SP-1505

RANK ORDER PATTERNS OF COMMON WORDS AS DISCRIMINATORS OF SUBJECT CONTENT IN SCIENTIFIC AND TECHNICAL PROSE

Everett M. Wallace

April 1964

SP-1505

SP *a professional paper*

RANK ORDER PATTERNS OF COMMON WORDS AS DISCRIMINATORS OF SUBJECT CONTENT IN SCIENTIFIC AND TECHNICAL PROSE

April 1964

Everett M. Wallace

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA

A-1162 REV. 4/64

ABSTRACT

There is a style of language characteristic of different subject areas which is particularly noticeable in scientific and technical writing. It is not only the unique vocabulary of a subject field which sets it apart from others, but also the different habits of writers in using the most common words. An experiment was devised to test whether these differences could be used for subject discrimination in addition to identification of unique vocabulary, particularly to determine whether or not author variation in style is sufficiently great to override the variation from field to field.

Fifty IRE abstracts in the field of electronic computers and fifty Psychological Abstracts were matched, one abstract at a time, one word type at a time, against two lists of words ranked in descending order of frequency as they occurred within two different sets of three hundred psychological and computer abstracts. All fully inflected forms of all function and content words were included in the rankings. Using the first 50 ranks only of the two lists, 93% of the abstracts were successfully discriminated. For the first 75 and 100 ranks, the success rates were 96% and 97%, respectively.

RANK ORDER PATTERNS OF COMMON WORDS AS DISCRIMINATORS OF SUBJECT CONTENT IN SCIENTIFIC AND TECHNICAL PROSE

Introduction

There is little reason to be satisfied with current information system designs either for dissemination or retrieval. The use of condensed representations in the form of class categories or index terms has limitations. Systems using such devices appear, inherently, to produce a great deal of "noise," as can be seen in the recent work on relevance/recall ratios. Whole text or "natural language" processing approaches appear to offer the greatest promise of improvement in retrieval systems. The designers of prose processing schemes, however, have encountered serious difficulties in building systems which are both practical and economical.

A major problem in working with natural language is the range of variation in linguistic behavior. The wide range of variation has been an obstacle to successful predictive generalization, whether applied to mechanical or human information storage and retrieval. One reason for the current difficulties is that we do not have a sufficiently precise knowledge of the stochastic parameters of language, particularly as it is used in different subjects and contexts. A second reason is that efforts directed at statistical techniques of linguistic analysis have concentrated upon the relatively infrequent verbal constructs.

It has been a common practice in building language processing programs to reduce the number of different entities which must be handled by excluding the most common articles, prepositions, conjunctions and auxiliary verb forms, and by combining inflected forms of common roots. Such procedures do result in the loss of a certain amount of information. Through reading the reports of G. Yule and G. Herdan and of F. Mosteller and D. Wallace in establishing the authorship of disputed works, I was led to consider ways in which this lost information could be recovered and used to supplement established methods. G. K. Zipf had already shown one way of using rank order distributions of words. Others have indicated that there is a considerable range of variation in the way individual authors use the most commonly occurring words in a language in different contexts.

There is a style of language characteristic of different subject areas which is particularly noticeable in scientific and technical writing. It is not only the unique vocabulary of a subject field which sets it apart from others, but also the different habits of writers in different fields in using common prepositions, nouns, and verbs. This is most clearly illustrated in mathematical writing, in which symbology is embedded in a highly stylized form of prose, sufficiently unlike ordinary language to be considered a distinct dialect. The growth of "dialects" in this sense is common to all subjects in varying degrees. The question is whether these behavioral differences are sufficiently distinctive to provide a basis for subject discrimination in addition to the identification of unique vocabulary.

One of the first considerations in estimating whether a practical discriminator could be built was whether or not author variation in style is sufficiently great to override the variation from field to field. An experiment was devised to test this proposition and to gather evidence for identification of statistical parameters and techniques useful for subject discrimination.

The Experiment

An experimental corpus was selected consisting of 350 Psychological Abstracts and 350 IRE Abstracts from the Transactions of the Professional Group on Electronic Computers (PGEC). The abstracts were available at System Development Corporation in machine-readable form.* This corpus was considered to provide an adequate reflection of author variation, in that the abstracts had largely been written by different persons, including authors of the papers abstracted.

Three hundred psychological abstracts and three hundred PGEC abstracts were taken from the corpus for establishment of population "profiles" of the two subject areas. The profiles consisted of two lists of the most frequent 100 words ranked in descending order of occurrence within the two sets of 300 abstracts. A System Development Corporation computer program called FEAT was used to provide the counts and listings. The Appendix presents a consolidated alphabetic list of the words in the two profiles, together with their rank numbers.

