General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

NASA Automatic Subject Analysis Technique for Extracting Retrievable Multi-terms (NASA TERM) System

by

Jack Kirschbaum and Robert E. Williamson
Informatics Incorporated
P. O. Box 8756
Baltimore/Washington International Airport, Maryland 21240

Computer systems can be effectively divided into three parts: information entry, information structuring, and information retrieval. In a document surrogate retrieval system, each document requires most of the following during entry: acquisition; cataloging, abstracting; indexing; generation of a machine-readable record; validation and editing; and preparation of records for retrieval. Information structuring, if present, supports all users by reducing the cost of searching with the use of the following: data compressions; reorganization of data for efficient access; and/or maintenance of auxiliary files which assist retrieval and publications. For every query, retrieval involves repeated steps which include: familiarization with user's needs; entry of request to the computer; access of the data base; and display of document and/or surrogates to the user.

As noted, entry cost is incurred for every document and retrieval cost is incurred for each query. Costs are divided between human and machine with entry and retrieval being mainly human costs. Since computer costs are decreasing and personnel costs are increasing, the cost/effectiveness of work is shifting in favor of having the machine accomplish more. Structuring costs are mainly computer costs with human costs limited to programming and vocabulary maintenance. Some structuring techniques have resulted in substantial decreases in entry and retrieval costs.

Natural language techniques work quite satisfactorily for those systems which are just concerned with information retrieval. For those systems that additionally require printed indexes of subject terms or possible thesaurus controlled terms, natural language indexing alone cannot accomplish the task. Satisfactory print terms of the appropriate caliber and number can be generated in a primarily automatic manner.

The NASA Scientific and Technical Information Facility is actively investigating methods of selecting multi-word index terms (phrases) for widely distributed abstract journals which will require minimal human effort to index each abstract while taking maximum advantage of computer index selections. Several experiments are planned to investigate the adequacy defined below relative to the present method that is based on professional indexers. The present computer algorithm generates index phrases from each abstract by: deletion of mundane words, look up of remaining word strings in the NASA Thesaurus (controlled) to determine the closest entry phrase, ranking the corresponding "use" phrase by the stored human assessment of the phrases usefulness, "boosting" that assessment by the assessment of all lower ranked phrases with word stems in common with a given phrase, reranking, and finally, the selection of the appropriate number of important phrases. Should there be too few adequately important terms, or too many of similar importance, the surrogate can be marked for manual consideration, resulting in modification of the thesaurus or of value which govern the print term selection. Lexicographic control of the thesaurus will maintain the system with respect to new terminology as well as typographic anomolies.

(NASA-CR-157398) NASA AUTOMATIC SUBJECT ANALYSIS TECHNIQUE FOR EXTRACTING RETRIEVABLE MULTI-TERMS (NASA TERM) SYSTEM (Informatics Information Systems Co.) 27 p HC A03/MF A01 CSCL 05B G3/82

N78-30992
Unclas
29868

NASA Automatic Subject Analysis Technique for Extracting Retrievable Multi-terms (NASA TERM) System

by

Jack Kirschbaum and Robert E. Williamson
Informatics Incorporated
P. O. Box 8756
Baltimore/Washington International Airport, Maryland 21240

1. Introduction

The NASA Automatic Subject Analysis Technique for Extracting Retrievable Multi-terms (NASA TERM) System is a computer coordinated indexing approach for publications and document surrogate retrieval systems from natural language abstracts. The NASA TERM System is in its embryonic state of development having recently completed its feasibility study phase. This paper will present some background information, the description of the planned NASA TERM System, and the itemization of research and development experiments.

2. Background

2.1 Parts of A Typical Document Retrieval System

There are two main types of information retrieval systems: data retrieval systems and document retrieval systems. The response of a data retrieval system has the pattern "X is Y"; the response of a document retrieval system has the pattern "the item Z discusses topic X". For example, a data retrieval system is built to answer questions like: "What is the melting point of titanium steel?"; whereas a document retrieval system is built to answer questions like: "Which documents discuss the melting points of titanium steels?" This difference is quite significant in that a data retrieval system must understand a user's requirement at a much deeper level and respond with facts; a document retrieval system need only supply the bibliographic reference (accession number) of pertinent documents.

Most data retrieval systems are based on sophisticated data bases loaded and searched by experts using highly structured queries. Document retrieval systems need not be so sophisticated; in fact, they may perform better if not overly complex. One pattern detected throughout an analysis of ongoing document retrieval systems is that sophisticated techniques help some users and hurt others; the problem is knowing which users will be helped and which hurt.

Most computer systems, including document retrieval systems, can be effectively divided into three parts, as shown in Table 1-1: information entry, information structuring, and information retrieval.

2.1.1 Information Entry

In a document retrieval system, each new document requires the performance

Phases Found in Many Document Retrieval Systems

ENTRY For each document individually	STRUCTURING For document or queries in general	RETRIEVAL For each query individual
Acquisition	Analysis of new terms and phrases	Conversion of entry text to internal form
Cataloging	Creation/extension of inverted file	Display of terms related terms in original query
Abstract review/preparation	Creation/extension of cluster Centroids	Display of search statistics; e.g., number of documents isolated
Selection of Print Terms		Display of citation and abstracts
Selection of Online Terms		Comparison of queries with centroids or documents
Entry of information to Computer		Concatenation and intersection of inversion lists- one operation per significant term in expanded query
Verification of entered data		
Conversion to storage format		

Table 1-1

2.1.1 Information Entry (cont'd)

of most of the following component steps during data entry: acquisition; cataloging; abstracting; indexing; entry of cataloging, abstracting and indexing information to the computer; computer verification of the adequacy of data about the document; conversion of entered data from entry standard "codes", etc. to retrieval or display standards; and, finally, the preparation of the document record for retrieval.

2.1.2 Information Structuring

Information structuring includes those steps applied to surrogates collectively. In some systems, there is no structuring at all; the data records produced by the information entry steps are read one after the other and "compared" to a user's request. Such a technique is exhorbitantly expensive if there are very many documents or queries. Information structuring is a "front money" cost that increases the base cost of a system. This investment is made in order to substantially decrease the cost of data searching and/or information entry; i.e., to decrease costs associated with every query and/or every document.

Typical steps in information structuring may include: (1) data compression such as the conversion of character strings to numbers to decrease the

amount of storage required; (2) reorganization of data from entry order to an order supportive of faster or less expensive searching; (3) generation and maintenance of auxiliary files which assist retrieval; and (4) publication of tools to be used for many searches.

2.1.3 Information Retrieval

Retrieval involves all steps accomplished for each individual query. These steps include expert familiarization with user's needs, entry of a request to the computer, access of the data base to locate names of documents which will, hopefully, be pertinent, and display of document names (and often surrogates) to the document retrieval specialist or to the user. The process is repeated for each query.

2.1.4 Economic Tradeoffs

As noted above, the cost of data entry is incurred for every new document; the cost of retrieval is incurred for every query. No structuring is required to support document retrieval; however, some structuring techniques have resulted in very substantial decreases in entry and/or retrieval costs. The primary management problem is to decide if an investment in structuring because of the expected payoff in reduced entry and retrieval costs is justified.

Costs can be divided into two types: machine and human. Structuring costs are mainly computer costs with human costs limited to such items as programming and to the addition of new terms and phrases to a thesaurus. Entry and retrieval costs are mainly human costs (with some machine costs in support). These costs are incurred for every document or individual query. The fact that computer costs are decreasing about 20% per year and personnel costs are rising about 20% per year means that the cost/effectiveness of work performed by the computer relative to work performed by the person is constantly shifting in the favor of having the computer do more and more.

