A Systematic Method for the Indexing and Retrieval of Patents Using a Faceted Approach

P. S. HUNTER
Central Laboratories, The Ever Ready Co.
(GB) Ltd., St. Ann's Road, London N15 3TJ, Great Britain

Received May 25, 1971

A thesaurus for an optical coincidence card system contains keywords which are divided into 8 facets or sections. Each facet or its subdivision contains only keywords that describe one aspect of the invention. By considering each section and subsection in turn, all the characteristics necessay for indexing and retrieving a patent are brought to mind; specific keywords are then chosen as appropriate. Thus patents are indexed and retrieved by a completely logical and standardized process.

A thesaurus of some 2000 keywords and an optical coincidence card retrieval system have been found useful for organizing a collection of scientific papers in an electric battery research establishment (I). Using the same equipment (supplied by Royal McBee Ltd.), a new system was set up which contains about 2000 British and United States patents granted within the last 20 years; new patents are added at the rate of about 3 per week. We hope to include soon our collection of foreign language patents in the system. This comprises some 1000 items and consists mainly of French, German, and Japanese patents plus some from Canada, Denmark, Netherlands, Norway, South Africa, Sweden, Switzerland, U.S.S.R., and a few from other countries. Patents of interest are found by consulting abstracting and indexing services (Chemical Abstracts Service for example), by taking ten of the subject groupings of the abridgements to British patents, and by instructing an external patent agent to supply abridgements or abstracts of relevant U.S., French, and Japanese patents, and titles of patents from the other countries listed above. This material is distributed throughout the establishment to all those interested. Any patent that is requested by a research and development scientist is put into the system. Indexing is carried out from the patent itself, and not from the abridgement. Whereas the Japanese abstracts are supplied in English, the French abridgements and the German titles are translated by an internal translator, as are the French and German patents themselves; other translation is carried out through an agency. Methods for utilizing foreign language material in the battery industry have been described previously.[2]

Before the present system was set up, the patent collection was arranged in country and in numerical order with only an author (inventor) index. Subject searches could only be carried out by going through the whole file or by using external patent indexes. An easily operated system was required that would select all patents whose principle objectives fell within a broad subject area; patents of only peripheral interest in the subject area were not wanted. For example, it may be necessary to retrieve all patents whose primary concern was seals in Leclanché cells. In short, high recall and high precision were needed in conjunction with broadly based searching. Hence, the philosophy behind the patent system is different from that behind the system for papers, where extremely detailed and/or narrow searching is possible. The depth of indexing is shallow for patents and they are not indexed exhaustively—only that which is new is recorded in detail. Keywords used for indexing patents are broader in scope than those for papers (e.g., MANGANESE DIOXIDE *vs.* MANGANESE DIOXIDE CRYSTALLINE TYPE-γ; SEPARATOR *vs.* SEPARATOR-FILM), and fewer of them are used when indexing (an average of ten for patents and 25 for papers). In this connection, a patent is usually concerned with a single invention or group of inventions, whereas a paper may cover a number of topics. The patent thesaurus contains fewer keywords (about 300). In order that patents can be indexed easily in a logical and standardized manner by a centrally based information scientist as well as by specialist research workers (*vs.* I), the thesaurus is divided into eight facets or sections, which are worked through systematically. To increase precision, the number of keywords that may be taken from each section is limited. Technically complex patents are indexed only by subject experts.

A further difference between the two systems is in their subject coverage; as patents are of a fundamentally practical nature, no keywords pertaining to theoretical concepts are included in this thesaurus. It does, however, cover equipment pertaining to batteries (flashlights for example) in addition to the batteries themselves and processes for making them covered by the first thesaurus. Another new feature is the inclusion of words relating to bibliographical details and legal aspects of patents, as found in their official headings (e.g., inventor and priority date). Because of the special nature of patents these 'Administration Keywords' are of particular importance to the system. Although strictly speaking, the name keyword should only be used in connection with the actual subject matter of a document; it is retained in this section in order to maintain uniformity throughout the thesaurus.

