VIII. THE EXTENDED SMART SYSTEM

Tom Evslin

1. Introduction

A new version of SMART has been designed as an extension and an improvement of the old system. Various limitations inherent in the old system have been removed, provisions have been made for additional options which are of interest in information retrieval, and several logical changes were made to increase the capacity of the system and to speed up processing.

2. Limitations Removed

The SMART system described in Information Storage and Retrieval, Report No. ISR-7, is designed to process not more than 500 documents of short abstract length. The number of documents which can be processed during a single run decreases as the number of words in the average document approaches the maximum text length of 1,000 words. In the new system, there is no limit to either the number of documents which can be processed in a single run, or the number of words included in a single document.

In the old system, the maximum number of word stems which can be included in the thesaurus is 1,000. Up to 30,000 noncommon stems may be included in the extended system. This number approximates more closely a useful technical vocabulary. In addition, 2,767 concept

numbers are reserved for common words, 32,000 numbers for mechanically generated word clusters, and 32,000 concept numbers each for authors, journals, and bibliographic entries.

3. New Options

A new program has been written to provide inexpensive semi-syntactic phrase searching. This program weights statistically found phrases according to the probability that the word elements actually constitute a phrase in the sentence where they are found to co-occur. Intervening punctuation, position in sentence, and syntactic values are taken into account by this program.

A new option is available which permits the user to specify that a looked-up document collection is on tape AG. Using this option, it is possible to run by submitting only new requests or by submitting only a few texts not previously looked up, etc.

As described above, provisions are made for concept numbers designating authors, journals, and citations. These concept numbers are added to the document vectors and correlated in the same way as concept numbers representing word stems or word clusters. The document vector format is described more fully in Part 4 of this section on Logical Changes.

In the old system, concepts which appear in the titles of documents can be weighted differently from those appearing in the body of the document. The weight is specified by the user at run time. In the extended system, the user may also specify separate weights for

statistical phrases, syntactic phrases, semi-syntactic phrases, word clusters, authors, journals, and citations. In addition, all of these classes may be weighted differently for occurrence in the title or body of a document.

4. Logical Changes

Experience has shown that syntactic phrase searching is extremely time consuming. In order to avoid unnecessary syntactic processing, statistical phrase searching is always executed prior to syntactic or semi-syntactic analysis. In this way it becomes possible to avoid the syntactic processing of any sentence which does not contain all the terms of at least one phrase from the criterion tree file. However, statistical phrase finding continues to be available as a separate option.

The generation of word clusters by means of term-sentence correlation is accomplished by a new program run prior to production runs of the main SMART system. This program, which is described more fully in Part 5 on Support Programs, produces a file of word clusters with generated concept numbers. These clusters are detected in documents and requests by the statistical phrase finder during the lookup, and the corresponding concept numbers become part of the document vectors.

In order to remove the limitations on the number of documents to be processed in a single run and the number of words which can be

included in a single document, it becomes necessary to store the document vectors on tape instead of keeping them all in core storage. Requests, however, will remain in internal memory as before.

Significant changes have been made in the format of document vectors. Each vector begins with a 12-word record of which the first word is either *TEXT or *FIND and the second and third words are the document identifier. The remaining nine words are not specified. The rest of the vector consists of an unlimited number of 260 word records, the last of which need not be exactly 260 words, and is terminated by a word of zeros. The following list summarizes the possible word formats.

Prefix	0:	either word stem, syntactic, semi-syntactic or statistical phrase
	4:	cluster
	5:	author
	6:	journal
	7:	bibliographic entry
	1,2,3:	not assigned
Decrement		contains a concept number (with a prefix of zero, 1-30,000 are reserved for noncommon word stems, 30,001-32,767 for common words)
Tag	0:	word stem from body
	1:	semi-syntactic phrase, body

	2:	statistical phrase, body
	3:	syntactic phrase, body
	4:	word stem, title
	5:	semi-syntactic phrase, title
	6:	statistical phrase, title
	7:	syntactic phrase, title
Address		contains a weight for the concept stored in the decrement

These items appear in logically sorted order on the tape.

Although it may seem that the changeover to tape would increase processing time, such an increase is not anticipated now. Previously, both the lookup and syntactic processing were done on a document-by-document basis with different chain links looping through core for each document. In the new system, these processes will be carried out serially for the entire collection with each chain link remaining in core until it is no longer needed. This alteration is expected to compensate for the increased tape time previously referred to.

