Division of Research
Graduate School of Business Administration
The University of Michigan

October 1984

AN EXTENSIONAL SEMANTIC ANALYSIS OF DOCUMENT INDEXING

Working Paper No. 395

David C. Blair

One of the persistant problems in indexing theory is maintaining the distinction between the indexing language (used to tag documents in a particular collection) and the English language (from which the indexing language is derived). There is an undeniable semantic relation between a term used as a descriptor in the indexing language and that same term used as a word in Standard English. But the semantic relation between these different usages of the same term is more like a family resemblance than an identity relationship. In other words, the dictionary definition of a word may be useful in determining the meaning of that word in Standard English, but it could be quite misleading when used to determine the meaning of that same word when used as a descriptor in an indexing language. This lack of close semantic correspondence between index terms and English words is what Maron and Kuhns refer to as "semantic noise:"

> It turns out that given any term there are many possible subjects that it could denote (to a greater or lesser extent), and, conversely, any particular subject of knowledge (whether broad or narrow) usually can be denoted by a number of different terms. This situation may be characterized by saying that there is 'semantic noise ' in the index terms. [pp 218-219]

This semantic noise is complicated by the fact that while there does exist a Standard English whose semantics is generally similar in a wide variety of contexts and glosses, there does not appear to be a similar standard indexing language. That is, if the same set of terms were used in two different document collections (eg, a psychology collection and a geology collection), the "meanings" (semantics) of these terms in these two collections (ie, the kinds of information they represent), would be quite different. So, in effect, instead of having two usages of one indexing language, we have two similar, yet distinct indexing languages. They are similar in that they use the same set of descriptors, but different because the same descriptor may (and usually does) have a different meaning in either language. One could think of these indexing languages as dialects of English.

The purpose of this paper will be to demonstrate a method which reduces the "semantic noise" of a given indexing language, and enables the semantic definition of terms to be made which are specific to that indexing language alone. There are two principal semantic aspects of indexing descriptors which this paper attempts to clarify: 1. The main semantic categories in the indexing language; 2. The

semantic relations between descriptors in the indexing language. Of course, all this information can be obtained from the Library of Congress list of subject headings or Roget's Thesaurus, but I maintain that this would be a misleading way of understanding the semantic relations between these indexing descriptors. The tacit semantic relations extant in an indexing language represent the indexing philosophy of that particular document collection, and this indexing philosophy will vary markedly from document collection to document collection.

FOUNDATIONS

The central issue here is how to make explicit this tacit indexing philosophy. To do this, two questions must be asked: 1. What are the "facts" of a document collection, and how would we describe them? (That is, what are the basic units or "things" which we can examine in a document retrieval system?) 2. What can we infer from these facts? What do these facts "mean?"

There are two major types of facts in a document retrieval system: 1. The set of documents in the collection. 2. The set of indexing descriptors actually assigned to the documents in the collection. The documents in the collection

can be counted (eg, the number retrieved at a given time) and can be distinguished one from the other (eg, document 1 is distinguishable from document 7). In terms of the present study this is all we can directly say about the documents; that they are countable and distinguishable. Nothing more can be said about the documents other than what is implied by the indexing descriptors assigned to them.

Like the documents in a collection, the indexing descriptors can be both counted and distinguished. More specifically, they can be counted and distinguished in several significant ways: 1. We can calculate the total number of times an individual discriptor is used in a document collection. This is the "breadth" of a given descrptor. 2. We can determine how many descriptors are assigend to each document. This is the "depth" of assignment for a given document. 3. Since we can distinguish different documents and distinguish different terms, we can therefore determine the frequency of cooccurrence of descriptors on documents within the collection. These, then, are the basic "facts" of a document retrieval system, and it is only from these fundamental, obsevable units that inferences will be made. Previously I had said that the purpose of this paper is to reduce the "semantic noise" of the indexing

language used in a document collection. More specifically, this paper will attempt to show how the basic "facts" of a retrieval system can be used to clarify the indexing philosophy of the system, and provide some further indication of the content of a document which a descriptor assigned to that document cannot give by itself (even if its meanings in Standard English are considered).

