrases from the query Abstracts database. The x-axis is the phrases that were used, and the y-axis is the 'distance', or measure of the similarity between phrases. If two phrases have the same co-occurrence profile with other phrases, the 'distance' will be very low.

   A dendogram is a structure that shows linkages among phrases. It does so by starting with a root that encompasses all the phrases. Then it splits into two groups (clusters) until all the phrases are contained in their own cluster. In Figure 4, the root at the bottom of the page encompasses all the phrases. The first split is into two large clusters. One cluster contains the phrases COAL, COALS, CARBON, CATALYST, CATALYSTS, and CONVERSION. The second cluster contains all the remaining phrases ENERGY, COMBUSTION, FUEL, EMISSIONS, GAS, ELECTRICITY, HEAT, WATER, HYDROGEN, OXIDATION, OXYGEN, CELL, CELLS, and BATTERIES.

2) A table that contains a quantitative measure of the similarity of adjoining phrases or phrase-cluster pairs. The similarity, or 'distance', of a phrase pair is obtained by matching the co-occurrence profiles of each phrase in the phrase pair against all other phrases in the matrix. Table 12, for example, is a table that contains the information portrayed in Figure 4. The distances shown on the dendogram are taken from the distances given in this table, thus the table is the numerical expression of the dendogram.

3) A taxonomy of a pre-specified number of groups of phrases. Table 13, for example, shows the groupings of phrases when four clusters were specified for the data portrayed in Figure 4.

High Level Taxonomy - Query-based Database

The 220 highest frequency phrases were used to form the symmetrical co-occurrence matrix using the Equivalence Index (Eij=Cij^2/Ci*Cj). Cij is the Abstract co-occurrence frequency of phrases i and j, Ci is the total Abstract occurrence frequency of phrase i, and Cj is the total Abstract occurrence frequency of phrase j. The resultant dendograms and associated data served as the basis for manually generating a hierarchical taxonomy. The first two levels are shown in Table 14.

The two clusters in the first hierarchical level (Direct Conversion, Thermal Conversion) are differentiated by the potential for direct energy conversion to electricity, and by the level of technology description. One cluster, Direct Conversion, contains direct conversion technologies such as BATTERIES, SOLAR CELLS, SOLID OXIDE FUEL CELLS, MAGNETIC ENERGY, PLASMA, AND FUSION, and describes these technologies at the detailed component or phenomenological level. The second cluster, Thermal Conversion, contains technologies that typically require an intermediate heat cycle step in the conversion of the fuel source energy into electricity, such as HEAT ENGINE, THERMODYNAMICS, HEAT EXCHANGER, STEAM,

NUCLEAR POWER PLANTS, WASTE HEAT, COMBUSTION, FLUE GAS, NATURAL GAS, CRUDE OIL, DIESEL ENGINE, GASOLINE, INTERNAL COMBUSTION ENGINE, and describes these technologies at the systems level. The Direct Conversion cluster reflects the more recent high technology advances in physics (especially plasma and solid state) and electrochemistry (especially solid state). The Thermal Conversion cluster reflects the more traditional thermodynamics-based approaches to energy conversion, and tends to be pursued more in the developing countries (on a relative emphasis basis) than the higher tech Direct Conversion cluster.

Each of the two first level clusters divides into two second level clusters. The Direct Conversion cluster divides into an Electromagnetic Storage and Conversion cluster (MAGNETIC FIELD, PLASMA, FUSION, MAGNETIC ENERGY STORAGE) and an Electrochemical Storage and Conversion cluster (BATTERIES, SOLAR CELLS, SOLID OXIDE FUEL CELLS). The Thermal Conversion cluster divides into a Combustion Cycle (fuel source, combustion process, combustion product) cluster (COMBUSTION, IGNITION, FUEL, OXIDIZER, SOOT, COAL, OIL, NATURAL GAS, DIESEL FUEL, FURNACES, BOILERS, DIESEL ENGINES, INTERNAL COMBUSTION ENGINES, SOOT, FLUE GAS, ASH, EXHAUST GASES, CARBON DIOXIDE, CARBON MONOXIDE, BENZENE, HYDROCARBONS), and a Systems and Thermodynamics cluster (ENERGY SOURCES, ENERGY PRODUCTION, ENERGY CONSUMPTION, ELECTRICITY PRODUCTION, RENEWABLE ENERGY SOURCES, ELECTRICAL ENERGY, GENERATORS, THERMAL ENERGY, HEAT TRANSFER, THERMODYNAMICS, HEAT ENGINES, HEAT EXCHANGERS, HEAT PUMP, GAS TURBINE, FUEL CYCLE). The generic components of the latter cluster cover all the energy technologies, but the technology-specific components focus on fossil fuel and nuclear, the traditional thermal conversion step technologies.

High-Level Taxonomy - Journal-based Database

The 220 highest frequency phrases were used to form the symmetrical co-occurrence matrix. The resultant dendograms and associated data served as the basis for manually generating a hierarchical taxonomy. The first two levels are shown in Table 15.

The first hierarchical level contains two clusters. One is a small tightly-knit group focused specifically on Lithium Batteries. The other is a large group covering the generic areas of Fossil Fuels and Renewable Energy. Because of this sharp differentiation in cluster size and focus, the Lithium Battery cluster will not be sub-divided further. Therefore, the second hierarchical level will consist of the first level Lithium Battery cluster, a Fossil Fuel cluster, and a Renewable Energy cluster.

The third hierarchical level will consist of the first level Lithium Battery cluster (LITHIUM ION BATTERIES, LITHIUM CELLS, LITHIUM SALTS), a sub-division of the Fossil Fuel cluster into component clusters, and a sub-division of the Renewable Energy cluster. The Fossil Fuel cluster is divided into Solid Fossil Fuel Cycle (RAW COAL, ANTHRACITE, QUARTZ REACTOR, COAL COMBUSTION, FLY ASH, EMISSIONS) Gaseous Fossil Fuel Cycle (NATURAL GAS, GASEOUS FUELS, GAS TURBINE, NITROGEN OXIDES, AROMATIC HYDROCARBONS), and Liquid Fossil Fuel Cycle (LIQUID FUELS, LIQUID

HYDROCARBONS, SHALE OIL, HEAVY OIL, CRUDE OIL, JET FUEL, DIESEL FUEL, DIESEL ENGINES, INTERNAL COMBUSTION ENGINES, FLUE GAS, GREENHOUSE GAS). The Renewable Energy cluster is divided into Solar (SOLAR RADIATION, SOLAR COLLECTOR, HEAT PIPE, SOLAR AIR HEATERS, SOLAR WATER HEATERS), Wind (WIND ENERGY, WIND TURBINES, Wood (FIREWOOD, SAWDUST, TIMBER), and Biomass (VEGETABLE OILS, RICE HUSK, MOLASSES, VEGETABLES).

The journal-based taxonomy emphases appear much different from those of the query-based taxonomy. For example, most of the direct conversion technologies in the query-based taxonomy do not appear in the high level journal-based technology. Even the nuclear technologies appear only peripherally. In addition, the detailed high frequency technical terms in the journal-based taxonomy (WHEAT STRAW, BROWN COAL, FLY ASH, COAL CHAR, STEAM, SUGAR CANE, DIESEL OIL, VEGETABLE OILS, SEWAGE SLUDGE, HEAT PUMP. FISH, SOLAR AIR HEATERS, VEGETABLES, TIMBER, FIREWOOD, RICE HUSK, MOLLASES, SAWDUST) have a more traditional focus in contrast to the high frequency technical terms (NUCLEAR POWER PLANTS, CATALYTIC COMBUSTION, SOLAR CELLS, SOLID OXIDE FUEL CELLS, MAGNETIC FIELDS, X-RAY DIFFRACTION, MAGNETIC ENERGY STORAGE) that appear in the query-based taxonomy. To take a specific technology comparison example, contrast the treatment of solar energy in the two databases. The query-based database focuses on direct conversion to electricity through solar cells and photovoltaics, whereas the journal-based database focuses on solar air and water heaters using solar concentrators, and solar coatings for thermal control. Finally, the journal-based taxonomy focuses on a number of hybrid-fuel systems with some lower technology components (BROWN COAL/ URANIUM/ GAS TURBINES [where the uranium is separated from the coal in a gas turbine], VEGETABLE OILS/ FUEL BLENDS/ DIESEL ENGINE [where the vegetable oils are mixed with the fossil-based oils in a diesel engine], SOLAR COLLECTOR/ FISH [where the solar energy is concentrated in a collector, and used to dehydrate fish (and other products)). Such hybrid systems were nowhere evident in the high level query-based taxonomy.

High-Level Taxonomy - Combined Query-Journal Database

The query and journal-based databases were combined. This total database contained over 20000 records. A sample database of 4000 records was extracted for this analysis.

The 220 highest frequency phrases were used to form the symmetrical co-occurrence matrix. The resultant dendogram and associated data served as the basis for manually generating a hierarchical taxonomy. The first three levels are shown in Table 16.

The first hierarchical level contains two clusters. The smaller cluster focuses on Energy Storage, and the larger cluster focuses on Power Sources and Converters. In the second hierarchical level, the Energy Storage cluster is sub-divided into Science and Development (measurement properties and instruments), and Systems and Applications. This latter category focuses solely on electrochemical components (ELECTROLYTE, CATHODE, ANODE, SEPARATOR), systems (BATTERIES), and applications (ELECTRIC VEHICLES), and at the high level, does

not contain any mechanical or magnetic systems or applications. Also, the latter category contains insufficient terms to justify a third hierarchical level.

In the second hierarchical level, the Power Sources and Converters cluster is sub-divided into a Fossil Energy cluster and a Renewable/ Long-Term Energy cluster.

For the third hierarchical level, the Storage-Science and Development second level category may be sub-divided into a micro category (SPECTROSCOPY, X-RAY DIFFRACTION, ELECTRON MICROSCOPY) and a macro category (ELECTRICAL CONDUCTIVITY, ELECTRICAL RESISTIVITY, HEAT CAPACITY, THERMAL CONDUCTIVITY, GLASS, POWDERS, METALS). The Sources and Converters-Fossil second level category may be sub-divided into three third level sub-categories: Sources (BITUMINOUS COAL, OIL SHALE, CRUDE OIL, GASES), Emissions (POLLUTANTS, TOLUENE, BENZENE, CARBON DIOXIDE, CARBON MONOXIDE, ATMOSPHERE), and Converters, which further subdivides into Direct Converters (FUEL CELLS, HYDROGEN ENERGY, NATURAL GAS, STEAM, ELECTRICITY), and Thermal Converters (COMBUSTION CHAMBER, FURNACE BOILER, DIESEL ENGINE, GAS TURBINE). The Sources and Converters-Renewable/ Long-Term second level category may be sub-divided into four third level sub-categories: Nuclear Sources (NUCLEAR, REACTORS, FUEL CYCLE), Non-nuclear Sources (RENEWABLE ENERGY SOURCES, WIND, SOLAR ENERGY), Direct Converters (MAGNETIC ENERGY, MAGNETIC FIELD, PLASMA), Thermal Converters (HEAT PUMP, HEAT EXCHANGER, THERMAL ENERGY).

The relative positioning of these sub-categories on the dendogram is interesting, and merits some description. The dendogram starts at one end describing various aspects of the coal source (COAL, LIGNITE). It gradually transitions into oil shale, which in turn transitions into oil-related terms (CRUDE OIL, PETROLEUM). The oil terms evolve into gas-related terms (GASIFICATION, GAS COMPOSITION), that translate smoothly into combustion-related terms (COMBUSTION CHAMBER, EMISSIONS, IGNITION). Various types of burners are included (FURNACE, BOILER, BURNER), and they metamorphosize into heat-cycle engines (DIESEL ENGINES, GAS TURBINES). Next is the direct conversion fuel cell, with primary focus on steam reforming of natural gas to produce the required hydrogen (HYDROGEN ENERGY, NATURAL GAS, STEAM, FUEL CELL). Next come a substantial list of fossil emissions (POLLUTANTS, BENZENE, TOLUENCE, CARBON DIOXIDE), so the direct converter fuel cell is bridging the divide between the heat cycle engines and their emissions. The emissions are followed by generic phrases relating to renewable sources and technologies (ENVIRONMENT, RENEWABLE ENERGY SOURCES, RENEWABLE ENERGY TECHNOLOGIES), paralleling the real-world promotion of renewable sources to reduce the impact of the fossil emissions. In the midst of these generic phrases is Wind. The specific renewable technologies that follow next are bounded by nuclear on one end (REACTORS, FUEL CYCLE, NUCLEAR) and fusion on the other end *MAGNETIC ENERGY, MAGNETIC FIELD, PLASMA), with solar energy in the middle (SOLAR ENERGY, SOLAR RADIATION, COLLECTORS). Although nuclear and fusion are typically not what people have in mind when discussing renewable sources, for all practical purposes they are discussed in the technical literature as potentially boundless energy supplies, and this is how they are treated by the clustering algorithm. The end of the fusion component may be interpreted as its direct

conversion capability to electricity, and this section is adjacent to the start of the storage section. Storage evolves from generic electrochemical conversion phraseology (CELLS, ELECTROLYTE, CATHODE, ANODE), that parallels electrochemical converter terminology, to electrochemical storage systems and applications (BATTERIES, ELECTRIC VEHICLES). The final storage section that bounds applications is the science that underlies mainly the electrochemical systems, evolving from micro experimental techniques (SPECTROSCOPY, X-RAY DIFFRACTION, ELECTRON MICROSCOPY), to materials and reactants (ALLOYS, LEAD, OXYGEN, SODIUM), to macro properties (ELECTRICAL RESISTIVITY, HEAT CAPACITY, THERMAL CONDUCTIVITY).

Low Frequency Phrase Relationships

The 220 highest frequency phrases and 8,036 lower frequency phrases, taken from the combined Query and Journal database, were used to form a co-occurrence matrix. The Inclusion index (Ii=Cij/Ci) was used to normalize the matrix elements because the numbers remain invariant with the distance from the origin. The resultant associated data served as the basis for finding relationships between low and high frequency phrases. In order for a phrase to be related to a cluster it must be either 1) very strongly related to at least one high frequency phrase in that cluster or 2) moderately strongly related to two or more high frequency phrases in that cluster. The following are typical examples of low frequency-high frequency phrase relationships.

Low Frequency Phrases unique to one higher frequency phrase

MERCURY, a low frequency phrase, is strongly related only to BITUMINOUS COAL, a high frequency phrase. Trace elements of Mercury can be found in Bituminous coal. Measurements of Mercury can help determine properties of activated carbon that is present in Bituminous coal.

TAXES, a low frequency non-technical phrase, is strongly related to FOSSIL FUELS, a high frequency phrase. It is suggested that the Government can help to slow the global climate change by imposing Taxes on the combustion of Fossil Fuels.

Low Frequency Phrases unique to a second tier cluster

ANTHROPOMORPHIC EMISSION, a low frequency phrase, is strongly related to GASES, EMISSIONS, CARBON DIOXIDE, ENERGY SOURCES, all high frequency phrases that occur in the second tier cluster entitled Fossil. Major problems with the changing atmosphere are discussed. Many countries are hindering efforts to stabilize potentially dangerous emissions from energy sources including Anthropomorphic emissions, Carbon Dioxide emissions and other gases.

Low Frequency Phrases unique to a first tier cluster

STEAM PRODUCTION, a low frequency phrase, is related to WATER, OIL, BED, HEAT TRANSFER, ENERGY, STORAGE, HEAT, ENERGY STORAGE, HOT WATER, all high frequency phrases that occur in the first tier cluster entitled Sources and Converters. WATER, OIL, BED and HEAT TRANSFER are all found in the second tier cluster Fossil, while

ENERGY, STORAGE, HEAT, ENERGY STORAGE and HOT WATER are all found in the second tier cluster Long-term/Renewable.

Low Frequency Phrases shared by all first tier clusters

WOOD SAMPLES, a low frequency phrase, is related to PHENOLS, COAL STRUCTURE, BITUMINOUS COAL, COAL, COKE, RENEWABLE SOURCES, ELECTRODE, ELECTRODES, ELECTRICAL RESISTIVITY, X-RAY DIFFRACTION all high frequency phrases. PHENOLS, COAL STRUCTURE, BITUMINOUS COAL, COAL, COKE and RENEWABLE SOURCES are in the first tier cluster Sources and Converters. ELECTRODE, ELECTRODES, ELECTRICAL RESISTIVITY and X-RAY DIFFRACTION are found in the first tier cluster Storage.

