Analysis and Display of Information". *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.* **1979**, *B35*, 2331-2339.
(26) Lesk, A. M. "Detection of 3-D Patterns of Atoms in Chemical Structures". *Commun. ACM* **1979**, *22*, 219-224.
(27) Esaki, T. "Quantitative Drug Design Studies. V. Approach to Lead Generation by Pharmacophoric Pattern Searching". *Chem. Pharm. Bull.* **1982**, *30*, 3657-3661.
(28) Ghose, A. K.; Crippen, G. M. "Geometrically Feasible Binding Modes of a Flexible Ligand Molecule at the Receptor Site". *J. Comput. Chem.* **1986**, *6*, 350-359.

An Expert System for Machine-Aided Indexing[†]

CLARA MARTINEZ,* JOHN LUCEY, and ELLIOTT LINDER

American Petroleum Institute, 156 William St., New York, New York 10038

Received December 5, 1986

The Central Abstracting & Indexing Service of the American Petroleum Institute (API–CAIS) has successfully applied expert system techniques to the job of selecting index terms from abstracts of articles appearing in the technical literature. Using the API Thesaurus as a base, a rule-based system has been created that has been in productive use since February 1985. The index terms selected by computer are reviewed by a human index editor, as are the terms selected by CAIS's human indexers. After editing, the terms are used for printed indexes and for online computer searching.

INTRODUCTION

The Central Abstracting & Indexing Service (CAIS) of the American Petroleum Institute (API) has produced indexes of patents and technical literature of interest to the petroleum and petrochemical industries since 1964. The documents to be indexed are selected from more than 150 journals in seven different languages and from about 50 recurring meetings and conferences. The abstractors on the CAIS staff select the documents according to specific rules and write an informative abstract of about 200 words. After being edited and proofread, the abstracts are used to prepare the alerting bulletins distributed to subscribers. At the same time, a copy of each abstract goes to the technical indexing staff for indexing.

Indexing is done by use of a controlled vocabulary. An indexer reads the abstract and tries to index all concepts using the valid terms in the API Thesaurus. The index is then edited to ensure accuracy and completeness (Figure 1).

Producing a high quality database using human indexers is expensive. For a long time CAIS had considered automated processing techniques to reduce the cost of indexing. Automated indexing as well as machine-aided indexing[1-12] was investigated.

In 1979, the API Special Task Force Comparing Automatic and Manual Indexing conducted a comparative study of retrieval of information from a system with automatic indexing versus that from the manually indexed API files. Later, the API Automatic Indexing Task Force was created to look into advanced techniques for information processing. These studies concluded that no existing system was suitable for our purposes. We therefore started our machine-aided indexing (MAI) project in 1982. It was felt that MAI would provide the flexibility necessary to handle the special features, i.e., the controlled vocabulary, roles, and links, of the API technical databases. The application of expert systems to a project of this type was also noted.[13]

The purpose of this paper is to present what we are doing at CAIS. It has been considered the first attempt to develop an automated indexing system that models human indexing behavior.[14]

KNOWLEDGE BASE

The MAI system is a rule-based expert system; that is to say, we are using the knowledge of our experts to create additional indexing rules and to modify or delete existing ones.

The machine scans the text of the abstract of a document and compares it with text fragments stored in the Knowledge Base. When a match occurs and certain specified conditions are met, the Knowledge Base supplies the appropriate index terms from the API Thesaurus. We refer to the package of a specific text fragment with any special conditions and the associated index terms as a "rule". The original Knowledge Base consisted of the index terms and cross-references in the 1982 edition of the API Thesaurus.

An example of a rule in the Knowledge Base is

TEXT: NAPHTHA
TERM: NAPHTHA

The text found in the abstract is followed by the term to be indexed. In this case the text and the term are the same. In another example

TEXT: LNG
TERM: LIQUEFIED NATURAL GAS

the rule is derived from a cross-reference in the API Thesaurus. In this case the term to be used is not the same as the text.