SEMANTIC CATEGORIES: RELATIVE BREADTH

The first step in making the indexing philosophy of a document collection explicit is the calculation of the "relative breadth" of each descriptor in the system. Given the set of descriptors actually used in the document collection, $\{T_1, T_2, T_3, \ldots, T_n\}$, and the aggregate of documents, D, in the collection, then the relative breadth of descriptor T_i is the value r_i, such that:

$$r_i = \frac{\text{the \# of times } T_i \text{ is used in D}}{\text{the number of times the most frequent descriptor is used in D}}$$

r_i thus reflects the number of times the descriptor T_i is used in D. For example, if r_i is the most frequently assigned descriptor in the collection then $r_i = 1$. If r_i is not the

most frequently used term, then it will equal something like, for instance, .65.

SPECIFICITY AND GENERALITY

How shall we interpret the relative breadth (r_i) of each descriptor in the collection? First of all, the values of r_i when ranked in order of magnitude, gives us an indication of the major and minor categories of the indexing language. (the major categories being those with the highest values for r_i, and the minor categories being those with the lowest values) More importantly, r_i gives us an indication of the "generality" ($r_i \longrightarrow 1$) or "specificity" ($r_i \longrightarrow 0$) of the descriptor whose relative breadth = r_i. But note that this specificity/generality is defined solely in terms of how the indexing descriptors are used in this system, and not in terms of the semantic estimation of their specificity/generality in Standard English. For example, in Standard English usage the descriptor "philosophy of science" would be considered quite general, while "tectonic plates" would be considered fairly specific (at least compared to "philosophy of science"). But if, in a given document collection, "philosophy of science" is used only 5 times, and "tectonic plates" is used 150 times, and their

resulting relative breadths (r_i) come out to be something like .08 and .85, respectively, then in terms of that collection "philosophy of science" is a very specific descriptor, while "tectonic plates" is a quite general one.

Now that we have a quantitative definition of specificity and generality for individual descriptors, we also have a method for estimating the specificity or generality of individual documents. For each document d_i in D construct the vector $\langle r_a, r_b, r_c, \ldots \rangle$ where $r_a, r_b, r_c, \ldots$ are the relative breadths of the descriptors $T_a, T_b, T_c, \ldots$ which are assigned to d_i. Next calculate the quantity,

$$((r_a)^2 + (r_b)^2 + \ldots.)^{\frac{1}{2}}$$

This quantity, S_i, is the norm (or length) of the vector composed of the relative breadths of the terms assigned to d_i, and represents the specificity/generality of the document d_i in terms of the descriptors $\{T_a, T_b, T_c, \ldots\}$ assigned to d_i. In a less rigorus form, S_i represents the approximate specificity of the document d_i. Note one important thing: the specificity/generality of d_i is contingent on the independant specificity/generality of the descriptors assigned (as reflected in the norm of the vector), and not on the number of descriptors assigned. In other words, a

given document with only one descriptor could be either quite specific, quite general or neither, depending on the overall frequency of use of the descriptor.

The value of r_i reflects a certain attitude which the indexers have towards how the descriptor T_i is used, and the value of S_i roughly reflects this same attitude in terms of the document d_i. In other words, if a document is indexed with several descriptors which are all quite general (have high r_i values) then in terms of that collection, that document would be quite general. Note also that the specificity/generality of a term (r_i) affects, but does not control the specificity/generality of the document (S_i) to which that term is assigned (except in the trivial case where T_i is the only descriptor assigned to the document). For example, if a user wanted the most "general" document which had the descriptor T_k assigned to it, then this estimation of the generality of these documents would not be a function of T_k alone, but of the relative generality of the other descriptors which were assigned to the documents to which T_k was assigned.