This relationship consists of three different types of links. First, Wood and Coal are linked by their structural properties. They can both be made to yield graphite. Second, similar applications link Electrode and Electrodes. Wood is used as a source of Coke for the production of graphite-like Electrodes. Finally, X-ray Diffraction and Electrical Resistivity are both experimental/diagnostic approaches and properties.

Low Frequency Multiple phrases strongly related to one higher frequency phrase

Two higher frequency phrases, DIESEL FUEL and X-RAY DIFFRACTION, were used as themes for a proximity analysis. Lower frequency phrases strongly related to these higher frequency phrases were identified.

Diesel Fuel

Phrases closely related to DIESEL FUEL may be divided into three categories: Fuel sources/ extraction processes; Combustion/ performance; Pollution/ remediation. Sources include HYSEE (hydrogenated soy ethyl ester), VEGETABLE OILS, COCONUT OIL, OIL METHYL ESTER, ETHANOL, FATS, BIOCRUDE, PLASMATRON, BIODIESEL, TETRADECANE, FUEL BLENDS, Combustion/ performance includes CFPP (cold filter plugging point), FLASH BOILING, PEROXIDES (additives), DME (additives), COMPRESSION IGNITION, CPD (additives), INJECTION TIMING, CETANE NUMBER, EXHAUST GAS TEMPERATURE, DROPLET COMBUSTION, CYLINDER PRESSURE Pollution/ remediation includes POC (particulate organic carbon), BIODEGRADATION OF PETROLEUM, N-2 FIXING, BLACK SMOKE, FILTER PLUGGING, FORMALDEHYDE, MTBE, THC, SOOT FORMATION, HYDROPEROXIDES, and ALDEHYDES.

X-Ray Diffraction

Phrases closely related to X-RAY DIFFRACTION may be divided into three categories: Target materials/ Phenomena studied/ Other diagnostics. Materials include: ACAC, BC2N, PM2, QUARTZ AND KAOLINITE, CU SI, COAL FE-BC, PYROPISSITE, SILICATE HYDRATES, XYLITIC LIGNITE, ZNO-BASED, MOS2, ASH MELTING, CAO LOADING, CDO, LA3AU4IN7, LI-MN-O, LITHIUM-SILICON, ND2FE114BNDELTA, OIL FLY ASH,

SPHERULITES, LIXNIO2, BLIND CANYON COAL, RUO6, CRYSTALLITES, VANADIUM OXIDE, LSGM, CARBAZOLE, NA2O, PRGAO3, ASH PARTICLES, TITANIUM, CHITOSAN, INORGANIC MATTER, QUARTZ, ALUMINOSILICATE, FE3O4. Phenomena include: L-C, DELITHIATED, ELECTRICAL CONDUCTION, ELECTROCHEMICAL CAPACITY, GMCFS, HIGHEST ELECTRICAL CONDUCTIVITY, INTERCALATED IN GRAPHITE, MECHANICALLY ACTIVATED, TERNARY PHASES, THERMODYNAMIC CRITERION, COMBUSTIBLE BURNOUT, FAST OXIDE ION, VOLTAGE PLATEAU, CALCINATION TEMPERATURE, SHS (self-propagating high-temperature synthesis), BCC, LITHIATION, CRYSTALLINATY, CRYSTALLINE PHASES, CRYSTALLOGRAPHIC, SSA (specific surface area), COMBUSTION REACTION. Other diagnostics include: RAMAN MICROSCOPY, X-RAY FLUORESCENCE, THERMOGRAVIMETRIC ANALYSIS, CHRONOPOTENTIOMETRY, TRANSMISSION ELECTRON MICROSCOPY, SCANNING ELECTRON MICROSCOPY.

Document Clustering

Document clustering is the grouping of similar documents into thematic categories. Different approaches exist [42-51]. The approach presented in this section is based on a partitional clustering algorithm [52-53] contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements. 32 individual clusters were chosen for the query-based database and the journal-based database. The 32 clusters for each type of database are presented in Appendix 2.

CLUTO also agglommorates the 32 clusters in a hierarchical tree (taxonomy) structure. The taxonomies for each of the two databases are presented here.

Query-based Database

Table 17 shows a four-level hierarchical taxonomy for the query-based database. The left-most column is the highest taxonomy level, and each column to the right is the next lowest level. The number of records in each category is shown in parenthesis.

The first level taxonomy can be sub-divided into two approximately equal categories: Power Generation/ Energy Storage, and Energy Conversion. Power Generation/ Energy Storage (4843) focuses on the systems aspects of energy generation and storage, while Energy Conversion (4527) focuses on the direct and indirect conversion of energy to electricity.

For the second level taxonomy, each first level category is divided into two sub-categories. Power Generation/ Energy Storage is divided into Fossil Remediation and Replacement Systems (1443 records, focusing on remediation of CO2 emissions from fossil plants, as well as renewable source systems to replace the CO2-emitting fossil plants), and Power Plant Heating and Storage Systems (3400 records, focusing on heating and energy storage systems, and nuclear power generation systems). Energy Conversion is divided almost equally into Direct Conversion

(2117 records, focusing on the direct conversion of energy sources to electrical power), and Thermal Step Conversion/ Combustion (2410 records, focusing on conversion with a thermal step (such as combustion)).

All second level categories are sub-divided to form eight third level categories, and the third level categories are sub-divided to form sixteen fourth level categories. The category headings for the third and fourth levels are sufficiently detailed that no further description is required.

Journal-Based Database Taxonomy

Table 18 shows a four-level hierarchical taxonomy for the journal-based database. The first level taxonomy can be sub-divided into two categories, Fossil Remediation and Replacement Systems, Turbine Conversion (6294 records, focusing partially on remediation of CO2 emissions from fossil plants, mainly on renewable source systems to replace the CO2-emitting fossil plants, emphasizing turbine conversion), and Fossil Generation and Storage (5860 records, focusing on fossil-based power plants and mainly battery storage systems).

For the second level taxonomy, each first level category is divided into two sub-categories. Fossil Remediation and Replacement Systems is divided into Solar Thermal (2623 records, focusing on solar collectors for heating and cooling applications), and CO2 Remediation and other Low Emission Replacement Systems, Turbine Conversion (3671 records, focused on CO2 emission reduction and other mainly renewable low emission power generating systems, emphasizing turbine conversion). Fossil Generation and Storage is divided into Fossil Generation (3970 records, focusing on fossil fuel sources and conversion technologies), and Batteries (1890 records, focusing on battery development).

All second level categories are sub-divided to form eight third level categories, and the third level categories are sub-divided to form sixteen fourth level categories. The category headings for the third and fourth levels are sufficiently detailed that no further description is required.

Comparison of Query and Journal-based Database Taxonomies

With the exception of the Journal of Power Sources, the journal query approach accessed generic energy related journals that, for the most part, focused on applied energy research. These journals reported on the numerous processes that utilize energy, and the potential that developed / developing energy sources / conversion methods could provide. Many of the contributors were from the developing countries, where those types of technologies could be readily produced and implemented.

This is substantially different from the articles retrieved from the specific phrase query, where the focus was well distributed among existing and developing primary sources of energy and the fundamental technology issues with converting these sources in various energy-requiring applications. The contributors reflected, on average, the more developed countries, that have the resources to both develop and implement these technologies.

The query taxonomy is more integrated structurally, and the major theme components tend to be complementary. The journal taxonomy is more disjoint, and thematic groupings are sometimes heterogeneous. The linkage between the documents in the query taxonomy is based on the query phrases, whereas the linkage between the documents in the journal taxonomy is their publication in discrete journals. Since the document clustering process is based on text similarity, and the query document linkage is query text similarity, the document clustering is more compatible with the query-based database. In addition, the query database taxonomy has much more of a high technology focus than the journal database taxonomy. The major technology differences that support this conclusion are presented here.

Nuclear

Nuclear power has modest representation in the query database compared to renewables and fossil, and no representation in the journal database. The reasons for low frequencies related to Nuclear are as follows.

There are three major journal types in the SCI that serve as sources of papers. First, there are the fundamental multi-discipline journals, such as Science and Nature. These journals would contain papers focused on the fundamental energy conversion phenomena. Because of the high tech nature of these journals, they would have a higher fraction of nuclear-related articles than are reflected in the Keyword analysis of the present study. These papers would have a higher probability of being accessed through phenomena-related terms, rather than the specific energy production and conversion terms in the query used to generate part of the overall database in this study.

The second journal type is generic power-oriented. These journals constituted the journal-derived component of the total database used in this study, and are listed in the Introduction. The journals in this category contain basic and applied research papers, but on average, as will be shown later, tend to emphasize fossil, electrochemical, and traditional renewables, with very modest representation of fusion, fission, MHD, and more exotic renewables.

The third journal type is specific power-oriented, and the thirty journals in this category are listed in Table 9. These journals were not added to the total database in full, as were the generic power-oriented, for the reasons provided in the database generation section. Their representation in the total database derived from their papers that were accessed by the query. Half of these journals were devoted to nuclear energy and power. It appears that the nuclear S&T community publishes mainly in the first and third types of journals, especially in their dedicated literatures for the more applied S&T.

Thus, the observation that nuclear documents are a small fraction of the fossil and renewables documents should not be interpreted that nuclear source S&T is not being performed or is not important. The proper interpretation is that when power source-related nuclear S&T is examined within the overall power source-related S&T, the high and low tech non-nuclear S&T performed globally dominate the higher tech nuclear S&T performed in a smaller number of the more developed countries. To obtain a more detailed picture of the advances in nuclear power S&T, a standard DT focused analysis of the literature would need to be performed. Detailed technical

terms would be used in the query, and the fifteen nuclear-specific journals listed in Table 9 could be added to form the total database.

Renewables

About twenty percent of the power systems in the query database are focused on renewables, whereas about forty percent of the sources in the journal database are focused on renewables. Additionally, the emphases on specific renewables are different between the two databases. For example, in solar energy, the query database emphasizes the higher tech solar electric (especially Photovoltaics targeted at higher direct electricity conversion efficiencies). The journal database emphasizes the lower tech non-direct electricity component of solar (desalinization, distillation, heating, refrigeration). In biomass, the query database had more generic representation (biomass, solid waste, sewage sludge, vegetable oils), while the journal database had higher representation in the traditional types of biomass (firewood, rice husks, wheat straw). Wind energy had low representation in both databases. Geothermal had very low representation in the journal database, and did not even display as a cluster in the query database.

Fossil

Fossil appears in two sections of the query database taxonomy. There is a modest effort on analysis of CO2 generation from fossil sources, and a more substantive contribution from fossil combustion techniques (catalytic combustion, engine droplet combustion). Combined, these two fossil components represent about thirty percent of the query database. The journal database taxonomy also represents fossil explicitly in two sections. There is a substantial section on fossil generation, and a smaller section on CO2 emissions from vehicles. Combined, these two fossil components represent about thirty-five percent of the journal database. The main difference between the two databases relative to fossil is that the journal database emphasizes source preparation and extraction, while the query database emphasizes the higher tech fuel combustion. Also, coal seems to have a much higher representation compared to oil in the journal database, whereas the representations are about equal in the query database. Natural gas had low representation in both databases relative to coal or oil.

Conversion

Nowhere are the structural differences between the query and journal databases better illustrated than in conversion. Energy conversion is identified as a separate thematic thrust at the highest taxonomy level of the query database, consisting of almost half the database records. In the journal database, energy conversion components can be found in solar thermal, low emission replacement systems, and fossil generation. Because of the lower tech focus of the journal database, the structure is determined more by specific systems than by advanced phenomena or processes, and conversion tends to be hierarchically identified under specific systems.

In the query database, the sub-categories within the conversion category emphasize the primary conversion phenomena, such as combustion, electrochemical, and magnetic field conversion. The systems aspects of the full conversion cycle, such as the final step in the conversion of

energy to electricity (e.g., turbines, power cycles), can be found within the specific power generation systems.

In the journal database, there is less emphasis on the higher tech direct conversion relative to the lower tech thermal step conversion. There is no category of magnetic field conversion, as exists in the query database. Additionally, both databases have a turbine conversion category. In the query database, the turbine conversion is closely associated with the higher tech nuclear power production category, whereas in the journal database, the turbine conversion is associated with the lower tech renewables category, most closely with the wind component. As mentioned under renewables, in the journal database, much of the solar conversion stops at the heating and cooling category, whereas in the query database, relatively more of the solar conversion is directly to electricity.

Storage

In the journal database, a separate second-level taxonomy category of batteries, containing about fifteen percent of total database articles, is identified. Many of these battery articles, and fuel cell articles in the journal database as well, result from the inclusion of the electrochemical-dominant Journal of Power Sources in the database. The main battery focus is divided between Nickel and Lithium batteries, with somewhat less effort devoted towards the traditional Lead-Acid batteries. No other types of storage are evident in the journal database, at least down to the fourth taxonomy level of resolution.

In the query database, energy storage is identified only at the third taxonomy level. The storage function is closely associated with control of power flow in systems. While batteries receive the primary emphasis, some work is reported in capacitors, especially electrochemical, and much less reported work in mechanical storage systems. The battery work appears focused toward vehicles, in concert with some hydrogen storage efforts for hydrogen-powered vehicles as well.

5. SUMMARY AND DISCUSSION

A query and journal-based hybrid process was used to retrieve records from the SCI for analysis. Generic energy or power-related terms were used for the query, relatively independent of any specific power supply, conversion, or storage system (e.g., ELECTRICITY PRODUCTION vs LIGHT-WATER REACTOR). This approach would retrieve documents that described technologies specifically related to power production, conversion, and storage. To retrieve documents related to power production, but where the author may not have used specific terminology relating the technology to power production in the write-up, the journal-based approach was added. The concept was to identify power source journals that were generic, not source specific, and add their articles to the phrase-based query database.

Even with the use of both approaches, one class of articles will not be retrieved. These are power source-related articles that do not contain the generic terms relating them to power sources, nor are published in a journal with a dedicated power source emphasis. Thus, an article on a new scientific phenomenon potentially related to power sources that was published in, for

example, Science or Nature would not appear in this retrieval. To retrieve such articles, a detailed technology-specific query, such as the type developed in past DT studies, is required.

Bibliometric analyses produced the EPS technical infrastructure. The most prolific EPS authors, journals, institutions, countries, cited authors/ journals/ paper were presented. There were 133 different countries listed. The dominance of a handful of countries was clearly evident (e.g., USA, Japan, England, India, Germany, Canada, France) but a series of small countries (Turkey, South Korea, Egypt, Greece, Taiwan) are also productive. The United States is more than twice as prolific as its nearest competitor (Japan), and is as prolific as its major competitors combined.

Two generic types of taxonomies were generated, a manually-based non-statistical approach, and a statistically-based clustering approach. The non-statistical approach was performed for a database of Keywords and a database of Abstracts. The statistical approach was performed for a database of Abstracts. For both the statistical and non-statistical approaches, the Abstract database was divided into its query-based and journal-based components, and taxonomies were generated for each component as well as the merged two-component database.

Overall, a hierarchical multi-level taxonomy can be generated to model the structure of electric power sources/ converters/ storage. The highest taxonomy level consists of three categories: Primary Energy Sources, Energy Converters, and Energy Storage Devices. Phrase frequency allocations to these categories (binning) suggest that Primary Energy Sources have more research activity than Energy Converters, and substantially more research than Energy Storage Devices. In an environment of increasingly scarce energy resources, developing new and affordable sources is of primary concern. Once the sources are defined, then focus on conversion and storage is appropriate. Additionally, energy needs to be converted to more usable forms before it can be stored in such forms. Therefore, substantially more research is performed on converters relative to storage.

Each of the categories in taxonomy level 1 can be subdivided into level 2 categories. Primary Energy Sources can be subdivided into Fossil Fuels, Renewable Energy/ Alternative Fuels, and Nuclear Fuels. Renewable Energy/ Alternative Fuels has a modestly higher level of activity than Fossil Fuels. In the past, substantial R&D was performed on Fossil Fuels, with relatively smaller amounts of research on renewable sources. Because of the foreseeable future decline in Fossil Fuel resources, and the perceived reduced environmental impacts of renewable sources, there are a wealth of opportunities for advancement in renewable sources research, and this is reflected in the relative levels of effort.

The technical emphases of Fossil Fuel research are primarily increasing efficiency and reducing emissions, with some emphasis on widening usage. The technical emphases of Renewable Energy/ Alternative Fuels are increased efficiency, reduced production and maintenance costs, increased commercial interest, and reduce environmental impact. The technical emphases of Nuclear Fuels research are safety, waste disposal, increased efficiency, and reduced life cycle costs.

