An Evaluation of Links and Roles as Retrieval Tools*

STANLEY M. COHEN,** CAROL M. LAUER, and BETTINA C. SCHWARTZ
Union Carbide Corporation, Linde Division, Tonawanda Laboratories, Tonawanda, New York
Received December 3, 1964

INTRODUCTION

Intuitively, Links and Roles have been considered to be valuable aids which effectively increase the efficiency and precision of retrieval operations. It is the purpose of this study to obtain some quantitative measure of their value. To accomplish this, actual inquiries were subjected to a number of different search methods. The differences in the methods involve the presence, absence, and the use in various combinations of Links and Roles. Parameters of relevance, recall, and productivity were determined and compared for all methods.

THE SYSTEM

To place this study in perspective, an outline of our system and a brief description of several of its special features are useful. The Information Indexing and Retrieval System[1] used by our laboratories was started in 1954 and now serves most of the various Union Carbide research, development, and engineering groups located at Tonawanda, N. Y. Both internal reports and items of special interest from the external literature are indexed by a staff of technically trained indexers, using the concepts of coordinate indexing. The file now contains approximately 10,000 indexed documents. Indexing is done from an authority list (glossary) of terms to which additions, and on which modifications, are made as needed. The average indexing depth is 31 terms per document and the average number of links per document is three. Term codes, role codes, link[2] designations, and document numbers are entered on worksheets by the indexers. The information on the worksheets is keypunched into standard IBM tabulating cards (Docuterm cards) which are then stored in an inverted file[3] arrangement. A sample Docuterm card is shown in Figure 1.

Retrieval questions are processed by extracting from the Docuterm file those decks which correspond to the subjects of interest to the inquirer. Each deck is sorted in order of document number on a sorter. Concept coordination is accomplished on a collator by matching Docuterm decks on document number, using any one of a number of

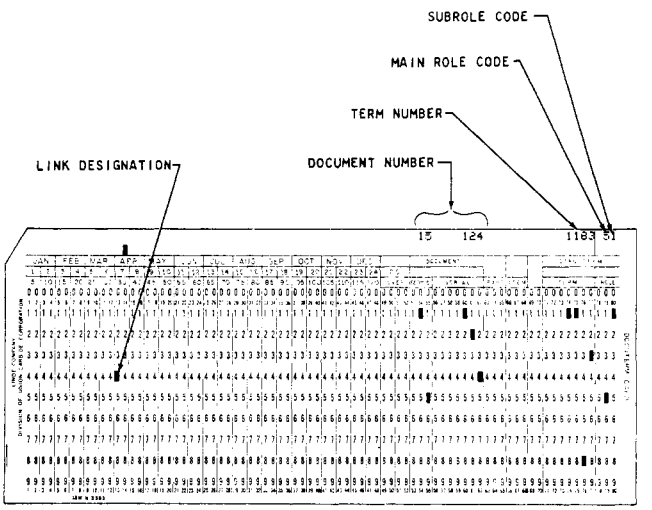

Figure 1.—Docuterm card.

possible logical operations[4] and search methods. The collating step matches Docuterm cards on the document level, after which an optical coincidence ("Peek-a-Boo") scheme is used to obtain matches at the link level. The inquirer is provided with a list of reports which satisfy the search requirements. Docuterm cards are then resorted into term number order and refiled.

Links.—All indexed items are divided into as many Links as needed to minimize the possibility of false retrieval. The number of Links used in the indexing of a document ranges from one to as many as 50 or more in some instances, depending upon the complexity of the subject matter. Each Link used is assigned a particular punch position on the Docuterm card. All Docuterm cards corresponding to the terms used in a single Link contain a common punch, which facilitates the visual matching step.

Roles.—Two digit Role indicators[5] are used in Linde's system. The Main Role (the first digit) and the Subrole (the second digit) operate independently in making distinctions in term meanings within the Link. The relationship between Main Role and Subrole is analogous to the relationship between street name and house number, in that the meaning of a term is defined, within broad limits by the Main Role, and to a more exact degree by the Subrole.

Some examples of the contextual distinctions made by Main Roles are: (1) independent variables are distin-

* Presented before the Division of Chemical Literature, 148th National Meeting of the American Chemical Society, Chicago, Ill., Sept. 3, 1964.

** Mr. Cohen is now associated with Engineering Index, Inc., New York, N. Y.

(1) F. R. Whaley, "A Deep Index for Internal Technical Reports," in "Information Systems in Documentation," J. H. Shera, *et al.*, Ed., Interscience Publishers, Inc., New York, N. Y., 1957.

(2) Document subdivisions now generally known as Links were referred to as Items in earlier works on the Linde system by Whaley.[1,3,5]

(3) F. R. Whaley, *Am. Doc.*, **12**, 101 (1961).