Certainly it must be understood that measurements of the specificity/generality of descriptors and (especially) documents is a very imprecise notion. A small difference

between two values (eg, .50 and .60) would not necessarily be significant. But when the differences are large (eg, .20 and .80) then S_i is a valuable heuristic for estimating specificity. The advantage of calculating S_i for the documents in the collection is that it provides an immediate short cut for searches in which the user asks for a "general" or "specific" document on a given topic. The retrieved set could then be the two or three documents (to which the descriptors indicated by the user are assigned) with the lowest/highest S_i. Alternatively, S_i could be a means of giving an ordering to the entire set of retrieved documents.

Given that the specificity/generality of terms is roughly equal to relative breadth, and that the specificity/generality of documents is, at best, a very imprecise notion, it may still be objected that the specificity of terms may in fact be misleading when generalized to documents. Suppose that in a geology library there is one book indexed with the descriptor "philosophy of science," and that that book is (in the estimation of a philosophy professor) a very general book on the philosophy of science. Since it is the only work in the collection indexed with this descriptor, the procedure I outlined above would label it a

very specific book on the subject. Who's right? Well, it depends on your point of view. From the perspective of a philosopher, the book is, in fact, quite general. But from the point of view of a geologist (ie, a typical user of this library) any book on philosophy of science would be a very specific book. [actually, the specificity of this hypothetical work, is really moot. A geologist who wanted a "specific" or "general" work on "philsoophy of science" would presumably go to the philosophy or main library to search for it. But it is still important to keep "philosophy of science" as a "specific" descriptor in itself, since it may affect the calculation of S_i when it is used with other terms on a given document] My basic arguments, then, for accepting relative breadth as an estimation of specificity/generality are: 1. the specificity/generality of a descriptor/document must be defined in terms of the collection in which these descriptors are used, not in the broader context of English semantics. 2. the number of times a descriptor is used in the collection, relative to the frequency of the most frequent descriptor, is an indication of the specificity/generality of that descriptor. That is, a descriptor which is assigned a great many times in a collection, describes a very large

set of documents which have something to do with the concept expressed by that descriptor. This descriptor, therefore, by describing such a large set, is not very specific.

SEMANTIC RELATIONS: DEFINITION OF A TOPIC

There are two ways a given descriptor, T_k, can be seen: 1. It is a descriptor (word) in the indexing language. 2. It is a word in the English Language. The word itself, regardless of which language it is in, can again be seen two ways: 1. As a word. 2. As the designation of a topic or subject area (a word with "semantic noise"). The definition of a word in either the indexing language or the English language is straightforward, but it is not entirely clear what we mean by a "topic"——especially in terms of the indexing language. For example, to understand the topic "statistics" one must have a familiarity with how the word "statistics" is used in standard English. This is a very subjective notion, but there is a certain agreement as to the topic "statistics" among speakers of English. An individual may understand "statistics to have something to do with mathematics, &/or sampling, &/or experimental design, &/or verification, &/or averaging, &/or sociology, &/or the Gallup Poll, etc. The agreement

upon what the topic "statistics" means is not important here (we are not lexicographers). What is important to understand is that the notion of a "topic" is contingent on the varieties of usage of the word in Standard English. That is, a "topic" designated by a word in a language is understood by the context(s) of that word in the language. Thus, since the topic T_k in the English language must be understood in the context of Standard English, it is reasonable to assume that the topic T_k in the indexing language must be defined in terms of the context(s) of T_k in the indexing language. But what is a "context" in the indexing language? In Standard English the context of a word was defined as the varieties of that word's usage. This appears to be a good rough definition of context (certainly more heuristic than "context" or "topic" by itself). It follows, then, that we must define the topic of T_k in the indexing language in terms of the usage of that descriptor in the indexing language. This usage of T_k consists, of course, in the assignment of T_k to documents in the collection, and the varieties of these assignments are distinguished by noting that the assignments are made to "different" documents (just like English words are used in "different" contexts). Since the contents

of these documents are not differentiated, the only way to distinguish their content is by noting the different descriptors assigned to them. The different contexts in which T_k may appear, therefore, are the groups of descriptors which are assigned to documents to which T_k is also assigned. For example, suppose that T_k is assigned to documents d_a,d_b, and d_c, and the descriptors assigned to each of these documents is as follows:

$$d_a = T_k, T_a, T_e, T_g$$
$$d_b = T_k, T_a, T_e, T_h$$
$$d_c = T_k, T_a, T_g, T_f$$