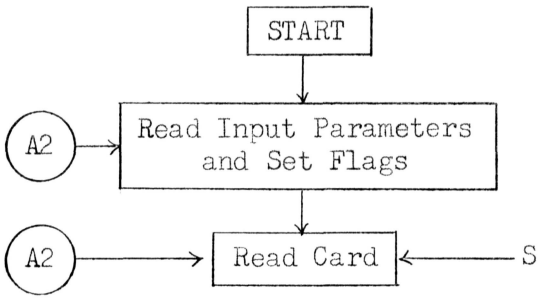
5. New Support Programs

A new null thesaurus generator has been written to produce the larger thesaurus acceptable to the new system. This new program also accepts an unlimited number of documents, each of unlimited length. The null thesaurus program accepts input parameters specifying the maximum number of noncommon concepts desired and/or the minimum number

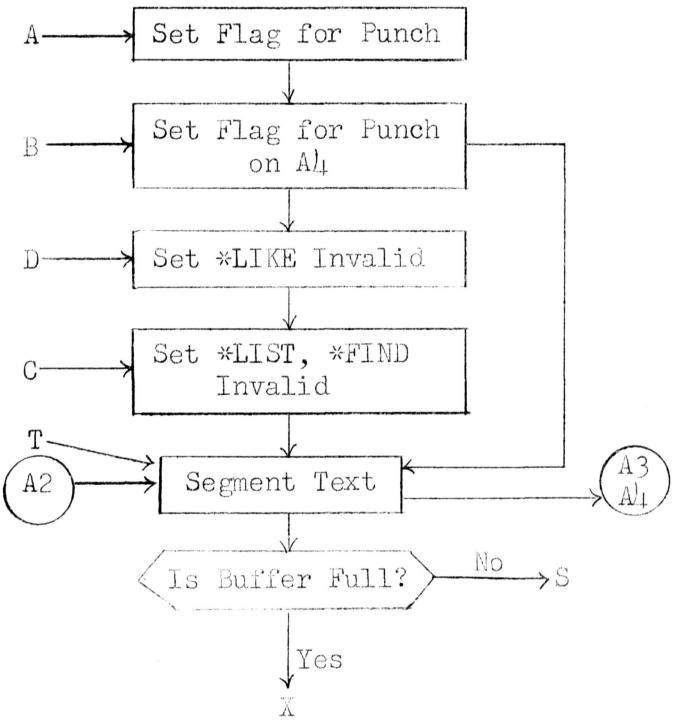
of times a word must appear in the collection in order to be included in the null thesaurus.

A program has been written to generate word clusters from a document collection and to produce a dictionary of these clusters. Documents are looked up in the thesaurus, and the term-sentence correlation is performed. Cluster coefficients are calculated for terms which co-occur in sentences. A concept number is generated for each cluster. The user supplies a cut-off coefficient and/or a maximum number of clusters to be included in the dictionary. This dictionary is used at lookup time to find phrases (word clusters) in texts and in search requests. The programming of all changes described in this section is presently under way.

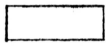
LINK 1



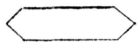
*LIST ⟶ A
*FIND ⟶ B
*LIKE ⟶ C
*TEXT ⟶ D
*ONLY ⟶ E
*TAPE ⟶ F
*STOP ⟶ F
*EØF ⟶ F



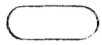
Legend:



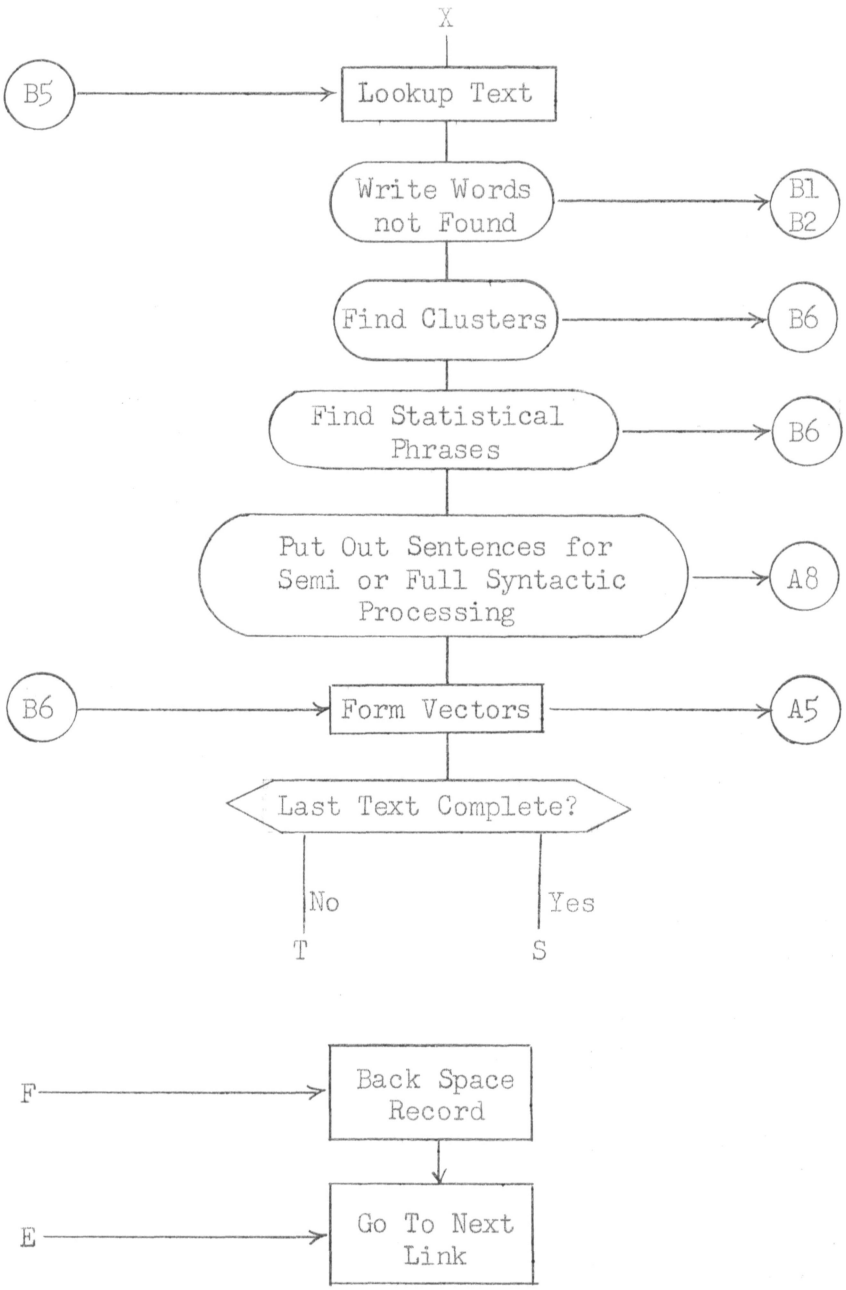
Logical Block in Main Flow of Control

