AN EXPERIMENT IN MECHANICAL STORAGE AND RETRIEVAL OF INFORMATION IN INTERNAL TECHNICAL REPORTS

R C Joshi
S C Rastogi
B Garudadwajan
Indian Institute of Petroleum, Dehra Dun

Observations are made on an information processing system designed and installed at the Indian Institute of Petroleum, based on the principle of keyword indexing coupled with the use of punched cards as storage medium and electric sorter for retrieval. Apart from versatility, speed and flexibility of the system, the low cost of the sorter is an advantage as compared to computers, whose costs are prohibitive for many organizations with a modest budget.

INTRODUCTION

A technical information system has been designed and installed at the Indian Institute of Petroleum (IIP), based on the principle of keyword indexing. For this purpose, the roles of A.I. Ch., E., and a thesaurus on cards prepared by the Coordination & Information Division (COIN) of IIP are used as tools to achieve syntactic and vocabulary controls, using punched cards as the medium of information storage and an electric sorter for retrieval of data and references against specific enquiries. The keywords are numerically coded in the thesaurus, and while indexing, they are converted into binary equivalents for punching. A combination of numerical code numbers and roles makes possible the association of keywords semantically in a number of ways describing the concept under search, thereby providing flexibility of retrieval by using more than one way of query formulation on the basis of likely interests of the users. Presently, the system is being utilised to index articles from scientific journals; statistical data pertaining to oil industry; bio-data of scientific and technical personnel of the Institute; and internal technical reports.

This article attempts to describe briefly our experience with the information processing system in the area of internal technical reports of IIP which embody the R & D effort ever since its inception.

SEQUENCE OF OPERATIONS

Card Design - Scanning/Digesting - Keywords Extraction - Role Assignment - Codification - Programming - Punching - Verification/Checking - Storage.

Card Design

The 80 column card has been designed as shown in fig. (1), using all the 12 rows of the card. Columns 1 to 17 record information in direct code. Columns 1 and 2 indicate the Division to which the report belongs and the author's Division respectively. The author is numerically coded and punched in columns 3 to 6. Column 7 indicates the month, while columns 8 to 11 give the year to which the report belongs. Columns 18 to 30 are left blank for future needs, if they arise.

Columns 31 to 75 are assigned for indexing the subject matter by keyword analysis and using binary code, and the indexing field is divided into 12 zones. Zones from 1 to 10 correspond to the ten roles devised by A.I. Ch. E. Two additional zones 11 and 12 consisting of 5 columns and 6 rows each, have been provided for indicating the concepts of place and time respectively to suit our purpose, e.g., numerical equivalent for Digboi from our thesaurus is 0.0.17.1. Accordingly, the numbers 17 and 1 will be punched in rows 0 and 1 respectively, in zone 11.

In zone 12 for time we utilize 4th row for date, 5th row for month, 6th and 7th rows for
the initial year and the 8th and 9th rows for the ending year of a specified year interval. A year not occurring in an interval is punched in the rows 8 and 9.

From our experience, the field required for punching space and time concepts could be smaller than that for the actual subject matter to be indexed and hence the roles XI and XII have been assigned one below the other, each with a 5 column and 12 row matrix. The design is illustrated below:

<table>
<thead>
<tr>
<th>Col.</th>
<th>Role I</th>
<th>Role II</th>
<th>Role III</th>
<th>Role IV</th>
<th>Role V</th>
<th>Role XI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows</td>
<td>Columns:</td>
<td>Columns:</td>
<td>Columns:</td>
<td>Columns:</td>
<td>Columns:</td>
<td>Columns:</td>
</tr>
<tr>
<td>31-38</td>
<td>39-46</td>
<td>47-54</td>
<td>55-62</td>
<td>63-70</td>
<td>71-75</td>
<td></td>
</tr>
</tbody>
</table>

Y X O Zone 1 Zone 2 Zone 3 Zone 4 Zone 5 Zone 11
1 2 3 4 Role VI Role VII Role VIII Role IX Role X Role XII
5 Columns: | Columns: | Columns: | Columns: | Columns: | Columns: |
| 31-38 | 39-46 | 47-54 | 55-62 | 63-70 | 71-75 |
6 Zone 6 Zone 7 Zone 8 Zone 9 Zone 10 Zone 12

**Fig. 1.**

**Role Assignment**

In order to identify the distinctive contexts of the concept-carrying keywords and appropriately retrieve desired information, roles are assigned to various keywords. The American Institute of Chemical Engineers have categorised the roles which are particularly suitable for chemistry and chemical engineering; and they are reproduced below:(2)

1. Input to a chemical reaction, physical production, operation, electrical or mathematical system.
2. Product, output, byproduct, coproduct from the reaction, production operation or system.
4. Special agent, catalyst, accelerator, stabilizer.
5. Solvent, media, environment, support.
6. Independent variable studied for its effect.
7. Dependent variable studied how it is affected.
8. Active concept, subject of study.
10. Device, material or method for accomplishment.