The above technical emphases strictly apply to the full conversion cycle, not to the source fuels alone. It is very difficult to separate the conversion from the fuels for specific systems in

research articles, since a research article on fuel sources (other than exploration or perhaps some stages of pre-processing) tends to incorporate some aspect of conversion.

Each of the categories in level two can be sub-divided into level 3 categories. Fossil Fuels was subdivided into Coal, Oil, and Natural Gas. The major sub-categories of Coal were constituents/ characteristics/ properties and pre-processing/cleansing/ combustion. The major sub-categories of Oil were constituents/ types, conversion processes, and by-products. The major sub-categories of Natural Gas were types, cleansing, and by-products. The relative magnitudes of research reflect the relative usage diversity of each type, the magnitude of perceived resources available, the energy potentially extractable per resource unit, and the perceived marginal utility of additional research for increased energy extraction. These conclusions are based on the published literature. If there is substantial proprietary research being done in one of these technology sub-areas relative to another sub-area (e.g., if the oil companies were doing substantially more proprietary research than the coal companies), then the total relative efforts among Coal, Oil, and Natural Gas would not be reflected by the numbers above.

Renewable Energy/ Alternative Fuels was subdivided into Solar Energy, Hydrogen, Biomass, Wind Energy, Geothermal Energy, and Hydropower. These five level 3 categories can be stratified into three groups. The largest group (Solar Energy, Hydrogen, and Biomass) has the common characteristics of non-site specificity and effective transportability. The next largest group (Wind Energy, Geothermal Energy) is constrained to geographical regions with favorable operating environments, but additional research is perceived as having the potential to produce substantial benefits at those sites. The smallest group (Hydropower), is also site constrained, but in addition is a mature technology. Hydropower articles address environmental issues (flood control, ecological damage) as much as technology improvement issues.

The major sub-categories of Solar Energy were conversion system characteristics, conversion system components, conversion system processes, and applications. Photovoltaics is classified under Converters. The major sub-categories of Hydrogen were materials/ compounds and conversion processes. The major sub-categories of Biomass were sources, types, and conversion processes. The major sub-categories of Wind Energy were converter systems and applications. The major sub-categories of Geothermal Energy were sources and applications, and the major sub-categories of Hydropower were environmental protection and applications.

Nuclear Fuels was subdivided into Fission and Fusion. The Fission component is a mature technology (proof-of-principle was demonstrated sixty years ago), and the research focuses on cost, safety, environmental, and health issues resulting from operational experiences. The Fusion component is in the proof-of-principle stage, and the research focuses on predicting/ demonstrating ignition and burn, as well as cost and size reduction, and maintenance and cleanup issues. Because of the nature of the query used (linked to power plant production issues), the Fusion papers are further under-represented relative to Fission papers due to the different levels of maturity and linkage to power production terminology.

Energy Converters can be divided into Thermal Converters, Direct Electric Converters, and Nuclear Converters. The research effort in Thermal Converters is significantly larger than in Direct Electric Converters because of the larger embedded operational base in Thermal

Converters (and therefore larger payoffs for small improvements), and the higher technology threshold required to perform research in Direct Electric Converters. Nuclear Converters phrase frequency is substantially smaller than either Thermal or Direct Electric, because of the type of query used, and the technology-specific nature of the dedicated journals in which Nuclear Converters is published frequently.

Thermal Converters can be subdivided into Engines and Turbines. There is more research effort on Engines because of the diversity of types and applications of Engines, as well as the pollution control issues unique to automotive engines, where a main target of pollution reduction research is improvement of the combustion process. The major Engine sub-categories include engine types, engine components, engine characteristics, conversion processes, conversion by-products, and engine fuels. The latter sub-category contained a number of examples of mixed fossil-alternative fuel combinations. The major Turbine sub-categories include fuels, turbine and conversion cycle types, and conversion processes. Acoustics, mixing, and combustion chemistry are focal research areas in the combustion chamber. Heat transfer at the blade, and the underlying flow-field and turbulence transition phenomena, tend to dominate the conversion section research.

Direct Electric Converters can be subdivided into Fuel Cells, Photo-voltaics, Thermoelectric, and MHD. Fuel Cells are researched most heavily because of wider diversity applications, higher efficiency potential, and higher power density. Photo-voltaics is researched more than Thermo-electrics because the light sources (sun, room lighting) required for input are readily available, compared to the requirement for high temperature heat sources for Thermo-electrics. In addition, the light sources are lower entropy than the heat sources, offering the potential for higher conversion efficiency, and the potential improvement in conversion efficiency for Photo-voltaics has been, and promises to be, substantially higher than for Thermo-electrics. MHD research is minimal due to technical difficulties caused by very high temperature gases operating in close proximity to super-cooled magnets.

Fuel Cell sub-categories include higher longevity and efficiency component technologies, diverse fuel cell types, candidate fuels, and component materials. Photo-voltaic sub-categories include conversion/ quantum efficiency improvement and cost reduction, with emphasis on: component materials; electrical properties; optical properties; fabrication techniques, and applications.

Energy Storage Devices can be divided into Electric and Mechanical. With no rotating parts and high energy density per unit weight, Electric storage is the preferred approach. Electric can be sub-divided into Battery, Capacitor, and Super-conducting Magnetic Energy Storage (SMES). Relative to batteries, capacitors have a virtually unlimited cycle life and rapid charging, but low energy density and high self discharge. Even the most promising capacitors, electrochemical super-capacitors, have an energy density an order of magnitude or more less than batteries. Further, their thin insulators limit voltages because of breakdown, and slow ionic liquid conduction limits discharge rate. For these reasons, battery research substantially outpaces capacitor research for energy storage. SMES differs from the other storage approaches in its ability to charge and discharge energy rapidly. The SMES technology is therefore suitable in

applications that require repeated pulses of large amounts of active power for a short duration of time. Because it is viewed presently as a niche technology, research level is limited.

Major battery sub-categories include Types, Components, Materials, Processes/ Phenomena, Properties, and Characteristics.

Major capacitor sub-categories include Structure, Fabrication, Materials,
Properties/ Characteristics/ Environment, Phenomena, Experiment, and System.

The SMES study emphasis appears focused on cost reduction through use of high temperature superconductors and optimized coil configurations. Systems studies and testing appear to receive more emphasis than research.

The document clustering results offered different perspectives on the query-based and journal-based databases.

Query-based Database

The first level taxonomy can be sub-divided into two approximately equal categories: Power Generation/ Energy Storage, and Energy Conversion. Power Generation/ Energy Storage (4843) focuses on the systems aspects of energy generation and storage, while Energy Conversion (4527) focuses on the direct and indirect conversion of energy to electricity.

For the second level taxonomy, each first level category is divided into two sub-categories. Power Generation/ Energy Storage is divided into Fossil Remediation and Replacement Systems (1443 records, focusing on remediation of CO2 emissions from fossil plants, as well as renewable source systems to replace the CO2-emitting fossil plants), and Power Plant Heating and Storage Systems (3400 records, focusing on heating and energy storage systems, and nuclear power generation systems). Energy Conversion is divided almost equally into Direct Conversion (2117 records, focusing on the direct conversion of energy sources to electrical power), and Thermal Step Conversion/ Combustion (2410 records, focusing on conversion with a thermal step (such as combustion)).

Journal-Based Database Taxonomy

The first level taxonomy can be sub-divided into two categories, Fossil Remediation and Replacement Systems, Turbine Conversion (6294 records, focusing partially on remediation of CO2 emissions from fossil plants, mainly on renewable source systems to replace the CO2-emitting fossil plants, emphasizing turbine conversion), and Fossil Generation and Storage (5860 records, focusing on fossil-based power plants and mainly battery storage systems).

For the second level taxonomy, each first level category is divided into two sub-categories. Fossil Remediation and Replacement Systems is divided into Solar Thermal (2623 records, focusing on solar collectors for heating and cooling applications), and CO2 Remediation and other Low Emission Replacement Systems, Turbine Conversion (3671 records, focused on CO2 emission reduction and other mainly renewable low emission power generating systems,

emphasizing turbine conversion). Fossil Generation and Storage is divided into Fossil Generation (3970 records, focusing on fossil fuel sources and conversion technologies), and Batteries (1890 records, focusing on battery development).

Comparison of Query and Journal-based Database Taxonomies

With the exception of the Journal of Power Sources, the journal query approach accessed generic energy related journals that, for the most part, focused on applied energy research. These journals reported on the numerous processes that utilize energy, and the potential that developed / developing energy sources / conversion methods could provide. Many of the contributors were from the developing countries, where those types of technologies could be readily produced and implemented.

This is substantially different from the articles retrieved from the specific phrase query, where the focus was well distributed among existing and developing primary sources of energy and the fundamental technology issues with converting these sources in various energy-requiring applications. The contributors reflected, on average, the more developed countries, that have the resources to both develop and implement these technologies.

The query taxonomy is more integrated structurally, and the major theme components tend to be complementary. The journal taxonomy is more disjoint, and thematic groupings are sometimes heterogeneous. The linkage between the documents in the query taxonomy is based on the query phrases, whereas the linkage between the documents in the journal taxonomy is their publication in discrete journals. Since the document clustering process is based on text similarity, and the query document linkage is query text similarity, the document clustering is more compatible with the query-based database. In addition, the query database taxonomy has much more of a high technology focus than the journal database taxonomy. The major technology differences that support this conclusion are presented here.

Nuclear

Nuclear power has modest representation in the query database compared to renewables and fossil, and no representation in the journal database. The reasons for low frequencies related to Nuclear are as follows.

There are three major journal types in the SCI that serve as sources of papers. First, there are the fundamental multi-discipline journals, such as Science and Nature. These journals would contain papers focused on the fundamental energy conversion phenomena. Because of the high tech nature of these journals, they would have a higher fraction of nuclear-related articles than are reflected in the Keyword analysis of the present study. These papers would have a higher probability of being accessed through phenomena-related terms, rather than the specific energy production and conversion terms in the query used to generate part of the overall database in this study.

The second journal type is generic power-oriented. These journals constituted the journal-derived component of the total database used in this study, and are listed in the Introduction.

The journals in this category contain basic and applied research papers, but on average, as will be shown later, tend to emphasize fossil, electrochemical, and traditional renewables, with very modest representation of fusion, fission, MHD, and more exotic renewables.

The third journal type is specific power-oriented, and the thirty journals in this category are listed in Table 9. These journals were not added to the total database in full, as were the generic power-oriented, for the reasons provided in the database generation section. Their representation in the total database derived from their papers that were accessed by the query. Half of these journals were devoted to nuclear energy and power. It appears that the nuclear S&T community publishes mainly in the first and third types of journals, especially in their dedicated literatures for the more applied S&T.

Thus, the observation that nuclear documents are a small fraction of the fossil and renewables documents should not be interpreted that nuclear source S&T is not being performed or is not important. The proper interpretation is that when power source-related nuclear S&T is examined within the overall power source-related S&T, the high and low tech non-nuclear S&T performed globally dominate the higher tech nuclear S&T performed in a smaller number of the more developed countries. To obtain a more detailed picture of the advances in nuclear power S&T, a standard DT focused analysis of the literature would need to be performed. Detailed technical terms would be used in the query, and the fifteen nuclear-specific journals listed in Table 9 could be added to form the total database.

Renewables

About twenty percent of the power systems in the query database are focused on renewables, whereas about forty percent of the sources in the journal database are focused on renewables. Additionally, the emphases on specific renewables are different between the two databases. For example, in solar energy, the query database emphasizes the higher tech solar electric (especially Photovoltaics targeted at higher direct electricity conversion efficiencies). The journal database emphasizes the lower tech non-direct electricity component of solar (desalinization, distillation, heating, refrigeration). In biomass, the query database had more generic representation (biomass, solid waste, sewage sludge, vegetable oils), while the journal database had higher representation in the traditional types of biomass (firewood, rice husks, wheat straw). Wind energy had low representation in both databases. Geothermal had very low representation in the journal database, and did not even display as a cluster in the query database.

Fossil

Fossil appears in two sections of the query database taxonomy. There is a modest effort on analysis of CO2 generation from fossil sources, and a more substantive contribution from fossil combustion techniques (catalytic combustion, engine droplet combustion). Combined, these two fossil components represent about thirty percent of the query database. The journal database taxonomy also represents fossil explicitly in two sections. There is a substantial section on fossil generation, and a smaller section on CO2 emissions from vehicles. Combined, these two fossil components represent about thirty-five percent of the journal database. The main difference between the two databases relative to fossil is that the journal database emphasizes source

preparation and extraction, while the query database emphasizes the higher tech fuel combustion. Also, coal seems to have a much higher representation compared to oil in the journal database, whereas the representations are about equal in the query database. Natural gas had low representation in both databases relative to coal or oil.

Conversion

Nowhere are the structural differences between the query and journal databases better illustrated than in conversion. Energy conversion is identified as a separate thematic thrust at the highest taxonomy level of the query database, consisting of almost half the database records. In the journal database, energy conversion components can be found in solar thermal, low emission replacement systems, and fossil generation. Because of the lower tech focus of the journal database, the structure is determined more by specific systems than by advanced phenomena or processes, and conversion tends to be hierarchically identified under specific systems.

In the query database, the sub-categories within the conversion category emphasize the primary conversion phenomena, such as combustion, electrochemical, and magnetic field conversion. The systems aspects of the full conversion cycle, such as the final step in the conversion of energy to electricity (e.g., turbines, power cycles), can be found within the specific power generation systems.

In the journal database, there is less emphasis on the higher tech direct conversion relative to the lower tech thermal step conversion. There is no category of magnetic field conversion, as exists in the query database. Additionally, both databases have a turbine conversion category. In the query database, the turbine conversion is closely associated with the higher tech nuclear power production category, whereas in the journal database, the turbine conversion is associated with the lower tech renewables category, most closely with the wind component. As mentioned under renewables, in the journal database, much of the solar conversion stops at the heating and cooling category, whereas in the query database, relatively more of the solar conversion is directly to electricity.

Storage

In the journal database, a separate second-level taxonomy category of batteries, containing about fifteen percent of total database articles, is identified. Many of these battery articles, and fuel cell articles in the journal database as well, result from the inclusion of the electrochemical-dominant Journal of Power Sources in the database. The main battery focus is divided between Nickel and Lithium batteries, with somewhat less effort devoted towards the traditional Lead-Acid batteries. No other types of storage are evident in the journal database, at least down to the fourth taxonomy level of resolution.

In the query database, energy storage is identified only at the third taxonomy level. The storage function is closely associated with control of power flow in systems. While batteries receive the primary emphasis, some work is reported in capacitors, especially electrochemical, and much less reported work in mechanical storage systems. The battery work appears focused toward vehicles, in concert with some hydrogen storage efforts for hydrogen-powered vehicles as well.

Value of DT and Bibliometrics

Advantages of using DT and bibliometrics for deriving technical intelligence from the published literature include:

- Large amounts of data can be accessed and analyzed, well beyond what a finite group of expert panels could analyze in a reasonable time period.
- Preconceived biases tend to be minimized in generating roadmaps.
- Compared to standard co-word analysis, DT uses full text, not index words, and can make more use of the rich semantic relationships among the words.
- It also has the potential of identifying low occurrence frequency but highly theme related phrases that are 'needles-in-a-haystack'.

Other co-occurrence methods matrix the higher frequency phrases against each other, and typically do not access the lower frequency phrases. Because DT builds dictionaries of phrases closely related to the theme phrase, it targets these low frequency phrases directly..

Combined with bibliometric analyses, DT identifies not only the technical themes and their relationships, but relationships among technical themes and authors, journals, institutions, and countries. Unlike other roadmap development processes, DT generates the roadmap in a 'bottom-up' approach. Unlike other taxonomy development processes, DT can generate many different types of taxonomies (because it uses full text, not key words) in a 'bottom-up' process, not the typical arbitrary 'top-down' taxonomy specification process. Compared to co-citation analysis, DT can use any type of text, not only published literature, and it is a more direct approach to identifying themes and their relationships.

The maximum potential of the DT and bibliometrics combination can be achieved when these two approaches are combined with expert analysis of selected portions of the database. If a manager, for example, wants to identify high quality research thrusts as well as science and technology gaps in specific technical areas, then an initial DT and bibliometrics analysis will provide a contextual view of work in the larger technical area; i.e., a strategic roadmap. With this strategic map in hand, the manager can then commission detailed analysis of selected abstracts to assess the quality of work done as well as identify work that needs to be done (promising opportunities).