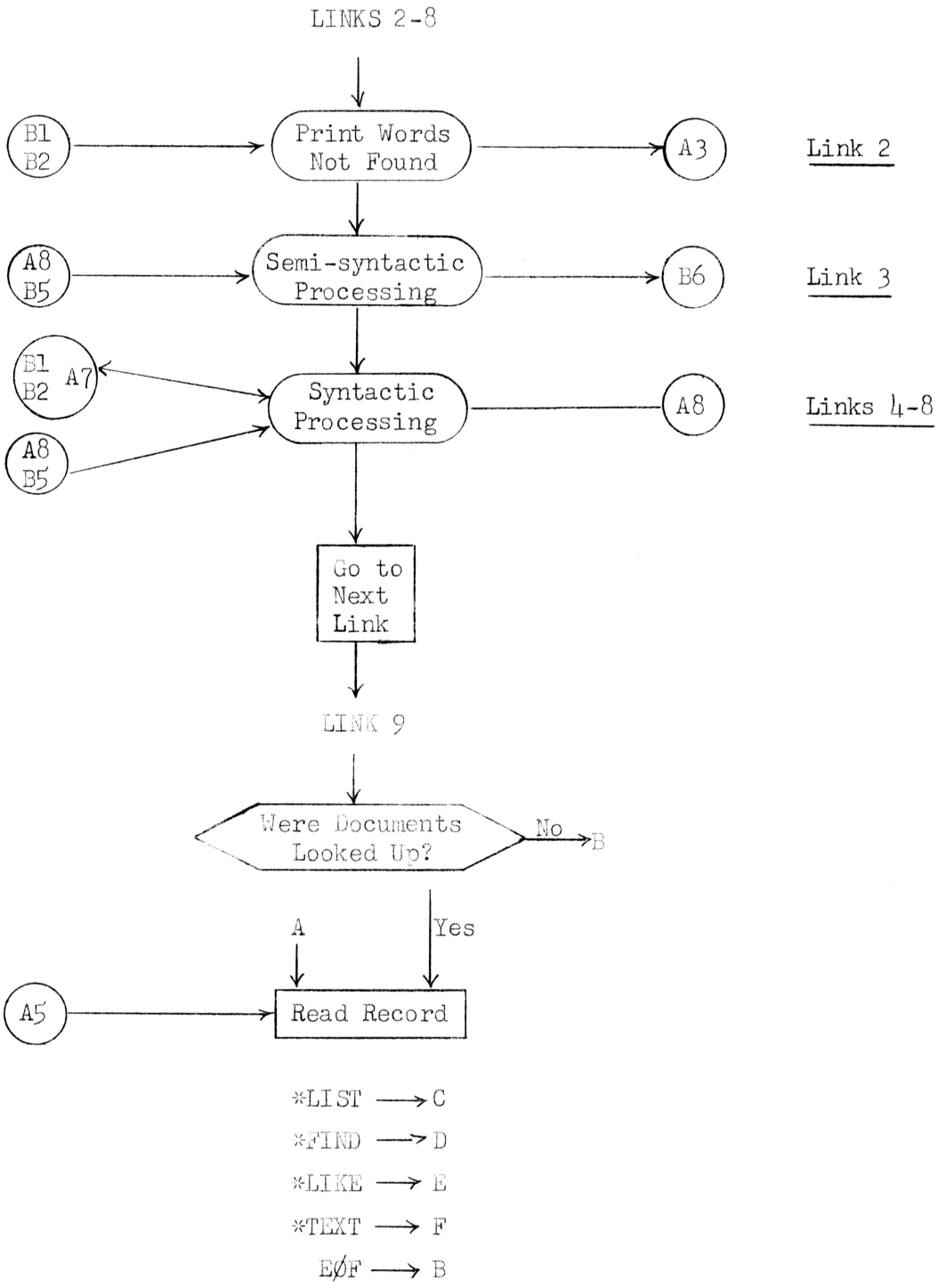


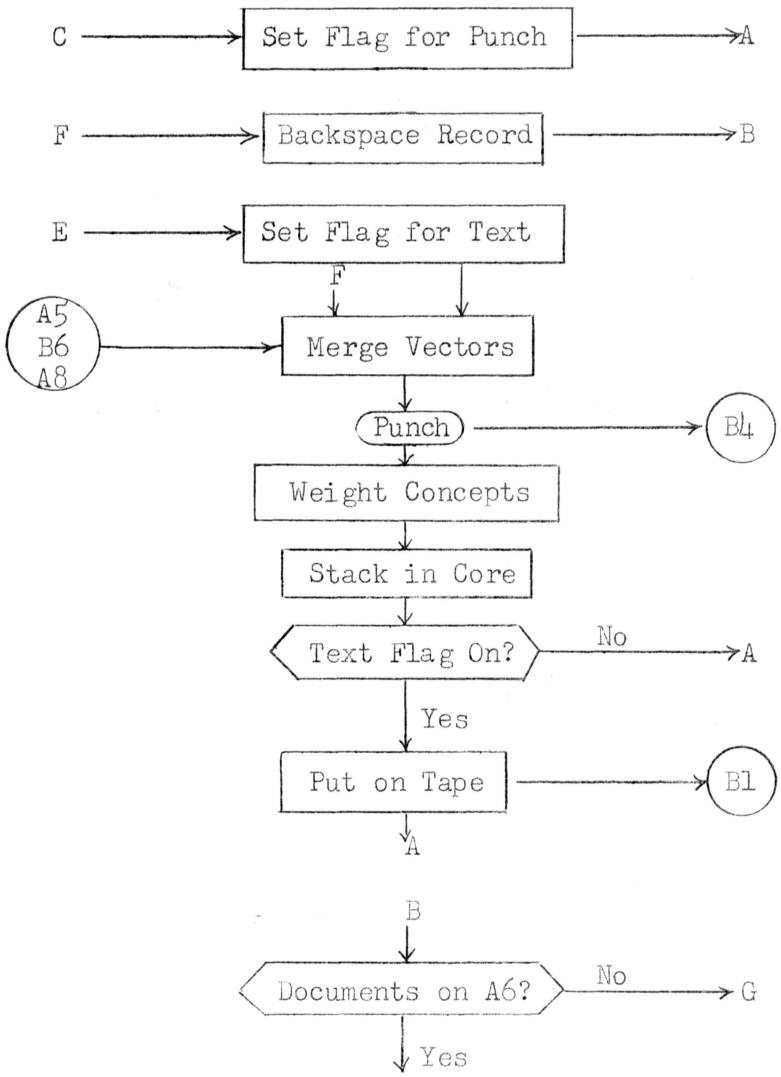