Many techniques have been described in the literature for the retrieval of information from patents,[3-6] but most are concerned with large collections, cover a wide subject area, use narrow terms, and index in depth or are designed for wide-spread use. What was wanted in the present case was to be able to handle a small collection, to index only to a shallow level, and above all to meet the information needs of a particular establishment. Hence, the system now in operation differs from those described previously.

The thesaurus was compiled by a small committee consisting of four subject specialists from the laboratories and an information scientist. The areas to be covered by the system were defined in consultation with research and development scientists from all parts of the establishment. Two approaches were used to select the concepts and keywords for the thesaurus. On the one hand, the thesaurus for papers was analyzed to see which parts of it were applicable to the retrieval of patents, while on the other hand, useful concepts were picked out from a representative sample of 200 patents. From time to time, draft thesauri were submitted to the full 'keydex' committee, who suggested modifications. The thesaurus was further modified after it had

[*] Present Address: School of Librarianship, The Polytechnic of North London, 207-225 Essex Rd., London, N.1. 3PN

been used by the subcommittee to keyword 50 patents. Before a final draft was produced, a sample of 200 patents were indexed, punched onto coincidence cards, and submitted to testing and subsequent modification by members of the establishment who had an interest in patents.

THE THESAURUS AND ITS USE

The thesaurus is divided into the following sections:

Section 1. Administration 'Keywords'
Section 2. Battery Equipment Keywords
Section 3. Battery Keywords
Section 4. Process Keywords
Section 5. Component Keywords
Section 6. Materials Keywords
Section 7. Method of Improvement Keywords
Section 8. Effect of Improvement Keywords

Sections 1 and 3 are further divided into subsections (see below). Within each section or subsection the terms are arranged in alphabetical order. Patents are particularly suited to this faceted approach, as they are divided into a series of logical sections (preamble, examples, claims, etc.). Indexing is carried out by working through the sections and subsections in the order given in the thesaurus, picking out not more than one keyword from each. Restricting the indexer to one keyword from each section means that he has to think very carefully what precisely is the area of the claimed improvement and what exactly is the real innovation. Hence he is better able to index the salient points of a patent than by casually scanning it. Limiting the number of keywords prevents retrieval of a large quantity of irrelevant material. In the case of a few patents, it is impossible to restrict the indexing to one word from each section, and the rule has to be broken. Nevertheless it is a very useful guideline. In subsections 3a and 3b which deal with cell types and cell systems, it is recognized that a patent may be concerned with more than one of them, and two keywords are allowed. It is usually inappropriate to choose words from every section. The general subject field covered by the patent is indexed by using words from sections 2 to 4, and the particular improvement by keywords from sections 5 to 8.

Searching is carried out similarly by working through the sections, selecting appropriate terms, or if the searcher is familiar with the system by going straight to the required cards. An alphabetical index to the whole scheme is not needed, as few ambiguities arise as to which section a particular keyword is in, and the thesaurus is small enough for difficulties to be sorted out easily.

Section 1 records information given in the official headings to the patents, and is subdivided as follows:

1a. Country of the patent 'Keywords' (e.g., GERMAN PATENT).
1b. Priority Date 'Keywords'.
  1b, 1. Decade 'Keywords' (e.g., DECADE 1970–1979).
  1b, 2. Year 'Keywords' (e.g., YEAR 1).
  1b, 3. Month 'Keywords' (e.g., APRIL).
  1b, 4. Day of the month 'Keywords' (e.g., 15th DAY).
1c. Country of original patent 'Keywords' (e.g., ORIGINATING IN JAPAN).
1d. Inventor 'Keywords' (e.g., INVENTOR EVER READY).
1e. Date received at central laboratories 'Keywords' e.g., RECEIVED BETWEEN 1.1.71 and 30.6.71).

As the priority date is indexed by the exact day, these terms may be used to trace equivalent patents. The priority date given in the examples is 15.4.71. If in a search for equivalent patents, a large number of answers is obtained, it may be possible to reduce the number of patents to be looked at by adding a 'Country of original patent' word and/or an inventor word. Before a new patent is indexed, a check is made to see if there is an equivalent already in the system. Where patents from an originating country with different priority dates have been combined to form a single patent in a second country, all the priority dates are indexed. This can lead to false drops, but these may be reduced as indicated above. As far as possible, equivalent patents are filed under the same number, but where two or more patents are divided out of an original, or where combination has taken place this may not be possible, and several files with cross references are used. This is not entirely satisfactory, but since all equivalent patents will probably be retrieved together in a subject search, this is not a major problem.

