A STUDY OF AUTOMATIC INDEXING

FOR PATENT EXAMINATION

Yohichi Ohmoli*

TR 77-317

* Author on leave from Japanese Government Patent Office.
This study has been supported by Science and Technology Agency of Japanese Government.

I. PATENT OFFICE PROBLEMS

I-1. Number of Patent Applications

The Japanese Government Patent Office is required by law to receive and process adequate applications for patents for inventions which are submitted to it by persons not only residing in Japan but also in all other countries. As all inventors of the world are eligible to receive Japanese patents, and as the field of invention is almost unlimited, it would be only reasonable to expect that many would take advantage of the opportunity. As a result of restoring International relations with many countries, and due to the evolution of technology after World War II, new technologies have been and require patents so that the number of applications has greatly increased. The rate at which they are received might be higher were we able to process them more promptly. At this time about two and three years elapses, in the average case, after the application is filed, before the patent is issued, and this is too long a time. In this decade, the staff is being increased in order to place the Patent Office in a position to give more prompt service and it is hoped that the time of pendency cay be reduced to two years or less.

I-2. Search for Patent

Since the search is made from the point of view of patentability, the test for the degree of pertinence of the subject matter of the application for the patent is governed by the established criteria for patentability. Hence the search is not only for disclosures of identical concepts but also for concepts which are analogous and similar according to these criteria. To find such similarities the search is made in effect on the basis of a class or a category which includes both the specified subject matter of the application and all related subject matter. It is thus of vital importance that the officials of the Patent Office know how to locate every patent and other publications, the disclosure of which is pertinent to the invention claimed in any application submitted by an inventor. This need to be fully advised of the nature of the disclosures of all publications is absolute and is not limited to a need to be familiar with the disclosure of publications printed in Japan.

The examiners are charged with the duty of considering the content of publications printed in any country and in any language. Our Office receives a great deal of publications every year, and we have now more than fifteen million publications in our library. In the event that a search conducted by a patent examiner is incomplete or imperfect and the examiner does not find and cite an earlier patent or publication the disclosure of which has an important bearing upon the invention for which protection is sought. A patent may be carelessly issued for an invention which is not, as a matter of fact, a new invention. Such a patent may be overturned when presented to a court for consideration in an infringement suit despite the presumption of validity which attaches to the considered judgments of Administrative Agencies.

A STUDY OF AUTOMATIC INDEXING FOR PATENT EXAMINATION

Yohichi Ohmoli*

Department of Computer Science
Cornell University
Ithaca, NY 14853

PREFACE

The present report is a basic material for discussion about automatic indexing for patent examination. I have completed this work at the Department of Computer Science in Cornell University. I am deeply indebted to Dr. Gerard Salton, the Chairman of the department, for offering every kindness to me and to my family. A number of members of the Japanese Government Patent Office have assisted me in coming to the U.S.A. Among them are Mr. Kiyotaka Sasaki, Mr. Hideo Uchiyama, Mr. Tadao Noguchi, Mr. Tsutomu Kusano. This manuscript was typed expertly and cheerfully by Linda Rask. I express my thanks to these people for their guidance and assistance.

---

* Author on leave from Japanese Government Patent Office.
This study has been supported by Science and Technology Agency of Japanese Government.

CONTENTS

II. RESEARCH ON MECHANIZATION OF PATENT SEARCHING

II-1. Brief History

The first trial of mechanized information retrieval for patent examination was held in 1947 by the classification group of the chemical division of the U.S. Patent Office [1]. Since then they have done a great deal of research work [2,3,4]. In 1961 the International Committee was established by the proposal of the Dutch Patent Office, the Committee was called ICIREPAT that stands for Committee of International Co-operation in Information Retrieval among Examining Patent Office. Every member country has developed IR systems respectively until 1963, [5] and the fourth General Meeting of the ICIREPAT in 1964 decided several rules which all the member countries are expected to observe [1]. The Japanese Patent Office has been developing several IR systems in accordance with the rules. Although the systems have been partly mechanized most depend on human ability, especially the indexing of documents which takes too much professional time of the examiners. It is generally believed that more than 80% of the professional time for the systems is spent in manual indexing.

