Jason Farradane
Peter Gulutzan
University of Western Ontario, Canada

A Test of Relational Indexing Integrity by Conversion to a Permuted Alphabetical Index

Farradane, J., Gulutzan, P.: **A test of relational indexing integrity by conversion to a permuted alphabetical index.**
In: Intern. Classificat. 4 (1977) Nr. 1, p. 20–25.
Relational indexing has been tested for any distortion of meaning in the course of the indexing by constructing a computer program for conversion of the diagrams to a permuted alphabetical index. The methods involve 'translation' of the diagrams into English sentences which include all words and interrelations in each entry, without any alternations of any kind to the words or their interrelations. The results have fully confirmed the expectation that relational indexing does not distort the meaning of the subject matter in any essential way.
(Authors)

1. Introduction

The method of information storage and retrieval based on relational indexing (1), (2) has now been tested in two large indexing projects, and has given very satisfactory results. The first project covered some 1000 documents (abstracts) on sugar technology; a condensed report was submitted to O.S.T.I., London, but, owing to various organizational difficulties, was never published in full; the results have however been briefly described in another paper (3). The second project (with Dr. P.A. Yates-Mercer) covered some 3000 documents in the field of metal properties (from abstracts in Metals Abstracts); the results were again very encouraging (4).

An important matter relating to any form of storage of information for retrieval is the degree of distortion of meaning which may arise in the method of coding; any such distortion will of course tend to increase the amount of noise or false drops arising during searches. Boolean connectors ('and', 'or' or 'not') are very limited in expressing the meaning between two terms, and will usually be incapable of expressing any exact interrelation. The nine relations (with possible negations of them) used in relational indexing (see also subsequent section), have been shown to express any desired meaning between two terms, and to be sufficient for all situations. One test for any distortion of meaning would be the reconversion of the coded subject back to the original language form by automated procedures, i.e. through discrete, purely mechanical steps. If this is possible, it also provides a simple and efficient means of producing permuted indexes, though this was not the primary aim of the investigation. For any good indexing system, the title of a document is usually inadequate, and the original statements to be indexed have been prepared from reading the whole paper, or the abstract in a good abstracts journal. In relational indexing procedure such a statement is converted to the two-dimensional index form of words interconnected by relations. For retrieval, a similarly analyzed question statement is matched, for both terms and relations, with the index entries, or parts of such entries. In the large-scale tests, the major sources of error were human (indexer or user), but there were some unidentified failures; the possibility of distortion of meaning in the indexing was not however then investigated.

2. "Relational Indexing" [1]

The basis of relational indexing is the expression of detailed information by the interposition, between the words of a given subject, of relational operators which are derived from the psychology of thinking. The standard operators replace the varied prepositions and even verbs by which the terms are connected in language. Thinking mainly involves two mechanisms: (1) three stages towards complete (or fixed) association, and (2) three stages towards complete distinction (or discrimination) between terms. Their combination yields nine *categories* of relation, as shown in the following table.

	Awareness	Temporary Association	Fixed Association
Concurrent	/θ Concurrent	/* Self-activity	/; Association
Not distinct	/= Equivalence	/+ Dimensional	/(Appurtenance
Distinct	/) Distinctness	/- Action	/: Functional dependence

The names given to the categories of relation are arbitrary and only for identification. The symbols are chosen for ease of typing, with some mnemonic value.

All the descriptors (nouns and verbs) representing a given subject are first determined, usually in the form of an English sentence; these descriptors are then interconnected by a suitable operator between pairs of words. More than one word may be connected to another, by different operators; if two words require the same operator, they are both connected by the one operator. A diagram is thus formed by chains of words with interposed operators; for clarity of representation, two-dimensional diagrams are often necessary, and then the 'direction' of the relation from a word above to a word below is, by convention, the same as from left to right. The operators have a direction in the sense that the second word (to the right, or below) is in some sense subordinate to the first, in the same way as subordination (greater intension) is implied in classification.

The meanings of the operators, derived from consideration of the pairs of 'mechanisms' involved, and also by analysis of informational situations, is most easily learned by examples.