The question of the effectiveness of computer processing versus human processing on a strict performance level also needs careful consideration. There is a very human tendency to presume that a computer is unable to perform certain tasks "which everyone knows only humans have the ability to perform." All claims that humans can perform given tasks better or more cost-effectively than computers, or vice-versa, should be considered suspect without specific scientific evidence. In general, one can say that most claims for humans being better than computers at document retrieval tasks are simply unproven; the reverse is also frequently the case for certain conditions. Informed decisions in specific instances appear to require explicit measurement of cost and effectiveness of all options; there are few, if any, "well accepted principles" in the area of document retrieval.

2.2 Current NASA Scientific and Technical Information Facility System

2.2.1 Missions

The Facility has the following primary missions:

(a) Acquisition and processing of the world's aerospace-related report literature.
(b) Processing of the world's aerospace-related open literature acquired from the American Institute of Aeronautics and Astronautics (AIAA).
(c) Announcement of current or selected literature to the aerospace communi via indexed abstract journals (i.e., Scientific and Technical Aerospace Reports -- commonly called STAR) and also via selective dissemination of information (SDI) techniques (i.e., Selected Current Aerospace Notices -- commonly called SCAN).
(d) Providing initial distribution automatically of full documents on microfiche available in the collection to qualified recipients and secondary distribution of the microfiche or blowbacks of the documents to others upon authorized requests.
(e) Providing bibliographic research services to qualified individuals or organizations who either request these services through the Facility or have online interactive access through terminals using NASA's REmote CONsole (RECON) System. RECON is a derivative of Lockheed's DIALOG System.
(f) Providing retrospective search services and bibliographies resulting therefrom, to qualified recipients.
(g) Provide library support services and products for NASA and NASA affiliated libraries for cost effective functions such as: online research through RECON; interlibrary network loan of books and periodicals; and preparation of catalog cards, book catalogs, shelf lists, acquisition lists, etc.
(h) Providing generalized support services for the dissemination of aeronautical and aerospace information to the public at large through technology utilization programs.

2.2.2 Indexing

Other than those activities pertaining to the generation and distribution of microfiche and miscellaneous support services, all of these missions are associated in one way or another with the development or use of indexes. From the time the NASA Facility began operation early in 1962 until the end of 1967, an indexing philosophy closely related to that of the "Uniterm" system was employed. Adjectival word-forms were permitted as index terms. The indexing was at first essentially "free" of any constraints; although in later years, more and more dependence upon a published guide to subject indexes was required.

In 1966, NASA determined that a change in philosophy was in order, so that system performance could be improved. NASA elected to prepare a thesaurus of aerospace terminology to be used as a vocabulary control authority. This NASA Thesaurus (NI) was prepared during the latter part of 1966 and 1967. As a base for this vocabulary, NASA adopted terminological conventions previously developed by the Engineers Joint Council and the Department of Defense.

The NASA Thesaurus was first published in December, 1967. Beginning with the accessions scheduled for announcement with the first issues of the 1968 volumes of the Scientific and Technical Aerospace Reports (STAR) and

International Aerospace Abstracts (IAA), the NASA Thesaurus was used as an indexing vocabulary control tool and has been used as such until the present time. The NASA Thesaurus is a dynamic publication, in that it is updated periodically, with the latest publication occurring in 1976. At that time, there were 15,060 postable terms and 3,343 nonpostable terms; however, with pseudoterms and other entry terms which provide multiple access to the NASA Thesaurus, there were 35,801 entry points.

2.2.3 Machine-readable Abstracts

In 1971, another major change in processing techniques was employed at the NASA Facility with the advent of the implementation of an online input capability. This change was accomplished in conjunction with a change in hardware capability from one that supports the IBM 1410 (in emulation) to an IBM 360 Operating System mode, which is still in place. With the advent of this processing change, natural language abstracts in machine-readable form were saved and have subsequently been a part of the NASA Facility data bases. Abstracts are now routinely incorporated into the data bases.

Abstracts for which NASA Thesaurus indexing is accomplished is available online as follows:

1. Scientific and Technical Aerospace Reports (STAR), November 1971 to the present time.
2. International Aerospace Abstracts (IAA), January 1972 to the present time.
3. Computer Program Abstracts (CPA), January 1973 to the present time.
4. Classified Scientific and Technical Aerospace Reports (CSTAR) (however, no classified data is available in the data base), January 1975 to the present time.

2.2.4 Searching

The abstracts which are available are all text searchable with the NASA RECON system, and the titles for all citations in these four files are also text searchable. Texting searching in the NASA system is based on a hierarchy of document records, sentences within the records and words within the sentences.

Searching in the NASA system can also be accomplished through the use of Personal Authors, Corporate Authors, NASA Thesaurus Subject Terms, Report Numbers, Contract Numbers, etc., as entry points. The NASA Thesaurus Subject Terms are generated by human indexers who determine six to twelve of the important ideas and concepts in a document and select the most descriptive terms from the NASA Thesaurus.

Major subject terms, selected to reflect the major concepts and research areas, are printed in the abstract journal in a subject index. To the user of the index, the published terms in combination with the title (or added title information, called a Title Extension) should permit a quick review of available material and assist in determining if a document is of further interest. Minor concepts are indexed for document retrieval only.

3. Central Algorithm

3.1 A Procedure for Selecting Print Terms from an Abstract

The primary problem for NASA, in moving toward a completely natural language abstract based system, is the need to prepare printed indexes. (There is no reason to presume that such a shift will do anything negative, and is very likely to do many things positive, for on-line retrieval). A consensus of expert technical opinion is that satisfactory print terms can be generated in a primarily automatic manner.

This section of this paper describes a suggested technique, NASA TERMS, which generates print terms in two parts. First, the idea of thesaurus is presented. Then a procedure is presented whereby an abstract can be processed "against" a thesaurus of the type described to produce a set of print terms.

3.2 A Print Term Generation Thesaurus

For maximal cost-effectiveness, on-line searchers require a thesaurus for the same reason indexers do -- human memory is fallible. Stored lists of related terms help the searcher specify synonyms and more accurately describe the subject in which he is interested. Since the printed index user does not have an on-line thesaurus to aid him, the producers of the index must map synonyms to a single posting term (and enable the user to determine the posting term). The computer is also capable of cost-effectively utilizing as many descriptive terms for a document as can be provided; printed indices are limited by economic considerations to 4 to 8 postings per document.

The proposed Print Term Selection Thesaurus is essentially the present NASA Thesaurus extended to include all of the "mappings" presently applied by the indexers "automatically," and extended to specify the "worth" of each posting term in the thesaurus, and also extended to indicate the areas of use of the Thesaurus Entry Phase (TEP).

The additional mappings required are what is normally known as entry postings, "see" postings or "use" postings. What is desired is to codify the rules used by indexers for selecting specific print terms. Initially, the thesaurus will need to be "educated." This can be done in several ways. The two most plausible are to add one or more existing thesauri which represent vocabularies that are generally available that contain special features worth having. Another is to process existing abstracts for which manually selected print terms are available. While the goal for the computer system is not to generate exactly the print terms manually specified, it is reasonable to expect the computer to generate most such print terms, a few others, and terms more specific or more general than the manual choice. Where differences exist, lexicographers, or the original indexers, are very likely to be able to specify a "path" by which the computer could determine that a given manually specified print term is useful -- if such a choice is indeed warranted.