* The abstracts were drawn from the experimental sets used originally by Borko for automatic classification and by Maron for automatic indexing.

Where occurrence frequencies of two or more words were equal, a word length criterion was applied such that the shorter word was given the higher rank. This was based on the assumption that, in general, short words are more prevalent than long. When word length as well as frequency were equal, the words were ranked in alphabetic order.

A version of the FEAT program was used to count and list the words in each of the 100 abstracts remaining in the experimental corpus of 700. Each abstract was matched, one word type at a time, against the two profiles of 100 rank-ordered words. The words in each abstract occurring in one or both of the two profiles were recorded, together with their rank numbers.

The purpose of this procedure was to segregate the abstracts into two files--psychological and PGEC abstracts, respectively. After considering a number of decision rules, the following criteria were adopted:

1. An abstract belongs to psychology if the number of words in common with the psychology profile is greater than the number in common with the PGEC profile, and conversely.

2. If the number of words in common in the abstract and the two profiles were equal, the sum of the rank numbers of those words on the two lists would be determined, and the abstract assigned to the profile with the smaller sum. If the sums were equal, no decision would be made.

Figures 1 and 2 illustrate the data recorded and the results of matching two abstracts against the first 50, 75, and the full 100 ranks of the two profiles. In both cases the number of words in the abstracts contained in the first 50 ranks of the two profiles is the same. Summing the rank numbers permits both abstracts to be correctly discriminated by the rule given.

The following table summarizes the results of matching the psychological and PGEC abstracts against the first 50, 75, and 100 ranks of the profiles:

	Number Correctly Discriminated for		
	50 Ranks	75 Ranks	100 Ranks
50 Psychological Abstracts	43	46	47
50 IRE PGEC Abstracts	50	50	50
Success Ratio	93%	96%	97%

All of the abstracts which were cast into the "wrong" category by this procedure were psychological abstracts. Examination of the abstracts contributing to the profiles suggests several reasons for this. The PGEC abstracts represent a more specialized subject matter than those from Psychological Abstracts. In general, the PGEC abstracts contain fewer word types used more frequently. Consequently the counts contributing to the PGEC profile are higher than those of psychology.

In examining the results it was found that, at the 100 rank level, 88% of the successfully discriminated abstracts were dependent on the 52 words that are unique to each profile, with 9% successfully decided through summing

PSYCHOLOGICAL ABSTRACT # 1 - 54 word types

Word in Abstract	Psych. Profile			PGEC Profile		
	50R	75R	100R	50R	75R	100R
a	6			3		
and	3			4		
be	17			13		
but	-	63		-	-	-
by	14			14		
first	-	74		-	-	-
have	-	56		-	69	
information	-	-		40		
in	4			7		
is	7			5		
of	2			2		
on	13			16		
the	1			1		
to	5			6		
were	18			-	-	-
with	9			17		
No. words in common	12	15	15	12	13	13
Rank no. sum	99			128		

Figure 1

IRE PGEC ABSTRACT # 1 - 15 word types

Word in Abstract	Psych Profile			PGEC Profile		
	50R	75R	100R	50R	75R	100R
are	1			9		
automatic	-	-	-	-	-	80
be	17			13		
considered	-	-	-	-	-	85
data	-	-	80	37		
may	50			-	-	91
of	2			2		
or	21			27		
that	12			19		
no. words in common	6	6	7	6	6	9
Rank no. sum	110	110		107	107	

Figure 2

the rank numbers. It was considered useful to investigate the discrimination to be obtained by the rank sum criterion alone, using only words common to the profiles.

There are 48 words in common on the profiles in the first 100 ranks. Figure 3 lists the words in common and their ranks. The mean difference of rank for these words is 17.4, with the lower ranks tending to larger differences than the higher ranks. As can be seen from the figure, function words predominate. The following table shows the results of matching the 100 abstracts against the list of 48 words common to the profiles and applying the rank sum criterion:

	Correct	Incorrect
50 Psychological Abstracts	36	14
50 IRE PGEC Abstracts	42	8
Percentage	78%	22%

Conclusions

The results of this experiment indicate that author variation in style imposes no serious obstacle to using patterns of common words as discriminators. Considering the length of the profiles, the small size of the sample contributing to the profiles, and the limited number of word types contained in individual abstracts, the success ratios are surprisingly high. It is uncertain, however, to what degree the results are biased by editorial conventions and style.