In order to create additional rules, a batch of about 1000 abstracts of documents indexed in 1982 was run against the rules in the API Thesaurus. The MAI indexing for these abstracts was then compared to the human indexing. This first round of the MAI was analyzed for good terms selected (HITS), good terms not selected (MISSED), and bad terms selected (NOISE). Statistical data were produced for the NOISE and MISSED terms, and lists were produced for the terms in descending order of frequency (Figure 2).

The results of the first round were HITS 40% and NOISE 38%.

It was found at that point that punctuation and plurals were causing problems. Handling the irregular plurals and eliminating most of the punctuation raised the percentage of HITS to 44% and lowered the NOISE to 21%. By reviewing samples of the abstracts in which the high-frequency terms were errors, we could determine what changes were needed in the

[†] Presented, in part, before the Division of Chemical Information, Symposium on Applications of Artificial Intelligence in Chemical Information, 191st National Meeting of the American Chemical Society, New York, NY, April 15, 1986.

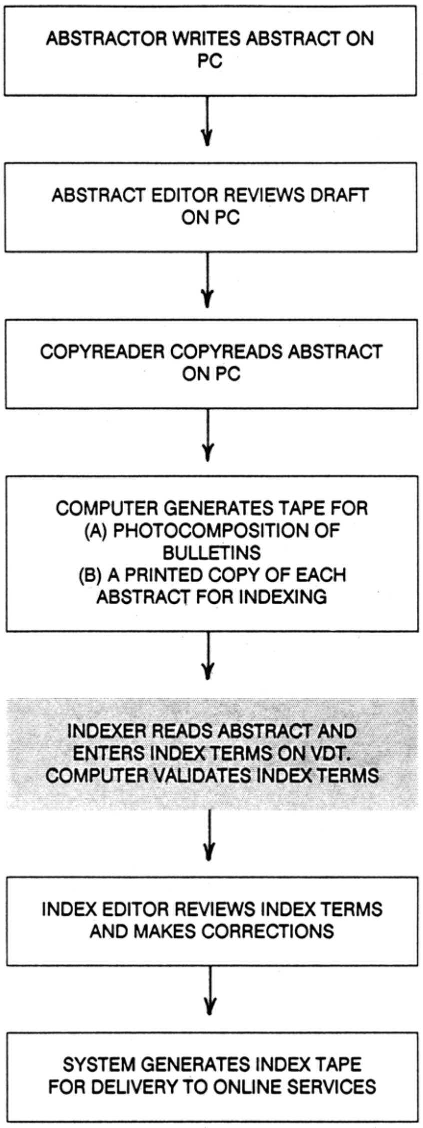

ABSTRACTOR WRITES ABSTRACT ON PC
ABSTRACT EDITOR REVIEWS DRAFT ON PC
COPYREADER COPYREADS ABSTRACT ON PC
COMPUTER GENERATES TAPE FOR (A) PHOTOCOMPOSITION OF BULLETINS (B) A PRINTED COPY OF EACH ABSTRACT FOR INDEXING
INDEXER READS ABSTRACT AND ENTERS INDEX TERMS ON VDT. COMPUTER VALIDATES INDEX TERMS
INDEX EDITOR REVIEWS INDEX TERMS AND MAKES CORRECTIONS
SYSTEM GENERATES INDEX TAPE FOR DELIVERY TO ONLINE SERVICES

Figure 1. CAIS text editing system. The MAI replaces the steps in the shaded box.

1	268	CONCENTRATION	268
2	211	CONTROL	479
3	198	COMPARISON	677
4	185	EFFICIENCY	862
5	130	ENERGY	992
6	127	TEMPERATURE	1119
7	122	REVIEW	1241
8	122	STATE	1363
9	121	MIXTURE	1484
10	112	HEAT	1596
11	110	EQUIPMENT TESTING	1706
12	108	PREDICTION	1814
13	102	LIQUID	1916
14	98	CYCLE	2014
15	95	HYDROGENATION	2109

Figure 2. List of NOISE terms.