The worth of a print term is a tool to enable the computer to determine which

of several terms in a hierarchy is desirable (as described in the next section). For example, print terms which occur frequently in STAR are likely to be of less use than more specific terms. In any hierarchy, if several terms of varying breadth are plausible, only the most specific is normally selected as the print term. Similarly, if several "brother" terms are indicated, only the general term which includes all of them is usually chosen as a print term.

A similar problem occurs with ambiguous terms, e.g., sizing (shaping) and sizing (surface treatment). In order to automatically disambiguate such terms, the thesaurus needs to include rules for selecting one form of a term rather than the other. Non-ambiguous terms have known usage patterns. The area specified by the majority of the non-ambiguous terms can be used to select the proper posting for an ambiguous term.

3.3 Types of Words

Although one normally thinks of a thesaurus as a place to find the proper posting term for a given entry term, the NASA TERMS thesaurus also contains indications for two classes of words. In general, words can be divided into three types: (i) those generally useful for retrieval; (ii) those occasionally useful directly or useful for determining proper retrieval phrases; and (iii) those never useful for retrieval.

Words never useful for retrieval are functor words such as articles, most prepositions, conjunctions, and all forms of the verbs "to be," "to have," etc. These words typically separate the content phrases that are useful for retrieval. There are not too many of these words; most systems declare under 200 words to be of this type. Most of these words are also very frequently used, consisting of upwards of 50% of running text. Thus, it is useful to store these words in the main memory of the computer when processing text.

Words occasionally useful are of three subtypes: functors, ambiguous, and qualifying. Functors like "and" and "of" are of no retrieval use in themselves. However, they must be considered correctly in order to properly handle contextually adjacent useful terms.

Occurrences of "and" will be deleted, obviously. Prior to deletion, however, surrounding phrases must be fully expanded. That is, constructs of the form "adjective noun and noun" must be rephrased as "adjective noun and adjective noun." (It is worth mentioning that this rephrasing will occasionally result in an error, i.e., the adjective does not have to modify the second noun.) Experience has shown that rephrasing leads to fewer errors.

Occurrences of "of" require special consideration so that phrases like "retrieval of information" and "angle of attack" are recognized as the same as the phrases "information retrieval" and "attack angle."

Certain words are ambiguous with one usage in the functor category and another usage useful for retrieval. The word "basic," as in the phrase "The basic use of ..." is ignorable. The chemical concept "basic" is not ignorable.

The fact that a word is ambiguous in this way is a manual decision. "Basic" can be retained if a significant number of the other words in nearby sentences have a chemical usage. This obviously leads to an error when "basic" is used in the general sense in text in a chemical area; e.g., consider the following sentence: The basic way to neutralize organic acids is with organic bases. (Such anomolies are easy to construct but rarely occur in practice.)

Qualifying words are not ambiguous in that they are used in a single sense. However, "high energy physics" is a phrase in which "high" must be retained; in most phrases, such as "in a high tail," "high" is not needed.

Words of this middle class must be kept and used when appropriate but can be ignored when not in one of their special situations. Generally, such words should not be used to divide phrases, nor should they be used to try to find phrases. For example, it is more efficient to look up "energ" (the stem of "energy") to find a "high energy physics" than to process all words following all occurrences of "high" for strings such as "energy physics."

Words not in the two preceding classes should be mapable into a valid thesaurus entry phrase. Any word not in the thesaurus is automatically listed for manual review. In addition, any phrase containing words which are present in other phrases but not in any phrase found in the thesaurus should be posted to a "strange usage" file for manual review.

Proximity requirements could also be used in the thesaurus, e.g., if "variable" or "changeable" occurs in the same sentence with the term "sweep" and the term "wings," post the term "variable sweep wings."

Use of stems can drastically shrink the size of a thesaurus; this is important more for human understanding of the thesaurus than for the reduction in computer storage possible. Normally, a stem can be given with no restriction on the possible suffixes, occasionally it may prove necessary to require one of standard sets of suffixes or even one of a specific list of suffixes. As an example, most any occurrence of some form of "varying," e.g., "variable," "varys," in the same context with "sweep wing" or "sweep wings." (Note that improper phrases can be constructed, e.g., "sweep wings cause varying turbulance patterns as speed increases." Such constructs are, fortunately, rare in actual text. How rare in NASA abstracts should be experimentally determined.)

3.4 The Basic NASA TERMS Algorithm

Table 3-1 presents the steps included in the NASA TERMS algorithm. Many are steps required in the present manual indexing system and will be retained essentially unchanged, in an automated term selection system. The following paragraphs highlight areas of the algorithm which may be unclear from the terse phraseology of the table. (It should be clearly understood that the table and what follows is not considered to be a systems design; rather it is just an outline. A detailed specification of the algorithm is a logical next step.)

NASA Automatic Subject Analysis Technique for Extracting Retrievable Multi-terms
(NASA TERM System)
A System Which Generates Print Terms from Natural Text

Table 3-1

1. Catalog document (as for normal system).

2. Keyboard document surrogate, i.e., title, abstract (if present), and other important descriptive text, e.g., section headings, figures and table titles -- between 200 and 500 words.

3. Isolate words in surrogate; convert variable length words into a fixed length (32 bit) number for the stem and a fixed length (8 bit) number for its suffix.

4. Based on a core-resident table, identify high-frequency, non-useful functor words.

5. Split surrogate into phrases based on sentence boundaries and high-frequency, unambiguous functor words, i.e., articles, prepositions, pronouns, and conjunctions.

6. Look up the first word of every phrase in thesaurus yielding part of speech for word and list of all entry phrases in the thesaurus which begin or include given word. (This step frequently results in the replacement of a word by the stem of that word.) Continue as needed to locate list treatment of all non-ignorable words. Selection of the best phrase for a given surrogate word may require contextual judgements.

7. The data from Step 6 is used to specify the majority of candidate print phrases (CPP) for the surrogate. Note that the conjunctions located in Step 4 must be considered to resolve typical adjective - "and" - adjective noun phrases, etc.

8. CPPs are ranked by the manually assigned weight obtained in Step 6 from the thesaurus. Typical weights are 50 for very high quality (specific) terms, 20 for average terms, 9 for single word terms and 4 for general single word terms. Terms found in titles are doubled in weight.

9. CPPs which occur more than once are replaced by a single CPP with a weight equal to the sum of the individual weights.

10. The weight of each CPP is boosted by the proportionate weight of any other CPP which contains any word stem(s) in common with the given ranking CPP.

11. The CPPs are ranked and the top four CPPs are used as print terms. The fifth CPP is also used if the difference in weight of that CPP and the next CPP exceeds the difference in weight of that CPP and the preceding CPP. Similarly for subsequent CPPs. No CPP is used if its weight is less than 30. and any CPP with a weight over 49 will be used. An attempt will be made to limit the number of CPPs to six. If two selected terms have a broader-narrower relationship in a thesaurus hierarchy, only the narrower is used as a print term unless the weight of the broader term exceeds 150% of the weight of the narrower term.

12. The document surrogate is added to the online file, be it clustered or inverted.

If the abstract is not long enough for satisfactory processing, say 200 words, the leading and/or terminating paragraphs of the document can be entered. It would probably be worth while to enter all section headings, illustration and table titles, etc., as these items would increase recall for many queries; nevertheless, the number of words entered is not excessive -- in respect to the payoff -- and the words can be quickly selected by clerical activity.

Common misspellings could be entered into the thesaurus and automatically corrected. Normally an error and activity report of automatically applied corrections would be prepared for a lexicographer to confirm that any given "spelling correction" is proper and not some unanticipated valid word or an unanticipated spelling error for which the wrong correction was supplied. Additionally, spelling, correction algorithms could be employed.