Rank Psych	Rank PGEC	Word	Rank Psych	Rank PGEC	Word	Rank Psych	Rank PGEC	Word
6	3	a	56	69	have	43	70	other
16	12	an	4	7	in	58	59	presented
3	4	and	91	64	into	52	88	problems
8	9	are	7	5	is	45	67	some
11	20	as	23	24	it	71	48	such
39	43	at	49	90	its	82	18	system
17	13	be	50	91	may	29	49	than
14	14	by	44	22	method	12	19	that
100	23	can	72	63	methods	1	1	the
80	37	data	30	66	more	33	44	these
34	21	discussed	90	53	new	19	25	this
84	54	each	94	50	number	46	62	time
10	8	for	2	2	of	5	6	to
96	68	function	13	16	on	36	61	two
69	94	general	65	41	one	20	11	which
64	58	has	21	27	or	9	17	with

Mean difference in Rank = 17.4

Figure 3

Rank Numbers of the 48 Words in Common in the First 100 Ranks of Psychological and IRE PGEC Abstract Profiles

The results also tend to support the idea that there is much useful information to be found in the high frequency area of word occurrence, and that frequency alone can provide a basis for subject discrimination of widely different fields, particularly when all word type occurrences of fully inflected forms are taken into account. Further work is required to establish the precision which may be expected of such a technique, especially if applied to fields more closely related than psychology and computers.

Potential Applications

A system designed to make use of common word patterns through a technique similar to that described in this paper would include a short table intended to combine the functions of an exclusion list with identification of broad subject areas. Such a quick initial segregation would reduce the search time required for matching against the particular vocabulary of those areas. Figure 4 illustrates the contrast between using a large dictionary with the familiar features of exclusion lists, root stripping and an extended search of a long table and the approach suggested here. The initial segregation would lead directly to a relatively short specialized dictionary or to a mis-match monitor. The thesaurus devices necessary to a large dictionary could be simplified, and the range of ambiguity inherent to terms used in many different fields would be narrowed. It is quite feasible to use specialized tables now, provided the texts are segregated by subject prior to input. This approach, however, looks forward to the application of optical readers for the transformation of printed text to machine readable form in systems that do not require the intervention of a human mind for prior subject classification.

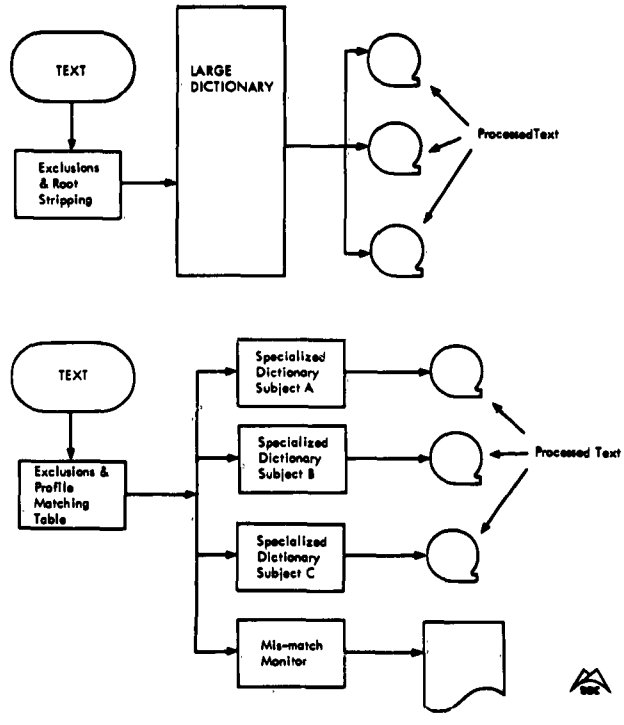

Figure 4

Schematic Flow Contrasting a Conventional Technique with Suggested Approach Using Common Word Patterns

(Page 12 blank)

REFERENCES

Borko, H. The Construction of an Empirically Based Mathematically Derived Classification System. Proc. Spring Joint Computer Conf., vol. 21, pp 279-289, 1962.

Herdan, G. Type-Token Mathematics. 'S-GravenHage, Mouton & Co., 1960.

Maron, M. E. Automatic Indexing: An Experimental Inquiry. J. ACM, vol. 8, no. 3, pp 404-417, 1961.

Mosteller, F. and D. L. Wallace Inference in an Authorship Problem. J. Am. Stat. Assn., vol. 58, no. 302, June 1963.

Yule, G. V. A Statistical Study of Vocabulary, Cambridge U. Press, 1944.

Zipf, G. K. Human Behavior and the Principle of Least Effort. Addison-Wesley, 1949.