(4) F. R. Whaley, *Spec. Libr.*, **53**, 65 (1962).

(5) F. R. Whaley, "Operational Experience With Linde's Indexing and Retrieval System," presented at the Conference on Information Retrieval at Poughkeepsie, N. Y., Sept. 1959; published by International Business Machines Corporation (E208040).

guished from dependent variables; (2) products are distinguished from reactants; (3) adsorbates are distinguished from adsorbents; (4) materials being coated are distinguished from coating agents.

Subroles are used to make still finer contextual distinctions, some examples of which are: (1) compounds used alone are distinguished from compounds which are parts of mixtures or formulations; (2) properties for which quantitative values are given, are distinguished from discussions of concepts which lack quantitative information; (3) the adjectival uses of a term are distinguished from cases in which the term has its free meaning.

Both Main Role and Subrole codes are assigned for all term entries. Main Roles have been used since the inception of the system, and Subroles were adopted in 1959, when our retrieval experience indicated the need for a method of differentiating between fine shades of term meaning.

Role Agreement Technique.—A type of search frequently requested at our Laboratory is that dealing with properties of materials. In indexing properties of materials, the indexer assigns the same Main Role code to both the property and the material terms. Similarly, a term used adjectivally is made to agree in Main Role with the term it modifies. As an example of the former, consider the chemical reaction

$$A + B \xrightarrow{C} D \text{ (where refractive index of D is given)}$$

If

Role 1 applies to reactants
Role 4 applies to catalysts
Role 5 applies to products

then the five terms would be indexed in the following Main Roles:

Term	Main Role
Chemical A	1
Chemical B	1
Chemical C	4
Chemical D	5
Refractive index	5

When running a search of the material-property type, the actual Main Roles of the terms involved are of minor importance; but the existence of agreement in Main Role between the material and property terms is useful. In the above example, the item could not be misconstrued to contain information on the refractive index of Chemicals A, B, or C, since the combination of any one of these three chemicals with refractive index would not yield the required Role Agreement. The combination of Chemical D and refractive index is the *only combination* which satisfies the Role Agreement requirement.

When the Role Agreement technique is employed in retrieval, the Main Role code is considered (for machining purposes) part of the document number. All Docuterm decks are then sorted and collated on the basis of this enlarged document number. Only cards which match in both document number and Main Role are considered hits. Cards which match in document number but not in Main Role (which otherwise would be considered hits) are rejected by the collator. The Role Agreement technique groups terms within a Link, as a Link groups terms within a document. This technique might also be called "Sublinking."

Search Methods.—To measure the value of Links and Roles, searches requested by members of our professional staff were subjected to a number (from four to seven) of different search methods (strategies). The differences in the seven methods involved the use (or non-use) of Links and Roles and various combinations thereof. All seven methods included the coordination (collation) of at least two Docuterm decks. The seven methods are described below.

Method 1. Links and Roles Not Utilized.—The appropriate Docuterm decks were first sorted and then collated on document number. The documents corresponding to matched pairs of Docuterm cards were considered hits.

Method 2. Links Utilized.—The Docuterm decks were sorted and collated on document number. Cards which matched in document number were then visually matched ("peek-a-booed") to obtain the hits at the Link level. *Note*: This procedure for using Links (in Method 2) is the same for all subsequent methods.

Method 3. Links and Main Roles Utilized.—Prior to the sorting of Docuterm decks into document number order, at least one Docuterm deck was sorted on Main Role code. The Docuterm cards which contained Main Role codes applicable to the search requirements were separated from the others and then processed as in Method 2.

Method 4. Links and Subroles Utilized.—Method 4 was the same as Method 3 except that Docuterm deck division was based on Subrole rather than on Main Role.

Method 5. Links, Main Roles, and Subroles Utilized.—Both Main Role and Subrole divisions of Docuterm decks were used.

Method 6. Links and Role Agreement Utilized.—The Role Agreement technique was explained earlier.

Method 7. Links, Role Agreement, and Subroles Utilized.—Method 7 was the same as Method 6 except that Docuterm deck division based on Subroles was used.

All questions asked of our Information Retrieval Group in a particular time period, except those which met one of the four following disqualifying conditions, were included in this study: (1) questions which did not involve Concept Coordination (collation); (2) questions which resulted in a number of hits (matches) so high as to render the determination of relevance impractical within our time limitations; (3) questions which resulted in no answers; (4) questions which were answered through the use of conventional library methods.

Each question was subjected to as many of the seven search methods as applicable. The determination of the applicability of search methods to specific questions was based upon the relationships between our indexing rules and the subject matter of the questions. Search Methods 1 and 2 were used for all questions, while Methods 3 through 7 were used only as they applied.