Decision

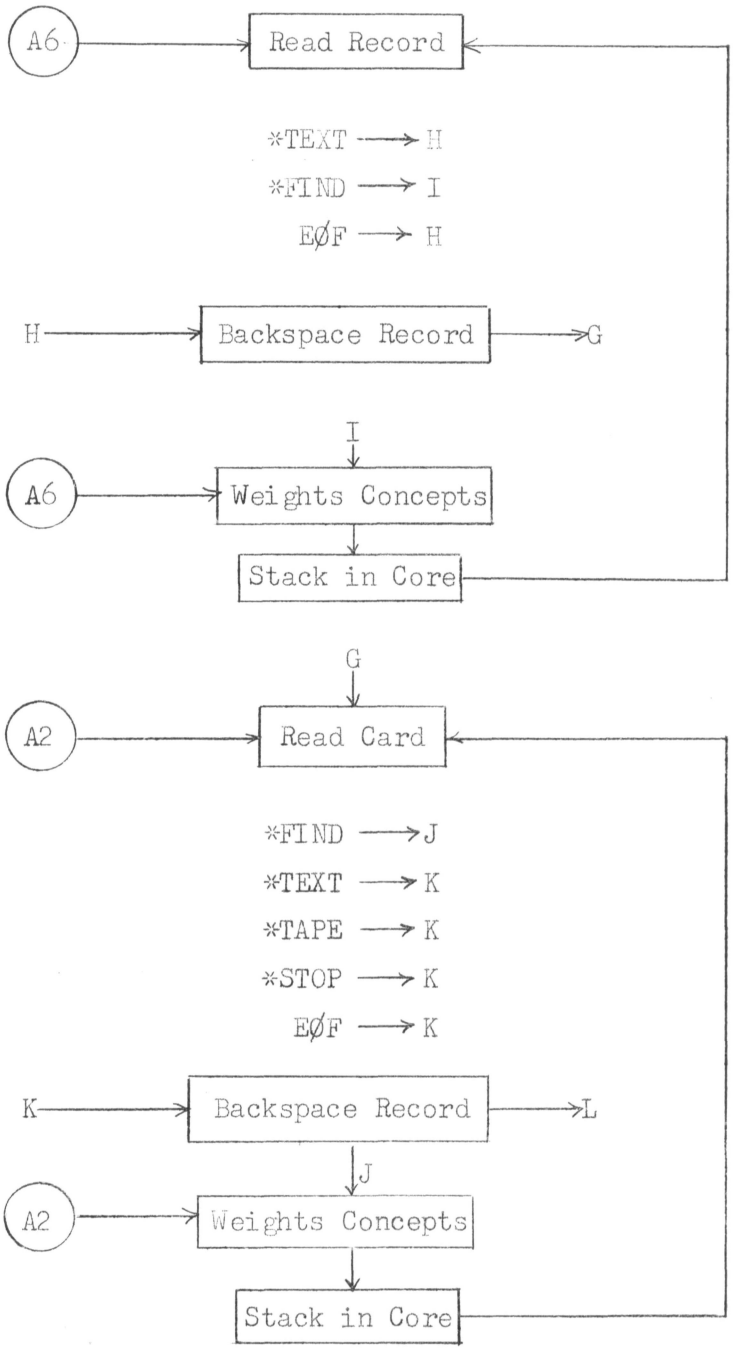