APPENDIX

The Profiles

The 300 Psychological Abstracts used to build the rank-ordered profiles for this experiment contained a total of 22,175 word occurrences of 4,587 word types. The 300 IRE PGEC abstracts contained 23,200 word occurrences of 3,678 word types. The mean number of word occurrences per abstract was 77.3 for PGEC versus 73.9 for Psychology. When broken into subsets, both samples exhibited a broad internal range of variation for the expectation that a given word would appear at a given rank, with the broader range appearing in the Psychological Abstract set.

The following table presents a consolidated alphabetic list of words occurring in the first 100 ranks of the IRE PGEC and Psychological Abstract Profiles, together with their rank numbers. A dash (--) is used instead of a rank number to indicate that the word does not occur in the first 100 ranks of one or other of the profiles.

Word Type	Rank Number Psych.	PGEC	Word Type	Rank Number Psych.	PGEC
a	06	03	circuit	--	46
all	99	--	circuits	--	34
an	16	12	computer	--	10
analog	--	42	computers	--	45
analysis	42	--	considered	--	85
and	03	04	control	--	56
any	--	65	counseling	87	--
are	08	09	data	80	37
as	11	20	described	--	15
at	39	43	design	--	36
author	66	--	development	38	--
automatic	--	80	differences	98	--
be	17	13	different	97	--
been	--	77	digital	--	26
behavior	27	--	discussed	34	21
between	22	--	during	75	--
binary	--	86	each	84	54
both	--	97	effect	37	--
but	63	--	effects	57	--
by	14	14	electronic	--	60
can	100	23	elements	--	87
change	92	--	equations	--	76

Word Type	Rank Number Psych.	PGEC	Word Type	Rank Number Psych.	PGEC
factors	68	--	one	65	41
findings	95	--	only	85	--
first	74	--	operation	--	55
for	10	08	operations	--	96
form	--	92	or	21	27
found	53	--	other	43	70
from	--	32	out	--	99
function	96	68	output	--	84
functions	--	47	part	73	--
general	69	94	perception	77	--
given	--	35	performance	83	--
group	32	--	personality	52	--
groups	76	--	possible	--	98
has	64	58	presented	58	59
have	56	69	problem	--	75
his	55	--	problems	51	88
human	81	--	program	--	51
in	04	07	programming	--	83
information	--	40	psychological	78	--
input	--	100	psychology	59	--
into	91	64	reinforcement	89	--
is	07	05	relationship	70	--
it	23	24	required	--	89
its	49	90	research	47	--
language	--	71	response	54	--
learning	31	--	results	35	--
logic	--	93	set	--	72
logical	--	52	shown	--	73
machine	--	30	social	25	--
magnetic	--	33	solution	--	79
may	50	91	some	45	67
means	--	74	storage	--	57
memory	--	28	study	28	--
mental	93	--	such	71	48
method	44	22	switching	--	39
methods	72	63	system	82	18
more	30	66	systems	--	38
network	--	95	technique	--	81
new	90	53	techniques	--	82
no	79	--	test	40	--
not	24	--	than	29	49
number	94	50	that	12	19
of	02	02	the	01	01
on	13	16	their	61	--

Word Type	Rank Number Psych.	PGEC	Word Type	Rank Number Psych.	PGEC
theory	26	--	used	--	29
these	33	44	using	--	78
this	19	25	various	86	--
time	46	62	visual	67	--
to	05	06	was	15	--
two	36	61	were	18	--
under	41	--	when	60	--
use	--	31	which	20	11
			with	09	17

UNCLASSIFIED

System Development Corporation,
Santa Monica, California
RANK ORDER PATTERNS OF COMMON WORDS AS
DISCRIMINATORS OF SUBJECT CONTENTS
IN SCIENTIFIC AND TECHNICAL PROSE.
Scientific rept., SP-1505, by
E. M. Wallace. 22 April 1964, 15p.,
6 refs., 4 figs.

Unclassified report

DESCRIPTORS: Information Retrieval.
Documentation.

States that there is a style of language characteristics of different subject areas which is particularly noticeable in scientific and technical writing. Also states that the unique vocabularly of a subject field and the different kinds of writers sets it apart from others.

UNCLASSIFIED

UNCLASSIFIED

Reports that an experiment was devised to test whether these language differences could be used for subject discrimination in addition to identification of unique vocabulary, particularly to determine whether or not author variation in style is sufficiently great to override the variation from field to field. Also reports that 50 IRE abstracts in the field of electronic computers and fifty Psychological Abstracts were matched against two lists of words ranked in descending order of frequency as they occured within two different sets of three hundred psychological and computer abstracts. States that using the first 50 ranks of the two lists of abstracts and words, 93% of the abstracts were successfully discriminated and for the first 75 and 100 ranks, the success rates were 96% and 97%, respectively.

UNCLASSIFIED