II-2. Problems of the Current Systems

Though it is possible to make up very precise (i.e. high recall, high precision) manual indexed information retrieval systems, the practice of manual indexing has several misgivings. Salton expresses these misgivings clearly in connection with the conventional indexing method [6].

> "First, it is not clear that all the complexity and refinements currently advocated--exemplified by the grouping of terms into categories or the assignment of relations between terms--are really beneficial. On the contrary, quite a few observers are convinced that too much professional time is needlessly wasted in cataloging and indexing and that current standards for bibliographic control and analysis are unnecessarily complicated. Second, it is argued that even if the indexing process were carried out accurately, and at the right level of detail, it is actually impossible to perform the procedure consistently since more than one indexer will necessarily be needed in practice. Thus, the same types of documents will be indexed differently by different indexers, and queries in a given subject area will most likely receive a treatment distinct from that afforded earlier to the documents covering the same subject area. This inconsistency necessarily affects retrieval performance and impairs the usefulness of the intellectual decisions that control the indexing process."

A comparison between manual indexing and the automatic indexing method developed by Salton [7], and a comparison between manual and machine aided indexing developed in Klingbiel [8] suggests strongly that these automated indexing methods compete in their effectiveness with manual indexing. This study will be given as material for discussion toward the mechanization of indexing methods for our information retrieval systems.

III. A STUDY OF AUTOMATIC INDEXING FOR PATENT EXAMINATION--AIPE

III-1. Language Problem

The Japanese language is quite unique among all the tongues of the world. The structure and characteristics of the Japanese language are quite different from those of English. This is not a paper for language processing, but for automatic indexing.

Therefore, these differences are not a serious problem. Furthermore, the Japanese Patent Office has a great deal of documents written in English. For this reason this study has been done in English. A study in Japanese will be done subsequently.

III-2. Targets of This Study

This study mainly targets

1. reduction professional time that is wasted in manual indexing
2. selecting of index terms from texts objectively

And the system must be able to deal with both narrow as well as wide technical fields, i.e. the system can be used as a small system and a large system. The system must have high recall ratio and high precision ratio which can compete with manual indexing. The system must be simple enough for practical use.

III-3. Document Length

The isolation of variables and the measurement of their impact on total system performance as accomplished by Salton [7] in the SMART system suggest that document abstracts are more effective for content analysis purposes than are document titles alone. Further improvement is possible naturally when abstracts are replaced by large text portions: however, the increase in effectiveness is not large enough to reach the equivocal conclusion that full text processing is always superior to abstract processing.

Titles and abstracts and categories which show technical field of each text will be read into the AIPE. The categories will help to make a system which can deal with wide technical fields i.e. mechanical engineering, electrical engineering, chemistry, etc. and narrow technical fields i.e. mechanical engineering only, electrical engineering only etc., simultaneously. Classifying technical document into categories is easy enough for the Patent Office or inventors. The categories should have a relation to the International Patent Classification which has eight categories. Recently all patent documents and nearly most technical papers have abstracts. All of them have, of course, titles. These three items surely help making practical systems.

III-4. The AIPE System

The system has three dictionaries and one table. The indexing cycle is completed by each text. A text is processed completely not divided into several parts. Each major component will be described before stepping into the AIPE process.

III-4-1. Throwaway Word Dictionary

The dictionary consists of single English words that are not suitable for index terms. Throwaway words are decided at first by human beings, and the words that are once decided as throwaway words are stored in the dictionary, they are processed automatically when they appear next time in the text. The words which appear for the first time in the text will be printed out as error items with some comment such as "Are they thrown away or not, Sir?", and waits human decision.