Then the contexts in which T_k appears are T_a,T_e,T_g; T_a,T_e, T_h; and T_a,T_g,T_f, which is nothing more than the distribution of co-occurrences of T_k with other descriptors in the collection. Thus the definition of the "topic" T_k in terms of the indexing language of the system in which it is used is as follows:

$$T_k = \langle z_a, z_b, z_c, \ldots \rangle$$

such that z_a,z_b,z_c,... corespond to descriptors T_a,T_b, T_c,... which have concurrent assignment with T_k (ie, they co-occur) on documents in the collection, and:

$$z_i = \frac{\text{number of times } T_k \text{ co-occurs with } T_i}{\text{maximum \# of times } T_k \text{ occurs with any descriptor}}$$

In the example above of documents d_a, d_b, d_c, the topic vector for T_k is:

$$T_k = \langle z_i \rangle \quad \text{s.t. } z_a, z_e, z_f, z_g, z_h = 1, .67; .33, .67, .33, \text{ respectively.}$$
(all other values of $z_i = 0$)

Topologically speaking, this vector defines the "region" on the topic T_k. As in the definition of specificity, the definition of the topic T_k is solely in terms of the retrieval system in which T_k is used, not in terms of the semantics of standard English.

It may be argued that the co-occurrence of descriptors can be purely coincidental, and that to define a topic in such arbitrary terms would be mistaken. Suppose every time "psychology" is used on a document in a collection the descriptor "statistics" is also used. It may be argued that the notions of "psychology" and "statistics" are independant, and that any relation between the two is tenuous at best. But there is a confusion here. It is true that in terms of Standard English, "psychology" and "statistics" have very little in common. But in an information retrieval system, this does not necessarily

concern us. The relation between "psychology" and "statistics" should be made solely in the indexing language---in terms of the retrieval system itself. In such a case, if "statistics" is used to index every document which is indexed by "psychology" then, in terms of that collection of documents, there is a strong and clear connection between the two descriptors. [Of course, the connection is not symmetrical. If statistics is used to index every document which is indexed by "psychology" then there is a strong dependance of "statistics" on "psychology." But unless most of the occurrances of "psychology" coincide with occurrances of "statistics," then "psychology" will not be strongly dependant on "statistics."]

DEFINITION OF TOPIC AND RETRIEVAL REQUESTS

The notion of a topic in terms of the indexing language can be a very powerful tool for ordering relatively unspecific requests (that is, requests which retrieve a large number of documents). For example, a user can request a set of documents "about" the topic designated by descriptor T_k. In such a case, the documents with the descriptor T_k are retrieved and this retrieved set is ordered according to the following procedure: 1. The topic vector for T_k

is recalled (for example, $T_k = \langle z_a, z_b, z_c, z_d \rangle$ where T_k co-occurs with descriptors T_a, T_b, T_c, and T_d) 2. The retrieved documents are ranked according to these categories: a. documents indexed with all terms in the topic region (T_k, T_a, T_b, T_c, T_d). b. The remainder of the documents are ranked according to the decreasing values of the results of the dot product of the document vectors and the topic vector for T_k. For example:

$T_k = \langle z_i \rangle$ s.t. $z_a, z_b, z_c, z_d = 1, .65, .4, .3$, resp.
(all other values of $z_i = 0$)

d_1 is a document indexed with descriptors T_a, T_d, T_k, and the corresponding document vector for $d_1 = \langle z_i \rangle$ s.t. $z_a, z_d, z_k = 1$, and all other values of $z_i = 0$. d_1's ranking number equals the dot product of the document vector and the topic vector for T_k: $\langle 1, .65, .4, .3 \rangle \cdot \langle 1, 0, 0, 1 \rangle = 1.3$

d_2 is indexed with T_b and T_k, and the corresponding document vector for $d_2 = \langle z_i \rangle$ s.t. $z_b, z_k = 1$, and all other values of $z_i = 0$. d_2's ranking number equals: $\langle 1, .65, .4, .3 \rangle \cdot \langle 0, 1, 0, 0 \rangle = .65$

d_3 is indexed with T_a, T_b, T_d, and T_k. The document vector for $d_3 = \langle z_i \rangle$ s.t. $z_a, z_b, z_d, z_k = 1$. d_3's ranking number works out to 1.95.

d_4 is indexed with T_a, T_b, T_c, T_k. The document vector for $d_4 = \langle z_i \rangle$ s.t. $z_a, z_b, z_c, z_k = 1$. d_4's ranking number works out to 2.05

The final ranking of the retrieved set:

d_4 (2.05)
d_3 (1.95)
d_1 (1.30)
d_2 (.65)

If the topic region for T_k is defined as the set of co-occurring descriptors $\{T_a, T_b, T_c, T_d\}$, then it would be possible (if necessary) to retrieve documents which were similar in subject matter to what T_k designates, but did not actually have T_k as a descriptor. This request would take the Boolean form $T_a \sqcap T_b \sqcap T_c \sqcap T_d$, and could be an alternative retrieval method if retrieving with T_k and ranking according to T_k's topic region proved unsatisfactory.

DEFINITION OF TOPIC AND BOOLEAN REQUESTS

When the retrieval request is of Boolean form, a topic region can be calculated for the combined descriptors and used to rank the documents retrieved. For example:

1. Request = $T_k \sqcap T_n$
Topic region for $T_k = \{T_a, T_b, T_c, T_d\}$
Topic region for $T_n = \{T_a, T_c, T_e, T_f\}$
Therefore the topic region for $T_k \; T_n$ =
$\{T_a, T_b, T_c, T_d\} \sqcap \{T_a, T_c, T_e, T_f\} =$
$\{T_a, T_c\}$

The topic vector for $T_k \sqcap T_n = \langle z_i \rangle$
s.t. $z_a, z_c, z_k, z_n = 1$. All other z_i values = 0
The retrieved set of documents are those indexed with both T_k and T_n, and they are ranked according to the values of the dot products of the document vectors of the retrieved documents and the topic vector for $T_k \sqcap T_n$ (similar to the document ranking method on p 16-17).

2. Request = $T_k \sqcup T_n$
If the topic regions are the same as above, then:
Topic region for $T_k \sqcup T_n$ =
$\{T_a, T_b, T_c, T_d\} \sqcup \{T_a, T_c, T_e, T_f\} =$
$\{T_a, T_b, T_c, T_d, T_e, T_f\}$
The topic vector for $T_k \sqcup T_n = \langle z_i \rangle$
s.t. $z_a, z_b, z_c, z_d, z_e, z_f = 1$. All other z_i values = 0
The retrieved set of documents are those indexed with T_k or T_n or both, and are ranked according to the values of the dot products of the document vectors of the retrieved documents and the topic vector for $T_k \sqcup T_n$ (same as above).

If searching under $T_k \sqcap T_n$ or $T_k \sqcup T_n$ proves unsatisfactory, then as an alternative, documents indexed with neither T_k or T_n can be retrieved and ranked according to how closely their document vectors approximate the topic vectors for either $T_k \sqcap T_n$ or $T_k \sqcup T_n$.