'Decade' and 'Year' words are useful in subject searches for selecting inventions made within a given time span. If patents taken out by a particular organization are required, 'Inventor' words may be employed; care must be taken to find out what names organizations patent under in different countries. The 'Date received at central laboratories' is included in the system as retrospective surveys are carried out from time to time, and, when updating these, it is desirable to know what has been added since the previous survey.

When the subject of a patent is a piece of equipment or a device pertaining to batteries, a keyword from section 2 is used (e.g., TORCHES AND LAMPS—nb. in the United Kingdom the word torches does not imply a flame, but is the equivalent of 'flashlights' in the United States.). If the type of battery used is important, this may be indexed by taking words from section 3.

The 'battery' keywords from section 3 are used if the general subject matter of the patent deals with batteries themselves. Several types of keyword are available to describe the broad area covered by a patent, and words may be taken from all of the subsections if required. The following extracts from this section give an indication of its scope:

3a (up to 2 words may be taken)	CELL TYPE FUEL CELL TYPE OTHER CELL TYPE PRIMARY	
3b (up to 2 words may be taken)	CELL SYSTEM LECLANCHE CELL SYSTEM NICKEL CADMIUM CELL SYSTEM ALKALINE (Use where a specific alkaline system is not stated, or is not in this list, or when the patent applies to many alkaline systems. Do not use for FUEL CELLS or for novel systems) CELL SYSTEM OTHER (Use only if the whole system is specified. Do not use for ALKALINE SYSTEMS or FUEL CELLS. May be used for novel cell systems)	
3c (only 1 word may be used)	CELL DESIGN CYLINDRICAL CELL DESIGN LAYER STACK	Only to be used if 1 design is particularly referred to
3d	SEALED CELL	
3e (only 1 word may be used)	ELECTRODE CONSTRUCTION PRESSED/POCKET (Not to be used for Leclanché systems) ELECTRODE CONSTRUCTION SINTERED	

3f (only 1 word may be used)	ELECTROLYTE AQUEOUS ELECTROLYTE SOLID	Use only with CELL SYSTEM OTHER, FUEL CELLS or when no system is specified

If the general subject matter of the patent deals with a process, then a keyword is taken from section 4 (e.g., EXTRUSION). Keywords may also be taken from sections 2 or 3 if the patent refers to a manufacture, as well as to a process (e.g., ASSEMBLING plus CELL SYSTEM LECLANCHE). For purely process patents, however, sections 2 and 3 will not be used. In general, words need only be taken from one of the sections 2 to 4, and in no case have words been taken from all three sections.

Sections 5 and 6 are used to describe in more detail the specific area in which the improvement is claimed. Section 5 lists specific components (e.g., BULB HOLDER, SEAL, ELECTROLYTE). The 'materials' keywords listed in section 6 (e.g., MANGANESE DIOXIDE, MnO_2/C MIX) may be used to index important materials mentioned in a process patent, or the material of a component from section 5, or in certain cases they may be used in their own right.

The actual method of improvement is indexed by taking words from section 7 (e.g., CHEMICAL TREATMENT, CONSTRUCTION/DESIGN). These words describe the practical means whereby an improvement is brought about, and the concepts they embody are often most easily picked out by looking at the claims section of a patent.

Keywords from section 8, on the other hand, are designed for indexing the aim of the improvement. Examples of these 'effect of improvement' keywords are DISCHARGE CHARACTERISTICS and PHYSICAL PROPERTIES. The aim of a patent is best found by looking towards the end of the preamble, and is often associated with the words: 'It is the object of the present invention to. . .' . When indexing, the keyword describing the most fundamental cause of the improvement is used—for example, if a patent claims that by controlling the particle size of manganese dioxide a cell with improved performance is obtained, then the keyword PHYSICAL PROPERTIES should be used, not DISCHARGE CHARACTERISTICS.