6. ACKNOWLEDGEMENTS

(*THE VIEWS IN THIS REPORT ARE SOLELY THOSE OF THE AUTHORS, AND DO NOT REPRESENT THE VIEWS OF THE DEPARTMENT OF THE NAVY OR ANY OF ITS COMPONENTS, UNIVERSITY OF MINNESOTA, DDL-OMNI, INC., OR NOESIS, INC. IN ADDITION, THE AUTHORS ACKNOWLEDGE THE CONTRIBUTIONS OF DR. RICHARD CARLIN, OFFICE OF NAVAL RESEARCH, FOR SPONSORING THIS EFFORT***.)***

7. APPENDIX 1 - POWER SOURCES QUERY

Phrase-Based Component

(BIOMASS ENERGY OR CONVENTIONAL ENERGY OR DISTRICT HEATING OR ELECTRICAL ENERGY OR ENERGY CONSUMED OR ENERGY RECOVERY OR ENERGY RESOURCE* OR ENERGY STORAGE OR HEAT ENGINE* OR HYBRID ENERGY OR MAGNETIC ENERGY OR POWER CONVERSION OR RENEWABLE SOURCE* OR SUSTAINABLE ENERGY OR (COGENERATION SAME (POWER OR HEAT)) OR (COMBUSTION SAME (ENERGY OR FUEL* OR POWER)) OR (ELECTRIC POWER SAME (RESEARCH OR TECHNOLOGY OR TURBOGENERATOR)) OR (ELECTRIC SAME (ENERGY CONSUMPTION OR FOSSIL FUEL* OR OUTPUT POWER OR POWER GENERATION OR POWER PRODUCTION OR TURBINE)) OR (ELECTRICAL SAME (EFFICIENCY OR ELECTRON MEDIATOR OR ENERGY SUPPLY OR FUEL* OR HEAT OR POWER DENSITY OR POWER GENERATION)) OR (ELECTRICITY SAME (BIOMASS OR ENERGY CONVERSION OR ENERGY SUPPLY OR ENERGY SYSTEM OR ENERGY TECHNOLOG* OR HEAT OR MICROBIAL FUEL* OR POWER GENERATION OR RENEWABLE ENERGY OR THERMAL)) OR (ENERGY CONSUMPTION SAME (BIOMASS OR POWER OR RENEWABLE ENERGY)) OR (ENERGY CONVERSION SAME RENEWABLE ENERGY) OR (ENERGY DISTRIBUTION SAME (ENERGY SOURCE* OR RENEWABLE ENERGY)) OR (ENERGY EFFICIENCY SAME POWER) OR (ENERGY SOURCE* SAME (ENERGY CONVERSION OR MOTOR* OR POWER GENERATION OR RENEWABLE ENERGY)) OR (ENERGY SYSTEM SAME POWER) OR (ENERGY TECHNOLOG* SAME (BIOMASS OR POWER OR RENEWABLE ENERGY)) OR (ENGINE SAME (ENERGY OR FUEL* OR POWER GENERATION OR POWER SYSTEM)) OR (FUEL* SAME (CYCLE OR ELECTRIC OR ELECTRIC ENERGY OR ELECTRIC POWER OR ELECTRON MEDIATOR OR ENERGY CONSUMPTION OR ENERGY SOURCE* OR ENERGY SYSTEM OR HEAT RECOVERY OR ION CONDUCTIVITY OR POWER DENSITY OR POWER GENERATION OR POWER PLANT* OR POWER PRODUCTION OR RENEWABLE ENERGY OR RESEARCH AND DEVELOPMENT OR STORAGE OR THERMAL ENERGY OR VEHICLE OR BIOMASS OR COMBUSTION OR ENERGY SOURCE* OR RENEWABLE ENERGY OR TURBINE)) OR (HEAT RECOVERY SAME POWER) OR (POWER DENSITY SAME ION CONDUCTIVITY) OR (POWER GENERATION SAME (COMBINED CYCLE OR EFFICIENCY OR ENERGY CONVERSION OR HEAT OR PLANT* OR RESEARCH OR TECHNOLOGIES)) OR (POWER PLANT* SAME (COMBINED CYCLE OR EFFICIENCY OR ELECTRIC OR ENERGY OR POWER GENERATION)) OR (RENEWABLE ENERGY SAME (BIOMASS OR CONVERSION OR POWER GENERATION OR RESEARCH OR SUSTAINABLE DEVELOPMENT)) OR (THERMAL ENERGY SAME (POWER OR RENEWABLE ENERGY OR RESEARCH AND DEVELOPMENT))) NOT (ACBL OR ACCIDENT OR ACCIDENTS OR ACOUSTICALLY OR ACTA METALLURGICA INC OR ACTINIDE* OR ACTIVATION ENERGY ASYMPTOTICS OR ADIABATIC SATURATION COOLING OR AEROSOL OR AGE OR AIDS OR ANIMALS OR ANNEALED OR ANTISOLVENT OR AQUIFERS OR ASH-CONCRETE OR ASHES OR ATHENS OR BANDWIDTH OR BEAMS OR BENIGN OR BIT OR BODY OR CABLES OR CALIBRATION OR CANCER OR CAPITA OR CCA OR CELLULAR OR CEMENT OR

CENT OR CHLORIDE OR CHLOROPHYLL OR CHROMOPHORE OR CIRCULATION OR CLAD OR CLOUD OR CLOUDS OR CONTAMINATION OR CORIOLIS OR CORONAL OR CRYOSTAT OR CURE OR CURING OR DAILY PEAK POWER OR DC DC CONVERTERS OR DEFORMATION OR DEICING OR DESALINATION OR DESALTING OR DESICCANT OR DETECTORS OR DISEASE OR DISTRICT HEATING SYSTEMS OR DRUG OR DUMP OR EHL OR ELASTIC ENERGY STORAGE OR ELPI OR EROSION OR EXCIMER OR FACTORY OR FAT OR FATE OR FATIGUE OR FEEDFORWARD OR FERMION OR FIREBALL OR FISH OR FLARES OR FLUXES OR FOOT OR FRACTAL OR FREE FATTY ACIDS OR FREEBOARD OR FUMIGATION OR FUZZY OR GALAXIES OR GATE OR GEOLOGIC OR GLASSY OR HAND AND FOOT OR HANDPIECE OR HEAL OR HEALTH OR HEAR OR HEAT PIPE HEAT OR HEAT TRANSFER EQUATION OR HEAT TREATMENT TEMPERATURE OR HMX OR HYDRAULIC OR HYDRAZINE OR HYPERSONIC CRUISE TRAJECTORIES OR ILL OR INCOME OR INJURY OR INSTRUMENTS OR INTERNET OR INVERTER OR ISFSI OR JUICE OR KERNEL OR KILN OR LABOR OR LAKE OR LAMBDA OR LAMP OR LANDER OR LEPTIN OR LIMESTONE OR LINE CONTROL SYSTEM OR LINGUISTIC OR LOGIC OR LUBRICANT OR LUNCH OR MAGNESIUM OR MANTLE OR MBMS OR MEAL OR MERCURY OR MESOPORES OR MILE OR MILK OR MINERALS OR MLO OR MMA OR MODULATION OR MONETARY OR MONEY OR MONOTONIC OR MOTHER OR MSF OR MUSCLE OR NEEDLES OR NERVE OR NEURAL OR NFL OR NITRIC OR NITROUS OR NOISE OR NORMAL SPECTRAL EMISSIVITY OR NTT OR NUMBER OF MULTIPLEXERS OR OPERATORS OR ORBITAL OR PAIN OR PARASITIC OR PATIENTS OR PCB OR PIPING OR PLUME OR POLICIES OR PONDS OR POOL OR PROTEIN OR PROTEINS OR RADIO OR RAT OR RATS OR RECONNECTION OR REPRODUCTIVE OR RETROFIT OR RIVER OR ROAD OR ROSE OR SAUTER MEAN DIAMETER OR SEDIMENTS OR SHEET OR SIGNATURES OR SILICA OR SKELETON OR SLAG OR SOFTWARE OR SOIL OR SOILS OR SOLVENTS OR SPATIAL OR SPAWNING OR STALAGMITE OR STAR OR STOVE OR STOVES OR SURVEY OR TAX OR THEORIES OR TIRES OR TISSUE OR TISSUES OR TRAFFIC OR TRANSFORMER OR TROPOSPHERE OR URBAN OR VITRO OR WELDING OR WOMEN OR WORKERS OR COMBUSTION DUST OR COMBUSTION MINERAL OR COMBUSTION SMOLDER OR (CONVERSION EFFICIENCY SAME LASERS) OR (ELECTRIC POWER SAME LIFE) OR (ELECTRICAL SAME (ANNEALING OR CIRCUIT OR ETCHING OR GROSS OR LIGHTING OR SPECIFIC OR WIDER)) OR (ELECTRICAL ENERGY SAME (CONCENTRATION OR POLLUTANT)) OR (ELECTRICITY SAME RECYCLING) OR (ENERGY SAME (ACCELERATION OR CONTROLLERS OR DISTURBANCE OR EQUIPARTITION OR FATTY OR FLAME OR HEART OR ISOTROPIC OR NETWORK OR NSPUDT OR PAYBACK OR PEI OR PENALTY OR SECTOR OR TREATMENT OR VELOCITY OR WAVES)) OR (ENERGY CONSUMPTION SAME PROGRAM) OR (ENERGY STORAGE SAME VIBRATIONAL) OR (ENERGY SUPPLY SAME (BOUNDARY OR DISTILLATION OR STORAGE)) OR (ENGINE SAME (ALGORITHM OR MODELS OR STABILIZATION)) OR (FUEL SAME (AEROSOL OR ALGORITHM OR HUMAN OR LEGISLATION OR NUMERICAL MODEL OR PAH OR PARTICULATE MATTER OR PLIF OR SIGNALS OR TROPOSPHERIC OR VIBRATION)) OR (FUELS SAME BUILDING) OR (HEAT STORAGE SAME HEAT PUMP) OR (POWER SAME (ABSORPTION OR ASH OR BUNDLE OR DOSE OR ECONOMY OR FAULT OR LASER

OR LEAKAGE OR LINE OR LOGIC OR MINOR OR MONITORING OR POLICY OR PROBABILISTIC OR RECTIFIER OR SMES OR SWITCHES)) OR (POWER GENERATION SAME (FRACTION OR HEAT RECOVERY OR PROBLEMS OR SELF-TUNING OR SIEMENS OR STAGE)) OR (POWER PLANTS SAME (CORROSION OR MECHANICAL OR PFBC OR SEPARATION OR SIMULATION)) OR (POWER SUPPLY SAME (CIRCUIT OR CIRCUITS OR SWITCHING)) OR (RENEWABLE ENERGY SAME FINANCIAL) OR (THERMAL ENERGY SAME (MEDIA OR PEAK OR PERCENT)))

Journal Title Component

FUEL
ENERGY FUELS
J. POWER SOURCES
ENERGY
ENERGY CONV. MANAG.
INT. J. ENERGY RES.
RENEW. ENERGY
J. INST. ENERGY
ENERGY SOURCES
PROG. ENERGY COMBUST. SCI.
RERIC INT. ENERGY J.

APPENDIX 2 – DOCUMENT CLUSTERS

Each Cluster is numbered (beginning with zero), and the number of documents in each cluster appears in parentheses at the beginning of every cluster. The most descriptive words (actually word stems) in each cluster are also shown in parentheses. Each word within the cluster is followed by a number that represents the percentage of intra-cluster similarity explained by the word. The theme of each cluster is represented by the initial high value keywords shown. The order of the clusters reflects the net cohesiveness (the intra-cluster similarity minus the inter-cluster similarity).

2A – QUERY-BASED DATABASE

Cluster 0, Size: 140, ISim: 0.073, ESim: 0.007
Descriptive: droplet 51.0%, sprai 5.3%, flame 2.3%, vapor 1.6%, liquid 1.4%, combust 1.2%, ignit 1.2%, fuel.droplet 0.9%, fuel 0.8%, burn 0.8%
Discriminating: droplet 36.4%, sprai 3.4%, energi 1.3%, power 1.2%, system 0.9%, flame 0.9%, vapor 0.9%, heat 0.7%, electr 0.7%, fuel.droplet 0.6%

Cluster 1, Size: 148, ISim: 0.056, ESim: 0.007
Descriptive: diesel 18.9%, blend 7.4%, oil 7.2%, diesel.fuel 6.6%, engin 5.9%, fuel 4.7%, diesel.engin 2.5%, exhaust 1.7%, emiss 1.4%, gasolin 1.0%
Discriminating: diesel 12.4%, blend 5.1%, diesel.fuel 4.7%, oil 3.6%, engin 1.7%, diesel.engin 1.6%, energi 1.3%, power 1.1%, system 1.0%, heat 0.9%

Cluster 2, Size: 148, ISim: 0.051, ESim: 0.008
Descriptive: batteri 36.6%, vehicl 13.8%, storag 1.8%, hydrogen 1.6%, system 1.4%, batteri.energi 1.3%, power 1.1%, technolog 0.9%, batteri.energi.storag 0.8%, energi.storag 0.7%

Discriminating: batteri 28.4%, vehicl 9.3%, heat 1.2%, batteri.energi 1.0%, combust 0.9%, temperatur 0.8%, magnet 0.7%, batteri.energi.storag 0.6%, ga 0.5%, lead.acid 0.5%

Cluster 3, Size: 165, ISim: 0.050, ESim: 0.008
Descriptive: catalyst 39.5%, catalyt 7.7%, activ 1.7%, nox 1.3%, combust 1.2%, catalyt.combust 1.2%, oxid 1.1%, reform 1.1%, reaction 0.9%, temperatur 0.8%
Discriminating: catalyst 31.6%, catalyt 5.9%, energi 1.2%, power 1.2%, catalyt.combust 1.0%, heat 0.8%, electr 0.8%, system 0.8%, magnet 0.7%, activ 0.7%

Cluster 4, Size: 184, ISim: 0.047, ESim: 0.006
Descriptive: magnet 22.1%, spin 2.6%, transit 2.6%, rho 2.4%, suscept 1.9%, alloi 1.8%, electr.resist 1.7%, resist 1.6%, antiferromagnet 1.6%, specif.heat 1.5%
Discriminating: magnet 11.7%, spin 1.8%, rho 1.7%, fuel 1.4%, transit 1.4%, suscept 1.4%, power 1.2%, antiferromagnet 1.2%, specif.heat 1.1%, combust 1.0%

Cluster 5, Size: 264, ISim: 0.050, ESim: 0.009
Descriptive: cell 25.0%, fuel.cell 22.1%, hydrogen 4.2%, fuel 3.4%, system 1.5%, power 1.3%, reform 1.2%, stack 1.0%, technolog 0.9%, plant 0.7%
Discriminating: fuel.cell 18.3%, cell 17.9%, hydrogen 2.0%, combust 1.1%, heat 0.9%, reform 0.9%, magnet 0.8%, stack 0.7%, model 0.7%, temperatur 0.6%

Cluster 6, Size: 235, ISim: 0.047, ESim: 0.007
Descriptive: heat.engin 8.0%, heat 7.9%, engin 6.6%, irrevers 5.4%, cycl 3.1%, carnot 2.7%, thermodynam 2.6%, maximum.power 2.5%, endorevers 2.3%, maximum 2.2%
Discriminating: heat.engin 6.0%, irrevers 3.9%, engin 2.3%, heat 2.2%, carnot 2.0%, maximum.power 1.8%, endorevers 1.7%, thermodynam 1.5%, fuel 1.4%, energi 1.3%

Cluster 7, Size: 253, ISim: 0.046, ESim: 0.007
Descriptive: nuclear 28.2%, nuclear.power 8.3%, nuclear.power.plant 2.8%, plant 2.6%, wast 2.4%, spent 2.4%, nuclear.fuel 2.0%, reprocess 1.7%, power 1.6%, spent.fuel 1.6%
Discriminating: nuclear 20.3%, nuclear.power 6.2%, nuclear.power.plant 2.2%, spent 1.7%, nuclear.fuel 1.5%, reprocess 1.3%, spent.fuel 1.1%, heat 1.1%, wast 1.1%, combust 1.1%

Cluster 8, Size: 153, ISim: 0.041, ESim: 0.006
Descriptive: laser 16.3%, puls 5.1%, optic 4.9%, pump 2.7%, diod 2.1%, beam 1.9%, effici 1.8%, output 1.5%, power 1.4%, power.convers 1.3%
Discriminating: laser 11.6%, optic 3.2%, puls 3.1%, fuel 1.5%, diod 1.4%, pump 1.4%, beam 1.2%, combust 1.0%, power.convers 0.8%, caviti 0.8%