The *concurrent relation* /θ expresses the recognition of the mere co-existence of the two terms, without

any other definite interrelation, e.g., chemistry/θ dictionary.

The *self-activity relation* /* expresses intransitive action, e.g., bird/*migrating. It also expresses the dative case situation, e.g., rabbit/*food (/- feeding), (feeding food *to* a rabbit).

The *association relation* /; expresses a fixed mental association, e.g., cathedral/; beauty, or the agent of a process, e.g., etching/; acid.

The *equivalence relation* /= expresses *some* degree of identity, e.g., leaves/= manure (leaves *as* manure).

The *dimensional relation* /+ expresses position, time or state, e.g., building/+ London, sleeping/+ night, or water/+ freezing point.

The *appurtenance relation* /(expresses the generic relation, e.g., genus/(species, or the whole-part relation, e.g., bicycle/(wheel, or fixed physical properties, e.g., wood/(density.

The *distinction relation* /) expresses difference alone, as for imitations or substitutes, e.g., pearl/) synthetic pearl.

The *action relation* /- expresses any action of the second term upon the first, e.g., anvil/- hammer (though the word for the action is usually included, e.g., anvil/- striking/; hammer. The second word is thus usually a verb, best used in the gerund form. The first word may also be a verb, e.g., decomposing/- retarding.

The *functional dependence relation* /:, also possibly to be considered as cause and effect, expresses the first word causing the second, or the second arising out of the first, e.g., author/: book, or compound/: derivative.

There are also some special, and more subtle, meanings of some of the relations. Whereas /(expresses fixed physical properties, /+ is used for temporary or variable properties, e.g. apples/+ 2 lbs. Where the physical properties are not intrinsic, but humanly calculated or inferred, the relation /; is used, e.g., steel/; thermal conductivity. Whereas /- denotes present action, past action (the results of which may be considered as 'fixed') is expressed by /;, e.g., potatoes/- washing for present action, but potatoes/; washing (washed) for past action. Future action is expressed by /θ. It will be realized that other subtle distinctions can be made, e.g., between an operation such as sugar/- crystallizing (being made to crystallize) and sugar /* crystallizing for the self-process.

The equivalents of Boolean connectors occasionally arise; these are not of the same type and imply the connection 'and' or 'or' between two terms as the *same* position (not between two terms in a chain), e.g.

alloy/({ copper / zinc (alloy containing copper *and* zinc), or

cake/([butter / margarine (cake containing butter *or* margarine).

The Boolean 'not' would be incorrect if applied to a noun (thing); it is correctly used only to negate a *relation*, and is expressed by a bar over the relational symbol, e.g., cake not containing butter becomes cake/(butter.

The application of these relations in more complex subjects will be seen in the diagrams for the examples indexed. The conversion to an alphabetically indexed form in natural language was undertaken to show that relation indexing did not distort the original meaning. The diagram form can be used directly for information retrieval.

3. Computerization

A computer program has now been written which converts the diagrams into the form of a permuted alphabetical index, each entry of which consists of complete sentences in English starting with each desired 'lead' term and using all the rest. The relational signs are translated into standard prepositions or, in one or two cases, prepositional phrases (e.g. effect on; affected by), by algorithm. The relational diagrams prepared by the indexer have, in previous work, always been checked by 'reading' them as a statement, and this was found quite easy after a little practice, but the unconscious means of doing this had not been analyzed. Careful comparison of the original statements and the relational diagrams for a number of documents of all kinds now revealed how such a standard list of appropriate translations of the relations into prepositions (in some cases 'no preposition') could be established. There are however eight possible interpretations of each relation, according to whether, in a triad of 'word'–'relation'–'word', the word in each position is a noun or a verb, so that the possible triads are N–R–N, N–R–V, V–R–N, and V–R–V (where N = noun, V = verb, and R = relation), and also according to the *direction* in which the triad is read (left to right, or right to left, to which, by convention, the downward or upward directions are equivalent) within the diagram. For this purpose, only words having a *direct dynamic meaning* are to be described as verbs; such words have, as far as possible, been written in the first place in the gerund form ending in 'ing'. In a few cases, common usage calls for a verbal noun, e.g. analysis (of), rather than analyzing. No other verbal implications are read into other words, and other words (other than occasionally needed adjectives) are all treated as nouns.

