A FIBONACCI-RELATED SERIES IN AN ASPECT OF INFORMATION RETRIEVAL

MICHAEL F. LYNCH
University of Sheffield, Western Bank, Sheffield, England S10 2TN

A continuing objective of research in the field of information science is a better understanding of the structure of subject indexes, and of methods of preparing and using them. Most of us depend on these tools for access to the steadily increasing flow of publications in science and technology, yet for the most part their preparation is still an art rather than a science. It was not a little surprising, therefore, to discover that a familiar linguistic device that is widely used in indexes, catalogs, and directories could be formalized, and that this formalization had connotations which included a Fibonacci-related series. The linguistic device is that of inversion of prepositional phrases, such as "England, Kings of," which is encountered in such diverse sources as back-of-the-book indexes and the Library of Congress catalog.

The process of inversion of phrases reaches its peak in complex subject indexes such as those to Nuclear Science Abstracts and Chemical Abstracts, the latter currently including about 300,000 scientific papers, books and patents each year. The magnitude of the task of publishing and searching such amounts of literature has called for the increasing application of computer technology during the past decade, and it was in the context of one such investigation that the process of inversion came to be more clearly scrutinized [1]. In these indexes, entries are made under a series of subject headings, which serve as the primary entry points for the user. The entries themselves consist of prepositional phrases, highly convoluted, but organized in such a way as to enable the reader to scan them rapidly and to extract the essential content during a rapid scan of the entry. The following example, taken from a recent index to Chemical Abstracts, illustrates the point (the numerical reference is the abstract number):

Coal
    flotation of, hydrocarbon agent activity in, oxygen compd. formation in relation to, 89893W.

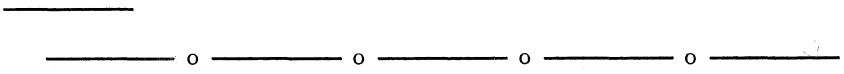
It is clear that, without particular training, the reader can reconstitute the sense of the original phrase as it was first conceived by the indexer. This is an intuitive process, not immediately formalizable. With computer techniques in view, however, it was necessary to define the procedure in symbol-manipulative terms. It was noted that the entries consisted, in the main, of sequences of phrases either beginning or ending with prepositions, and it was this which provided the necessary clue. In the case of an entry such as "England, Kings of," it is clear that the natural order would read "Kings of England," while if the entry read

"Kings, of England," no alteration in sequence would be required. So too with highly complex index entries, provided that the constituent phrases can be suitably identified. Fortunately, this delimitation is provided by the sequence of commas within the entry, which usually serve to separate the component phrases from one another. Thus, extending the rule which gives us "Kings of England," we can say that if we take the component phrases of an entry in sequence, then, according as the phrase begins with a preposition (or connective such as "and"), or ends with one, it is to be placed so as either to precede the subject heading or to follow it, as the case may be. Applying this to each component in turn, and adding successive phrases at one end or the other of that part of the sequence built up always produces the intended result, i.e., the normal form of the description as originally derived by the indexer. In practice, the rule cannot be applied to all entries, since commas may also occur in the normal form of the expression; however, for those entries in which each component phrase either begins or ends with a preposition or other function word, the rule is absolutely consistent, and is illustrated by its application to the example noted above:

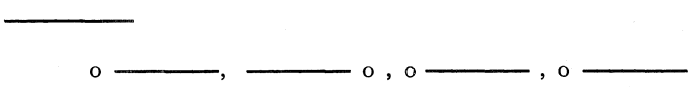
"oxygen compd. formation in relation to hydrocarbon activity
in flotation of coal."

While interesting, this formalization has not yet been widely utilized in computer studies of index structure. Its usefulness seemed to us to lie rather in the fact that its obverse offered the possibility of taking natural language title-like phrases, and automatically producing an index of high quality from them. This reverse transformation, from natural language phrase to index entry, presented particular problems, since it became apparent that it produced not a single result, but rather a variety of possible forms of entries, that is, that while the transformation from entry to the normal form of the description is single-valued, the transformation from normal format to entry is many-valued. This became clear while the selection rule for entry production was being elaborated — a process which the indexer carries out intuitively, and which has now been termed articulation.

It is useful at this point to consider a simple model for these transformations. The model necessarily ignores certain complexities which are encountered in practice, notably those due to the proportion "of," as illustrated below. It consists of a formalized descriptive phrase composed of a sequence of nouns or noun phrases separated by function words:



An entry consists of an articulated form of these, in the following fashion:



in which the pairs of function words/nouns form the components of the entry. The selection rule is as follows. A noun or noun phrase is selected to act as a subject heading from any

position in the sequence. As a result, equal numbers of nouns/noun phrases and function words remain. The entry may then be formed by successive selection of components from positions adjacent to the subject heading, either to the right or to the left of it, a kind of decision tree resulting from the multiplicity of choices that are open. The following example illustrates the point:

rains on plains in Spain

Heading: Plains

1st Component:

Plains	Plains
rains on,	in Spain

2nd Component:

Plains	Plains,
rains on, in Spain	in Spain, rains on

Heading: Spain

1st Component:

Spain	Spain
plains in	rains on

2nd Component:

Spain	Spain
plains in, rains on	rains on plains in

The complication caused by the preposition "of" can be illustrated by the following example:

"production of indexes by computer;"

when "indexes" is selected as the subject heading, two entries are provided by the simple model:

Indexes	Indexes
by computer, production of,	production of, by computer

Of these, only the second is acceptable, the first seeming ill-formed, due to separation of the phrase "production of" from the noun which it qualifies directly. In practice, this can be accommodated by simple additional rules.