Cluster 9, Size: 297, ISim: 0.040, ESim: 0.007
Descriptive: renew 19.5%, renew.energi 14.0%, wind 7.4%, energi 6.7%, sourc 3.3%, energi.sourc 2.5%, solar 2.3%, technolog 2.0%, renew.energi.sourc 1.7%, energi.technolog 1.2%
Discriminating: renew 14.6%, renew.energi 10.7%, wind 4.9%, energi.sourc 1.7%, sourc 1.5%, energi 1.5%, renew.energi.sourc 1.3%, combust 1.0%, heat 1.0%, energi.technolog 0.9%

Cluster 10, Size: 176, ISim: 0.040, ESim: 0.007
Descriptive: switch 8.9%, convert 8.4%, voltag 7.3%, circuit 4.4%, current 3.9%, reson 3.9%, puls 2.3%, frequenc 2.1%, control 1.9%, capacitor 1.8%
Discriminating: switch 6.6%, convert 5.5%, voltag 4.7%, circuit 3.1%, reson 2.7%, current 1.7%, fuel 1.5%, puls 1.3%, capacitor 1.3%, frequenc 1.2%

Cluster 11, Size: 189, ISim: 0.040, ESim: 0.007
Descriptive: bed 21.4%, fluidiz 7.8%, combust 4.9%, fluidiz.bed 4.7%, rdf 2.0%, pcdd 1.6%, wood 1.4%, chlorin 1.4%, particl 1.3%, ga 1.2%
Discriminating: bed 16.8%, fluidiz 6.4%, fluidiz.bed 3.8%, rdf 1.7%, pcdd 1.3%, power 1.3%, energi 1.2%, chlorin 1.1%, combust 1.1%, system 1.0%

Cluster 12, Size: 226, ISim: 0.039, ESim: 0.007
Descriptive: reactor 12.7%, fusion 8.4%, tritium 5.9%, core 4.9%, fuel 2.8%, neutron 2.3%, fuel.cycl 2.0%, cycl 2.0%, design 1.5%, plutonium 1.4%
Discriminating: reactor 8.8%, fusion 6.7%, tritium 4.7%, core 3.7%, neutron 1.8%, fuel.cycl 1.4%, combust 1.2%, plutonium 1.1%, heat 1.1%, energi 0.9%

Cluster 13, Size: 247, ISim: 0.038, ESim: 0.007
Descriptive: biomass 30.1%, crop 4.1%, forest 3.9%, product 3.6%, biomass.energi 2.0%, harvest 1.8%, wood 1.8%, land 1.6%, energi 1.2%, agricultur 1.0%
Discriminating: biomass 22.2%, crop 3.3%, forest 3.0%, biomass.energi 1.6%, harvest 1.4%, land 1.2%, product 1.1%, wood 1.1%, heat 1.0%, power 0.9%

Cluster 14, Size: 441, ISim: 0.037, ESim: 0.006
Descriptive: magnet 40.1%, field 8.1%, magnet.field 4.6%, magnet.energi 3.8%, coil 1.6%, superconduct 1.3%, plasma 1.1%, energi 1.1%, current 1.0%, bear 0.7%
Discriminating: magnet 28.6%, field 4.8%, magnet.field 3.5%, magnet.energi 2.8%, fuel 1.5%, coil 1.2%, combust 1.1%, heat 1.0%, superconduct 0.8%, power 0.8%

Cluster 15, Size: 160, ISim: 0.035, ESim: 0.005
Descriptive: acid 8.0%, enthalpi 7.8%, delta 6.0%, mol 3.4%, 298 3.1%, compound 2.8%, standard 1.9%, molar 1.5%, format 1.4%, delta.degre 1.4%
Discriminating: enthalpi 5.2%, acid 4.9%, delta 3.7%, 298 2.2%, mol 2.1%, compound 1.4%, power 1.1%, molar 1.0%, delta.degre 1.0%, standard 1.0%

Cluster 16, Size: 331, ISim: 0.032, ESim: 0.008
Descriptive: engin 20.9%, inject 4.3%, cylind 3.3%, ignit 2.8%, fuel 2.7%, combust 1.8%, spark 1.6%, hydrogen 1.6%, diesel 1.5%, exhaust 1.4%
Discriminating: engin 14.7%, inject 3.2%, cylind 3.0%, ignit 1.9%, energi 1.6%, spark 1.4%, power 1.2%, heat 1.0%, electr 1.0%, magnet 0.9%

Cluster 17, Size: 340, ISim: 0.031, ESim: 0.007
Descriptive: cell 12.0%, electrod 12.0%, electrolyt 4.7%, membran 3.1%, cathod 2.7%, electrochem 2.7%, fuel.cell 2.4%, anod 2.1%, current 1.5%, oxid 1.3%
Discriminating: electrod 10.4%, cell 7.3%, electrolyt 4.0%, membran 2.6%, cathod 2.3%, electrochem 2.2%, anod 1.7%, combust 1.3%, fuel.cell 1.1%, heat 1.1%

Cluster 18, Size: 197, ISim: 0.029, ESim: 0.006
Descriptive: film 3.7%, cell 3.4%, polym 3.4%, effici 2.8%, light 2.4%, convers.effici 2.4%, convers 2.0%, layer 1.7%, devic 1.7%, charg 1.3%
Discriminating: polym 2.2%, film 2.2%, convers.effici 1.7%, fuel 1.5%, light 1.5%, combust 1.1%, heat 1.1%, system 1.1%, cell 0.9%, energi 0.8%

Cluster 19, Size: 497, ISim: 0.031, ESim: 0.008
Descriptive: turbin 12.4%, plant 7.7%, steam 6.7%, ga.turbin 5.2%, ga 4.6%, cycl 4.0%, power 3.0%, combin 2.1%, combin.cycl 2.1%, power.plant 2.1%
Discriminating: turbin 11.3%, steam 6.1%, ga.turbin 4.9%, plant 4.7%, cycl 2.0%, combin.cycl 2.0%, ga 1.9%, combin 1.3%, power.plant 1.2%, energi 1.0%

Cluster 20, Size: 171, ISim: 0.029, ESim: 0.007
Descriptive: wast 9.7%, product 5.1%, oil 5.0%, recycl 3.7%, recoveri 2.2%, process 2.0%, energi.recoveri 1.8%, msw 1.8%, solid.wast 1.2%, environment 1.2%
Discriminating: wast 7.0%, recycl 3.0%, oil 2.7%, product 2.1%, msw 1.6%, energi.recoveri 1.5%, recoveri 1.4%, power 1.2%, solid.wast 1.1%, heat 1.1%

Cluster 21, Size: 391, ISim: 0.030, ESim: 0.008

Descriptive: co2 17.9%, emiss 10.7%, carbon 2.5%, vehicl 2.4%, fossil 2.3%, fossil.fuel 1.9%, co2.emiss 1.7%, fuel 1.5%, coal 1.3%, dioxid 1.2%
Discriminating: co2 15.9%, emiss 7.2%, fossil 1.7%, co2.emiss 1.5%, vehicl 1.4%, fossil.fuel 1.4%, carbon 1.2%, heat 1.2%, dioxid 1.0%, greenhous 1.0%

Cluster 22, Size: 419, ISim: 0.026, ESim: 0.008
Descriptive: heat 10.3%, storag 4.0%, transfer 3.6%, heat.transfer 3.6%, pcm 3.5%, thermal 2.7%, temperatur 2.2%, phase 2.1%, melt 1.8%, model 1.6%
Discriminating: heat 4.9%, pcm 3.8%, heat.transfer 3.2%, transfer 2.6%, storag 1.9%, fuel 1.9%, melt 1.5%, combust 1.4%, power 1.2%, fluid 1.2%

Cluster 23, Size: 342, ISim: 0.026, ESim: 0.008
Descriptive: solar 12.9%, heat 6.1%, system 4.5%, energi 1.8%, exergi 1.8%, thermal 1.7%, cost 1.6%, storag 1.5%, electr 1.5%, pump 1.4%
Discriminating: solar 11.2%, heat 1.9%, fuel 1.8%, exergi 1.6%, combust 1.4%, chp 1.3%, solar.thermal 1.1%, magnet 1.0%, system 1.0%, collector 1.0%

Cluster 24, Size: 286, ISim: 0.024, ESim: 0.006
Descriptive: conduct 10.4%, electr.conduct 4.4%, dope 4.2%, degre 2.6%, ion 2.1%, temperatur 1.7%, electr 1.5%, structur 1.5%, oxygen 1.3%, oxid 1.3%
Discriminating: conduct 7.2%, electr.conduct 3.7%, dope 3.5%, ion 1.4%, power 1.3%, combust 1.1%, fuel 1.1%, degre 1.1%, system 1.0%, ionic 1.0%

Cluster 25, Size: 372, ISim: 0.025, ESim: 0.008
Descriptive: coal 8.3%, combust 7.5%, particl 6.9%, char 4.6%, nitrogen 3.3%, oil 1.8%, fuel 1.8%, furnac 1.6%, pyrolysi 1.4%, burn 1.2%
Discriminating: coal 6.7%, particl 5.8%, char 4.8%, combust 3.2%, nitrogen 2.9%, power 1.6%, energi 1.6%, system 1.4%, furnac 1.3%, pyrolysi 1.2%

Cluster 26, Size: 540, ISim: 0.024, ESim: 0.008
Descriptive: combust 8.8%, flame 7.9%, combustor 3.4%, air 2.8%, fuel 2.3%, flow 2.1%, burner 2.0%, pressur 1.6%, model 1.4%, jet 1.4%
Discriminating: flame 7.9%, combust 4.7%, combustor 3.3%, burner 1.9%, energi 1.9%, power 1.7%, air 1.4%, jet 1.4%, nozzl 1.3%, flow 1.2%

Cluster 27, Size: 405, ISim: 0.022, ESim: 0.006
Descriptive: alloi 5.4%, film 5.1%, resist 5.0%, treatment 3.4%, heat.treatment 3.3%, properti 2.6%, temperatur 2.1%, electr 1.8%, heat 1.8%, materi 1.7%
Discriminating: alloi 4.4%, film 4.0%, resist 3.7%, heat.treatment 3.0%, treatment 2.8%, fuel 1.6%, properti 1.5%, electr.resist 1.3%, combust 1.2%, system 1.2%

Cluster 28, Size: 337, ISim: 0.021, ESim: 0.006
Descriptive: energi 9.6%, resourc 8.5%, energi.resourc 3.7%, consumpt 2.4%, countri 2.1%, econom 2.1%, energi.consumpt 1.6%, technolog 1.6%, world 1.2%, market 1.2%
Discriminating: resourc 7.1%, energi.resourc 3.4%, energi 3.4%, countri 1.5%, consumpt 1.5%, energi.consumpt 1.2%, combust 1.2%, econom 1.1%, heat 1.0%, temperatur 1.0%

Cluster 29, Size: 365, ISim: 0.021, ESim: 0.008
Descriptive: reaction 18.6%, oxid 3.5%, combust 2.8%, rate 2.5%, temperatur 1.7%, oxygen 1.3%, kinet 1.3%, product 1.0%, degre 1.0%, ga 0.9%
Discriminating: reaction 18.4%, oxid 2.2%, power 1.8%, rate 1.1%, magnet 1.1%, electr 1.1%, engin 1.0%, kinet 1.0%, system 1.0%, heat 1.0%

Cluster 30, Size: 598, ISim: 0.018, ESim: 0.007
Descriptive: power 7.7%, system 6.3%, control 3.4%, cost 3.2%, gener 2.9%, electr 2.6%, wind 2.5%, power.system 1.7%, electr.power 1.6%, util 1.5%

Discriminating: power 3.6%, system 2.5%, wind 2.1%, control 2.1%, power.system 1.9%, cost 1.8%, heat 1.7%, combust 1.6%, fuel 1.6%, electr.power 1.3%

Cluster 31, Size: 353, ISim: 0.014, ESim: 0.007
Descriptive: energi 2.3%, plasma 2.0%, model 1.9%, storag 1.5%, mechan 1.4%, energi.storag 1.3%, electr 0.9%, state 0.9%, wave 0.8%, time 0.8%
Discriminating: fuel 2.0%, combust 1.6%, plasma 1.6%, magnet 1.1%, heat 1.1%, engin 1.1%, plant 1.0%, ga 0.8%, system 0.8%, cell 0.8%

2B – JOURNAL-BASED DATABASE

Cluster 0, Size: 163, ISim: 0.076, ESim: 0.005
Descriptive: wind 61.1%, wind.energi 6.9%, energi 2.7%, speed 2.0%, wind.power 1.6%, wind.speed 1.3%, wind.turbin 1.2%, turbin 0.9%, power 0.7%, gener 0.6%
Discriminating: wind 38.1%, wind.energi 4.4%, coal 2.0%, heat 1.2%, speed 1.1%, wind.power 1.0%, fuel 0.8%, wind.speed 0.8%, temperatur 0.8%, wind.turbin 0.7%

Cluster 1, Size: 147, ISim: 0.069, ESim: 0.005
Descriptive: ash 49.7%, fly 9.9%, fly.ash 8.7%, coal.ash 2.2%, coal 1.8%, deposit 0.7%, slag 0.7%, ash.sampl 0.6%, boiler 0.5%, temperatur 0.5%
Discriminating: ash 31.7%, fly 6.5%, fly.ash 5.7%, energi 2.0%, coal.ash 1.4%, system 1.2%, heat 1.0%, solar 0.9%, model 0.8%, cell 0.7%

Cluster 2, Size: 221, ISim: 0.058, ESim: 0.005
Descriptive: renew 29.3%, renew.energi 21.8%, energi 14.5%, resourc 3.6%, sourc 2.1%, renew.energi.sourc 1.4%, energi.sourc 1.3%, geotherm 1.1%, technolog 1.1%, energi.resourc 0.8%
Discriminating: renew 19.6%, renew.energi 14.8%, energi 3.4%, coal 2.1%, resourc 2.0%, heat 1.1%, model 1.0%, renew.energi.sourc 1.0%, temperatur 0.9%, carbon 0.8%

Cluster 3, Size: 319, ISim: 0.059, ESim: 0.006
Descriptive: cell 22.9%, fuel.cell 22.7%, fuel 13.1%, power 1.6%, system 1.3%, stack 1.1%, sofc 1.0%, molten.carbon 1.0%, molten 1.0%, mcfc 0.9%
Discriminating: fuel.cell 16.0%, cell 12.6%, fuel 5.4%, coal 2.3%, energi 1.3%, heat 1.2%, solar 0.7%, stack 0.7%, sofc 0.7%, molten.carbon 0.7%

Cluster 4, Size: 236, ISim: 0.053, ESim: 0.005
Descriptive: collector 26.3%, solar 11.4%, solar.collector 4.7%, plate 4.6%, flat 3.7%, flat.plate 3.3%, heater 2.4%, air 2.4%, air.heater 2.2%, solar.air 2.2%
Discriminating: collector 17.9%, solar 4.2%, solar.collector 3.2%, plate 2.6%, flat 2.5%, coal 2.3%, flat.plate 2.2%, air.heater 1.5%, solar.air 1.5%, heater 1.5%

Cluster 5, Size: 471, ISim: 0.052, ESim: 0.004
Descriptive: lead 20.4%, batteri 14.4%, acid 10.2%, lead.acid 9.5%, acid.batteri 6.0%, lead.acid.batteri 5.9%, valv 1.2%, regul 1.2%, valv.regul 1.1%, posit 1.0%
Discriminating: lead 12.6%, batteri 6.9%, lead.acid 6.3%, acid 5.5%, acid.batteri 3.9%, lead.acid.batteri 3.9%, coal 2.2%, energi 1.9%, heat 1.3%, system 1.0%

Cluster 6, Size: 337, ISim: 0.051, ESim: 0.005
Descriptive: bed 24.3%, fluidiz 14.1%, fluidiz.bed 12.0%, combust 4.5%, bed.combust 1.8%, combustor 1.7%, coal 1.5%, fluidiz.bed.combust 1.4%, n2o 1.3%, circul 1.3%
Discriminating: bed 15.9%, fluidiz 9.8%, fluidiz.bed 8.3%, energi 2.3%, combust 1.6%, bed.combust 1.3%, system 1.2%, combustor 1.1%, heat 1.0%, fluidiz.bed.combust 1.0%

Cluster 7, Size: 254, ISim: 0.046, ESim: 0.005

Descriptive: electrolyt 21.3%, lithium 8.5%, polym 6.3%, ethylen 3.5%, polym.electrolyt 3.4%, carbon 3.1%, poli 2.2%, propylen 1.6%, conduct 1.6%, salt 1.5%
Discriminating: electrolyt 13.0%, polym 4.0%, lithium 3.8%, ethylen 2.3%, coal 2.3%, polym.electrolyt 2.2%, energi 2.2%, poli 1.4%, heat 1.3%, model 1.1%