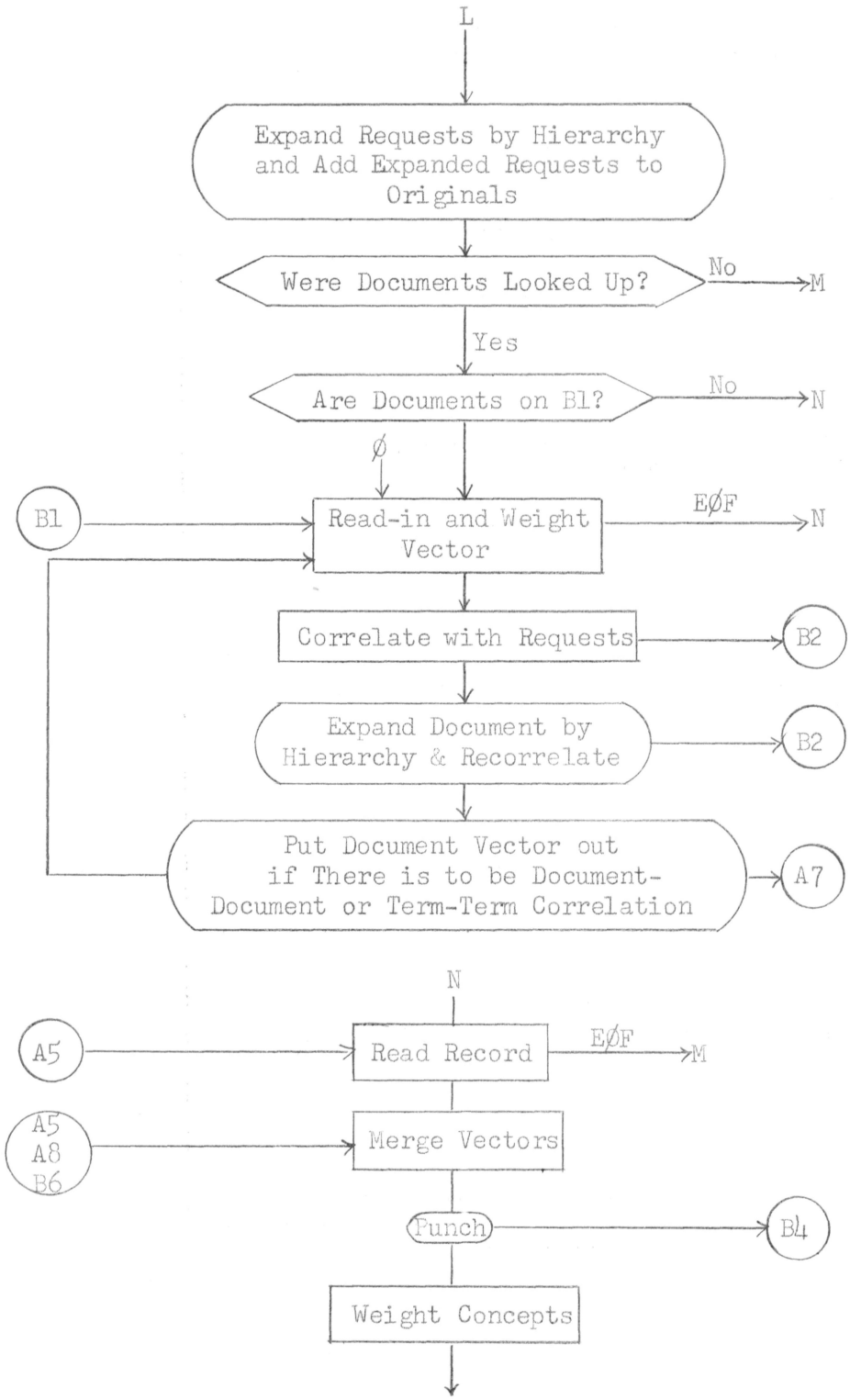
Logical Block to be Executed if Needed or Requested