Knowledge Base. Our efforts were concentrated on the highly posted terms. Figure 3 shows the Knowledge Base creative loop.

FAILURE ANALYSIS

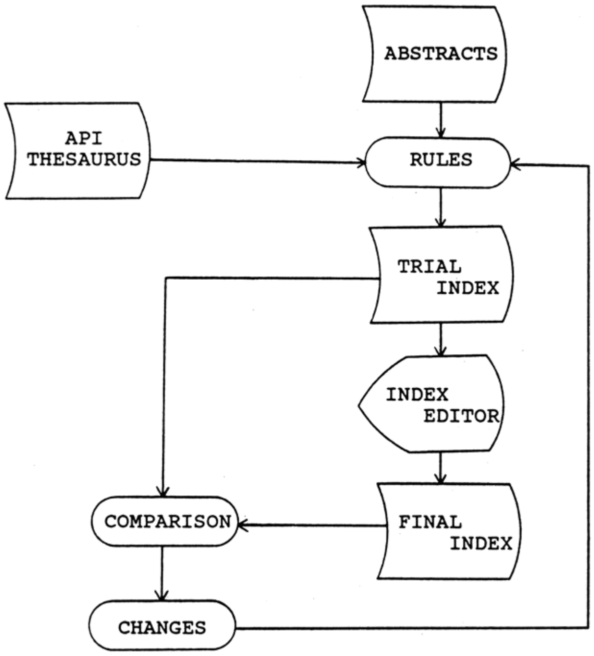


Figure 3. Knowledge Base creative loop.

...at substantially the same speed as
the air stream...

...mark the first time that gas carriers
have been converted to....

...transportation of crude from its Huntington
Beach, Calif., offshore terminal...

...developed by Land & Marine Engineering
Ltd.,...

Figure 4. Sample phrases producing NOISE.

Analysis of the MISSED terms allowed us to identify a term for indexing by the following guidelines. (1) In the case of co-occurrence of two words in an abstract, e.g., "converted", "conversion", or "converting" and "tanker" appear in an abstract, the term to be used is MODIFICATION. (2) Terms may be implied in a phrase; e.g., for "gas pipeline", use NATURAL GAS. (3) A word may be in an abstract under a specified subject section of our abstract bulletin; for, e.g., "supervisory" in the Pipeline Communication–Control section, use SUPERVISORY COMPUTER CONTROL.

As we advanced in the analysis, the rules became more complex. For example

SOLUTION
If "solution cavern", use ELUTION plus CAVERN.
If "aqueous solution", use WATER link INORGANIC SOLVENT.
If "solution mining", use ELUTION plus MINING.

Otherwise, ignore "solution" because, in the API Thesaurus, "solution" is a phase material and should not be used for, e.g., the "solution of an equation". (Note that our rules, like the API databases themselves, are often subject oriented.)

Some examples of the NOISE problems were the following (Figure 4): (1) common words, e.g., "can", which is a valid term in the API Thesaurus used to index the type of container; (2) terms with different meanings, e.g., STREAM, which is used in the API system for a river but not for a process stream; (3) words like "time" used in expressions such as "at the same time", "first time", etc.; (4) words that when capitalized are part of a company name, e.g., "engineering", or a geographical name, e.g., "beach"; (5) words that could be an acronym or a company name if capitalized, e.g., TOTAL (the French company) or DOE (the U.S. Department of Energy; this results when the system automatically removes the final "s" from "does").