It was also found, by careful tests of 'reading' a large number of diagrams, that standard rules could be devised for the order in which the words should be read, starting at any required 'lead' word in the diagram.

A typical diagram of a complex subject is:

Research
/;
Projects /: Documents /- Copying /; Proposal
/; /; /=
Supporting Copyright /- Infringement
/;
U.S.A. /(Government

This subject was: "A proposal that copying for research should not be an infringement of copyright of documents from projects supported by the government of the U.S.A." (J. Amer. Soc. Inf. Sci., 1974, *25*, 145).

The rules are that reading should go from the first (lead) word towards the left (if possible), and on reading the last word in that direction jump back to the right of the starting word and proceed to the right as far as possible. If any words occur above or below any word reached in the main line, then, after reading the main line word, the word(s) below are read, then the word(s) above, before proceeding along the line, in either direc-

tion. If there is a chain of words above or below, the whole chain is read before continuing along the line. If a ring of words occurs, this is always to be read clockwise, except that if the lead word is in the top (main line) of a ring, the main line is read from right to left before proceeding round the ring. When a last unused term in a ring is reached, the reading jumps to the right of the ring.

Some diagrams contain certain other complexities. If there are multiple relations to one word, e.g.

the secondary terms are read from bottom to top (with any associated chains), the diagrams being suitably prepared in the first instance. If two or more words are connected by *one* relation to a first word, e.g. $A/(\begin{matrix}B\\C\end{matrix}$, this means A containing B with C, but the Boolean connectors (which do *not* occur between words *along* the chain) are written as:

$A/(\left\{\begin{matrix}B\\C\end{matrix}\right.$, which is the Boolean 'and' (A contains B and C)

or $A/(\left[\begin{matrix}B\\C\end{matrix}\right.$, which is the Boolean 'or' (A contains B or C).

These are treated as 'associated' words in the program.

Another special case is when adjectives are needed; adjectives may be significant for the indexing, in which case they are printed immediately after the lead term; if non-significant, they are held to the end of the entry, and then inserted, starting with a capital letter and followed by '–,' to take the reader back to the beginning. The usual type of alphabetical index is produced, with the lead term in the first line, and subsequent terms, indented, in the second line; significant adjectives are followed by a comma. Elsewhere, adjectives will precede the word concerned.

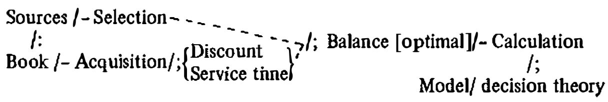
4. Connection table

The problem was now whether these processes and rules of permutation were generalizable and algorithmizable, and to find an input format which, like the diagrams, would be capable of being prepared, and of being 'read' by the indexer, for checking. The diagrams present an obvious analogy with structural formulae of organic chemical compounds; such chemical diagrams are commonly represented by means of a 'connection-table' which lists each atom and its appropriate triads of atom-bond-atom. This method was found particularly suitable for the relational indexing diagrams. The connection tables are easy to prepare from the diagrams, are easy to check, and form satisfactory computer input. The indexer is in fact doing a little of the work which might otherwise have been quite complex in getting the computer to take note of multiple connections and rings. The indexer's task is quite simple. The words in the diagram are first listed, preferably in order of interconnection across the diagram, from left to right, with an indication in the next column as to whether each word is a noun or a verb. Each triad present, of word-relation-word, is then entered in successive lines of the table, starting the triads at the left end of the main line of the diagram. The words are represented by the numbers which the words receive against the list position. The relations are given numbers in the order of the usual relational indexing presentation, i.e.