All inflected forms are held as unique words. For instance the dictionary holds as unique lexical items the words co-ordinate, co-ordinated, co-ordinating, co-ordination. Possessive forms have been distinguished from the plural in such instances as girls, girl's. The apostrophe in men's is dropped and mens is carried. The idea is essentially that of Klingbiel [9]. All verbs, adverbs and all prepositions except "of" are thrown away. The number of throwaway words will be less than 10,000 at the level of one million words having indexed [9] and will increase nearly 23 per ten thousand words [10]. This dictionary is common to all categories. As an example of this dictionary look at Fig. 1. A complete dictionary can be made from the common word dictionary of the SMART and the Recognition dictionary of the MAI which is designed by Klingbiel.

III-4-2. <u>Word-stem Dictionary</u>

The word-stem dictionary is a kind of categorically classified thesaurus which consists of a list of word stems. They are produced automatically being assisted by the Suffix Table. (Refer to III-4-4) The dictionary will be constructed by using the words included in a typical document collection but not included in the throwaway dictionary. Each distinct word-stem will be furnished with a different number. A typical sample from a stem dictionary is shown in Fig. 2. According to the SMART experiment, a sample set of documents abstracts of some 50,000 total coverage words would typically produce a full stem dictionary of about 2,000 distinct word stems [7]. But the dictionary for the AIPE will be less than half that of the SMART because it does not contain

A	All	Any
About	Allow	Appendices
Accessible	Almost	Application
Accumulation	Already	Approach
Achieved	Also	Appropriate
Across	Alternative	Are
Actual	Although	Arise
Adequate	Always	As
Advance	Among	Assumption
Advanced	An	At
After	Analysis	Available
Again	Analyzed	Away
Aided	And	
Aims	Another	
	- - - - - - -	
	- - - - - - -	
W	What	Who
Was	When	Whose
We	Where	Why
Well	Whereas	Will
Went	Which	With
Were	While	Within
Without	X	You
Would	Y	Your
Written	Yet	Z

Fig. 1. Throwaway Word Dictionary

STEM	SUFFIX	SEQUENCE NUMBER	STEM	SUFFIX	SEQUENCE NUMBER
MOCULE	S	2099	MANIPUL	ATION	2112
SOURCE		2100	OSCILL	ICAL	2113
THICK		2101	PREV	IOUS	2114
TRUNC	ATION	2102	RECCRD		2115
WAVE		2103	RELAX	ATION	2116
WHEREB	Y	2104	REPCRT	ED	2117
WIR	ING	2105	REVERS	ED	2118
RASE		2106	SHCW		2119
CENT		2107	TREE	S	2120
DEPCSIT	ED	2108	TUNNEL		2121
FUNCTION	AL	2109	ANISOTROP	Y	2122
GRAPH		2110	CARRI	ER	2123
MAGNETIZ	ATION	2111	COMPOS	ITION	2124
			DENS	ITY	2125

Fig. 2. A Example of Word Stem Dictionary
(Partly Derived from [7])

throwaway words. A word-stem dictionary will be applied to each category.

III-4-3. <u>Phrase Dictionary</u>

A phrase is much more useful for subject identification than a single word alone. Such phrase dictionaries would then normally include pairs, triples, or quadruples of word concepts, corresponding in a written text to the more likely noun and prepositional phrases which may be expected to indicate the subject content in a given topic area. All components of phrases must be contained also in the stem dictionary. The phrases are determined by human and when there are new phrases, they are printed out as errors with a comment as same as we have done in the throwaway dictionary.

A typical example of phrase dictionary is shown in Fig. 3.

PHRASE CONCEPT	COMPONENT CONCEPTS			
543	544	608	-0	-0
282	280	281	-0	-0
282	306	281	-0	-0
280	69	648	-0	-0
280	69	215	-0	-0
694	1285	1284	-0	-0
291	265	290	-0	-0
291	265	496	-0	-0
422	646	185	-0	-0
640	309	290	-0	-0
294	21	293	-0	-0
393	21	635	-0	-0
393	635	106	-0	-0
294	21	245	-0	-0
300	306	2	281	-0

Fig. 3. Example of Phrase Dictionary

III-4-4 Suffix Table

Typically each suffix is listed with a sequence number. A representative suffix table for English suffixes may contain 368 entries [11]. But the suffix table for the AIPE is smaller than this representative one because the suffix table for the AIPE contains only noun and adjectival forms. An example of a suffix table is shown in Fig. 4. in which entries are listed from a longer one.