There were text scanning problems due to (1) misspelling, i.e., errors that get past our proofreaders; (2) spaces, e.g., "under water" and "underwater"; (3) the way abstracts are

```
TEXT:ZINC DIALKYLDITHIOPHOSPHATE
LINK:
TERM:ZINC
TERM:SINGLE STRUCTURE TYPE
TERM:SATURATED CHAIN
TERM:SULFUR CONTAINING ACID
TERM:SULFUR CONTAINING ESTER
TERM:PHOSPHORUS CONTAINING ACID
TERM:PHOSPHORUS CONTAINING ESTER
TERM:OXYGEN ORGANIC
TERM:OTHER OXYGEN ESTER
TERM:ORGANIC SALT
TERM:COMPOUNDS
ENDL:
```

Figure 5. MAI rule using Chemical Aspects.

written (Rules were needed for "IR absorption spectrometry", "IR absorption spectra", and "absorption IR spectroscopy". To avoid this multiplicity of rules, proximity was programmed: If IR is adjacent to "absorption" and "spectr?" (where ? indicates a truncation) is within three words of "absorption", the term to use is INFRARED SPECTROSCOPY); (4) chemical nomenclature.

INDEXING CHEMICALS

Petroleum is a complex mixture of hydrocarbons with organic sulfur, oxygen, nitrogen, and metallic impurities. It also provides feedstock for the petrochemical industry.

Therefore, to index a database dealing with petroleum, there should be a way to index chemical compounds. In the API technical index system, chemical compounds are indexed by use of a fragmentation system employing standard index terms that represent molecular structural features. We call these terms Chemical Aspects. Indexing entails assigning appropriate Aspects to each compound according to established rules.

The names of the most commonly occurring chemical compounds are valid index terms (Chemical Index Terms). It was easy for the MAI to identify these compounds, e.g., METHANE. In other cases there were cross-references in the API Thesaurus, e.g., for "ethanol", use ETHYL ALCOHOL. However, if a Chemical Index Term does not exist for a compound, the compound must be indexed by Aspects.

To tie together the structural features of each compound and to prevent false coordination when a document that contains several chemical compounds is indexed, links are used with the Aspects. Zinc dialkyl dithiophosphate, a common additive for lubricants, has to be indexed by Chemical Aspects because of the undefined alkyl. We developed a rule for this compound using the Chemical Aspects and linking capability (Figure 5). (Note: All terms in between LINK and ENDL are assigned the same link.)

Another special feature of the API technical databases is the use of roles. Roles are assigned to Chemical Aspects, Chemical Index Terms, and Materials when they are starting materials (role A) and products (role P) of chemical reactions that are intended to produce recoverable products. For "Carbon monoxide hydrogenation", a common phrase found in documents classified in our abstract bulletins under the section Synthesis Gas, a rule was entered to index CARBON MONOXIDE with role A, elemental HYDROGEN with role A, and REDUCTION REACTION for the hydrogenation reaction (Figure 6).

Rules have been entered not only for names of compounds but also for formulas. In the above example a rule was entered for "CO hydrogenation" as well. In the case of CO, to establish a difference between carbon monoxide and the symbol for cobalt, we set the condition that all letters must be capitalized. To establish a difference between the symbol for cobalt and the abbreviation for "company", we have to be able to use the period for the abbreviation.

```
TEXT:CARBON MONOXIDE HYDROGENATION
SECT:SYNTHESIS GAS
LINK:
TERM:CARBON MONOXIDE
ROLE:A
ENDL:
LINK:
TERM:HYDROGEN
ROLE:A
TERM:ELEMENT
ROLE:A
ENDL:
TERM:REDUCTION REACTION
```

Figure 6. MAI rule using roles.