/θ 1	/* 4	/; 7
/= 2	/+ 5	/(8
/) 3	/- 6	/: 9

Certain additional information is now entered. Associated words (words occurring at the same position) are entered in a column marked A, followed by 1 for 'and', 2 for 'or', or 3 for 'with'. For each of such words the appropriate triad will have been recorded, and all the other words at the position will be entered as associated words to each triad (line of the table). Words required in the lead position in the eventual indexing are noted by a figure 1 in the lead column against the words. If a word is a compound of a noun and an adjective, or a verb and an adjective or adverb, a significant adjective is entered after a slash (in place of a comma, which is needed in the normal use) placed after the word in the first word list, and a non-significant adjective is put after the word, in square brackets. Two further instructions were found necessary. After more than two associated words at one position, or after a ring has been completed, it is useful to instruct the computer to execute a jump to the next appropriate triad, so a GO TO column is provided in which the appropriate next line number is entered. Finally owing to the peculiarities of English style, the standard preposition is not always satisfactory; it would not be wrong, but would read clumsily. For example, one would say 'I am staying *on* the top floor *at* a hotel *in* the town'; each preposition has the same implication of position (the relation /+), but habit prescribes different prepositions in different circumstances. Similarly, the statement 'analyzing a mineral' (which takes no preposition) has to become 'analysis *of* a mineral', if the commoner verbal noun form is used (the choice is arbitrary). A 'written-in preposition' space (**WIP**) is therefore provided in each line on either side of the relation. An asterisk * indicates 'no preposition'. The final connection-table of the diagram given above is then as in Fig. 1. It may be noted that "/; supporting" means the past tense, so 'supported' is entered in the table.

Another example, which has adjectives and associated words, is shown in fig. 2. The relational diagram is:



All the word-relation-word triads are entered in the table in the order left-to-right, or top-to-bottom, as they appear in the indexing diagram, and this fixes the noun-or-verb order of the two words, whichever direction the triad is to be 'read' by the computer.

5. The program

The most accessible computer on the campus of the University of Western Ontario is a DECsystem10, a respectable 'timesharing' machine with on-line access, geared to

N

REFERENCE NUMBER: 25145

RELATIONAL INDEXING
CONNECTION TABLE

WORD No.	V N or V 1 or 2	S WORD	LINE No.	G A-GO-TO	L W-1 LEAD 0-1	W WORD-1	A 1 = "and" 2 = "or" 3 = "with" AW-a	AW-b	AW-c	P WIP-1	R RELATION	P WIP-2	L W-2 LEAD 0-1	W WORD-2	A 1 = "and" 2 = "or" 3 = "with" AW-a	AW-b	AW-c	G B-GO-TO
1	1	projects	1			1				*	7	*		2				
2	2	supported	2			2					7	*		3				
3	1	government	3			4					8		1	3				
4	1	USA	4	1		1					9		1	5				
5	1	documents	5			5					6		1	6				
6	2	copying	6	7		7					7	for	1	6				
7	1	research	7			6					11			8				
8	2	infringement	8			9					6	of		8				
9	1	copyright	9	4		5					7		1	9				
10	1	proposal	10			6				:	7			10				
11			11															
12			12															
13			13															
14			14															
15			15															
16			16															
17			17															
18			18															
19			19															
20			20															

Fig. 1

N

REFERENCE NUMBER: 25227

RELATIONAL INDEXING
CONNECTION TABLE

WORD No.	V N or V 1 or 2	S WORD	LINE No.	G A-GO-TO	L W-1 LEAD 0-1	W WORD-1	A 1 = "and" 2 = "or" 3 = "with" AW-a	AW-b	AW-c	P WIP-1	R RELATION	P WIP-2	L W-2 LEAD 0-1	W WORD-2	A 1 = "and" 2 = "or" 3 = "with" AW-a	AW-b	AW-c	G B-GO-TO
1	1	sources	1		1	1					6			2				
2	2	selection	2			2					7			3				
3	1	balance [optimal]	3	5		3					6	of		4				
4	2	calculation	4			4					7	for	1	5				
5	1	model/ decision theory	5			6	17				7	of		3				
6	1	discount	6			7	16				7	of		3				
7	1	service time	7	9		8					7	for	1	6	17			
8	2	acquisition	8	9		8					7	for	1	7	16			
9	1	book	9	10		9					6	of	1	8				
10			10	1		1					9		1	9				
11			11															
12			12															
13			13															
14			14															
15			15															
16			16															
17			17															
18			18															
19			19															
20			20															