The number of hits and the processing times were recorded for each search method used in a question. Processing time included time spent sorting, collating, and visually matching cards. Documents corresponding to all hits were examined to determine their relevance, and the number of relevant documents was recorded for each method used in a question.

The following parameters were then calculated for each method used in a question:

Relevance

$$\text{Relevance} = \frac{\text{number of relevant documents found}}{\text{number of documents found (hits)}} \times 100$$

For each of the more sophisticated search methods (methods using Links and/or Roles) used in a question, the ratio of its relevance to the relevance of Method 1 (the least sophisticated method) was calculated.

Relative Relevance

$$\text{Relative relevance} = \frac{\text{relevance } (n)}{\text{relevance } (1)}$$

where (n) refers to any one of the seven search methods, and (1) refers to Method 1.

Relative Recall

$$\text{Relative recall} = \frac{\text{number of relevant documents found in Method } n}{\text{total number of relevant documents found in Method 1}} \times 100$$

Because there was no practical way of measuring "absolute recall," it was decided to use relative recall; *i.e.*, the number of relevant documents found in the more sophisticated search methods (Methods 2 through 7) were compared with the number of relevant documents in Method 1, instead of being compared with the total number of relevant documents in the entire collection, the latter being impractical to measure.

Since Method 1 is the method of least constraint, the number of documents found in all subsequent methods was necessarily the same as the number of documents found in Method 1 or some fraction thereof.

Productivity

$$\text{Productivity} = \frac{\text{number of relevant documents found}}{\text{processing time (min.)}}$$

Productivity Ratio

$$\text{Productivity ratio} = \frac{\text{productivity } (n)}{\text{productivity } (1)}$$

Table I contains a compilation of averages of the above parameters for all search methods.

Table I
Summary of Results

Search method	No. of questions	Av. relevance, %	Av. relative relevance	Av. relative recall, %	Av. productivity, answers/min.	Av. productivity ratio
1	33	51.4	1.00	100.0	0.857	1.00
2	33	75.5	1.73	98.0	0.549	0.720
3	15	89.3	3.32	96.1	0.717	1.54
4	10	93.0	1.91	90.0	1.02	0.927
5	13	96.5	3.89	90.8	0.819	1.78
6	19	79.7	1.56	88.5	0.610	0.689
7	14	94.0	1.85	80.7	0.803	0.789

RESULTS AND DISCUSSION

The relevance values obtained by use of the more sophisticated search methods (methods involving the use of Links and/or Roles) were substantially higher than those obtained in the least sophisticated search method (Method 1). The search methods which involved the maximum utilization of Links and Roles (Methods 5 and 7) resulted in the highest relevances. In almost all cases, the relevance of Method 5 was equal to or greater than the relevance of Method 3, which in turn was equal to or greater than the relevance of Method 2, etc. In other words, relevance increased as the use of Links and Roles increased (see Figure 2).

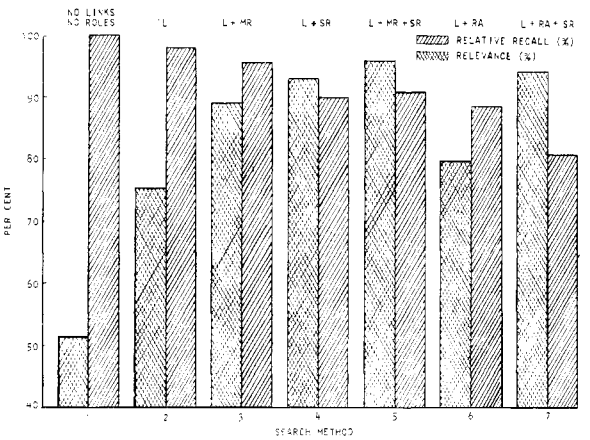

Figure 2.—Average relevance compared with average relative recall: L, Links; MR, Main Role; SR, Subrole; RA, Role Agreement.

The introduction of search methods which use Links and Roles tends to decrease relative recall. Links alone (Method 2) have only a small effect on this decrease, but some answers were lost as a result of using Roles. This effect is greatest in Method 7, where about 20% of the relevant answers found in Method 1 were lost. It appears that the use of Roles opens additional "degrees of indexing freedom"; this results in a somewhat higher level of indexing inconsistency. The average loss of answers (as compared to Method 1) in the methods involving Roles is about 10%, and is primarily the result of these indexing inconsistencies.

The price one pays for the higher relevances obtainable through the use of Links and Roles is a decrease in recall. As shown in previous work by Cleverdon,[6] the tendency toward an inverse relationship between relevance and recall is evident in Figure 2.