Cluster 8, Size: 317, ISim: 0.039, ESim: 0.004
Descriptive: radiat 18.8%, solar.radiat 9.5%, solar 7.1%, data 4.7%, global 3.4%, daili 2.6%, monthli 2.4%, hourli 2.2%, averag 1.6%, measur 1.5%
Discriminating: radiat 12.3%, solar.radiat 6.4%, coal 2.2%, solar 2.0%, data 2.0%, global 1.8%, daili 1.7%, energi 1.6%, monthli 1.6%, hourli 1.4%

Cluster 9, Size: 420, ISim: 0.039, ESim: 0.005
Descriptive: lithium 32.2%, ion 5.8%, batteri 4.7%, lithium.ion 3.9%, cell 2.8%, materi 2.3%, cathod 2.1%, intercal 1.8%, graphit 1.8%, electrochem 1.6%
Discriminating: lithium 21.3%, ion 3.6%, lithium.ion 2.7%, coal 2.4%, energi 2.1%, batteri 1.3%, intercal 1.2%, cathod 1.2%, fuel 1.1%, graphit 1.1%

Cluster 10, Size: 308, ISim: 0.035, ESim: 0.006
Descriptive: heat 30.7%, pump 13.9%, heat.pump 9.2%, system 4.6%, storag 1.4%, water 1.4%, cool 0.8%, energi 0.8%, pump.system 0.8%, thermal 0.8%
Discriminating: heat 16.7%, pump 10.6%, heat.pump 7.4%, coal 2.7%, fuel 1.0%, cell 0.9%, carbon 0.9%, model 0.7%, batteri 0.7%, oil 0.7%

Cluster 11, Size: 395, ISim: 0.035, ESim: 0.005
Descriptive: co2 35.0%, emiss 11.6%, co2.emiss 2.8%, dispos 1.7%, carbon.dioxid 1.7%, dioxid 1.6%, greenhous 1.6%, carbon 1.6%, ocean 1.6%, atmospher 1.2%
Discriminating: co2 23.4%, emiss 7.3%, co2.emiss 2.0%, coal 1.8%, energi 1.5%, heat 1.4%, dispos 1.2%, ocean 1.1%, carbon.dioxid 1.0%, greenhous 1.0%

Cluster 12, Size: 316, ISim: 0.034, ESim: 0.005
Descriptive: refriger 10.5%, cycl 8.5%, thermodynam 5.7%, heat 5.5%, engin 4.9%, irrevers 4.0%, absorpt 2.2%, heat.engin 2.2%, finit 2.1%, system 1.9%
Discriminating: refriger 7.5%, cycl 4.5%, thermodynam 3.8%, irrevers 2.9%, engin 2.8%, coal 2.5%, energi 1.9%, heat.engin 1.7%, finit 1.4%, absorpt 1.3%

Cluster 13, Size: 386, ISim: 0.031, ESim: 0.004
Descriptive: catalyst 41.3%, hydrogen 6.2%, catalyt 2.6%, support 1.7%, al2o3 1.4%, activ 1.3%, iron 1.3%, hydrocrack 1.0%, reaction 1.0%, zeolit 0.9%
Discriminating: catalyst 29.1%, hydrogen 3.2%, energi 2.5%, catalyt 1.5%, heat 1.3%, system 1.3%, al2o3 1.0%, support 1.0%, solar 1.0%, cell 0.8%

Cluster 14, Size: 358, ISim: 0.031, ESim: 0.005
Descriptive: transfer 15.2%, heat 13.0%, heat.transfer 12.0%, flow 3.3%, convect 2.4%, wall 1.6%, fluid 1.2%, conduct 1.0%, tube 0.9%, model 0.9%
Discriminating: transfer 10.8%, heat.transfer 9.0%, heat 4.8%, coal 2.5%, energi 1.8%, convect 1.8%, flow 1.5%, wall 1.0%, carbon 0.9%, cell 0.9%

Cluster 15, Size: 489, ISim: 0.029, ESim: 0.005
Descriptive: oil 44.6%, shale 7.8%, crude 4.0%, oil.shale 3.4%, crude.oil 3.3%, pyrolysi 2.1%, heavi 0.8%, process 0.8%, product 0.7%, kerogen 0.7%
Discriminating: oil 31.1%, shale 5.9%, crude 2.8%, oil.shale 2.6%, energi 2.4%, crude.oil 2.4%, coal 1.5%, system 1.3%, heat 1.1%, solar 1.0%

Cluster 16, Size: 239, ISim: 0.029, ESim: 0.005
Descriptive: fuel 28.3%, vehicl 11.3%, engin 4.0%, diesel 3.7%, electr.vehicl 3.0%, electr 2.9%, combust 1.7%, diesel.fuel 1.1%, batteri 1.0%, ignit 0.7%

Discriminating: fuel 16.1%, vehicl 8.9%, diesel 2.5%, coal 2.5%, electr.vehicl 2.4%, engin 2.3%, energi 1.8%, heat 1.3%, solar 0.9%, model 0.8%

Cluster 17, Size: 255, ISim: 0.029, ESim: 0.005
Descriptive: particl 17.1%, combust 12.4%, coal 5.4%, pulver 4.7%, pulver.coal 3.5%, coal.particl 2.3%, size 2.0%, coal.combust 1.5%, furnac 1.4%, ga 1.2%
Discriminating: particl 12.7%, combust 7.3%, pulver 4.0%, pulver.coal 3.0%, energi 2.7%, coal.particl 1.9%, system 1.2%, coal.combust 1.2%, solar 1.2%, cell 1.0%

Cluster 18, Size: 274, ISim: 0.028, ESim: 0.005
Descriptive: carbon 25.7%, co2 9.7%, activ 8.2%, activ.carbon 6.1%, dioxid 2.0%, carbon.dioxid 1.9%, adsorpt 1.8%, oxid 1.4%, surfac 1.1%, methan 0.7%
Discriminating: carbon 15.0%, activ 4.9%, activ.carbon 4.8%, co2 4.7%, energi 2.6%, coal 2.0%, heat 1.2%, system 1.2%, carbon.dioxid 1.2%, adsorpt 1.2%

Cluster 19, Size: 412, ISim: 0.027, ESim: 0.004
Descriptive: coal 20.9%, extract 7.2%, solvent 5.6%, liquefact 3.2%, coal.tar 2.8%, pitch 2.5%, tar 2.4%, swell 2.4%, pyridin 1.9%, argonn 1.2%
Discriminating: coal 7.6%, extract 4.5%, solvent 3.5%, energi 2.6%, coal.tar 2.2%, liquefact 2.1%, swell 1.8%, pitch 1.7%, tar 1.6%, pyridin 1.4%

Cluster 20, Size: 315, ISim: 0.027, ESim: 0.005
Descriptive: char 20.6%, coal 15.6%, pyrolysi 4.3%, coke 4.2%, gasif 2.6%, coal.char 2.4%, rate 1.6%, reactiv 1.5%, nitrogen 1.5%, temperatur 1.0%
Discriminating: char 17.0%, coal 5.0%, coke 3.0%, energi 2.6%, pyrolysi 2.3%, coal.char 2.1%, gasif 1.6%, system 1.6%, solar 1.1%, cell 1.0%

Cluster 21, Size: 434, ISim: 0.025, ESim: 0.005
Descriptive: electrod 18.8%, nickel 8.2%, cell 6.2%, discharg 3.7%, electrochem 2.6%, zinc 1.9%, electrolyt 1.8%, alloi 1.8%, imped 1.4%, cathod 1.3%
Discriminating: electrod 13.3%, nickel 6.2%, coal 2.7%, energi 2.3%, discharg 2.3%, cell 2.1%, heat 1.5%, zinc 1.4%, electrochem 1.3%, system 1.1%

Cluster 22, Size: 437, ISim: 0.025, ESim: 0.005
Descriptive: solar 33.0%, water 5.1%, system 3.4%, cooker 2.0%, solar.energi 1.5%, solar.water 1.4%, energi 1.3%, design 1.2%, solar.cell 1.2%, thermal 1.1%
Discriminating: solar 22.0%, coal 2.9%, water 2.0%, cooker 1.7%, fuel 1.2%, solar.water 1.2%, carbon 1.0%, solar.energi 0.9%, solar.cell 0.9%, batteri 0.8%

Cluster 23, Size: 311, ISim: 0.023, ESim: 0.004
Descriptive: film 9.6%, electrochem 6.7%, rai 3.8%, diffract 3.3%, structur 3.0%, thin 2.6%, rai.diffract 2.5%, oxid 2.2%, thin.film 2.2%, synthes 2.0%
Discriminating: film 6.7%, electrochem 3.9%, coal 2.5%, rai 2.4%, diffract 2.3%, energi 2.3%, rai.diffract 1.8%, thin 1.8%, thin.film 1.6%, synthes 1.4%

Cluster 24, Size: 283, ISim: 0.023, ESim: 0.005
Descriptive: build 13.6%, cool 9.9%, air 8.4%, thermal 5.2%, design 2.3%, condit 1.9%, evapor 1.6%, ventil 1.5%, comfort 1.5%, air.condit 1.2%
Discriminating: build 10.1%, cool 7.0%, air 4.5%, coal 2.8%, thermal 2.0%, ventil 1.2%, comfort 1.2%, evapor 1.1%, solar 1.0%, fuel 1.0%

Cluster 25, Size: 815, ISim: 0.021, ESim: 0.005
Descriptive: energi 46.3%, consumpt 3.5%, energi.consumpt 2.1%, product 1.0%, system 1.0%, sector 0.9%, countri 0.8%, sourc 0.8%, suppli 0.8%, effici 0.7%
Discriminating: energi 30.2%, consumpt 2.6%, coal 2.6%, energi.consumpt 1.8%, temperatur 1.1%, heat 1.1%, cell 0.9%, carbon 0.8%, oxid 0.7%, sector 0.6%

Cluster 26, Size: 368, ISim: 0.021, ESim: 0.005
Descriptive: model 29.6%, predict 4.0%, mathemat 2.9%, equat 2.0%, dimension 1.9%, flow 1.9%, mathemat.model 1.7%, comput 1.6%, simul 1.5%, numer 1.3%
Discriminating: model 19.4%, predict 2.8%, coal 2.3%, mathemat 2.3%, energi 2.1%, dimension 1.5%, mathemat.model 1.4%, equat 1.3%, model.predict 1.1%, carbon 1.0%

Cluster 27, Size: 433, ISim: 0.020, ESim: 0.004
Descriptive: asphalten 10.1%, bitumen 6.3%, chromatographi 3.4%, fraction 3.4%, aromat 2.8%, extract 2.7%, hydrocarbon 1.9%, liquid 1.8%, residu 1.6%, solvent 1.4%
Discriminating: asphalten 8.1%, bitumen 5.0%, chromatographi 2.6%, energi 2.5%, fraction 2.1%, aromat 1.9%, coal 1.6%, system 1.4%, heat 1.3%, extract 1.3%

Cluster 28, Size: 590, ISim: 0.020, ESim: 0.005
Descriptive: power 21.0%, plant 7.4%, gener 6.1%, turbin 3.7%, system 3.6%, ga 2.7%, power.gener 2.1%, electr 1.9%, power.plant 1.6%, ga.turbin 1.5%
Discriminating: power 15.0%, plant 5.4%, gener 3.4%, turbin 3.1%, coal 2.2%, power.gener 1.8%, energi 1.5%, ga.turbin 1.3%, power.plant 1.3%, heat 1.2%

Cluster 29, Size: 755, ISim: 0.019, ESim: 0.005
Descriptive: coal 34.4%, bitumin 3.2%, rank 3.0%, bitumin.coal 1.9%, lignit 1.9%, sampl 1.8%, sulfur 1.5%, low 1.4%, degre 0.9%, volatil 0.9%
Discriminating: coal 20.0%, energi 2.9%, bitumin 2.5%, rank 2.4%, system 1.7%, bitumin.coal 1.4%, lignit 1.3%, solar 1.2%, fuel 1.1%, cell 1.1%

Cluster 30, Size: 508, ISim: 0.015, ESim: 0.004
Descriptive: electr 7.4%, demand 4.3%, cost 3.2%, program 3.0%, photovolta 3.0%, paper 2.3%, countri 2.2%, econom 2.1%, util 2.0%, manag 1.9%
Discriminating: electr 3.8%, demand 3.1%, coal 2.7%, photovolta 2.0%, program 1.8%, cost 1.8%, heat 1.6%, manag 1.5%, util 1.4%, countri 1.4%

Cluster 31, Size: 588, ISim: 0.013, ESim: 0.005
Descriptive: reaction 6.0%, degre 5.4%, temperatur 4.7%, reactor 2.9%, oxid 2.5%, compound 2.0%, pressur 2.0%, ga 2.0%, high 1.6%, kinet 1.6%
Discriminating: reaction 4.3%, energi 3.7%, degre 3.1%, coal 2.7%, reactor 1.8%, temperatur 1.5%, solar 1.4%, system 1.3%, compound 1.3%, flame 1.3%

8. REFERENCES

[1] Kostoff RN. Text mining for global technology watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003.
[2] Hearst MA. Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
[3] Zhu DH, Porter AL. Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change. 2002. 69 (5): 495-506.
[4] Losiewicz P, Oard D, Kostoff RN. Textual data mining to support science and technology management. Journal of Intelligent Information Systems. 2000. 15. 99-119.

[5] Kostoff RN, Eberhart HJ, Toothman DR. Database Tomography for information retrieval. Journal of Information Science. 1997. 23:4.
[6] Greengrass E. Information retrieval: An overview. National Security Agency. 1997. TR-R52-02-96.
[7] TREC (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
[8] Swanson DR. Fish Oil, Raynauds Syndrome, and undiscovered public knowledge. 1986. Perspect Biol Med. .30: (1). 7-18.
[9] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. 1997. Artif Intell, 91 (2). 183-203.
[10] Kostoff RN. Stimulating innovation. International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 2003.
[11] Gordon MD, Dumais S . Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science. 1998. 49 (8): 674-685.
[12] Goldman JA, Chu, WW, Parker, DS, Goldman, RM. Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 1999. 38. 96-101.
[13] Kostoff, RN. Bilateral asymmetry prediction. Medical Hypotheses. 61:2. 2003. 265-266.
[14] Kostoff RN, Green KA, Toothman DR, and Humenik, JA. Database Tomography applied to an aircraft science and technology investment strategy. Journal of Aircraft. 2000. 37:4. 727-730.
[15] Kostoff RN, Shlesinger M, Malpohl G. Fractals roadmaps using bibliometrics and database tomography. Fractals. 2003 (Dec).
[16] Viator JA, Pestorius FM . Investigating trends in acoustics research from 1970-1999. 2001. Journal of the Acoustical Society of America. 109 (5): 1779-1783 Part 1.
[17] Kostoff RN, Shlesinger M, Tshiteya R. Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 2004 (Jan).
[18] Davidse RJ, Van Raan AFJ. Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. Scientometrics 1997. 40:2 . 171-193.
[19] Kostoff RN, Del Rio JA, García EO, Ramírez AM, Humenik JA. Citation mining: integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science and Technology. 2001. 52:13. 1148-1156.
[20] Narin F. Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. 1976. Contract NSF C-627. NTIS Accession No. PB252339/AS.
[21] Garfield E. History of citation indexes for chemistry - a brief review. JCICS. 1985. 25(3). 170-174.
[22] Schubert A, Glanzel W, Braun T. Subject field characteristic citation scores and scales for assessing research performance. Scientometrics. 1987. 12 (5-6): 267-291.