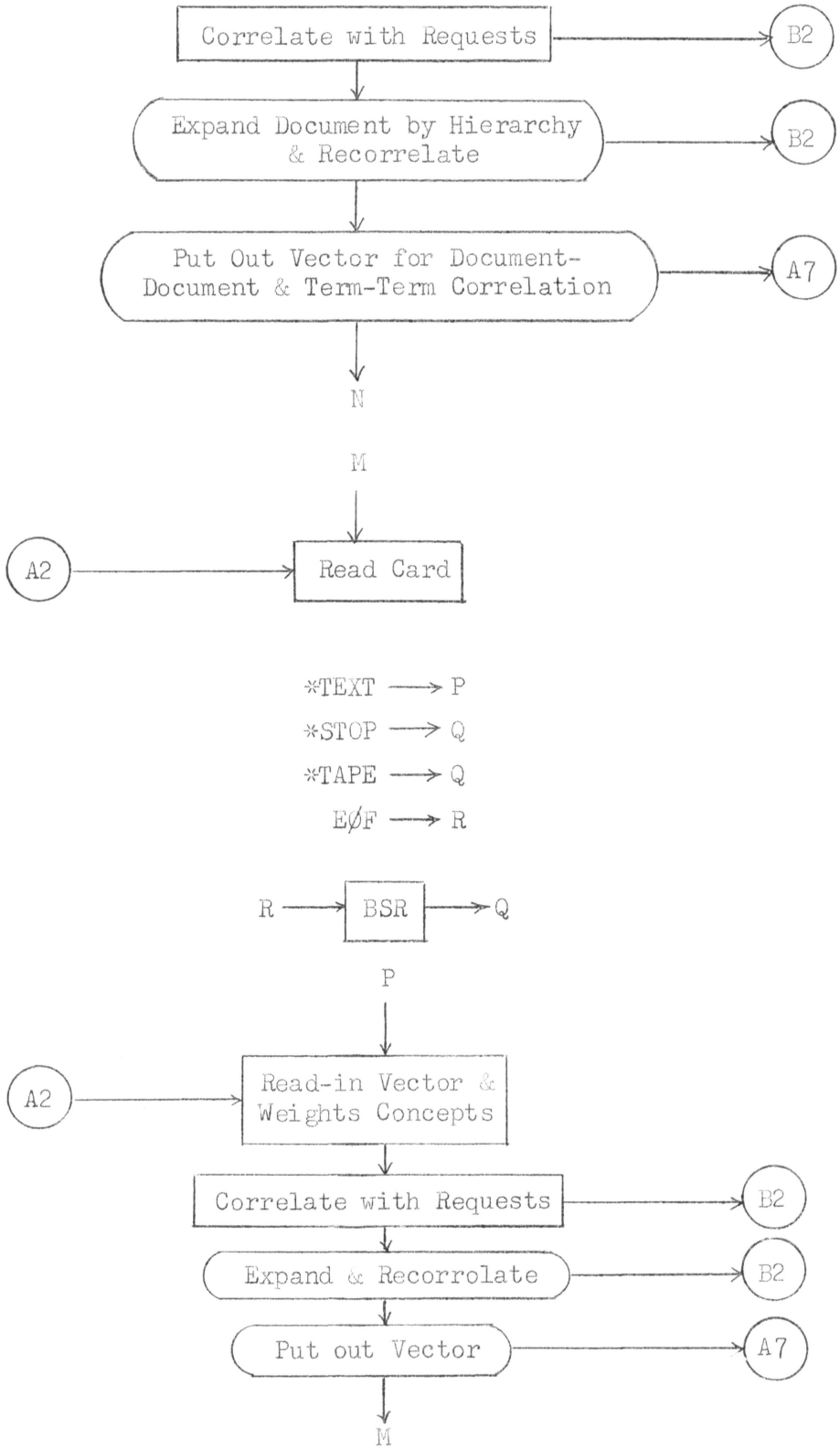

X
B5
Lookup Text
Write Words not Found
B1 B2
Find Clusters
B6
Find Statistical Phrases
B6
Put Out Sentences for Semi or Full Syntactic Processing
A8
B6
Form Vectors
A5
Last Text Complete?
No
T
Yes
S
F
Back Space Record
E
Go To Next Link

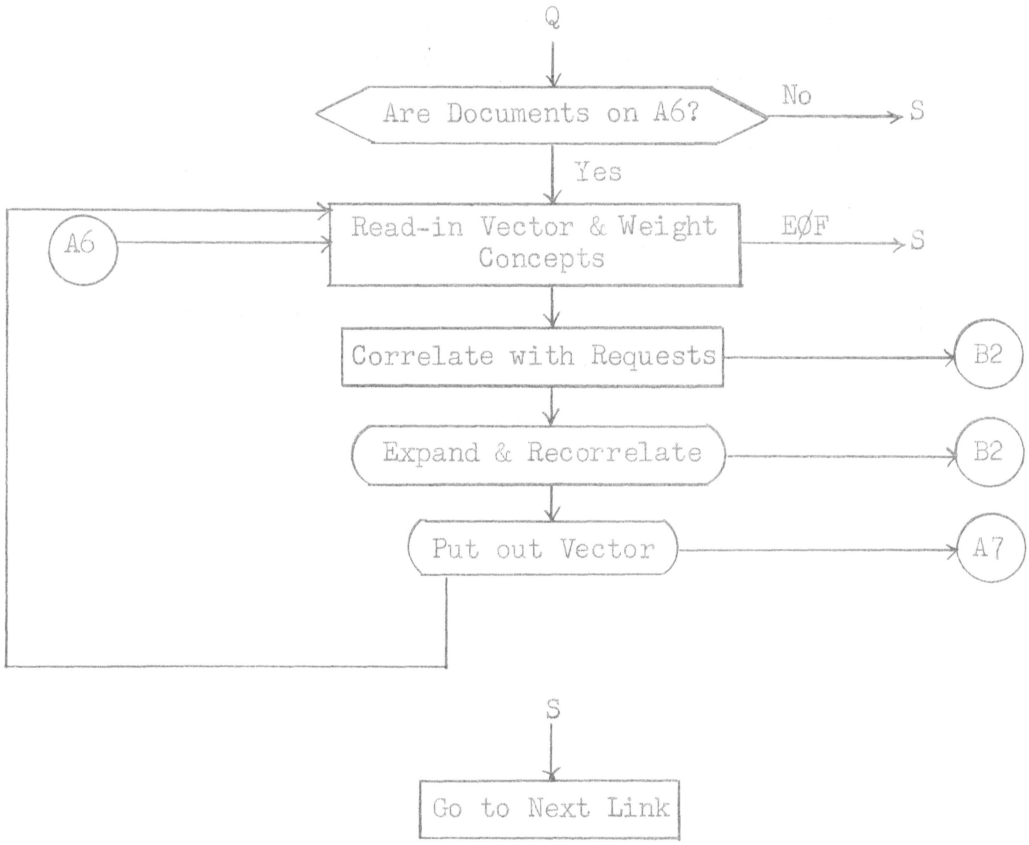

LINKS 2-8
B1
B2
Print Words Not Found
A3
Link 2
A8
B5
Semi-syntactic Processing
B6
Link 3
B1
B2
A7
A8
B5
Syntactic Processing
A8
Links 4-8
Go to Next Link
LINK 9
Were Documents Looked Up?
No
B
A
Yes
A5
Read Record
*LIST → C
*FIND → D
*LIKE → E
*TEXT → F
EØF → B