Again, in practical terms, economic factors, both of production and of size of the resulting index for users, do not permit the inclusion in a printed index of all of the variant forms of entry which the model permits. Further characteristics of printed subject indexes, including the use of indentation to enhance the ease of scanning of the printed display, have enabled us to adduce further rules which are now incorporated within a useful program suite for the automatic production of printed subject indexes [2, 3]. The advantages of this technique are that the indexer need concern himself solely with providing an accurate and consistent

record of the content of the subject matter of the document being indexed, and can economize on the time needed to make an entry under each heading in articulated form, which is required in the traditional index-production method.

It is nonetheless interesting to pursue the implications of the simple model somewhat further, particularly in terms of the great variety of variant entries which can be formed from a single title-like phrase describing the subject content of an article or book. It is clear that if the first noun or noun phrase of a longer description is chosen as the subject heading, only a single form of entry is possible. Taking the earlier example:

"rains on plains in Spain"

when "rains" is selected as the heading, only a single form of entry is possible, i.e.,

Rains
on plains in Spain .

This is termed an invariant phrase. When the last noun, Spain, is chosen, either of the nouns preceding it may form the first component of the entry, while if a noun occurring at an intermediate position is selected, the first component can be formed from any of the nouns preceding it or from the one following it. Using a different symbolism, in which the components are denoted by alphabetical symbols, a sequence of three nouns can, in theory, give rise to the following entries:

A · B · C		
A	B	C
BC	AC	AB
	CA	BA

A sequence of four noun phrases, A. B. C. D can produce a greater variety:

A	B	C	D
BCD	ACD	ABD	ABC
	CAD	BAD	BCA
	CDA	BDA	CAB
		DAB	CBA
		DBA	

Tabulating these graphically for phrases of lengths 1 to 4 provides the following possibilities:

No. of headings	Phrase	Possible Entries			
1	A	A			
2	A. B.	A_B	B_A		
3	A. B. C.	A_{BC}	B_{CA}	C_{BA}	
4	A. B. C. D.	A_{BCD}	B_{ACD} CAD CDA	C_{ABD} BAD BDA DAB DBA	D_{ABC} BCA CAB CBA

Replacing now the particular articulated arrangements by the numbers of variant entries possible under each heading in turn, we obtain the following table:

n	No. of entries under n^{th} heading							Total
	1	2	3	4	5	6	7	
1	1							1
2	1	1						2
3	1	2	2					5
4	1	3	5	4				13
5	1	4	9	12	8			34
6	1	5	14	25	28	16		89
7	1	6	20	44	66	64	32	233

This series proves to be of more than casual interest. Not only are the row sums the alternate terms of the Fibonacci series, the internal structure of the table also provides an algorithmic extension, other than by an exhaustive examination of all the possibilities provided by the selection rule. Thus any entry in the table may be computed by taking the entry above it, and adding to it the entry immediately to the left of it and all those on the left-hand diagonal of the latter.

Finally, a general expression for computing the row sums for each value of n takes the following form:

$$a_n = \frac{1}{\sqrt{5}}\left[\left(\frac{\sqrt{5}+1}{2}\right)^{2n-1} - \left(\frac{1-\sqrt{5}}{2}\right)^{2n-1}\right] = F_{2n-1}\,.$$

ACKNOWLEDGEMENTS

Financial assistance from the Office for Scientific and Technical Information, London, in support of this work is gratefully acknowledged, as also the capable help of J. E. Ash, J. H. Petrie, I. J. Palmer and M. J. Snell in the computational work. Dr. I. J. Good provided the expression for the row sum series.

REFERENCES

1. M. F. Lynch, "Subject Indexes and Automatic Document Retrieval: The Structure of Entries in Chemical Abstracts Subject Indexes," J. Documentation, 22 (1966), pp. 167-185.
2. J. E. Armitage, M. F. Lynch, J. H. Petrie and M. Belton, "Experimental use of a Program for Computer-Aided Subject Index Production," Information Storage and Retrieval, 6 (1970), pp. 79-87.
3. M. F. Lynch, J. H. Petrie, "A Program Suite for the Production of Articulated Subject Indexes," Computer Journal (in the press).

LETTER TO THE EDITOR

Dear Editor:

Professor Dr. Tibor Šalát of Bratislava has pointed out two corrigenda to my article on arithmetic progression, April, 1973, Fibonacci Quarterly, pp. 145-152.

In the proof of Lemma 2.2, one may not assume that ad and $c/(a,c)$ are relatively prime. After the second display in the proof, proceed as follows:

$$(i - i')ad \equiv (j' - j)bc \pmod{c}$$

$$(i - i')ad \equiv 0 \pmod{c}.$$

Since $(c,d) = 1$, we get $(i - i')a \equiv 0 \pmod{c}$. Division by (a,c) yields

$$(i - i')(a/(a,c)) \equiv 0 \pmod{c/(a,c)},$$

hence

$$i - i' \equiv 0 \pmod{c/(a,c)}.$$

On page 151, insert a "1 -" before $\prod$ in the second, third, and fourth displays.

How far can Theorem 4.1 be generalized to other polynomials?

Sherman K. Stein
University of California,
Davis, Calif. 95616