1001	ARIZABILITY	------	
1002	ANTIALNESS	1250	YL
1003	ARISATIONS	1251	A
1004	ARIZATIONS	1252	E
1005	ENTIALNESS	1253	I
	------	1254	O
	------	1255	S
	------	1256	Y

Fig. 4. Example of Suffix Table

To simplify the use of the algorithm, noun suffixes may be entered in the plural as well as singular forms, and adjectival suffixes may also be listed. Verb suffixes should be included only in adjectival forms such as "-ed", "-ing". As this is a fixed table, it does not need to change occasionally after once set.

III-4-5. The AIPE Process

The process is represented by a simple flowchart.

Fig. 5. The AIPE Process

Note that the text processed by the AIPE is converted to term vectors which would make easy numerical processing after indexing such as storing, searching, etc. The process will be explained with an example. The example given in this study is derived from summaries published in <u>The Journal of Spacecraft and Rocket</u> Vol. 13, 1976. All periods, colons, semicolons, and end of field marks are replaced by code 1, the preposition "of" is replaced by code 2, and throwaway words are replaced by code 1. All the codes which happen continuously after a code 1 are deleted until an algorithm in use finds the next non-throwaway word.

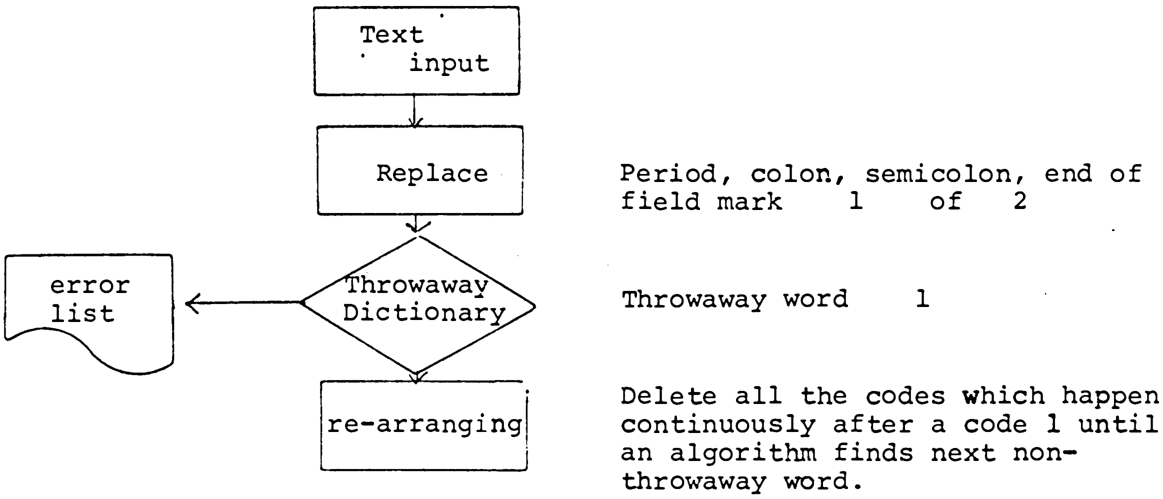

Fig. 6. Beginning Part of the AIPE Processing

Solid propellant pulsed plasma propulsion system design-- William J. Guman

```
Solid propellant pulsed plasma propulsion system design.
                                          1      1     1

Fundamental definition, a few semiempirical correlations of
1           1         1 1 1   1             1            2

experimental data, and two design constraints of solid
1            1     1   1   1      1           2

propellant-pulsed plasma thrusters are used to illustrate the
                                   1   1    1  1          1

design analysis of such an electric propulsion system.  The
1      1        2  1    1                      1     1  1

semiempirical relations presented have been generated from
1             1         1         1    1    1         1

thruster data covering impulse bits extending from 2.7 dyne-sec
         1    1                     1         1    1   1    1

(6 lb-sec) to 31mN-sec (7mlb-sec) and a specific impulse up
 11   1     1  1 1  1   1 1  1    1   1 1                1

to 5100 sec.  They are descriptive to within about 8%.
1  1    1  1  1    1   1           1  1      1     111
```

CATEGORY B

CATEGORY B means the field of mechanical engineering and this is same as that of the International Patent Classification.