EVALUATION OF THE SYSTEM

The performance of the system may be judged by reference to two specimen searches. In the first, patents which described improvements in the design of seals in Leclanché cells were sought. The keyword cards CELL SYSTEM LECLANCHE, SEAL and CONSTRUCTION/DESIGN were superimposed for the search. Of the 17 patents retrieved, 14 were of interest. While one of the nonrelevant answers resulted from a punching error, the other two referred to sealing by means of a sheath and not by a seal as such. To determine how many relevant patents were not retrieved by the search, a series of very broad searches were carried out using the system (for example by using only one keyword card). It was found that six patents had been missed: three of these occurred because the CELL SYSTEM LECLANCHE keyword had not been used (the patents did not mention Leclanché cells, but they obviously referred primely to them). The other three patents failed to be retrieved because SEAL had been omitted at the indexing stage. In one of these cases, the omission was clearly an indexing slip as the word 'seal' appeared in the title of the patent; the other two patents were only of slight interest as they were only partially concerned with seals. These omissions highlight the need for careful indexing even when a systematic indexing process is used. The search results indicate a recall of about 70% and a precision of about 80%

The second search was for patents which describe an improvement in discharge characteristics resulting from changes relating to manganese dioxide. Six separate searches were carried out each using two keyword cards: one of the 'materials' keywords—MANGANESE DIOXIDE or MnO_2/C MIX was used in conjunction with an 'effect of improvement' keyword (DISCHARGE CHARACTERISTICS or PHYSICAL PROPERTIES or STORAGE LIFE). The number of patents retrieved is shown below (the number of relevant answers is given in brackets).

	MANGANESE DIOXIDE	MnO_2/C MIX
DISCHARGE CHARACTERISTICS	4(4)	1(1)
PHYSICAL PROPERTIES	10(4)	3(1)
STORAGE LIFE	1(0)	0(0)
Total =	19(10)	

Four relevant patents were not retrieved by the above searches. Of these, one relating to the capacity of cells at low temperatures was of only marginal interest (the 'effect of improvement' keyword used was WORKING TEMPERATURE). The remaining three were keyworded ELECTROLYTE and not MANGANESE DIOXIDE or MnO_2/C MIX (one of these was of only marginal interest). Here the indexer was not at fault as it was difficult to decide whether the improvement was concerned primely with the electrolyte or with manganese dioxide. The results of the searches show the importance of using alternative 'effect of improvement' keywords when retrieving patents. This is because only one 'effect of improvement' keyword is used when indexing (i.e., that relating to the most fundamental cause of the improvement). There are two alternatives here, either one uses many keywords and risks a large number of nonrelevant items being retrieved, or one uses a few keywords and has to carry out more than one search. You cannot have it both ways. The present search shows, however, that greater precision is obtained when using the most obvious keyword, i.e., DISCHARGE CHARACTERISTICS, than when using a keyword dealing with a related cause of the improvement *vs.* PHYSICAL PROPERTIES. In this case, the unwanted patents described manganese dioxides but not their discharge characteristics. The patent retrieved using the STORAGE LIFE keyword likewise mentioned only physical deterioration not discharge characteristics.

In general the system operates with recall ratios of between 60 and 100%, and when the searches are broad precision ratios of up to 100% occur. However, because the majority of the keywords in the thesaurus are wide in scope, on those rare occasions when narrow searches need to be carried out, a large number of nonrelevant items may be retrieved. Failure to reach 100% performance with respect to the above parameters occurs because a keyword has been left out or used in error at the indexing stage, or because there is a difference in interpretation of the significant points of a patent or of the scope of a keyword between the indexer and the searcher. Scope notes are provided, but it is impossible to cover every eventuality. Difficulty also occurs when the keywords in the thesaurus do not exactly fit the search, or when only patents on the borderline of relevance are retrieved—here it is difficult to modify the search to get what is wanted, as the patents retrieved do not suggest useful additional keywords for searching on.