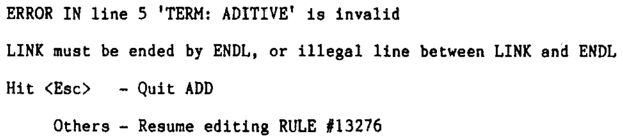
```
TEXT:PHENOL          TEXT:PHENOLS
TERM:PHENOL          LINK:
                     TERM:BENZENE RING
                     TERM:MONOHYDROXY
                     TERM:COMPOUNDS
                     ENDL:
```

Figure 7. Rules for a specific compound and for the generic class of compounds to which it belongs.

```
COND:

1-INITIAL CAP
2-AFTER INITIAL CAP WORD
3-FOLLOWED BY INITIAL CAP
4-PRECEDED BY INITIAL CAP WORD
5-ALL CAPS
6-INITIAL LOWER CASE
7-FOLLOWED BY LOWER CASE
8-PRECEDED BY LOWER CASE
9-IN TITLE
10-IN FIRST 2 SENTENCES
11-IN LAST SENTENCE
12-FOLLOWED BY NUMBER
13-PRECEDED BY NUMBER
14-IN T&S
```

Figure 8. Some of the conditions available for formulating rules.



```
ERROR IN line 5 'TERM: ADITIVE' is invalid

LINK must be ended by ENDL, or illegal line between LINK and ENDL

Hit <Esc>    - Quit ADD

      Others - Resume editing RULE #13276
```

Figure 9. Error messages displayed for an invalid term and for a wrong format.

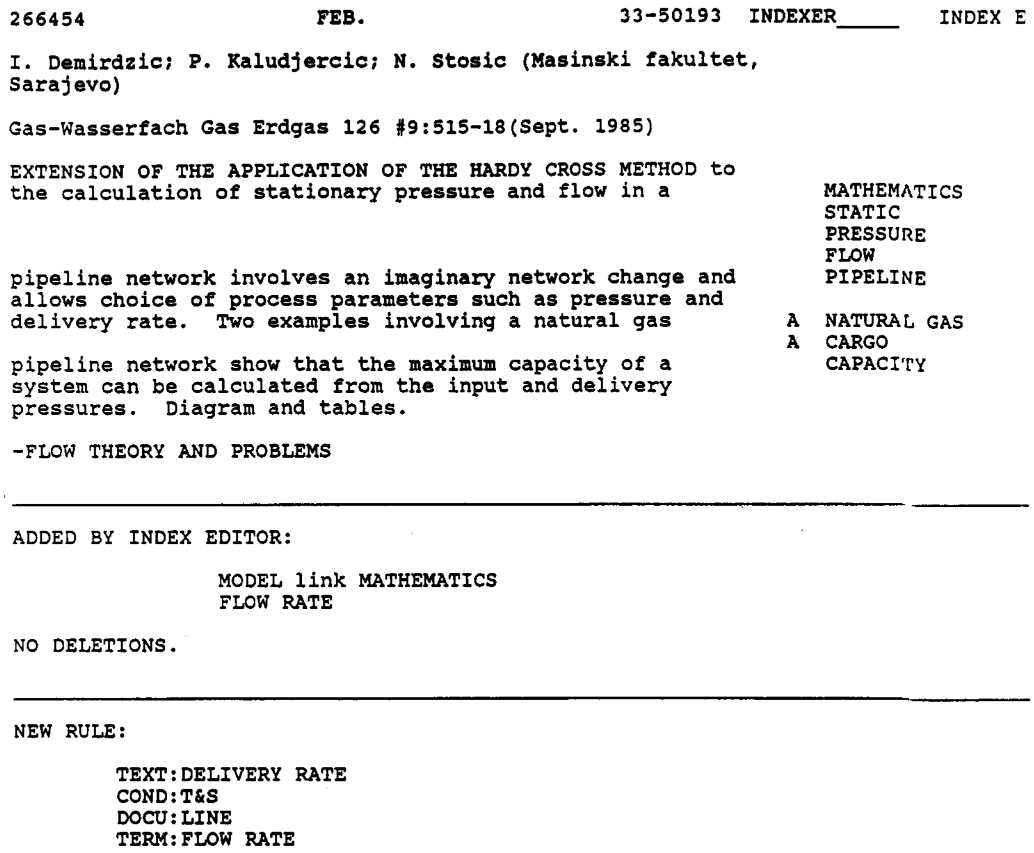
So far we have been able to solve some problems with the chemicals. For example, to differentiate "phenol", the specific compound, and "phenols", the class of compounds, we have had to include a separate rule for each text because the MAI handles regular plurals by dropping a single "s" at the end of a text (Figure 7).