Through the use of the number corresponding to a given stem, the set of phrases in which each word can appear is obtained from the thesaurus. The computer will determine any phrases present in the document which correspond to requirements specified in the thesaurus. For each such phrase, the thesaurus print term which corresponds is added to the document record. Words of some importance for which no satisfactory entry phrase can be determined are included in an "exceptions report" for manual processing.

It is important to realize the necessity of including some manual processing. Words which are new must be added to the thesaurus. Words which have occurred in phrases not anticipated in the thesaurus should be looked at again. Unusual documents should be examined. All of these actions should result in improvements in the system. Such improvements (i.e., additions) will be very frequent when the system is young. The number of improvements will decrease with time but never to the zero level. It is anticipated that new terminology and typographical problems (not all of which are simple keyboard errors) will keep several lexicographers and indexers busy full-time at the Facility. The main intent of automatic processing is to record every manual decision when that decision is made so that there is rarely the need to make the same decision repeatedly.

In order to prove that such a process can generate suitable print terms, several experiments are recommended. As each is "passed," it will be reasonable to incur the cost of the next. It would be unwise to invest time and money in advanced, complex experiments until all parties are convinced that the result is likely to be an acceptable indication of usefulness or non-usefulness of the technique under test.

Phrases which contain word roots that appear in other phrases are additionally boosted by the proportionate weight of those other roots. This technique will highlight long, and thus highly specific, phrases that are not likely to be repeated in an abstract. Where two phrases, typically single word phrases, differ only in suffix, the more frequently occurring form is used.

3.5 Thesaurus Entry Phrases Have Normal "Areas of Use"

All NASA abstracts are published in one of 75 categories. The NASA Thesaurus

contains about 2,000 distinct hierarchies and most entry phrases are present in a (single) hierarchy. These two ideas can be used to define an "area of use" for each entry phrase. The proper category for a phrase can be determined by "looking" at the category used for abstracts for which that phrase is a descriptor. Naturally this process is a one-time batch run which would augment the record for each phrase with the number of times that phrase was used in each area. Most phrases are used exclusively within three or four categories and any phrase which was used to describe only one or two abstracts in a given category can be dropped from that category. With these "areas of use" known, ambiguous entry phrases can be properly classified (based on the consensus of the "areas of use" of unambiguous phrases). This technique might also allow selection of a more specific phrase such as "wing root", rather than the general term "root" when a surrogate uses only the term "root" at some point.

The mapping of "root torsion" into "wing roots" (and "torsional stress") in the first example (Exhibit A), will strike many as far fetched for an automatic system. However, the mapping is straight-forward when one considers what should be done with the word "root." In the thesaurus, it appears as: root mean square error, roots, roots of equations, plant roots, and wing roots. If one assigns entries to general categories, i.e., "most likely to be used in area...," then the above entries, except for the ambiguous "roots," have definite areas. The use in this sentence is from the same subject area as "wing roots"; hence, "wing roots" is the proper index point.

3.6 Examples of NASA TERMS for Automatically Generating Print Terms

This section shows the way a computer could generate print terms for several abstracts. The process is manually done, i.e., it has not yet been programmed. However, every attempt has been made to be fair; no decision has been made that is not believed to be easily programmed. The computer selected print terms were obtained without knowledge of the NASA expert's choices which are included for comparison. Note please, duplication of the expert's choices is not a goal; an equally useful set for STAR is the goal.

Four simulations are presented. The first two are doctoral dissertation abstracts from STAR. The third is an author written report abstract from STAR. The fourth is an AIAA prepared abstract from IAA. Each simulation is presented as one exhibit in several parts. The first part shows the surrogate as published. The second part shows the RECON display for the surrogate. The third part of each exhibit shows the words in each title and abstract which are totally ignorable (underlined) or ignorable except when used in entry phrases. These words and sentence ends separate surrogate entry phrases (SEPs). The fourth part of each exhibit shows (in varying degrees) how each SEP is processed. At a minimum, each candidate print phrase (CPP) selected is shown with the weight assigned that SEP. In many cases, the reasoning behind the selection is indicated. Any SEPs the system cannot handle well are noted and would appear on a list of such SEPs for manual review. The fifth part of each exhibit lists the CPPs in weight order showing all components of the total weight and showing which CPPs are actually selected as print terms by the NASA TERMS algorithm. The sixth part of each exhibit lists the manually chosen print and non-print terms. Terms selected or identified by the NASA TERMS and those produced manually are

so identified. Other terms are commented upon. Please note again that duplication of manual choices is not a goal -- only the selection of terms equally (or more) useful for retrieval is a reasonable goal.

N77-29090 Georgia Inst. of Tech., Atlanta
A METHOD OF COMPUTING THE POTENTIAL FLOW ON THICK WING TIPS Ph.D. Thesis
Pradeep Rao 1976 174 p
Avail: Univ Microfilms Order No. 77-7352

An iterative procedure to compute detailed velocity and pressure distributions on the surface of thick wing tips is developed using potential flow theory. The method uses a two dimensional surface vorticity distribution as an initial approximation. Therefore, the two dimensional problem is first formulated in the form of an integral equation using vorticity as the surface singularity which is solved by the elementary vortex distribution technique. A comparison of the flow computed on a circular cylinder with the exact analytical results provides a measure of accuracy. The two dimensional noncirculatory and circulatory flow is computed for NACA basic thickness form airfoils. Dissert. Abstr.

First Surrogate as Published (STAR)

Exhibit A-1

77N29090 ISSUE 20 PAGE 2621 CATEGORY 2 76/00/00 174 PAGES
UNCLASSIFIED DOCUMENT
A METHOD OF COMPUTING THE POTENTIAL FLOW ON THICK WING TIPS PH.D. THESIS
A/RAO, P.
GEORGIA INST. OF TECH., ATLANTA. AVAIL UNIV. MICROFILMS ORDER NO. 77-7352
/*POTENTIAL FLOW/*PRESSURE DISTRIBUTION/*VELOCITY MEASUREMENT/*WING TIPS/ AIRFOILS/ FLOW THEORY/ TWO DIMENSIONAL FLOW/ VORTICITY
NDA DISSERT. ABSTR.
ABS AN ITERATIVE PROCEDURE TO COMPUTE DETAILED VELOCITY AND PRESSURE DISTRIBUTIONS ON THE SURFACE OF THICK WING TIPS IS DEVELOPED USING POTENTIAL FLOW THEORY. THE METHOD USES A TWO DIMENSIONAL SURFACE VORTICITY DISTRIBUTION AS AN INITIAL APPROXIMATION. THEREFORE, THE TWO DIMENSIONAL PROBLEM IS FIRST FORMULATED IN THE FORM OF AN INTEGRAL EQUATION USING VORTICITY AS THE SURFACE SINGULARITY WHICH IS SOLVED BY THE ELEMENTARY VORTEX DISTRIBUTION TECHNIQUE. A COMPARISON OF THE FLOW COMPUTED ON A CIRCULAR CYLINDER WITH THE EXACT ANALYTICAL RESULTS PROVIDES A MEASURE OF ACCURACY. THE TWO DIMENSIONAL NONCIRCULATORY AND CIRCULATORY FLOW IS COMPUTED FOR NACA BASIC THICKNESS FORM AIRFOILS.