Expression of ideas is made by role-keyword combination, and there are many possible ways to express an idea. The efficiency of the retrieval system lies in its manoeuvrability through a multiplicity of these routes to find the same information from its store.

**Codification/Programming**

For the sake of simplicity, brevity and elegance, the numerical codes of different keywords are binarily converted and punched one below the other up to a maximum depth of
six in any role-zone. In a particular row of a given zone, a punch in the first position is binarily equivalent to 20 = 1, a punch in the 2nd position is 21 = 2, that in the 3rd indicates 22 = 4; and so on; and the last (8th) position 27 = 128. Thus any number from 1 to 255 can be punched in a particular row. Further, a concept code isolated up to a depth of 6 can be converted into the binary system in a zone determined by its corresponding role. The isolate 12, 3, 16, 4, 7, 5, for example, can be coded neatly one below the other in a role zone as illustrated below:

12, 3, 16, 4, 7, 5 can be coded either in the upper half of the card (comprising rows from Y to 3) or in the lower half (comprising rows from 4 to 9).

In the first row, 12 = 22 + 23 (4th and 5th positions)
In the second row, 3 = 20 + 21 (1st and 2nd position)
In the third row, 16 = 24 (5th position)
In the fourth row, 4 = 22 (3rd position)
In the fifth row, 7 = 20 + 21 + 22 (1st, 2nd and 3rd positions)
In the sixth row, 5 = 20 + 22 (1st and 3rd positions)

Let us take a practical example.

"Sulphur as a contaminant in Jet fuels"

'Sulphur' whose numerical equivalent from the "term file" is 12.3.78; and its Role as a contaminant is 3.

According to card design (page 120) the number will be punched in columns 47 to 54 and rows Y to 3 (positions ticked):

Using coding sheets, a programme is prepared covering different concepts embodied. For economy of storage and elimination of the problem of matching, we resort to "superimposition of keywords" for punching keywords occurring in the same role zone; and reduction of noise, if any, is achieved in the final sort by eliminating cards with punches in contiguous positions.

Punching/Verification/Storage

Cards are punched according to programme prepared, using manually or electrically operated punching machine, checked for errors, if any, and stored in cabinets designed for the card size (18.8 x 8.4 cms.).

RETRIEVAL OPERATIONS

Observations with 2 queries are described here:

Query (a):

Is there any report on the "Production of Fatty Acids from Paraffin Waxes"?

Analysis: Paraffin waxes - Role I
(12.3.1.18.1)
Fatty Acids - Role II
(12.3.50)

Retrieval Chart
List of reports available noted from the cards sorted out:

1. Synthetic fatty acids from paraffin wax (Report No. 7, 105)

2. Manufacture of fatty acids by oxidation of paraffin wax - an economic evaluation (Report No. 3, 142)

Comments

(1) Though the observations are carried out up to 7th run for sorting the concept "paraffin wax - as a raw material", it can be seen that we get available information after 2nd run and there is no need to sort further.

(2) It is not necessary to search for the concept "Fatty Acids - Production".

(3) Total time taken for machine operation is (1 minute 47 seconds + 1 minute for feeding and arranging in between) about 2 minutes only.

Query (b): Is there any report on the "Performance Evaluation of Lubricating Oils"?

Retrieval Chart

<table>
<thead>
<tr>
<th>Rows</th>
<th>Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4 47 48 50 55 56 57 58</td>
<td></td>
</tr>
<tr>
<td>5 47 48 49 55 56 57</td>
<td></td>
</tr>
<tr>
<td>6 47 52 55 60</td>
<td></td>
</tr>
<tr>
<td>7 55 57 58</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

Machine Operation (Represented graphically No. 2)

<table>
<thead>
<tr>
<th>Run</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>8th</th>
<th>9th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Taken</td>
<td>1 Min.</td>
<td>44 Secs.</td>
<td>44 Secs.</td>
<td>41 Secs.</td>
<td>38 Secs.</td>
<td>25 Secs.</td>
<td>18 Secs.</td>
<td>11 Secs.</td>
<td>4 Secs.</td>
</tr>
<tr>
<td>Total No. Cards/Card</td>
<td>590/</td>
<td>269/</td>
<td>262/</td>
<td>254/</td>
<td>237/</td>
<td>149/</td>
<td>102/66</td>
<td>66/15</td>
<td>15/15</td>
</tr>
</tbody>
</table>

(Continued)
EXPERIMENT IN MECHANICAL STORAGE AND RETRIEVAL

<table>
<thead>
<tr>
<th>Run</th>
<th>10th</th>
<th>11th</th>
<th>12th</th>
<th>13th</th>
<th>14th</th>
<th>15th</th>
<th>16th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Taken</td>
<td>4 Secs.</td>
<td>4 Secs.</td>
<td>4 Secs.</td>
<td>3 Secs.</td>
<td>3 Secs.</td>
<td>3 Secs.</td>
<td>3 Secs.</td>
</tr>
<tr>
<td>Total No. of Cards/Cards sorted.</td>
<td>15/15</td>
<td>15/15</td>
<td>15/12</td>
<td>12/12</td>
<td>12/12</td>
<td>12/12</td>
<td>12/4</td>
</tr>
</tbody>
</table>