Certainly, some difficult problems remain to be solved. We will be developing software to deal with (1) "derivaties of" a compound and (2) lists of hyphenated prefixes to a base compound, e.g., *m*- *o*-, and *p*-xylene.

KNOWLEDGE BASE MAINTENANCE

The MAI system was developed on the computer available to us at the time, which was a Hewlett-Packard 1000 series. We are now converting to IBM PC-AT's. The Knowledge Base is now maintained on an AT. The system allows for the following functions: (1) enter a new rule; (2) modify or delete an existing rule; (3) browse the knowledge base, e.g., retrieve and display the rules.

To facilitate entering the data, the function keys on the terminal are used. For example, we can copy part of a rule when similar rules are being entered into the system, or we can get a display of the conditional relations available to formulate a rule. So far we have the following types of conditions: SECT, an abstract bulletin section in which abstracts are classified; SENS, a character string appearing in the same



```
266454              FEB.                    33-50193  INDEXER_____    INDEX E

I. Demirdzic; P. Kaludjercic; N. Stosic (Masinski fakultet,
Sarajevo)

Gas-Wasserfach Gas Erdgas 126 #9:515-18(Sept. 1985)

EXTENSION OF THE APPLICATION OF THE HARDY CROSS METHOD to
the calculation of stationary pressure and flow in a          MATHEMATICS
                                                               STATIC
                                                               PRESSURE
                                                               FLOW
pipeline network involves an imaginary network change and      PIPELINE
allows choice of process parameters such as pressure and
delivery rate.  Two examples involving a natural gas         A NATURAL GAS
                                                             A CARGO
pipeline network show that the maximum capacity of a           CAPACITY
system can be calculated from the input and delivery
pressures.  Diagram and tables.

-FLOW THEORY AND PROBLEMS

ADDED BY INDEX EDITOR:

                MODEL link MATHEMATICS
                FLOW RATE

NO DELETIONS.

NEW RULE:

          TEXT:DELIVERY RATE
          COND:T&S
          DOCU:LINE
          TERM:FLOW RATE
```

Figure 10. API abstract and the terms selected by the MAI system.

sentence; DOCU, a character string appearing in the title or text of the abstract; PROX, a character string appearing within three words before or after and within the same sentence; COND, apply rule only if one of several conditions is met (Figure 8); ELSE, specifies an alternative set of rules for the same text if previous set fails to be satisfied.

The conditions are limits imposed on the rules. The negation of each condition may be stipulated.

To check a rule we can search the Knowledge Base using single words, text phrases, or valid terms in the API Thesaurus. The use of Boolean operators with terms or search statement numbers is permitted, as is truncation.

The valid terms contained in the API Thesaurus are stored in the memory of the computer. When we modify or add a rule to the Knowledge Base, the MAI system performs two different types of validation: for the validity of terms and for the format of the rules (Figure 9). Also, when we attempt to enter a rule for a given text, any rule already existing for that text is displayed on the screen.

MAI PROCEDURE

The machine scans the text of each abstract looking for a match with the Knowledge Base. The selection of phrases from the abstract for the matching process starts at the beginning of the abstract, including the title. The first phrase selected is up to 56 characters long. No stopword list is used.