RECON Display of First Surrogate

Exhibit A-2

A METHOD OF COMPUTING THE POTENTIAL FLOW ON THICK WING TIPS

An interative procedure to compute detailed velocity and pressure distributions on the surface of thick wing tips is developed using potential flow theory. The method uses a two dimensional surface vorticity distribution as an initial approximation. Therefore, the two dimensional problem is first formulated in the form of an integral equation using vorticity as the surface singularity which is solved by the elementary vortex distribution technique. A comparison of the flow computed on a circular cylinder with the exact analytical results provides a measure of accuracy. The two dimensional noncirculatory and circulatory flow is computed for NASA basic thickness form airfoils.

Dissert. Abstr.

Phrase Isolation
Exhibit A-3

Text Phrase Reduction
Exhibit A-4

computing ⟶ computation, weight 9
potential flow ⟶ potential flow, weight 20
wing tips ⟶ wing tips, weight 20
iterative procedure ⟶ iterative procedures, weight 20
compute ⟶ computation, weight 9
velocity distributions ⟶ velocity distribution, weight 20
pressure distributions ⟶ pressure distribution, weight 20
wing tips ⟶ wing tips, weight 20
potential flow theory ⟶ potential flow, weight 20
two dimensional surface vorticity distribution ⟶ two dimensional boundary layer flow, weight 50, & vorticity, weight 9
initial approximation ⟶ initial approximations, weight 20
two dimensional problem ⟶ two dimensional, weight 9
integral equation ⟶ integral equations, weight 20
vorticity ⟶ vorticity, weight 9
surface singularity ⟶ surface properties, weight 9; singularity (mathematics), weight 9
elementary vortex distribution technique ⟶ vorticity, weight 20
flow computed ⟶ flow computation, weight 20
ci ular cylinder ⟶ circular cylinder, weight 20
analytical results ⟶ (phrase not found, posted for manual review)
two dimensional noncirculatory and circulatory flow ⟶ two dimensional flow, weight 20 & noncirculatory flow, weight 20, & circulatory flow, weight 20
NACA basic thickness form airfoils ⟶ NACA, weight 20, & thickness form airfoils ⟶ airfoil profiles, weight 20

NASA TERMS Selections
Exhibit A-5

Print Terms	Occurrence Weights	Sum	Co-Occurring Boost	Total
Two dimensional boundary layer flow	50	50	53	103
Potential flow	2*20,20	60	33	93
Two dimensional flow	20	20	51	71
Flow computation	20		60	80
*Circulatory Flow	20		46	66
*Noncirculatory flow	20		46	66
Non-Print Terms				
Wing tips	2*20,20	60		60
Vorticity (including vortices)	9,9,20	38		38
Computation	2*9,9	27	7	34
Two dimensional	9		22	31
Velocity distributions	20		10	30
Pressure distribution	20		10	30
Integral Equations	20			20
Air foil profiles	20			20
Initial approximations	20			20
Iterative procedures	20			20
Circular cylinder	20			20
NACA	20			20
Surface properties	9			9
Singularity (mathematics)	9			9

*Dropped as print terms despite weight due to "broader/narrower" relationships.

Manually Selected Terms
Exhibit A-6

Print Terms	Comments
Potential Flow	Common selection
Pressure Distributions	Usefulness as a print term is questionable
Velocity Measurements	Incorrect term - should be "Velocity Distribu
Wing Tips	Common selection
Search Terms	
Airfoils	Important and underrated b both techniques
Flow Theory	Too important not to be used as a print term
Two Dimensional Flow	
Vorticity	
Computer selected print terms not above	
Two dimensional boundary layer flow	Requires recognition of parallel way of describing.

N77-29089 California Univ., Los Angeles
THE COUPLED FLAP-LAG-TORSIONAL AEROELASTIC STABILITY OF HELICOPTER ROTOR BLADES IN FORWARD FLIGHT Ph.D. Thesis
Manuel Reyna-Allende 1976 295 p
Avail. Univ. Microfilms Order No 77-8530

A set of coupled flap lag torsional equations of motion capable of simulating general hingeless rotor blade configurations are derived for the case of a rotor blade having moderate deflections. The final equations of motion are represented by a system of coupled, nonlinear partial differential equations. The equations are capable of simulating rotor blades having: (1) precone; (2) droop; (3) built in twist; (4) distributed torsion; (5) root torsion (or pitch link flexibility); (6) blade root offsets; (7) and offsets between the elastic axis, aerodynamic center and the blade cross sectional center of mass. Quasisteady aerodynamic loads are used and the effects of stall and compressibility are neglected. Reversed flow is included in the representation of the airloads.
Dissert. Abstr.

Second Surrogate as Printed (STAR)
Exhibit B-1

77N29089 ISSUE 20 PAGE 2620 CATEGORY 2 76/00/00 295 PAGES
UNCLASSIFIED DOCUMENT
THE COUPLED FLAP-LAG-TORSIONAL AEROELASTIC STABILITY OF HELICOPTER ROTOR BLADES IN FORWARD FLIGHT
PH.D. THESIS
A/REYNA-ALLENDE, M.
CALIFORNIA UNIV., LOS ANGELES. AVAIL UNIV. MICROFILMS ORDER NO. 77-8530
/*AERODYNAMIC STABILITY/*EQUATIONS OF MOTION/*HELICOPTERS/*ROTOR BLADES/ AERODYNAMIC LOADS/ AIRCRAFT CONFIGURATIONS/ NONLINEARITY/ PARTIAL DIFFERENTIAL EQUATIONS
ABA DISSERT. ABSTR.
ABS A SET OF COUPLED FLAP LAG TORSIONAL EQUATIONS OF MOTION CAPABLE OF SIMULATING GENERAL HINGELESS ROTOR BLADE CONFIGURATIONS ARE DERIVED FOR THE CASE OF A ROTOR BLADE HAVING MODERATE DEFLECTIONS. THE FINAL EQUATIONS OF MOTION ARE REPRESENTED BY A SYSTEM OF COUPLED, NONLINEAR PARTIAL DIFFERENTIAL EQUATIONS. THE EQUATIONS ARE CAPABLE OF SIMULATING ROTOR BLADES HAVING (1) PRECONE; (2) DROOP; (3) BUILT IN TWIST; (4) DISTRIBUTED TORSION; (5) ROOT TORSION (OR PITCH LINK FLEXIBILITY); (6) BLADE ROOT OFFSETS; (7) AND OFFSETS BETWEEN THE ELASTIC AXIS, AERODYNAMIC CENTER AND THE BLADE CROSS SECTIONAL CENTER OF MASS. QUASISTEADY AERODYNAMIC LOADS ARE USED AND THE EFFECTS OF STALL AND COMPRESSIBILITY ARE NEGLECTED. REVERSED FLOW IS INCLUDED IN THE REPRESENTATION OF THE AIRLOADS.

RECON Display of Second Surrogate
Exhibit B-2

THE COUPLED FLAP-LAG TORSIONAL AEROELASTIC STABILITY OF HELICOPTER ROTOR BLADES IN FORWARD FLIGHT

A set of coupled flap lag torsional equations of motion capable of simulating general hingeless rotor blade configurations are derived for the case of a rotor blade having moderate deflections. The final equations of motion are represented by a system of coupled, nonlinear partial differential equations. The equations are capable of simulating rotor blades having: (1) precone; (2) droop; (3) built in twist; (4) distributed torsion; (5) root torsion (or pitch link flexibility); (6) blade root offsets; (7) and offsets between the elastic axis, aerodynamic center and the blade cross sectional center of mass. Quasisteady aerodynamic loads are used and the effects of stall and compressibility are neglected. Reversed flow is included in the representation of the airloads.