[23] Narin F, Olivastro D, Stevens KA. Bibliometrics theory, practice and problems. Evaluation Review. 1994. 18(1). 65-76.
[24] Kostoff RN, Eberhart, HJ, Miles, DA. System and method for Database Tomography. U.S. Patent Number 5440481. 1995.
[25] Kostoff RN, Eberhart HJ, and Toothman DR. Database Tomography for information retrieval. Journal of Information Science. 1997; 23(4): 301-311.
[26] Kostoff RN. Database Tomography for technical intelligence. Competitive Intelligence Review. 1993; 4(1): 38-43.
[27] Kostoff RN, Eberhart HJ, and Toothman DR. Hypersonic and supersonic flow roadmaps using bibliometrics and Database Tomography. JASIS. 15 April 1999; 50(5): 427-447.
[28] Kostoff RN, Eberhart HJ, and Toothman DR. Database Tomography for technical intelligence: a roadmap of the near-earth space science and technology literature. Information Processing and Management 1998; 34(1): 69-85.
[29] Kostoff RN, Eberhart HJ, Toothman DR, and Pellenbarg R. Database Tomography for technical intelligence: comparative roadmaps of the research impact assessment literature and the journal of the American chemical society. Scientometrics. 1997; 40(1): 103-138.
[30] Kostoff RN, Braun T, Schubert A, Toothman DR, and Humenik JA. Fullerene roadmaps using bibliometrics and Database Tomography. Journal of Chemical Information and Computer Science. 2000; 40(1): 19-39.
[31] Kostoff RN, Green KA, Toothman DR, and Humenik J. Database Tomography applied to an aircraft science and technology investment strategy. Journal of Aircraft. 2000; 37(4): 727-730.
[32] Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA. Electrochemical power source roadmaps using bibliometrics and Database Tomography. Journal of Power Sources. 2002. 110:1. 163-176.
[33] Ding J, Berleant D, Nettleton D, Wurtele E. Mining Medline: abstracts, sentences, or phrases. Pacific Symposium on Biocomputing. 2002. 326-337.
[34] SCI. Science Citation Index. Institute for Scientific Information. Phila., PA. 1999.
[35] Kostoff RN, The underpublishing of science and technology results. The Scientist 1 May 2000; 14(9): 6-6.
[36] Kostoff RN. Science and technology innovation. Technovation 1999; 19(10): 593-604.
[37] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 1997; 91(2): 183-203.
[38] Hearst MA. Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26, 1999. 1-9.
[39] Garfield E. History of citation indexes for chemistry - a brief review. JCICS. 1985; 25(3): 170-174.
[40] Kostoff RN. The use and misuse of citation analysis in research evaluation. Scientometrics 1998; 43:1: 27-43.
[41] MacRoberts M, MacRoberts B. Problems of citation analysis. Scientometrics 1996; 36(3): 435-444.

[42] Cutting DR, Karger DR, Pedersen JO, Tukey JW. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92). 1992. 318-329. [43] Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD'98). 1998. 73-84.
[44] Hearst MA. The use of categories and clusters in information access interfaces. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic Publishers. 2000.
[45] Karypis G, Han EH, Kumar V. Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining. 1999. 32(8). 68--75.
[46] Prechelt L, Malpohl G, Philippsen M. Finding plagiarisms among a set of programs with JPlag. Journal of Universal Computer Science. 2002. 8(11). 1016-1038.
[47] Rasmussen E. Clustering Algorithms. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms. 1992. Prentice Hall, N. J.
[48] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques. Technical Report #00--034. 2000. Department of Computer Science and Engineering. University of Minnesota.
[49] Willet P. Recent trends in hierarchical document clustering: A critical review. Information Processing and Management. 1988. 24:577-597.
[50] Wise MJ. String similarity via greedy string tiling and running Karb-Rabin matching. ftp://ftp.cs.su.oz.au/michaelw/doc/RKR_GST.ps, 1992. Dept. of CS, University of Sidney.
[51] Zamir O, Etzioni O. Web document clustering: A feasibility demonstration. In: *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98). 1998. 46-54.
\[52] Karypis G. CLUTO—A clustering toolkit. http://www.cs.umn.edu/˜cluto.
[53] Zhao Y, Karypis G. Criterion functions for document clustering: Experiments and analysis. Machine Learning, in press.

FIGURES

FIGURE 1 – COUNTRY-COUNTRY CO-OCCURRENCE MATRIX

	# Records	5285	2269	1358	1196	1141	997	813	603	586	559	498	474	464	382	353
# Records	Country	USA	JAPAN	ENGLAND	INDIA	GERMANY	CANADA	FRANCE	Australia	PEOPLES R CHINA	ITALY	SPAIN	TURKEY	RUSSIA	SWEDEN	NETHERLANDS
5285	**USA**	5285	84	59	27	62	85	47	30	56	28	25	9	20	8	29
2269	**Japan**	84	2269	14	11	11	26	10	19	19	5	2	2	5	2	3
1358	**England**	59	14	1358	6	21	7	20	11	10	14	24	16	2	8	11
1196	**India**	27	11	6	1196	8	4	2	1	1	5	1				1
1141	**Germany**	62	11	21	8	1141	10	15	7	1	10	8	6	8	9	13
997	**Canada**	85	26	7	4	10	997	13	6	10	2	2	6	3	2	2
813	**France**	47	10	20	2	15	13	813	1		17	30		14		9
603	**Australia**	30	19	11	1	7	6	1	603	11		1	1	1	3	2
586	**Peoples R China**	56	19	10	1	1	10		11	586					4	5
559	**Italy**	28	5	14	5	10	2	17			559	6	1	1	6	7
498	**Spain**	25	2	24	1	8	2	30	1		6	498		1	1	5
474	**Turkey**	9	2	16		6	6		1		1		474		2	2
464	**Russia**	20	5	2		8	3	14	1		1	1		464	2	7
382	**Sweden**	8	2	8		9	2		3	4	6	1	2	2	382	3
353	**Netherlands**	29	3	11	1	13	2	9	2	5	7	5	2	7	3	353

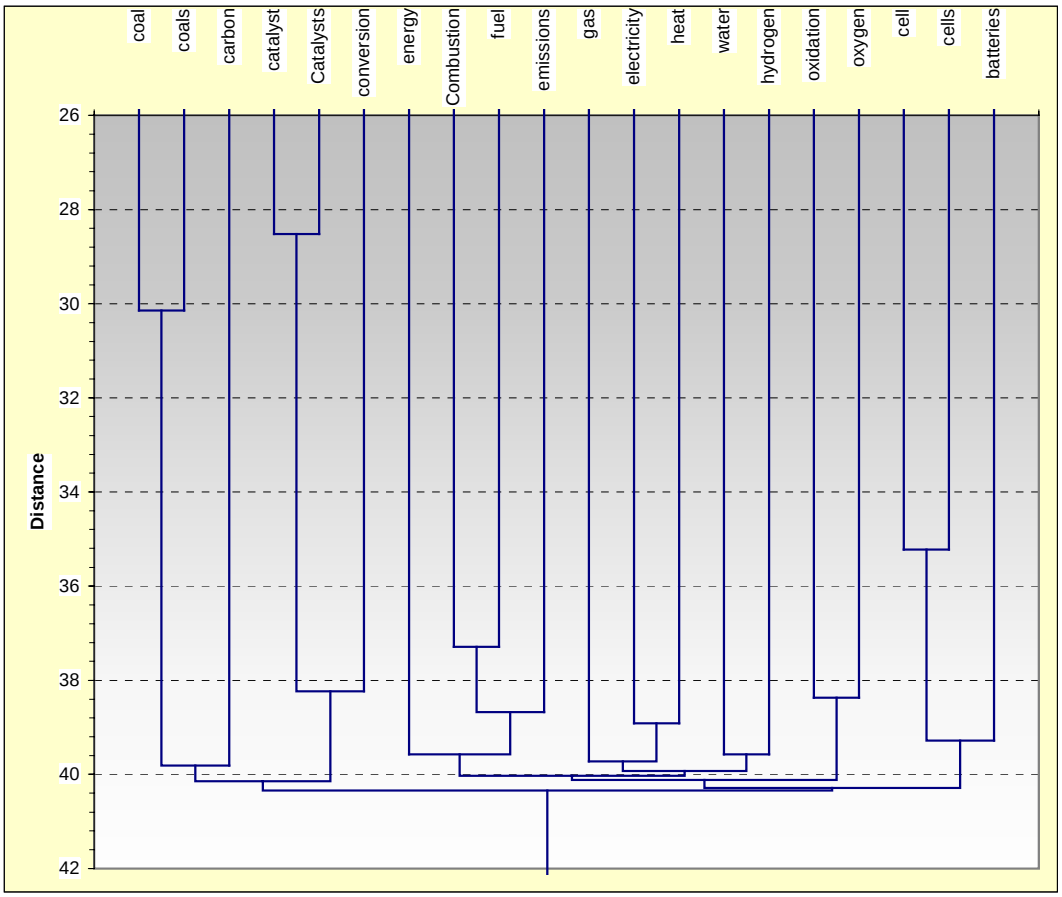
FIGURE 2 – COUNTRY-TIME MATRIX

COUNTRY	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
USA	471	456	587	532	505	566	552	521	500	433
JAPAN	132	137	154	144	267	227	363	259	270	251
ENGLAND	79	93	112	157	143	159	130	158	146	132
INDIA	119	85	94	130	111	128	113	144	124	114
GERMANY	102	95	110	106	103	107	103	148	136	83
CANADA	72	85	95	92	124	116	116	84	107	91
FRANCE	52	44	62	79	92	92	88	93	129	64
AUSTRALIA	37	54	54	55	38	73	54	60	59	73
PEOPLES R CHINA	23	22	33	29	44	70	57	106	107	79
ITALY	22	27	48	47	61	57	59	82	70	65
SPAIN	20	26	23	51	49	54	71	57	77	60
TURKEY	12	16	26	29	46	63	57	56	78	83
RUSSIA		15	32	36	43	56	61	43	64	35
SWEDEN	21	16	33	39	27	60	40	46	41	52
NETHERLANDS	14	26	35	45	34	44	37	45	32	29
SOUTH KOREA	15	13	7	11	23	24	38	42	78	53
EGYPT	16	12	27	37	27	32	39	36	23	38
SAUDI ARABIA	14	11	16	29	21	41	12	41	37	24
POLAND	9	11	20	37	29	25	23	37	28	28
GREECE	11	13	16	21	17	26	26	35	27	28
TAIWAN	12	12	13	21	18	35	26	23	18	29
ISRAEL	14	14	27	11	19	18	20	24	27	17
SCOTLAND	13	7	13	18	13	19	22	32	24	21
FINLAND	16	14	11	14	23	23	17	26	19	20
BRAZIL	3	12	5	3	6	16	23	34	33	30

FIGURE 3 – COUNTRY-JOURNAL MATRIX

JOURNAL	USA	JAPAN	ENGLAND	INDIA	GERMANY	CANADA	FRANCE	AUSTRALIA	CHINA	ITALY	SPAIN	TURKEY	RUSSIA	SWEDEN	NETHERLANDS
Fuel	0.157	0.126	0.305	0.092	0.147	0.211	0.23	0.337	0.175	0.147	0.44	0.198	0.207	0.171	0.183
J. Power Sources	0.151	0.3	0.16	0.109	0.374	0.135	0.398	0.19	0.305	0.203	0.08	0.002	0.239	0.122	0.228
Energy Fuels	0.27	0.211	0.047	0.015	0.056	0.16	0.126	0.153	0.056	0.04	0.269	0.05	0.033	0.137	0.041
Energy Conv. Manag.	0.07	0.181	0.069	0.296	0.043	0.097	0.05	0.033	0.133	0.168	0.031	0.214	0.109	0.072	0.219
Renew. Energy	0.033	0.041	0.181	0.096	0.104	0.031	0.081	0.151	0.047	0.176	0.088	0.074	0.065	0.11	0.082
Energy	0.091	0.062	0.025	0.082	0.078	0.056	0.027	0.019	0.128	0.053	0.047	0.133	0.054	0.152	0.068
Int. J. Energy Res.	0.022	0.016	0.054	0.197	0.024	0.087	0.025	0.041	0.077	0.061	0.016	0.079	0.022	0.065	0.018
Energy Sources	0.04	0.01	0.014	0.063	0.017	0.14	0.012	0.017	0.023	0.013	0.005	0.219	0.022	0.019	0.009
J. Eng. Gas. Turbines Power-Trans. ASME	0.043	0.018	0.012	0.001	0.015	0.011	0.002	0.000	0.002	0.024	0.000	0.000	0.000	0.011	0.005
J. Inst. Energy	0.009	0.003	0.088	0.004	0.006	0.009	0.006	0.021	0.019	0.005	0.000	0.01	0.011	0.03	0.05
Int. J. Hydrog. Energy	0.016	0.008	0.003	0.017	0.047	0.027	0.002	0.000	0.009	0.032	0.003	0.002	0.109	0.011	0.005
J. Propul. Power	0.033	0.006	0.002	0.002	0.004	0.004	0.000	0.006	0.005	0.003	0.000	0.000	0.065	0.011	0.009
Biomass Bioenerg.	0.013	7E-04	0.011	0.011	0.006	0.016	0.002	0.000	0.000	0.011	0.005	0.000	0.000	0.076	0.032
Combust. Sci. Technol.	0.016	0.005	0.016	0.003	0.019	0.004	0.008	0.004	0.002	0.021	0.000	0.000	0.022	0.000	0.014
Combust. Flame	0.016	0.004	0.008	0.000	0.009	0.009	0.01	0.004	0.007	0.011	0.005	0.000	0.011	0.000	0.014
Sol. Energy	0.004	0.002	0.005	0.011	0.026	0.003	0.006	0.023	0.009	0.019	0.01	0.019	0.000	0.008	0.018
IEEE Trans. Magn.	0.017	0.007	0.001	0.001	0.026	0.001	0.014	0.002	0.002	0.011	0.000	0.000	0.033	0.004	0.005

FIGURE 4 – SAMPLE DENDOGRAM



TABLES

TABLE 1 - DT STUDIES OF TOPICAL FIELDS

TOPICAL AREA	NUMBER OF SCI ARTICLES	YEARS COVERED
1) NEAR-EARTH SPACE (NES)	5480	1993-MID 1996
2) HYPERSONICS (HSF)	1284	1993-MID 1996
3)CHEMISTRY (JACS)	2150	1994
4) FULLERENES (FUL)	10515	1991-MID 1998
5) AIRCRAFT (AIR)	4346	1991-MID 1998
6) HYDRODYNAMICS (HYD)	4608	1991-MID 1998
7) ELECTROCHEM POWER (ECHEM)	6985	1991-MID-2001
8) RESEARCH ASSESSMENT (RIA)	2300	1991-BEG 1995
9) ELECTRIC POWER SOURCES (EPS)	20835	1991 – LATE 2000

TABLE 2 – MOST PROLIFIC AUTHORS

(present institution listed)

AUTHOR NAME	INSTITUTION	COUNTRY	# PAPERS
WU C	U. S. NAVAL ACADEMY	USA	71
KANDIYOTI R	UNIVERSITY LONDON	UK	69
TIWARI GN	INDIAN INST TECHNOLOGY	INDIA	62
DINCER I	KING FAHD UNIV	SAUDI ARABIA	61
GARG HP	INDIAN INST TECHNOLOGY	INDIA	49
KANDPAL TC	INDIAN INST TECHNOLOGY	INDIA	48
SNAPE CE	UNIV NOTTINGHAM	UK	43
WILLIAMS A	UNIV LEEDS	UK	42
ISHIKAWA M	YAMAGUCHI UNIV	JAPAN	41
KUMAR S	INDIAN INST TECHNOLOGY	INDIA	39

TABLE 3 – JOURNALS FROM QUERY-DERIVED COMPONENT OF DATABASE CONTAINING MOST PAPERS

JOURNAL NAMES	# PAPERS
J. ENG. GAS. TURBINES POWER-TRANS. ASME	200
INT. J. HYDROG. ENERGY	186
J. PROPUL. POWER	140
BIOMASS BIOENERG.	134
COMBUST. SCI. TECHNOL.	121
BRENNST.-WARME-KRAFT	119
IEEE TRANS. MAGN.	108
COMBUST. FLAME	103
ENERGY POLICY	102
SOL. ENERGY	98
APPL. ENERGY	90
COMBUST. EXPLOS.	88
J. APPL. PHYS.	82
SOLID STATE ION.	75
FUSION TECHNOL.	71
J. ELECTROCHEM. SOC.	67
IEEE TRANS. ENERGY CONVERS.	62
JSME INT. J. SER. B-FLUIDS THERM. ENG.	58
APPL. THERM. ENG.	57
IEEE TRANS. POWER SYST.	55

TABLE 4 – PROLIFIC INSTITUTIONS

INSTITUTION NAMES	COUNTRY	# PAPERS
INDIAN INST TECHNOL	INDIA	415
CSIC	SPAIN	186
PENN STATE UNIV	USA	172
RUSSIAN ACAD SCI	RUSSIA	164
TOHOKU UNIV	JAPAN	163
ARGONNE NATL LAB	USA	142
CSIRO	AUSTRALIA	137
KING FAHD UNIV PETR & MINERALS	SAUDI ARABIA	137
UNIV LEEDS	UK	127
UNIV TOKYO	JAPAN	122