Figure 3 shows the average *relative relevance* of each of the seven search methods. Method 6 had the lowest average relative relevance of the six methods involving Links and/or Roles, but still had relevances which averaged 1.56 times those of Method 1 for the same question. The relevances of Method 5 averaged almost four times the relevances of Method 1 for the same question.

In Question No. 4, Method 1 yielded a relevance of 13.8% (far below the average of 51.4% for Method 1). When Question 4 was processed by Methods 5 and 7, the resulting relevances were 100 and 66.7%, respectively. This example (and similar ones) indicates that, regardless

(6) C. W. Cleverdon, "Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems," ASLIB Cranfield Research Project, Cranfield, England, Oct. 1962.

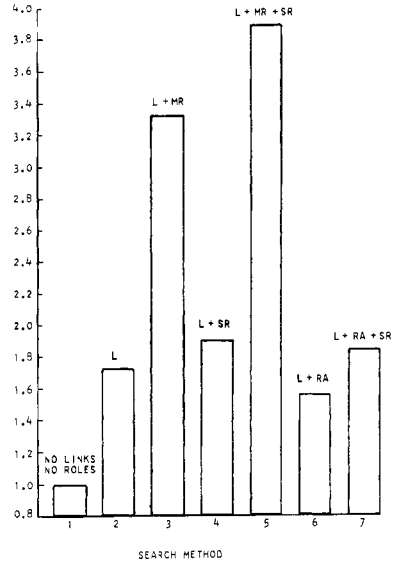

Figure 3.—Average relative relevance
(key to abbreviations in caption to Figure 2).

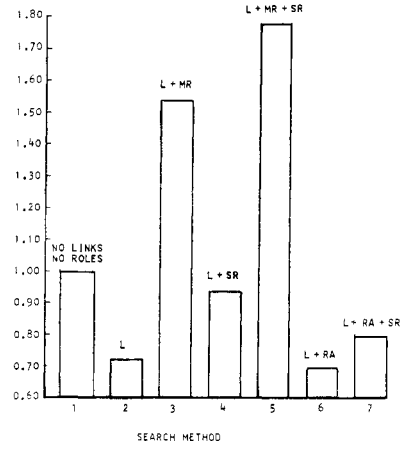

Figure 4.—Average productivity ratios
(key to abbreviations in caption to Figure 2).

of the actual relevance level of a given question, the relevances of the more sophisticated search methods are generally several times higher than those of a search method not involving the use of Links and/or Roles.

The productivity (relevant answers/minute) for individual questions varies widely and is dependent upon such factors as (1) the size of Docuterm decks, (2) the number of hits, and (3) the search logic employed. Therefore, comparisons of the ratios of the productivities of Methods 2–7 to the productivity of Method 1 (for the same question) is more meaningful than comparisons of the actual productivities of the various methods. Figure 4 shows the average productivity ratios for each search method.

The use of Links involves a manual step in the Linde retrieval procedure. This tends to have a negative effect on productivity. The use of Roles in retrieval reduces the number of Docuterm cards and therefore saves machining time. However, one or two extra sorter passes are required to make use of Roles. The main factor affecting the relative productivities of search methods with and without the use of Links and Roles is the extent to which the size of the original Docuterm decks are reduced as a result of division by Role. If the machining time saved by use of the smaller decks is sufficient to overcome that of the extra steps required for the use of Links and Roles, productivity is increased. The productivities of Method 5, which involves the use of Links, Main Roles, and Subroles, average 1.78 times the productivities of Method 1 for the same questions.

It is important to point out that, in our system, for any given search, a number of options are available with respect to the use (or non-use) of Links and/or Roles. For example, if an inquirer requests a "leave no stone unturned" approach to a search, we must be primarily concerned with obtaining high recall. We would then probably select Search Method 2, which uses Links but not Roles, and is characteristically high in recall. In this type of search we would usually be willing to sacrifice some of the higher relevances of the search methods using Roles to gain the higher recall. The other extreme is the situation in which the inquirer is satisfied to receive a few good (relevant) answers. In this case, we would be less concerned with obtaining high recall and more interested in obtaining high relevance; we would therefore select either Search Method 5 or 7, both of which utilize Links, Main Role, and Subrole. This flexibility of method adds significantly to the intrinsic value of Links and Roles as retrieval tools, in that the selection of a search method for a particular question can be based upon the stated objectives of user. In this way the most beneficial aspects of all seven search methods are utilized as individual search requirements indicate their use.

CONCLUSIONS

The use of Links and Roles in retrieval tends to increase relevance and to decrease recall. In the majority of searches, the time saved for users (in scanning lists of reports) as a result of the decidedly higher relevance characteristic of search methods involving the use of Links and Roles, more than compensates for slight losses of pertinent information and/or increases in processing time which occur in some cases.