AUTOMATIC WEIGHTING OF ASSIGNED INDEXING DESCRIPTORS

The method of calculating topic regions for index terms (the set of co-occurring descriptors) and topic vectors (the values of z_i associated with the descriptors in a topic region) can now be used to automatically calculate weights for terms which have already been assigned to a document. The procedure is as follows: 1. The indexer assigns a set of descriptors to a document (eg, $d_o = \{T_d, T_m, T_p\}$. 2. The topic regions for T_d, T_m, and T_p are called: (eg)

$$T_d = T_m, T_o, T_p, T_r, T_s$$
$$T_m = T_d, T_n, T_p$$
$$T_p = T_a, T_d, T_m, T_r$$

3. The corresponding topic vectors are derived: (eg)

$$T_d = \langle .7, 1, .1, .5, .3 \rangle \quad [z_m, z_o, z_p, z_r, z_s]$$
$$T_m = \langle .1, .6, 1 \rangle \quad [z_d, z_n, z_p]$$
$$T_p = \langle 1, .6, .8, .2 \rangle \quad [z_a, z_d, z_m, z_r]$$

4. The individual weights of T_p, T_m, and T_d are calculated for document d_o: a. To determine the weight of T_d on d_o find the average of the z values for T_m and T_p in T_d's topic vector. ie, the weight of T_d on $d_o = \frac{z_m + z_p}{2} = \frac{.7 + .1}{2} = .4$. The weight of T_m on d_o is the average of the

z values for T_p and T_d <u>in T_m's topic vector</u>. Thus the weight of T_m =

$$\frac{z_p + z_d}{2} = \frac{1 + .1}{2} = .55$$

And finally, the weight of T_p on d_o is the average of the z values for T_d and T_m <u>in T_p's topic vector</u>. The weight of T_p =

$$\frac{z_d + z_m}{2} = \frac{.6 + .8}{2} = .7$$

The resulting weights of the descriptors assigned to d_o =

$$d_o = \left\{T_d(.4),\ T_m(.55),\ T_p(.7)\right\}$$

The basic assuption behind this weighting method is that the topic region for a given descriptor is the semantic definition of the descriptor in a given document retrieval system. In such a case the weight of a descriptor (say, T_k) on a document——ie, the degree to which T_k applies to a document——is a function of how strongly the other descriptors on that document are related to T_k. If the other descriptors on a document do not correlate highly (have low z values in T_k's topic vector), then it seems reasonable to claim that T_k is not

highly correlated with the document and should recieve a correspondingly low weight.

The advantage of this automatic weighting procedure is that it keeps the subjective decisions of the indexer down to one (ie, whether to assign a term or not) while increasing the descriptive power of term assignment by indicating what the assignment of certain terms to a document implies about that document.

A MEASUREMENT OF RETRIEVAL EFFECTIVENESS

The measurement of effectiveness of a retrieval system has generally been guided by the touchstones of precision and recall (the related notion of fallout is also sometimes used). The calculations of precision and recall are full of difficulties. While it is not the purpose of this paper to offer an extended criticism of these two notions, I would like to point out some of the salient difficulties involved as a preliminary to discussing my own measure of retrieval effectiveness.

The first difficulty is with calculating these measures. Although the determination of precision can be done relatively easily by asking the user which retrieved documents he believes are relevant (fatigue factors aside,

and user co-operation assumed), the estimation of recall is not quite so straightforward. That is, while the user can determine which retrieved documents are relevant to his need/request*, who is going to determine which documents of the large unretrieved set of documents are relevant to this request? [*for simplicity, I won't distingquish between relevance to need and relevance to request, although the difference is certainly non-trivial] Since it is unlikely that the user would have the time or inclination to peruse the entire unretrieved collection, the assessment of whether certain unretrieved documents are relevant must be made by some other person(s). In other words, some "experts" are trying to tell the user what is relevant to his need/request. I am sceptical whether any two individuals' judgement of relevance can be compared meaningfully in this manner at all. But beyond this, even if we have accurate and reliable means for determining precision and recall, there appears another question: do precision and recall measure what we really want to measure when measuring the effectiveness of a retrieval system? More specifically, it is often the case that a user doesn't want all the documents which are relevant to his reques†/need. In other words it would be perfectly consistant for him to say,