Phrase Isolation
Exhibit B-3

Text Phrase Reduction
Exhibit B-4

Coupled flap lag torsional aeroelastic stability ⟶
  (coupled ignored, posted for manual review)
  flaps (control surfaces), weight 20 (lag not in thesaurus and ignored, posted for manual review)
  torsional stress, weight 20
  aeroelasticity, weight 9
  stability, weight 9
Helicopter rotor blades ⟶ rotary wings, weight 20
Forward Flight ⟶ Flight (Forward dropped as unimportant), weight 9
coupled flap lag torsional equations of motion ⟶ coupled flap lag torsional - see above equations of motion, weight 20
simulating ⟶ hingeless rotor blade
  simulation (general dropped), weight 20
  rigid rotor, weight 20
  rotary wings (configuration dropped), weight 20
rotor blade ⟶ rotary wings, weight 20
deflections ⟶ deflection, weight 9
equations of motion ⟶ equations of motion, weight 20
coupled, nonlinear partial differential equations ⟶ nonlinear equations, weight 20
  partial differential equations, weight 20
equations ⟶ equations, weight 9

Text Phrase Reduction
Exhibit B-4
(continued)

simulating rotor blades ——→ simulation, weight 20, rotary wings, weight 20 (precone, droop and built in twist are all lost since not in Thesaurus), posted for manual review
distribution torsion ——→ torsional stress (distribution dropped), weight 20
root torsion ——→ wing roots, weight 20; torsional stress, weight 20
(pitch link flexibility) ——→ (too strong to ignore, human assistance requested)
pitch attitude control ——→ longitudinal control, weight 20
Blade root offsets ——→ rotary wings, weight 20, & wing roots (offsets dropped), weight 20
elastic axis ——→ (too strong to ignore, human assistance requested)
aerodynamic center ——→ aerodynamic configurations, weight 20
blade cross sectional center of mass ——→ rotary wing, weight 20, & cross sections, weight 9, & center of gravity, weight 20
quasisteady aerodynamic loads ——→ aerodynamic loads (quasisteady dropped), weight 20
stall ——→ stalling, weight 9
compressibility ——→ compressibility, weight 4
Reversed flow ——→ reversed flow, weight 20
airloads ——→ (new word, if second occurrence, request manual assistance)

NASA TERMS Selections
Exhibit B-5

Print Terms	Occurrence Weights	Sum	Co-occurring Root Boost	Tot Wei
rotary wings	2*20,20,20,20,20,20	140	15	15
torsional stress	2*20,20,20,20	100		10
wing roots	20,20	40	47	8
rigid rotor	20	20	47	6
flaps	2*20,20	60		6
equations of motion	20,20	40	16	5
Non-print Terms				
nonlinear equations	20	20	22	4
partial differential equations	20	20	21	4
simulation	20,20	40		4
equations	9	9	24	3
aerodynamic loads	20		7	2
aerodynamic configuration	20		7	2
center of gravity	20			2
reversed flow	20			2
longitudinal control	20			2
flight	2*9	18		1
stability	2*9	18		1
aeroelasticity	2*9	18		1
cross sections	9			
stalling	9			
deflections	9			
compressibility	4			

Manually Selected Terms
Exhibit B-6

Print Terms	Comments
aerodynamic stability	Of questionable utility
equations of motion	Selected by both
helicopters	Too general but on target
rotor blades	Bad choice - refers to turbines not helicopters

Search Terms

aerodynamic loads
aircraft configurations
nonlinearity
partial differential equations

Computer terms not manually selected

Rotary wings	Excellent descriptor
Wing roots	Too highly rated; okay as a search term not as a print term
Rigid rotor	Highly descriptive, perhaps too specific for a print term, however.
Torsional stress	Too descriptive not to include.

N77-29097* National Aeronautics and Space Administration. Langley Research Center, Langley Station, Va
LOAD DISTRIBUTION ON A CLOSED-COUPLED WING CANARD AT TRANSONIC SPEEDS
Blair B. Gloss and Karen E. Washburn Aug. 1977 11 p refs
(NASA-TM-74053) Avail: NTIS HC A02/MF A01 CSCL 01A

A wind tunnel test where load distributions were obtained at transonic speeds on both the canard and wing surfaces of a closely coupled wing canard configuration is reported. Detailed component and configuration arrangement studies to provide insight into the various aerodynamic interference effects for the leading edge vortex flow conditions encountered are included. Data indicate that increasing the Mach number from 0.70 to 0.95 caused the wing leading edge vortex to burst over the wing when the wing was in the presence of the high canard.
Author

Third Surrogate as Published (STAR)
Exhibit C-1

77N29[illegible] ISSUE 20 PAGE 2622 CATEGORY 2 NASA-TM-74053
77/08/[illegible] 11 PAGES UNCLASSIFIED DOCUMENT

LOAD DISTRIBUTION ON A CLOSED-COUPLED WING CANARD AT TRANSONIC SPEEDS

A/GLOSS, B. B.: B/WASHBURN, K. E.
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION. LANGLEY RESEARCH CENTER, LANGLEY STATION, VA. AVAIL.NTIS HC A02/MF A01

/*CANARD CONFIGURATIONS/*LOAD DISTRIBUTION (FORCES)/*TRANSONIC SPEED/*WINGS/ AIRCRAFT STRUCTURES/ FLOW VISUALIZATION/ MACH NUMBER/ VORTICES/ WIND TUNNEL TESTS

ABA AUTHOR

ABS A WIND TUNNEL TEST WHERE LOAD DISTRIBUTIONS WERE OBTAINED AT TRANSONIC SPEEDS ON BOTH THE CANARD AND WING SURFACES OF A CLOSELY COUPLED WING CANARD CONFIGURATION IS REPORTED. DETAILED COMPONENT AND CONFIGURATION ARRANGEMENT STUDIES TO PROVIDE INSIGHT INTO THE VARIOUS AERODYNAMIC INTERFERENCE EFFECTS FOR THE LEADING EDGE VORTEX FLOW CONDITIONS ENCOUNTERED ARE INCLUDED. DATA INDICATE THAT INCREASING THE MACH NUMBER FROM 0.70 TO 0.95 CAUSED THE WING LEADING EDGE VORTEX TO BURST OVER THE WING WHEN THE WING WAS IN THE PRESENCE OF THE HIGH CANARD.

RECON Display of the Third Surrogate
Exhibit C-2

Title: Load distribution on a Closed-coupled Wing Canard at Transonic Speeds

Abstract: A wind tunnel test where load distributions were obtained at transonic speeds on both the canard and wing surfaces of a closely coupled wing canard configuration is reported. Detailed component and configuration arrangement studies to provide insight into the various aerodynamic interference effects for the leading edge vortex flow conditions encountered are included. Data indicate that increasing the Mach number from 0.70 to 0.95 caused the wing leading edge vortex to burst over the wing when the wing was in the presence of the high canard.

Phrase Isolation
Exhibit C-3

Jack Kirschbaum

Text Phrase Reduction
Exhibit C-4

Load: Load distribution (forces) -- closest match, weight 20, Load Factors/ Loads (forces), and 31 others which include the root "load."