An average search takes five minutes to superimpose and read off the coincidence cards, and a further 15 minutes to pull out the papers themselves. The system is used two or three times per week. As in the previous system the number of answers may be increased or decreased by deleting or adding keyword cards. Because the file is small and contains only material of interest to the establishment in which it operates, retrospective patent searches are quicker than with *Chemical Abstracts.* Moreover, within the establishment's limited field of interest, the system contains more material, and perhaps of more importance all the patents retrieved are available close at hand. The system is easier to use and gives greater precision in searches for British specifications than the British patent office classification scheme. This is because the keywords are tailor made for the type of searches which are carried out. However, if an exhaustive search is needed, particularly in a field of interest new to the establishment, the patent office scheme and *Chemical Abstracts* should also be used, as the internal system may not contain all relevant patents in this new field. Derwent patent publications and the official indexes and abridgements to foreign patents are useful when searching for overseas patents, but the latter publications are often difficult and tedious to use, and it is hoped that the inclusion of patents from other countries in the present system will make these searches easier.

The effectiveness of the presently described patent retrieval system or, for that matter, of any retrieval system depends ultimately on the contents of the file. That which is not there cannot be retrieved. In the particular case under consideration, it depends on the efficiency of the establishment in selecting patents of current interest, and of foreseeing changes in interest.

ACKNOWLEDGMENT

The author wishes to thank the directors of the Ever Ready Co (G.B.) Ltd. for permission to publish this paper, and is grateful for the support received from his former colleagues when setting up the above system.

LITERATURE CITED

(1) Hunter, P. S., "Information Retrieval in the Technology Based Industries," *Chem. Ind. (London),* **1971,** pp. 667-70.

(2) Hunter, P. S., "The Foreign Language Barrier—Stumbling Block or Stepping Stone?" *Inform. Sci.* **4,** 65-70 (1970).

(3) McGarvey, A. R., "Uniterm Index to U. S. Chemical Patents —User Evaluation," *J. Chem. Doc.* **8,** 23-5 (1968).

(4) Silk, J. A., "A Notation-Based Fragment Code for Chemical Patents," *Ibid.,* **8,** 161-5 (1968).

(5) McDonnell, P. M., "Technical Information Management in the U. S. Patent Office," *Ibid.,* **9,** 220-3 (1969).

(6) Rasmussen, L. E., and Van Oot, J. G., "Operation in Dupont's Central Patent Index," *Ibid.,* **9,** 201-6 (1969).

Computer Aided Bibliographies for Personal or Group Use

R. J. CEDERGREN
Département de biochimie, Université de Montréal, Case Postale 6128, Montréal, Québec

Received June 3, 1971

A computer-based personal bibliographic system is described. Input via punched card containing the main author, keywords, reference, a number corresponding to the reprint list, and a commentary yields outputs consisting of alphabetized listings of authors and keywords.

Recently computer technology has been developed to aid the scientist in handling the mass of printed information. For example, the Canadian National Library[1] offers a Selective Dissemination of Information (SDI) based on tapes obtained from Chemical Abstracts Service and the Institute for Scientific Information. The computer search may be done on the basis of keywords, author, or references cited by the author. The availability of these alerting systems incites the development of equally rapid and work-saving personal or group reference systems.

More recently a number of computer-based systems have been developed. One type involves a computer search of stored material along the same lines as the commercial computer searches outlined above.[2-5] Unfortunately, this system requires that a computer be available at the time one is looking for a reference. When only large computer installations are available, the time needed to make the search and receive the output must be considered in the over-all efficiency. These types of search are, however, very useful for preparing a bibliography for a new member of a research group, for example. In larger groups where a small computer is available on demand, tremendous amounts of literature can be facily handled to great advantage by this type of search.

A second type of computer treatment involves the printing of lists of references generally arranged alphabetically.[6] This method offers the following advantages: no need for a computer after the list is made, the ease of reference recovery by scanning a printed list, and no need for complex peripheral equipment. There are, however, certain disadvantages—namely, the computer and printing time in preparing the lists, which necessarily limits the number of entries to perhaps a few thousand references (thereby limiting its usefulness to smaller research groups).

This paper describes a system based on the second type of treatment. This system has been used for the past year and a half by our research group comprising 5 to 7 people. Because of our interests in a rather prolific field, t-RNA structure, sequence and evolution, the development of the described system was judged a necessity.