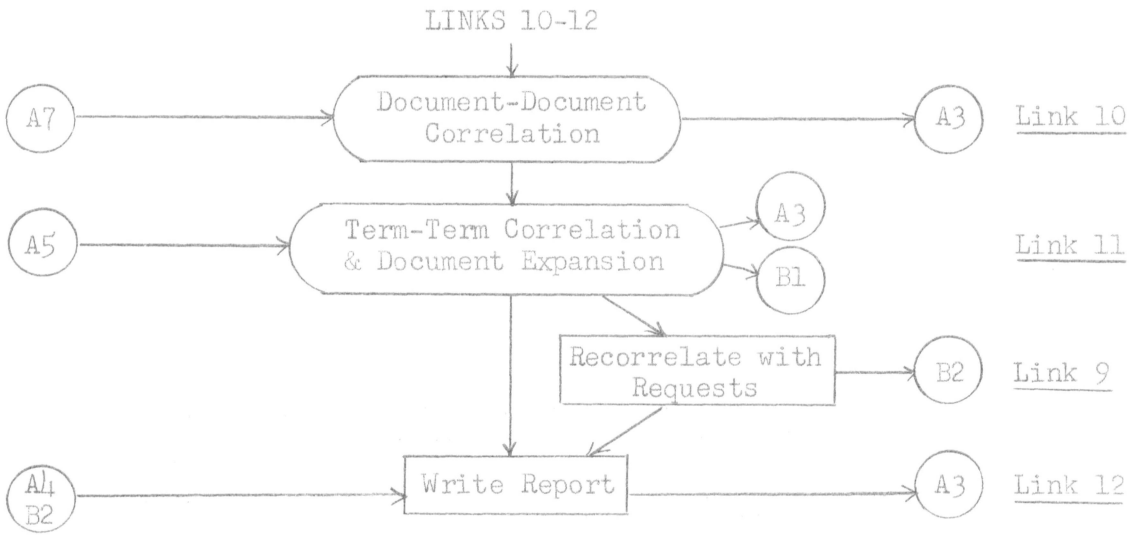

C
Set Flag for Punch
A
F
Backspace Record
B
E
Set Flag for Text
F
A5
B6
A8
Merge Vectors
Punch
B4
Weight Concepts
Stack in Core
Text Flag On?
No
A
Yes
Put on Tape
B1
A
B
Documents on A6?
No
G
Yes


A6
Read Record
*TEXT → H
*FIND → I
EØF → H
H
Backspace Record
G
I
A6
Weights Concepts
Stack in Core
G
A2
Read Card
*FIND → J
*TEXT → K
*TAPE → K
*STOP → K
EØF → K
K
Backspace Record
L
J
A2
Weights Concepts
Stack in Core


L
Expand Requests by Hierarchy and Add Expanded Requests to Originals
Were Documents Looked Up?
No
M
Yes
Are Documents on B1?
No
N
Ø
B1
Read-in and Weight Vector
EØF
N
Correlate with Requests
B2
Expand Document by Hierarchy & Recorrelate
B2
Put Document Vector out if There is to be Document-Document or Term-Term Correlation
A7
N
A5
Read Record
EØF
M
A5
A8
B6
Merge Vectors
Punch
B4
Weight Concepts


Correlate with Requests
B2
Expand Document by Hierarchy & Recorrelate
B2
Put Out Vector for Document-Document & Term-Term Correlation
A7
N
M
A2
Read Card
*TEXT → P
*STOP → Q
*TAPE → Q
EØF → R
R
BSR
Q
P
A2
Read-in Vector & Weights Concepts
Correlate with Requests
B2
Expand & Recorrolate
B2
Put out Vector
A7
M


Q
Are Documents on A6?
No
S
Yes
A6
Read-in Vector & Weight Concepts
EØF
S
Correlate with Requests
B2
Expand & Recorrelate
B2
Put out Vector
A7
S
Go to Next Link


LINKS 10-12
A7
Document-Document Correlation
A3
Link 10
A5
Term-Term Correlation & Document Expansion
A3
B1
Link 11
Recorrelate with Requests
B2
Link 9
A4
B2
Write Report
A3
Link 12