"Yes, all 50 of these documents which you retrieved for me are relevant to my request, but I only want this one [exit]." Every information system contains a great deal of redundant information, and althogh redundant information may be in fact highly relevant, it may not be required by the user. Certainly too much redundant, though relevant, information would make for poor retrieval efficiency even though precision and recall were both 100%. Clearly, then, the measurement of the effectiveness of retrievals should be based on some other measure. My proposal is that the measurement of effectiveness for a retrieval system should be whether a set of retrieved documents, ranked according to their estimated relevance to the user's need as expressed in a formal request, coincides with the user's preference order of the retrieved documents. For example, suppose the system retrieves ten documents and ranks them as numbers one through ten in decreasing estimated relevance. If the user considers document number one the best,and doesn't want the other nine, then the system should be given the highest mark for retrieval effectiveness. Likewise, if the user wants all ten documents,and prefers them in order one through ten, then the system should still be given the highest

mark for effectiveness. A statistical measure such as Spearman's formula for rank correlation could be used to calculate the goodness of fit between the system's ranking of the retrieved documents and the user's ranking of the retrieved documents:

$$r = 1 - \frac{6\Sigma D^2}{N(N^2-1)}$$

D = the difference between ranks of corresponding values of X and Y.

N = the number of pairs of values (X,Y) in the data (set of retrieved documents).

The document ranking procedure for the measurement of retrieval effectiveness is the same as the one outlined in the previous sections covering pages 16 - 18.

SEMANTIC DISTANCE BETWEEN TOPICS

Given two topic vectors such as $T_d = \langle z_a, z_b, z_c, z_d, z_e \rangle$ and $T_p = \langle z'_a, z'_b, z'_c, z'_d, z'_e \rangle$ the quantitative estimation of the semantic distance between the topics represented by T_d and T_p is merely the distance between their respective topic vectors:

$$\text{Semantic distance} = [(z_a - z'_a) + (z_b - z'_b) + \ldots + (z_e - z'_e)]$$

[N.B. all other values for z_i in these vectors = 0]

Of course it is easy to see that if the two topic vectors being compared are equivalent the semantic distance between them will be zero. Thus the closer the semantic distance between two topic vectors approaches zero, the closer the two topics are to being synonymous (in terms of that collection).

****

It is easy to see how many of the measures outlined here could be extended to produce alternative searching procedures. Although this may be a worthwhile exercise, the purpose of this paper has only been to provide the fundamental framework for the quantification of semantic noise in indexing terms. In such a case, if the arguments of this paper are accepted, then these alternative or extended mathematical formulations can only be thought of as subsidiary to the formulations worked out in this paper.

BIBLIOGRAPHY

Cooper, William S. "On higher level association measures," JASIS, v 22, n 5, Sept-Oct. 1971, p 354-355.

Gries, David. Compiler Construction for Digital Computers, John Wiley and Sons, Inc., New York, 1971.

Kleene, Stephen C. Mathematical Logic, John Wiley and Sons, Inc., New York, 1967.

Maron, M.E. and J.L. Kuhns. "On relevance, probabalistic indexing and information retrieval," Journal of the Association for Computing Machinery, v 7, n 3, July, 1960, p 216-244.

Siegal, Sidney. Non-Parametric Statistics for the Behavioral Sciences, McGraw-Hill, New York, 1956.

Tarski, Alfred. Logic, Semantics, Metamathematics, Oxford at the Clarendon Press, 1956.

Zadeh, L.A. "Fuzzy Sets," Information and Control, v 8, 1965, p 338-353.

------"The concept of a linguistic variable and its application to approximate reasoning," Dept. of Electrical Engineering and Computer Science, Electronics Research Laboratory, UC, Berkeley, Memorandum no. ERL-M411, 15 Oct. 1973.

Zunde, P. and M. Dexter. "Indexing Consistency and Quality," JASIS, v 20, n 3, July 1969, p 259-267.