TABLE 5 – PROLIFIC COUNTRIES

COUNTRY	#PAPERS	POPULATION (MILLIONS)	GROSS DOMESTIC PRODUCT ($BILLIONS)	#PAPERS/ POPULATION	#PAPERS/ GROSS DOMESTIC PRODUCT
USA	5285	278	9963	19.01079	0.530463
JAPAN	2269	127	3150	17.86614	0.720317
ENGLAND	1358	60	1360	22.63333	0.998529
INDIA	1196	1030	2200	1.161165	0.543636
GERMANY	1141	83	1936	13.74699	0.58936
CANADA	997	31	775	32.16129	1.286452
FRANCE	813	59	1448	13.77966	0.561464
AUSTRALIA	603	19	445	31.73684	1.355056
PEOPLES R CHINA	586	1284	4500	0.456386	0.130222
ITALY	559	58	1273	9.637931	0.43912
SPAIN	498	40	720	12.45	0.691667
TURKEY	474	66	444	7.181818	1.067568
RUSSIA	464	145	1120	3.2	0.414286
SWEDEN	382	9	197	42.44444	1.939086
NETHERLANDS	353	16	388	22.0625	0.909794
SOUTH KOREA	316	48	765	6.583333	0.413072
EGYPT	294	68	247	4.323529	1.190283
POLAND	256	39	328	6.564103	0.780488
SAUDI ARABIA	248	23	232	10.78261	1.068966
GREECE	225	11	182	20.45455	1.236264

TABLE 6 – MOST CITED AUTHORS

(cited by other papers in this database only)

AUTHOR	TOPIC	INSTITUTION	COUNTRY	#CITES
SOLOMON PR	COAL PYROLYSIS	ADV FUEL RES INC	USA	510
PAVLOV D	LEAD-ACID BATTERIES	BULGARIAN ACAD SCI	BULGARIA	420
BEJAN A	THERMODYNAMICS	DUKE UNIV	USA	405
AURBACH D	LITHIUM BATTERIES	BAR ILAN UNIV	ISRAEL	367
LARSEN JW	COAL PYROLYSIS	LEHIGH UNIV	USA	355
MOCHIDA I	CARBON APPLICATIONS	KYUSHU UNIV	JAPAN	292
OHZUKU T	LITHIUM BATTERIES	OSAKA CITY UNIV	JAPAN	274
SUUBERG EM	COAL PYROLYSIS	BROWN UNIV	USA	245
NISHIOKA M	COMBUSTION	NAGOYA UNIV	JAPAN	233
WU C	THERMODYNAMICS	US NAVAL ACADEMY	USA	230
DUFFIE JA	SOLAR HEATING	UNIV WISCONSIN	USA	221
VANKREVELEN DW	POLYMERS	AKZO RES AND ENGRNG	NETHERLANDS	206
DEVOS A	THERMODYNAMICS	STATE UNIV GHENT	BELGIUM	198
SUZUKI T	COAL PYROLYSIS	KYOTO UNIV	JAPAN	196
PAINTER PC	COAL PROPERTIES	PENN STATE UNIV	USA	194
LI CZ	COAL PYROLYSIS	UNIV LONDON IMPER COLL	UK	193
SABBAH R	COMB THERMODYNAMICS	CNRS	FRANCE	190
HEROD AA	COAL COMBUSTION	UNIV LONDON IMPER COLL	UK	190
CHEN JC	THERMODYNAMICS	XIAMEN UNIV	CHINA	185
HUFFMAN GP	FOSSIL COMBUSTION	UNIV KENTUCKY	USA	184

TABLE 7 – MOST CITED PAPERS

(total citations listed in SCI)

AUTHOR	YEAR	JOURNAL	VOLUME	SCI CITES	TOTAL CITES
CURZON FL	1975	AM J PHYS	V43	154	366
CARNOT ENGINE EFFICIENCY AT MAXIMUM POWER OUTPUT					
MILLER JA	1989	PROG ENERG COMBUST	V15	90	825
MODELING NITROGEN CHEMISTRY IN COMBUSTION					
SOLUM MS	1989	ENERG FUEL	V3	83	170
SOLID STATE NMR OF ARGONNE PREMIUM COALS					
VORRES KS	1990	ENERG FUEL	V4	82	153
ARGONNE PREMIUM COAL					
FONG R	1990	J ELECTROCHEM SOC	V137	68	346
LITHIUM INTERCALATION INTO CARBON					
LARSEN JW	1985	J ORG CHEM	V50	59	125
STRUCTURE OF BITUMINOUS COALS					
SOLOMON PR	1990	ENERG FUEL	V4	59	143
ARGONNE PREMIUM COAL ANALYSIS					
IINO M	1988	FUEL	V67	56	112
COAL EXTRACTION					
OHZUKU T	1990	J ELECTROCHEM SOC	V137	54	336
MANGANESE DIOXIDE IN LITHIUM NONAQUEOUS CELL					
NISHIOKA M	1990	ENERG FUEL	V4	51	80
AROMATIC STRUCTURES IN COALS					

TABLE 8 – MOST CITED JOURNALS
(cited by other papers in this database only)

JOURNAL	TIMES CITED
FUEL	15013
J ELECTROCHEM SOC	6600
ENERG FUEL	6317
J POWER SOURCES	4238
SOL ENERGY	2957
COMBUST FLAME	2611
SOLID STATE IONICS	1922
J CHEM PHYS	1752
CARBON	1686
J APPL PHYS	1654
J PHYS CHEM-US	1652
FUEL PROCESS TECHNOL	1573
ELECTROCHIM ACTA	1558
COMBUST SCI TECHNOL	1523
J AM CHEM SOC	1511
ENERGY	1466
IND ENG CHEM RES	1426
ANAL CHEM	1412
J CATAL	1371
NATURE	1358

TABLE 9 – SPECIFIC POWER-ORIENTED JOURNALS FROM SCI

JOURNAL OF THE AMERICAN OIL CHEMISTS SOCIETY
OIL SHALE
ENERGY EXPLORATION AND EXPLOITATION
PETROLEUM SCIENCE AND TECHNOLOGY
CHEMISTRY AND PETROLEUM ENGINEERING
SEKIYU GAKKAISHI
PETROLEUM CHEMISTRY
PIPELINE GAS JOURNAL
BIOMASS AND BIOENERGY
SOLAR ENERGY
SOLAR ENERGY MATERIALS AND SOLAR CELLS
JOURNAL OF SOLAR ENERGY ENGINEERING
PROGRESS IN PHOTOVOLTAICS
JOURNAL OF WIND ENGINEERING AND INDUSTRIAL AERODYNAMICS
JOURNAL OF NUCLEAR MATERIALS
NUCLEAR ENERGY-JOURNAL OF THE BRITISH NUCLEAR ENERGY SOCIETY
ANNALS OF NUCLEAR ENERGY
NUCLEAR ENGINEERING INTERNATIONAL
PROGRESS IN NUCLEAR ENERGY
NUCLEAR SCIENCE AND ENGINEERING
FUSION TECHNOLOGY
FUSION ENGINEERING AND DESIGN
NUCLEAR FUSION
PLASMA PHYSICS AND CONTROLLED FUSION
JOURNAL OF FUSION ENERGY

TABLE 10 - DIFFERENCES BETWEEN THE JOURNAL QUERY AND PHRASE QUERY DATABASES

PHRASE	FREQUENCY	
	JOURNAL	QUERY
COAL	9451	1029
GAS	3865	3557
BIOGAS	220	0
FLUE GAS	284	145
OIL	2491	1040
FURNACE	521	301
BOILER	533	255
BIOMASS	743	1237
FIREWOOD	31	7
RICE HUSK	60	25
WIND	1060	571
GEOTHERMAL	187	108
HYDROPOWER	37	29
SOLAR	3249	1334
SOLAR COLLECTOR(S)	213	69
PHOTOVOLTAIC(S)	60	286
FUSION	106	381
PLASMA	92	540
TRITIUM	13	240
TOKAMAK	0	59
MAGNETIC ENERGY	9	402
MAGNETIC FIELD	39	301
MAGNETOHYDRODYNAMIC	10	32
SUPERCONDUCTIVITY	0	31
FISSION	34	98
URANIUM	33	176

TABLE 11 – ABSTRACT TAXONOMY – NON-STATISTICAL CLUSTERING

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
PRIMARY ENERGY SOURCES (23422)	FOSSIL FUELS (9509)	COAL (4753)	CONSTITUENTS/ CHARACTERISTICS/ PROPERTIES/ PRE-PROCESSING/ CLEANSING/ COMBUSTION
		OIL (3148)	CONSTITUENTS/ TYPES, BY-PRODUCTS, CONVERSION PROCESSES
		NATURAL GAS (1608)	TYPES, CLEANSING, BY-PRODUCTS
	RENEWABLE ENERGY/ ALTERNATIVE FUELS (12874)	SOLAR (4285)	CONVERSION SYSTEM CHARACTERISTICS/ COMPONENTS/ PROCESSES, APP
		HYDROGEN (3917)	MATERIALS/ COMPOUNDS, CONVERSION PROCESSES
		BIOMASS (2701)	SOURCES, TYPES, CONVERSION PROCESSES
		WIND (1063)	CONVERTER SYSTEMS, APPLICATIONS
		GEOTHERMAL (844)	SOURCES, APPLICATIONS
		HYDROPOWER (64)	ENVIRONMENTAL PROTECTION, APPLICATIONS
	NUCLEAR FUELS (1039)	FISSION (712)	COST, SAFETY, ENVIRONMENT, HEALTH
		FUSION (327)	IGNITION/ BURN, MAINTENANCE, COST/ SIZE REDUCTION
ENERGY CONVERTERS (17481)	THERMAL CONVERTERS (12514)	ENGINES (7543)	TYPES, COMPONENTS, CHARACTERISTICS, CONVERSION PROCESSES, CONVERSION BY-PRODUCTS, FUELS
		TURBINES (4971)	FUELS, TURBINE TYPES, CONVERSION CYCLE TYPES, CONVERSION PROCESSES
	DIRECT ELECTRIC CONVERTERS (4441)	FUEL CELLS (3154)	LONGEVITY, COMPONENT EFFICIENCY TYPES, FUELS, MATERIALS
		PHOTOVOLTAICS (1096)	EFFIC, COST, MATERIALS, FABRICAT, ELECTRO-OPTICAL PROPERTIES
		THERMOELECTRIC (106)	
		MHD (85)	
	NUCLEAR CONVERTERS (526)		
ENERGY STORAGE DEVICES (2901)	ELECTRICAL STORAGE (2774)	BATTERY (2400)	TYPES, COMPONENTS, MAT'LS, PROCESSES, PROPERTIES, CHARACT.
		CAPACITOR (334)	STRUCTURE, FABRICAT, MAT'LS, PROP, PHENOM, EXPERIMENT
		SMES (40)	COST REDUCTION, SYSTEMS STUDIES, TESTING
	MECHANICAL STORAGE (127)		

TABLE 12 – CLUSTER FORMATION STEPS

joining Cluster 1	Size 1	With Cluster 2	Size 2	Distance
Catalyst	1	Catalysts	1	28.519621
Coal	1	Coals	1	30.14870681
Cell	1	Cells	1	35.22156621
Combustion	1	Fuel	1	37.29106612
Catalyst	2	Conversion	1	38.23295
Oxidation	1	Oxygen	1	38.3740922
Combustion	2	Emissions	1	38.67338425
Electricity	1	Heat	1	38.918252
Cell	2	Batteries	1	39.28160721
Water	1	Hydrogen	1	39.57577802
Energy	1	Combustion	3	39.57768894
Gas	1	Electricity	2	39.72063788
Coal	2	Carbon	1	39.8118834
Gas	3	Water	2	39.92700574
Energy	4	Gas	5	40.03441637
Energy	9	Oxidation	2	40.12124741
Coal	3	Catalyst	3	40.14114663
Energy	11	Cell	3	40.28600632
Coal	6	Energy	14	40.34089979

TABLE 13 – FOUR CLUSTER TAXONOMY

Cluster #	Phrases
1	Coal
1	Carbon
1	Coals
2	Energy
2	Combustion
2	Fuel
2	Emissions
2	Gas
2	Water
2	Hydrogen
2	Electricity
2	Heat
2	Oxidation
2	Oxygen
3	Catalyst
3	Conversion
3	Catalysts
4	Cell
4	Cells
4	Batteries

TABLE 14 – TWO LEVEL TAXONOMY – QUERY-BASED DATABASE

Electrical Power Sources–Query			
Direct Conversion		**Thermal Conversion**	
Electromagnetic Storage and Conversion	**Electrochemical Storage and Conversion**	**Combustion Cycle**	**Systems and Thermodynamics**

TABLE 15 – TWO LEVEL TAXONOMY – JOURNAL-BASED DATABASE

Electrical Power Sources-Journal			
Lithium Batteries		**Fossil Fuels and Renewable Energy**	
		Fossil Fuels	**Renewable Energies**

TABLE 16 – THREE LEVEL TAXONOMY – COMBINED QUERY/ JOURNAL DATABASE

Electrical Power Sources-Query/Journal										
Energy Storage				**Power Sources and Converters**						
Science and Development		**Systems and Applications**		**Fossil Energy**			**Renewable / Long-term Energy**			
Micro	Macro			Sources	Emissions	Converters	Nuclear Sources	Non-Nuclear Sources	Direct Converters	Thermal Converters

TABLE 17 – FOUR LEVEL TAXONOMY – QUERY DATABASE

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
POWER GENERATION/ ENERGY STORAGE (4843)	FOSSIL REMEDIATION AND REPLACEMENT SYSTEMS (1443)	BIOMASS AND RENEWABLE GENERATION (1052)	WIND AND SOLAR GENERATION (297)
			BIOMASS GENERATION (755)
		CO2 EMISSIONS FROM FOSSIL GENERATION (391)	CO2 EMISSIONS FROM FOSSIL GENERATION (391)
	POWER PLANT HEATING AND STORAGE SYSTEMS (3400)	NUCLEAR POWER GENERATION (976)	NUCLEAR AND FUSION (479)
			STEAM TURBINE PLANT (497)
		HEATING AND ENERGY STORAGE (2424)	HEAT ENGINE STORAGE (996)
			POWER SYSTEM CONTROL AND BATTERY STORAGE (1428)
ENERGY CONVERSION (4527)	DIRECT CONVERSION (2117)	MAGNETIC FIELD CONVERSION (625)	MATERIAL MAGNETIC PROPERTIES (184)
			MAGNETIC FIELD STRUCTURES (441)
		ELECTROCHEMICAL AND PHOTOCHEMICAL CONVERSION (1492)	MATERIAL ELECTRICAL PROPERTIES (691)
			FUEL CELLS AND PHOTOVOLTAICS (801)
	THERMAL STEP CONVERSION (2410)	CATALYTIC COMBUSTION (1251)	CATALYTIC REACTIONS (690)
			COAL PARTICLE BED COMBUSTION (561)
		ENGINE DROPLET COMBUSTION (1159)	DROPLET COMBUSTION (680)
			DIESEL ENGINE COMBUSTION (479)

TABLE 18 – FOUR LEVEL TAXONOMY – JOURNAL DATABASE

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
FOSSIL REMEDIATION AND REPLACEMENT SYSTEMS, TURBINE CONVERSION (6294)	SOLAR THERMAL (2623)	HEATING AND COOLING MODELING (1633)	HEAT TRANSFER MODELING (1009)
			HEAT PUMP SYSTEMS (624)
		SOLAR COLLECTORS (990)	SOLAR COLLECTOR SYSTEMS (673)
			SOLAR RADIATION DATA (317)
	CO2 REMEDIATION AND OTHER LOW EMISSION REPLACEMENT SYSTEMS, TURBINE CONVERSION (3671)	POWER PLANT PRODUCTION, TURBINE CONVERSION, WIND, PHOTOVOLTAICS, GEOTHERMAL (2444)	ENERGY CONSUMPTION AND PRODUCTION (1036)
			WIND, TURBINE CONVERSION, PHOTOVOLTAICS, BIOMASS, AND GEOTHERMAL POWER (1408)
		FUEL CELLS AND CO2 EMISSIONS (1227)	CO2 EMISSIONS FROM VEHICLES (669)
			VEHICLE FUEL CELLS (558)
FOSSIL GENERATION AND STORAGE (5860)	BATTERIES (1890)	LITHIUM AND NICKEL (1419)	NICKEL BATTERIES (745)
			LITHIUM BATTERIES (674)
		LEAD-ACID BATTERIES (471)	LEAD-ACID BATTERIES (471)
	FOSSIL GENERATION (3970)	COAL (3048)	COAL EXTRACTION, LIQUEFACTION, GASIFICATION, PYROLYSIS (2325)
			FLUIDIZED BED CATALYSIS (723)
		OIL (922)	MULTIPLE OIL SOURCES (489)
			ASPHALTENE STRUCTURE AND PROPERTIES (433)