Closed-coupled: not found

Closed:
Close packed lattices
Closed Basins
Closed Circuit Television
Closed Cycles
Closed Ecological Systems
Closed Faults
Closed Loop Systems
Closing
Closure Law
Closures
: No satisfactory usage found

All occurrences of "CLOSE..." in NASA Thesaurus Access Vocabulary

Coupled:
Charge Coupled Devices
Coupled Modes
Couplers
Antenna Couplers
Couples
Coupling
Coupling Circuits
Coupling Coefficients
Cross Coupling
Gyroscopic Coupling
Microwave Coupling
Optical Coupling
Spin-spin Coupling
Thermodynamic Coupling
Couplings
: No satisfactory usage

Wing:
Fan in Wing Aircraft
Fixed Wing Aircraft
(through 64 such phrases, none involving CANARD)
Wings, best match, weight 9

Canard: Canard Configurations, weight 50

Transonic:
Transonic Aircraft/supersonic aircraft
Transonic Aircraft Technology Program/TACT Program
Transonic Flight
Transonic Flow
Transonic Flutter
Transonic Inlets/Supersonic Inlets
Transonic Nozzles
Transonic Speed -- term matches surrogate, weight 20
Transonic Turbines/Supersonic Turbines
Transonic Wind Tunnels
Transonics/Transonic Flow

Wind: (44 entries including) Wind Tunnel Tests, weight 50

Load: (33 entries including) Load Distribution (Forces), weight 50

Transonic, canard, wing -- see above

Jack Kirschbaum

Text Phrase Reduction
Exhibit C-4
(continued)

Surfaces:	(68 entries including)
	Surface layers
	Surface Properties
	Surfaces -- best match, weight 9
	Control surfaces
	Tail Surfaces
	Lifting Surfaces/Surfaces & Lift Devices & Lifting Bodies
Closely Coupled Wing Canard Configuration -- see above	
Component:	Components (and 7 weight 20 specific uses), weight 9
Configuration:	(16 entries including)
	Hammerhead Configuration
	Configuration Management
	Configurations
	Aerodynamic configurations
	Aircraft configurations
	Body-Wing Configurations
	Body-wing and Tail Configurations
	Canard Configurations - strengthened since previously selected, weight 50
	Launch Vehicle Configurations
Aerodynamic:	(26 entries including) Aerodynamic Interference, weight 20
Leading:	(15 entries for LEAD and)
	Leadership
	Leading Edge Slats
	Leading Edge Sweep
	Leading Edges -- best match, weight 9
	Sharp Leading Edges
Vortex:	(17 entries including) Vortex Flow/Vortices, weight 9
Mach:	Mach Cones
	Mach Inertia Principle
	Mach Number -- closest match, weight 8 without numbers
	Critical Mach Number
	Mach Zehnder Interferometers
	From 0.70 with "Mach" induced number evaluation: Subsonic speeds, weight 20
	To 0.95, with "Mach" ...: Transonic speeds, weight 20
Wing, Leading Edge, Vortex -- see above	
Burst:	Bursts -- best fit, weight 9
	Meteor Bursts
	Radio Bursts
	Solar Radio Bursts
	Type 2,3,4 or 5 Bursts
Wing, Wing - Canard -- see above	

NASA TERMS Selections
Exhibit C-5

Terms	Occurrence Weights	Sum	Co-occurring Boosts	Total
Canard Configurations	2*50,50,50,50,50	300		300
Transonic speeds	2*20,20,20	80	10	90
Wings	2*9,9,9,9,9,9	63		63
Load distribution	2*20,20	60		60
Wind Tunnel Tests	50	50		50
Subsonic Speeds	20	27		47
Aerodynamic Interference	20	20		20
Leading Edges	9,9	18		18
Vortices	9,9	18		18
Burst	9	9		9
Surfaces	9	9		9
Components	9	9		9

Manually Selected Terms
Exhibit C-6

Print Terms	Comments
Canard Configurations	Common to both
Load distribution (forces)	Common to both
Transonic speeds	Common to both
Wings	Common to both
Non-Print Terms	
Aircraft structures	Why?
Flow visualization	How?
Mach Number	Plausible
Vortices	Reasonable, common to both
Wind Tunnel Tests	Should be a print term considering usefulness for retrieval

A78-11362 Solar electric-energy market penetration. R. K. Sarin and K. Nair (Woodward-Clyde Consultants, San Francisco, Calif.). In: International Solar Energy Society, Annual Meeting, Orlando, Fla., June 6-10, 1977, Proceedings. Sections 26-38. (A78-11212 01-44) Cape Canaveral, Fla., International Solar Energy Society, 1977, p. 28-13 to 28-17.

A Bayesian approach was employed to forecast the solar electric market penetration by the years 1990 and 2000. The study identified a multitude of factors, including relative cost of competitive energy systems, government incentives, future environmental regulations, and new technologies, that would affect the solar market share. The judgments of several experts from utility companies, government agencies, and research laboratories were utilized in a systematic manner to quantify the probability distributions of future solar market share as a function of the various factors. The likelihood of the occurrence of these factors was also assessed, and the solar market share was forecasted for the most-likely future scenarios.
(Author)

Fourth Surrogate as Published (IAA)
Exhibit D-1

78A11362 ISSUE 1 PAGE 79 CATEGORY 44 77/00/00 5 PAGES
UNCLASSIFIED DOCUMENT
SOLAR ELECTRIC-ENERGY MARKET PENETRATION
A/SARIN, R. K.; B/NAIR, K. B/(WOODWARD-CLYDE CONSULTANTS, SAN FRANCISCO, CALIF.)
IN INTERNATIONAL SOLAR ENERGY SOCIETY, ANNUAL MEETING, ORLANDO, FLA., JUNE 6-10, 1977, PROCEEDINGS. SECTIONS 26-38. (A78-11212 01-44) CAPE CANAVERAL, FLA., INTERNATIONAL SOLAR ENERGY SOCIETY, 1977, P. 28-13 TO 28-17.
/*ELECTRIC POWER/*MARKET RESEARCH/*SOLAR ENERGY/*TECHNOLOGICAL FORECASTING/ BAYES THEOREM/ CLEAN ENERGY/ COST INCENTIVES/ ENERGY TECHNOLOGY/ PROBABILITY DISTRIBUTION FUNCTIONS
ABA (AUTHOR)
ABS A BAYESIAN APPROACH WAS EMPLOYED TO FORECAST THE SOLAR ELECTRIC MARKET PENETRATION BY THE YEARS 1990 AND 2000. THE STUDY IDENTIFIED A MULTITUDE OF FACTORS, INCLUDING RELATIVE COST OF COMPETITIVE ENERGY SYSTEMS, GOVERNMENT INCENTIVES, FUTURE ENVIRONMENTAL REGULATIONS, AND NEW TECHNOLOGIES, THAT WOULD AFFECT THE SOLAR MARKET SHARE. THE JUDGMENTS OF SEVERAL EXPERTS FROM UTILITY COMPANIES, GOVERNMENT AGENCIES, AND RESEARCH LABORATORIES WERE UTILIZED IN A SYSTEMATIC MANNER TO QUANTIFY THE PROBABILITY DISTRIBUTIONS OF FUTURE SOLAR MARKET SHARE AS A FUNCTION OF THE VARIOUS FACTORS. THE LIKELIHOOD OF THE OCCURRENCE OF THESE FACTORS WAS ALSO ASSESSED, AND THE SOLAR MARKET SHARE WAS FORECASTED FOR THE MOST-LIKELY FUTURE SCENARIOS.

RECON Display of the Fourth Surrogate
Exhibit D-2

Title: Solar electric-energy market presentation.