If no match is found to the first phrase, it is altered in several ways and compared again to the Knowledge Base after each alteration. A sample is as follows:

phrase	**alteration**
long-range predictions assess	original phrase
long range predictions assess	eliminate hyphens
long-range predictions asses	eliminate s from end of last word
long range predictions asses	eliminate s from end of last word and eliminate hyphens
long-range predictions	eliminate last word
long range predictions	eliminate hyphens
long-range prediction	eliminate s from end of last word
long range prediction	eliminate s from end of last word and eliminate hyphens

The Knowledge Base contains

TEXT: LONG RANGE PREDICTION

TERM: PREDICTION

so the term PREDICTION is selected for indexing. The next phrase selected for comparison to the Knowledge Base starts with the next word. In our example above, "range" would be the first word in the next phrase.

The process advances through the entire abstract selecting terms. The selected terms are entered into the CAIS Index Editing computer system so that the indexer/editor will see them displayed on his indexing screen when he starts to edit the index. The abstract is also printed for use by the indexer/editor with the terms selected printed next to the line from which they were selected. Duplicate indexing is eliminated by the computer.

Figure 10 shows an abstract and the terms selected by the MAI. (The letter A on the left of NATURAL GAS and CARGO denotes a link.) No terms selected by the MAI were deleted by the editor. The editor added the term MODEL linked to MATHEMATICS and the term FLOW RATE and suggested a new rule: if "delivery rate" is mentioned in the same document as "line" or "pipeline" in the Transportation & Storage section, then the term to use is FLOW RATE.

RESULTS

The first group of abstracts analyzed was from the Transportation & Storage section because it contained the

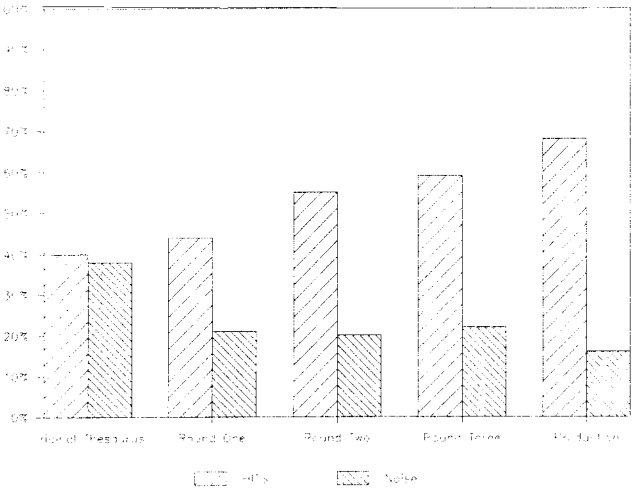


Figure 11. HITS and NOISE for the different rounds of Transportation & Storage.

Table I. HITS, MISSED, and NOISE Terms on the First Round

API section	HITS, %	MISSED, %	NOISE, %
Transportation & Storage	40	60	38
Health & Environment	41	59	18
Petroleum Substitutes	44	56	17
Petroleum Refining	49	51	16

easiest vocabulary to handle. Figure 11 shows the results of the different rounds after statistics and samples of the NOISE and MISSED terms were reviewed. After three rounds, in February 1985, the MAI was put into production.

Evaluation of the MAI system had heretofore been done by comparing the MAI output to the previous indexing of the CAIS staff. The production system is evaluated by comparing the unedited MAI output to the edited version.

When the MAI rules are used in production, one gets about 10% more in HITS and 10% less in NOISE than in the tests of human vs machine indexing. This is due to the following circumstances:

(1) In the API Technical Information System, abstracts of documents are indexed with the most specific terms available in the API Thesaurus. The API Thesaurus has a hierarchical structure, and every time a specific term (Narrower Term) is used, the generic term (Broader Term) is automatically posted to the record. Whenever the MAI selected a Narrower Term for a record but the human indexer used the Broader Term, even though it was present in the MAI-indexed record, the Broader Term was considered a MISSED term.

(2) Inconsistency due to the subjective nature of human indexing has been found. Tests run at CAIS revealed a 75% overall consistency between indexer/editor pairs.

On this basis, then, after four months in production, the HITS increased 5% (from 68% to 73%) and the NOISE decreased 5% (from 16% to 11%).