Solid propellant pulsed plasma propulsion 1

solid propellant pulsed plasma thrusters 1

electric propulsion 1

thruster 1

impulse bits 1

impulse 1

CATEGORY B

All words are stemmed according to the suffix table and stem dictionary. A number of spelling rules are necessary. [7] [12]

Next are derived from [7].

1. The word matches exactly a dictionary entry.
2. It matches a dictionary entry with a final "e" dropped and a suffix beginning with a vowel added.
3. It matches a dictionary entry plus a suffix.
4. It matches a dictionary entry with a final "y" changed to "i" and a suffix added.
5. It matches a dictionary entry, with a final consonant double and a suffix added.

When several possible matches are found, the match involving the longest stem is preferred; within stems of the same length, preference is in numerical order as above. Thus, if "cop", "cope", and "copy" are all stems in the dictionary, and all normal English suffixes are included in the suffix list, "cops" is found from "cop" under rule 3; "copes" or "coping" is found from "cope" under rule 2; "copying" from "copy" under rule 3; "copies" from "copy" under rule 4; "copper" from "cop" under rule 5. Other morphological features of English are not recognized. Such word pairs as "mouse" and "mice", or "court-martial" and "courts-martial" must be entered explicitly in the dictionary if both members are to be recognized. Special rules exist which specify that all stems must be at least three letters long (to avoid, for example, finding "wing" from "we" under rule 2 or "inning" from "in" under rule 5).

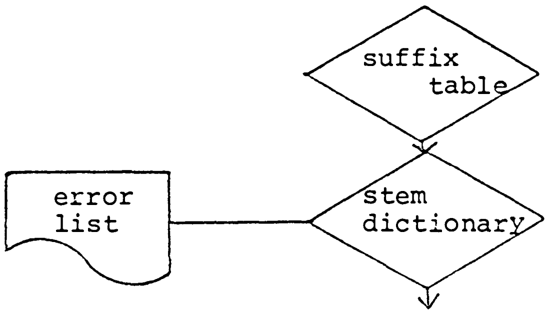

Fig. 7. Stemming Process

solid propell pulse plasma propul 1

solid propell pulse plasma thrust 1

electr propul 1

thrust 1

impulse bit 1

impulse 1

All stems are replaced to concept numbers.

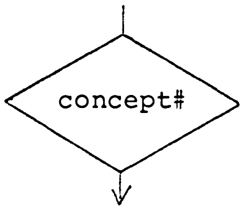

Fig. 8. Replacing to Concept Number

111	235	205	312	274	1
111	235	205	312	472	1
334	274	1			
472	1				
492	713	1			
492	1				

Phrase codes are detected according to phrase dictionary from the beginning of the first concept number to the first code 1, from the first code 1 to the second code 1, and so on to the last code 1. The algorithm finds phrases between a code 1 and the next code 1.

In this example the possible number of phrases that the machine has to examine between a code 1 and next code 1 is at most

$$\binom{5}{5} + \binom{5}{4} + \binom{5}{3} + \binom{5}{2} = 1 + 5 + 10 + 10$$
$$= 26$$

26 is relatively smally number compare with the case which has not 1 code. 5 words phrase is not detected because there is no phrase that is consisted of 5 words. Next the 5 possibilities of 4 words phrase are detected, then the possibility of 4 words phrases and so on. If one of three words phrase is found, it is replaced by phrase number. After that, only the rest of the two words are examined whether it is a phrase or not. The possible number of phrase that machine has to examine is only one. In this case, after 205, 312, 274 is detected only 11, 235 is simply examined whether this is a phrase or not. If 205, 312, 274 was not a phrase, all ten possibilities of three words phrases are examined, then two words phrases are examined. After 11, 235 is detected, only $\binom{3}{2}$ = 3 possibilities of two words phrases are examined. As there are no phrase in the rest of words in this case, finally 205, 312, 274 are printed out, as a error with a

comment that is similar to we saw in the throwaway dictionary that is "Are there any phrases in pulse plasma propul, Sir?"