List of reports available noted from the cards sorted out:-

1. Lubricant Performance Evaluation by Caterpillar 1-G Test Technique (Report No. 44/361).
2. Lubricant Performance Evaluation by Kirloskar AV-1 Test Technique (Report No. 44/351).
3. Lubricant Performance evaluation by Kirloskar AV-1 Test Technique (44/331).
4. Lubricant Performance Evaluation Pette AV-1 Test Technique (44/352).
5. Lubricant Performance Evaluation by Kirloskar AV-1 Test Technique (44/345).
7. Lubricant Performance Evaluation by Kirloskar AV-1 Test Technique (44/347).
8. Diesel Engine Performance with Adulterated and Sub-Standard Fuels (44/210).

Comments

(1) Runs from 1st to 7th pertain to the search for the concept "Lubricating Oils"; from 8th to 11th for "Performance"; from 12th to 15th for "Evaluation" and 16th run is for noise reduction.

(2) It became necessary to search for all the concepts analysed in the query.

(3) After 15th run a card not conforming to the retrieval chart, by seeing through the punched positions, was taken and a position 6/51 different from the retrieval chart was selected for the 16th run. Total cards sorted were 4 and rejected 8. The punches of the rejected cards were identical and since "6/51" position is not in the retrieval chart, the rejected ones are those where we get our information. This operation is called "noise reduction" or "elimination" by the principle of sorting for contiguous positions.

(4) Taking into account 5 minutes for feeding, etc., total time taken for machine operation comes to about 10 minutes.

Many row/column combinations for sorting are possible. With the help of retrieval chart and visual inspection of a few punch positions in the cards, sorting routes at random can be selected resulting in fewer operations and overall economy of time. Two such routes in respect of query(b) are given below and illustrated in graph No.3.

Route No. 2.

<table>
<thead>
<tr>
<th>Operation No.</th>
<th>Row/Column</th>
<th>No. of cards fed</th>
<th>No. of cards sorted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>6/60</td>
<td>590</td>
<td>75</td>
</tr>
<tr>
<td>2nd</td>
<td>6/62</td>
<td>75</td>
<td>48</td>
</tr>
<tr>
<td>3rd</td>
<td>7/58</td>
<td>48</td>
<td>17</td>
</tr>
<tr>
<td>4th</td>
<td>7/57</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>5th</td>
<td>7/55</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>6th</td>
<td>5/57</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>7th</td>
<td>4/57</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>8th</td>
<td>6/47</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

Route No. 3

| 1st           | 7/58        | 590              | 107                 |
| 2nd           | 6/60        | 107              | 28                  |
| 3rd           | 5/57        | 28               | 17                  |
| 4th           | 6/52        | 17               | 8                   |
| 5th           | 6/55        | 8                | 8                   |
| 6th           | 5/55        | 8                | 8                   |
MERITS AND DEMERITS

Merits:

1. In the example "sulphur as a contaminant in Jet Fuels" the integrating concepts are completely isolated by placing "sulphur" in Role-3, Jet Fuels in Role-5 & 9 and if the document contains devices like chromatography, adsorption or catalysts used in a particular medium, they can be indicated in the respective zones on a single card, while separate entry is needed in a catalogue card system for separate concepts.

2. The number of cards and ultimately the storage space is reduced to a minimum.

3. The need for maintaining the cards subjectwise or alphabetically is completely dispensed with.

4. The coordination process in cards for matching components of each integrative concept proved to be quite tedious and time consuming, which are cut short by the process of superimposition of concept constituents and binary coding.

5. Manual sorting is easy and economical if the file size is small, say upto 1000. But when the cards swell up to 10,000 and more, mechanical sorting is quicker and less tedious.

6. Computers are not necessary even for large bulk processing of cards.

Demerits:

1. Initial capital cost of about Rs. 20-25 thousands is needed for the sorter assembly.

2. Considerable insight and care is needed during the extraction of keywords role assignment, programming, punching and verification. A wrong punch means "no retrieval".

CONCLUSION

By appropriate design of software, the simple electrical sorter can form the basis of a versatile, fast and flexible, information handling system. The low cost of the electrical sorter is an advantage as compared to computers whose costs are prohibitive for many organisations. The use of simple indexing and retrieval procedure described above can be considered by organisations desiring to achieve an economical, flexible, and simple information system.

ACKNOWLEDGEMENT

The authors thank Dr. R. K. Srivastava, Head, Coordination and Information Division, IIP, and Dr. M. G. Krishna, Director, IIP, for their valuable guidance and encouragement to write this paper.

REFERENCES


GRAPH NO. 2
Reduction Factor: Query (b)

GRAPH NO. 3
Reduction Factor: Query (b) - Two more Approaches.