Fig. 2

academic use. It was decided to program in COBOL, despite that language's awkwardness when textual data must be manipulated character by character, because it was felt that such a standard industry language would be more generally applicable to the varying situations and environments which might be faced, and this program is also being considered as a forerunner to a full information retrieval program. The problems involved were not simple, firstly in finding an adequate input format, and secondly in converting what seems intuitively simple in the writing of English into mechanical procedures, without any alterations in the terms used or in the relations between them; this last condition is essential for judgment of the integrity of the relational indexing. Detailed discussions were needed between one of us (J. F.), as indexer and 'analyst', and the other, as programmer, for elucidating the necessary procedures and devising the detailed flow-chart. Even so, problems of special cases (more than one of the same type of associated word present, both an associated word and adjectives held over for the end of the entry, and acronyms ending in a period, which the program would otherwise take as the end of a sentence, etc.) arose several times in the course of tests. The final program provides the indexer with readable input, so that it can be checked by him (as well as by the machine for input typing errors); it also incorporates all necessary mechanisms for 'smoothing out' the punctuation of the English in the printout, and is reasonably efficient. An example of the input format is shown in Fig. 3.

```
n=25145
v=1;s=projects
v=2;s=supported
v=1;s=government
v=1;s=USA
v=1;s=documents
v=2;s=copying
v=1;s=research
v=2;s=infringement
v=1;s=copyright
v=1;s=proposal
w=1;p=*;r=7;p=*;w=2
w=2;r=7;p=*;w=3
w=4;r=8;l=1;w=5
g=1;w=1;r=9;l=1;w=5
w=5;r=6;l=1;w=6
g=7;w=7;r=7;p=for;l=1;w=6
w=6;r=11;w=8
w=9;r=6;p=of;w=8
g=4;w=5;r=7;l=1;w=9
w=6;p=:;r=7;w=10
```

Fig. 3

The program operates on the connection table in the following way: Starting at the first unused lead-word from the top, the line is read from this word across to the other word, which may be in either direction as necessary; at each step a tally is kept of the words so 'used'. Subsequent lines to be read must always contain one already-used word and one unused word. The lines are first searched up the table, and each is read from the used word to the unused word, the relation being interpreted according to the direction of reading. The list of relation words is held in the working-storage section of the program. Written-in prepositions override the standard relation which follows in the order of reading. The computer skips any line which contains either no used word, or two used words. On reaching the top of the table, the reading continues downwards. On reaching the bottom, the table is searched upwards for previously unused lines, and finally downwards again. All terms are thus eventually used once. Whenever the top left word of the table is reached as the *second* word read in the triad, i.e. from right to left, the program inserts a period followed by two spaces, and starts the next word with a capital letter. If GO TO is indicated in the table, the instruction is followed (in the first two directions of search), and the direction of reading up or down the table, as produced by the GO TO instruction, is continued to the top or bottom, as necessary, disregarding the previous direction of search. Non-significant adjectives or words 'associated' with the *lead* term are held in temporary storage for entry at the end of the final index statement. Such a final adjective is followed by "–,", to indicate that the sentence is to be continued at the beginning (in accordance with well-known alphabetical indexing practice). Associated words to the lead word are entered at the end, followed by "and –,", "or –," or "with –,", as suitable, but the "–," is suppressed if there is also an adjective to follow at that position; further rules deal with the case when there is more than one such associated word. Words associated with non-lead terms are inserted in their suitable places in the sentence of the entry, with the Boolean connector term before the word. When the first entry has been completed, the program repeats the process for the next lead word, until all have been used. Finally, all entries are sorted alphabetically, first for the lead word, and secondly for the sublead word (second word of the entry, disregarding prepositions). Proper punctuation and use of capital letters is introduced as necessary. The output places the lead word on a first line, and subsequent words on a second line (or more), with appropriate indentation. Cross references can be introduced in connection-table form. The width of the printout is held to a narrow format of 50 characters. Two computer 'pages' from the output from the indexing of 90 documents (articles in J. Amer. Soc. Inf. Sci.) are reproduced in Fig. 4.