If human decided it is a phrase, it is stored in the dictionary as a phrase. This phrase is not printed out next time. Making the phrase dictionary would be hard for first time but it will be getting easier and easier because the words which are printed out will be getting fewer and fewer.

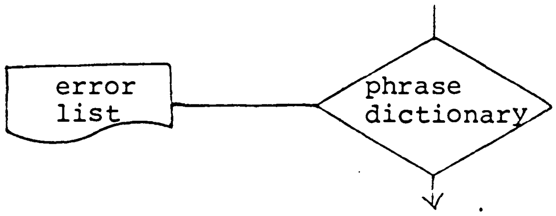

Fig. 9. Phrase Look Up Processing

	111	235	⟶	11
205	312	274	⟶	72
205	312	472	⟶	73
	334	274	⟶	25
	492	713	⟶	35

Then codes are

11	72	1
11	73	1
25	1	
472	1	
35	1	
492	1	

As the punctuations expressed in code 1 are no-longer necessary, they are deleted, and all the concept numbers are rearranged in a numerical order.

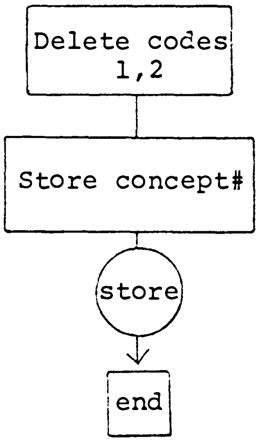

Fig. 10. Final Process of the AIPE

concept number	frequency

11	2	25	1	35	1	72	1	73	1	472	1	492	1

III-4-6. Discussion

It is difficult to predict the size of phrase dictionaries precisely but each phrase dictionary will be smaller than the throwaway dictionary. The practicality of the AIPE will come to light after a series of experiments.

References

1. Hatsumei Soran (Superintendence of inventions) 1973 "Tokkyo joho no kikai kensaku" (Information retrieval for patent examination) Japanese Governement Patent Office.

2. Bailey, M.F., Lanham, B.E., Leibowitz, J. "Mechanized Searching in the U.S. Patent Office" Journal of the Patent Office Society, 35, August 1953.

3. Lanham, B.E., Leibowitz, J., Koller, H.R. "Advances in Mechanization of Patent Searching" Journal of the Patent Office Society, 38, December 1956.

4. Andrews, D.D., Frome, J., Koller, H.R., Leibowitz, J., Pfeffer, H. "Recent advances in Patent Office searching: Steroid compounds and ILAS" Advances in Documentation and Library Science Vol. 12 Information Systems in Documentation, Interscience Publisher Inc., N.Y. 1957.

5. Information Retrieval--Patent Office, The report of the third annual meeting of the ICIREPAT, Spartan Books, Inc., Baltimore 1964.

6. Salton, G. Dynamic Information and Library Processing, Prentice-Hall, Inc., Englewood Cliffs, N.Y.

7. ed. Salton G., The SMART Retrieval System, Experiments in Automatic Document Processing, Prentice-Hall, Inc., Englewood Cliffs, N.J.

8. Klingbiel, Paul H., Rinker, C. Cathrine. "Evaluation of Machine Aided Indexing" Information Processing and Management Vol. 12, Pergamon Press 1976.

9. Klingbiel, Paul H. "Machine-Aided Indexing of Technical Literature" Information Storage and Retrieval Vol. 9, Pergamon Press 1973.

10. Klingbiel, Paul H. "A Technique for Machine-Aided Indexing" Information Storage and Retrieval Vol. 9, Pergamon Press 1973.

11. "Storage of table of fixed list of suffixes" The SMART System, Nov. 19, 1976, unpublished.

12. Salton, G. Automatic Information Organization and Retrieval McGraw-Hill 54485P 1968.