Many of the rules entered for the Transportation & Storage abstracts applied to those in our Health & Environment, Petroleum Substitutes, and Petroleum Refining abstract bulletins. Therefore, only one round was analyzed for these sections. The percentages of HITS and NOISE obtained on the first round for each section are shown in Table I. The percentages after the analysis and rule entry of the first round are shown in Table II.

Table II. HITS and NOISE When Documents Were Run against Updated Knowledge Base

API section	HITS, %	NOISE, %
Health & Environment	53	17
Petroleum Substitutes	53	15
Petroleum Refining	52	14

The indexing produced by the MAI is consistent and editable. Feedback from the editors is used to add new rules to the Knowledge Base in order to improve the quality of the MAI output. The Knowledge Base currently contains about 14000 specific text rules. More sophisticated technology, e.g., the identification of clusters of Index Terms, will be helpful in solving the remaining problems.

As a byproduct of the MAI project, a front-end system for the API technical databases has been envisioned. The rules developed for indexing can be used as well to assist in the retrieval process.[15]

REFERENCES AND NOTES

(1) Gray, W. A.; Harley, A. J. "Computer Assisted Indexing". *Inf. Storage Retr.* **1971**, *7*(4), 167–174.

(2) Klingbiel, P. H. "Machine-Aided Indexing of Technical Literature". *Inf. Storage Retr.* **1973**, *9*(2), 79–84.

(3) Dillon, M.; Gray, A. S. "FASIT: A Fully Automatic Syntactically Based Indexing System". *J. Am. Soc. Inf. Sci.* **1983**, *34*(1), 99–108.

(4) Salton, G.; McGill, M. J. "Introduction to Modern Information Retrieval". McGraw-Hill: New York, 1983; Chapter 3, p 52.

(5) Field, B. J. "Towards Automatic Indexing I. Relationship between Free—and Controlled—Language Indexing and the Automatic Generation of Controlled Subject Headings and Classifications". INSPEC Report No. R75/20. Institution of Electrical Engineers: London, 1975; 70 pp.

(6) Doszkocs, T. E. "Automatic Free-Text to Controlled Vocabulary Indexing". Presented at the 3rd National Online Meeting, New York, 1982.

(7) Jones, K. Spark; Bates, R. G. "Research on Automatic Indexing 1974–1976". Computer Laboratory, University of Cambridge, 1977.

(8) McGill, M. J.; Noreault, T. "Syracuse Information Retrieval Experiment (SIRE): Rationale and Basic Systems Design". School of Information Studies, Syracuse University, 1977.

(9) Hunt, B. L.; Snyderman, M.; Payne, W. "Machine-Assisted Indexing of Scientific Research Summaries". *J. Am. Soc. Inf. Sci.* **1975**, *26*(4), 230–236.

(10) Earl, L. L. "Experiments in Automatic Extracting and Indexing". *Inf. Storage Retr.* **1970**, *6*(4), 313–334.

(11) Harding, P. "Automatic Indexing and Classification for Mechanised Information Retrievel". BLADD Report No. 5723. Institution of Electrical Engineers: London, 1982; 109 pp.

(12) Field, B. J. "Towards Automatic Indexing: Automatic Assignment of Controlled Language, Indexing, and Classification from Free Indexing". *J. Doc.* **1975**, *31*(4), 246–265.

(13) "Monitor Survey of the Information Industry: Expert Systems". *Monitor* **1982**, No. 19, 4–8.

(14) Fidel, R. "Toward Expert Systems for the Selection of Search Keys". *J. Am. Soc. Inf. Sci.* **1984**, *37*(1), 37–44.

(15) Brenner, E. H.; Lucey, J. H.; Martinez, C. L.; Meleka, A. "API's Machine-Aided Indexing Project". *Sci. Technol. Libr.* **1984**, *5*(1), 49–62.