6. Results

The results have fully confirmed the expectation that relational indexing does not distort the meaning of the subject matter in any essential way. A small number of the entries show slightly stilted English, without ambiguity of meaning, and this could be overcome by refinements in the program. As an alphabetical index, it appears easily readable and searchable, and it is of course as complete as required, according to the number of terms marked as lead-words. This format of alphabetical index, and its return of reading from held-over words, is considered the best for ease of locating and reading

Goffman
epidemic theory of growth of literature. And contagion theory by Menzel comparison.
24343
Government
of USA supported projects. : documents copying for research not as infringement of copyright : proposal.
25145
Hash coding
model for storing in computer of data. And retrieving by non-unique search key.
25232
High-energy physics
see Physics, high-energy.
IBM research laboratory
in San Jose, California interactive Negotiated Search Facility experiments for assisting subject indexing.
24009
Index
see also Subject index.
Index terms
for documents retrieval performance affected by weighting or clustering of -,
24246
Indexing
see also Subject indexing.
of articles in serials. Affected by duplicating by Chemical Abstracts Service or BIOSIS or Engineering Index. Abstracting or -,
24025
documents and questions for information retrieval performance affected by levels of exhaustivity of -,
24313
automated, of documents. By probabilistic model of distribution of words.
25312
automated, of documents. By significant key phrases selecting by algorithm.
25237
automated, of documents. Model. Access control with -,
25162
questions and documents for information retrieval performance affected by levels of exhaustivity of -,
24313
for teaching at University of Strathclyde in information science. By students. Collective -,
24054
Information
on physics, high-energy, exchanging by international invisible college identification by sociometric techniques.
25113
studying by Shannon communication theory and Cherry semiotics.
24242
Information technology
development affected by cutting of budget for 1975 of NSF-OSIS. Information science and -,
25077
Information theory
derived measures for measuring semantic information content of documents and surrogates. Coding theory and -,
24300
International Documentation in Chemistry
TOSAR system of graph-theoretic representation of relations between concepts and concepts.
25287
Invisible college
international, exchanging information on high-energy physics. Identification by sociometric techniques.
25113
J.C.I.
see Journal Citation Index.
Journal
see Serial.
Journal Citation Index
derived from Science Citation Index for procedure of clustering scientific serials.
24425
Keyword
see also Descriptor.
Keywords
of Uniterm system of information retrieval for small information centers in National Institute for Road Research in South African CSIR. Specifiers geographical terms and -,
24180
KWIC index
documents retrieval evaluation with subject headings from subject index of Engineering Index.
24282
Language
natural, metalanguage for systematic research on human communication and its construction by Computer Assisted Language Analysis System (CALAS) at Ohio State University.
25058
Languages
foreign, literature. Problems experience in multidisciplinary laboratory : Oakridge National Laboratory.
25080
for programming interactive systems for information retrieval with data bases of library, scientific & business data : DIRAC designing.
24287
Lawrence Livermore Laboratory
in University of California : system of single generalized file from data bases merging.
24045

Fig. 4

entries, but a different lay-out could equally well have been produced. The sublead words are those directly related to the lead word and are logically those most *usefully* reached after finding the lead word. All the terms, and their interconnections, are reproduced in each entry. In very complex diagrams, some tests have been made of using only parts of the diagram for certain leads words, and this seems a feasible procedure in some cases in order to reduce the length of the eventual entries. This method of alphabetical indexing will be compared with other existing methods in another paper (J. Docum., in the press).

References

(1) Farradane, J.: Concept Organization for Information Retrieval. In: Inform. Stor. Retr. 3 (1967) p. 297–311.
(2) Farradane, J., Russell, J. M., Yates-Mercer, P. A.: Problems in Information Retrieval: Logical Jumps in the Expression of Information. In: Inform. Stor. Retr. 9 (1973) p. 65–77.
(3) Farradane, J. E. L.: Analysis and Organization of Knowledge for Retrieval. In: Aslib Proc. 22 (1970) p. 607–616.
(4) Yates-Mercer, P. A.: Relational Indexing applied to the Selective Dissemination of Information. In: J. Docum. 32 (1976) No. 3, p. 182–197.

1 (Editor's note) This section was originally not included in the paper. We asked the author for it on behalf of all those newcomers to the field, who have not as yet encountered Prof. Farradane's operator matrix and its explanation in the literature.