Abstract: A Bayesian approach was employed to forecast the solar electric market penetration by the years 1990 and 2000. The study identified a multitude of factors, including relative cost of competitive energy systems, government incentives, future environmental regulations, and new technologies, that would affect the solar market share. The judgments of several experts from utility companies, government agencies, and research laboratories were utilized in a systematic manner to quantify the probability distributions of future solar market share as a function of the various factors. The likelihood of the occurrence of these factors was also assessed, and the solar market share was forecasted for the most-likely future scenarios.

Phrase Isolation
Exhibit D-3

Text Phrase Reduction
Exhibit D-4

Solar Electric-energy market penetration:

Solar energy -- since "solar is not a valid term and the only "solar electric" is "solar electric propulsion" which failes the subject area test. Due to the large number of solar phrases, the phrase solar electric energy will be reported as a candidate phrase for manual review, weight 20.

Electric power -- "electric" alone is not an entry phrase; therefore, use "electric power" via "electrical energy" as best match, weight 20

Market -- Marketing, weight 9

Penetration -- Penetration weight 9, flagged for manual review since use is not in document having substance in normal areas of use, specifically "geology" or "metalic materials"

Bayesian: Bayes theorem via "Bayesian Statistics" the only term including "Bayesian," weight 20

Forecast: Technology Forecasting, weight 20

Solar electric market penetration: (see above) - posted for manual revision as a recurring phrase not in the thesaurus.

Cost: Costs, weight 20

Energy: Electric Power, in preference to "Energy," weight 20

Government incentives: Government, weight 9 - Cost incentives, weight 20, rather than "incentives," weight 9

Environmental regulations: Environmental control, weight 20, via "regulations" being "used" to "control"

Technologies: Technology forecasting, weight 20, in preference to Technologies, weight 4

Solar market share: Solar energy, weight 20, in preference to "solar, weight 9; Marketing, weight 9

Utility companies: Utilities, weight 20

Government: Governments, weight 20

Research laboratories: Research Facilities, weight 20; Laboratories, weight 9

Probability distributions: Probability Distribution Functions, weight 20

Solar Market Share: see above

Likelihood: Maximum Likelihood Estimates, weight 20, with posting to review as over specific but only valid choice

Solar Market Share: See above.

Forecasted: See above.

Most-likely: posted to new-word list.

NASA TERMS Selections
Exhibit D-5

Print Terms	Occurrence Weights	Co-occurring Boosts	Total
Solar energy	2*20,20,20,20,20		120
Electronic Power	2*20,20,20		80
Technological Forecasting	20,20,20		60
Marketing	2*9,9,9,9,9		54
Non-Print Terms			
Cost Incentives	20	10	30
Penetration	2*9,9		27
Costs	20	7	27
Environmental Control	20		20
Utilities	20		20
Probability Distributions	20		20
Maximum Likelihood Estimate	20		20
Research Facilities	20		20
Bayes Theorem	20		20
Government	9,9		18
Laboratories	9		9

Manually Selected Terms
Exhibit D-6

Print Terms:	Comments:
Electric Power	Common selection.
Market Research	Good term, not in reach of present algorithm computer chose the general term 'marketing' which is inferior.
Solar Energy	Common selection.
Technological Forecasting	Common term
Non-print Terms:	
Bayes Theorem	Also computer located
Clean energy	Unclear why selected
Cost incentives	Also computer located
Energy technology	General phrase, computer term and manual term are more specific.
Probability Distribution Functions	Also computer located.

4. Findings and Recommendations

During the preparation of this paper, a few experimental and operational systems that deal with either computer aided indexing or natural language indexing have been identified as being of interest. Amongst these are DDC, SSIE, IBM/BROWSER, Wright Patterson AFB/CIRCOL, and SMART. Except for the CIRCOL study by WESTAT, none of these operational systems has been subject to comparative analysis. The MEDLARS system is currently based on human indexing and has undergone several analytical studies; all were designed to find out where the system was failing - not to compare alternatives. In all cases, the systems are functioning at a presumed satisfactory level of cost and performance -- although no one has really made a significant attempt to evaluate either adequacy of performance relative to less (or more) costly alternatives.

Based primarily on the ongoing operational success of DDC, SSIE, CIRCOL and BROWSER and the systematic studies performed by Lancaster, Cleverdon, and Salton (and others), we believe that the publication of computer generated indexes from abstracts and information retrieval functions could be accomplished at significantly lower cost with negligible overall changes to effectiveness.

BIBLIOGRAPHY OF MAJOR REFERENCES

- Cleverdon, C. W., Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, 1962.
- Cleverdon, C. W., et. al., Factors determining the performance of indexing systems. Vol. 2, Cranfield, 1966.
- Cleverdon, C. W., A comparative evaluation of searching by controlled language and natural language in an experimental NASA data base, European Space Agency, Space Documentation Service, Draft Report, April, 1977.
- Hargrave, C. W. and Wall, E., Retrieval Improvement Effected by Use of A Thesaurus, ASIS Conference Proceedings, Philadelphia, Pa. 1971.
- Hunt, B. L., Snyderman, M. and Payne, W., Machine-assisted indexing of scientific research summaries, Journal of the American Society for Information Science, July-August 1975, pp. 230-236.
- King, D. W., et. al., Comparative evaluation of the retrieval effectiveness of descriptor and free-text search systems using CIRCOL (Central Information Reference and Control On-Line), Vol. 1, Westat Research, Inc., 1971.
- Klingbiel, Paul H., Machine-Aided Indexing of Technical Literature, Information Storage and Retrieval, February 1973.
- Klingbiel, P. H., Machine Aided Indexing Reports DDC-TR-69-1, DDC-TR-71-3 and DDC-TR-71-7 (AD 969-200 (1969), AD 721-875 (1971) and AD 733-800 (1971)).
- Lancaster, F. W. and Fayen, E. G., Information Retrieval On-Line, John Wiley & Sons, Inc., 1973 (Information Sciences Series).

BIBLIOGRAPHY OF MAJOR REFERENCES
(continued)

L1 Lancaster, F. W., Vocabulary control for information retrieval, Information Resources Press, 1972.

o The Principles of MEDLARS, U. S. Department of Health, Education and Welfare, Public Health Service, National Institute of Health, National Library of Medicine, 1970.

N1 NASA Thesaurus - NASA SP-7050 (1976 Edition), National Aeronautics and Space Administration, Washington, D.C.

o Operating Manual - NASA Facility Operations, Chapter 2, Section 4 - Abstracting/Indexing Section, Analysis Department, Revised June 23, 1977, Part 1.

o Salton, G., Automatic Text Analysis, Gesellschaft fuer Informatik, 1970, pp. 421, N71-25981.

o Salton, G., Automatic Information organization and retrieval, McGraw-Hill Book Company, 1968.

o Salton, G., The SMART retrieval system: Experiments in automatic document processing, Prentice-Hall, Inc., 1971.

o Siroonian, H. A., Pandex, Special Libraries 58:10 (December 1967), pp. 728-730.

o Schurfeld, H., et. al., A method for the automatic indexing, storing, and retrieving of full-text documents, IBM Germany, Patent Department, 1971, N71-25951.

o Williams, J. H., Jr. and Perriens, M. P., Automatic Full Text Indexing and Searching System, IBM Federal Systems Division, Gaithersburg, MD, 1968.

o Williams, J. H., Jr., Functions of a man-machine interactive information retrieval system, Journal of American Society of Information Science, Vol. 22, 1971, pp. 311-317.

o Williamson, Robert E., Real-Time Information Retrieval, PhD Thesis, Cornell University, 1974.

o Williamson, Robert E., An N-Log-N Single Pass Clustering Algorithm (to be published) COMPSAC 1977 Proceedings of IEEE Computer Society, Chicago, Ill.