Dagobert Soergel
University of Maryland, College Park, Md., USA

Automatic and Semi-Automatic Methods as an Aid in the Construction of Indexing Languages and Thesauri[1]

Soergel, D.: **Automatic and semi-automatic methods as an aid in the construction of indexing languages and thesauri.**
In: Intern. Classificat. 1 (1974) No. 1, p. 34–39

Development of a framework for the description and classification of statistical and graph-theoretical methods for the determination of terms and concepts and of relationships between and among them. Discussion of the problem of terms versus concepts in this context, the 'units of text' to be used for these methods, total count versus unit-wise count, and practical problems of data collection. Brief characterization of methods for the use of 1) frequency data in descriptor selection, 2) co-occurrence data to determine terminological relationships between terms and classificatory relationships between concepts, 3) binary relationships detected in the previous step to construct global classificatory structures. (Author)

0. Introduction

The relationships that exist between terms based on their meaning result in certain statistical patterns of occurrence and co-occurrence of these terms in text. Conversely, we should be able to conclude from observed statistical patterns of the occurrence and co-occurrence of terms what these conceptual relationships are. This idea is the basis for automatic methods in thesaurus construction. These automatic methods can assist in, but not replace, the intellectual effort needed for the construction of an indexing language or a thesaurus. "Statistics should not take precedence over human judgment in the evaluation of vocabulary, but these studies and other provide the basis for some useful decisions." In other words, the identification of terms and relationships by automatic methods should be considered as a kind of pre-processing of open-ended sources, especially abstracts and full-text documents resulting in a list of terms and potential relationships between these terms. The results of this pre-processing are then used, along with other sources in the further steps of thesaurus building. Fully automatic thesaurus building may be attractive as an idea, but it is not feasible.

There are two levels of complexity or sophistication at which automatic or semi-automatic methods can be applied. On the first level we deal with frequency and co-occurrence data for terms and/or concepts that at one point or other have been picked or assigned by a human editor. These data are then used to select preferred terms and descriptors and to detect classificatory relationships among them. On the second level of complexity we deal with the text of documents, abstracts, or search request statements submitted by users and the terms must be isolated in the text before further processing can begin. This paper deals with both levels of complexity.

The paper starts with an exploration of the units from which frequency and co-occurrence data may be gathered and of different kinds of data collections and counting (1). It proceeds to methods by which promising descriptor candidates can be identified from frequency patterns (2) and how relationships between terms can be detected from co-occurrence patterns (3). Both sections are concerned with "local" information. Section 4 turns to the automatic derivation of classification schemes, i.e., "global" structures, from co-occurrence data or from the indication of relationships between terms. One should keep in mind that the following considerations hold also for updating of indexing languages and thesauri and not only for initial construction.

1. Definitions

Some introductory classifications and definitions are in order to create a basis for the description of methods.

1.1 Counting terms versus counting concepts

This is a most important distinction, even though it is often overlooked. Frequency and co-occurrence data for terms can be used for developing the terminological structure, i. e. for selecting the preferred term from a class of synonyms and quasi-synonyms (where the synonymity is known beforehand) and, on the second level, for the detection of synonyms. Frequency and co-occurrence data for concepts can be used for developing the classificatory structure and for selecting descriptors. (In this paper, descriptor is defined as a term or notation that, in a 1 – 1 relationship, designates a concept to be used in search request formulations and document representations.)

The frequency of occurrence of a concept be computed as the sum of the frequencies of all the terms designating that concept. In many studies this point is overlooked, and term frequencies are used where concept frequencies would be appropriate. (A related and somewhat tricky point is the following: Suppose we have a concept *A* and three narrower concepts *B, C,* and *D*. If *A, B, C, D* are all seldom used, we may not consider them to be good descriptor candidates. However, if we do not use *B, C,* and *D* as descriptors and say "USE BT *A*" instead, we have to sum up all frequencies to obtain the new frequency of *A*. This *new* frequency may then suggest that *A* should

1 This article is a slightly differing version of Chapter H (incl. also Sect. F. O. 4.4) of Prof. Soergel's book: Indexing languages and thesauri: construction and maintenance. Los Angeles, Calif.: Melville/New York: Wiley 1974. ca. 600 p.

in fact be a descriptor, or it may still be so low that we should rather say "USE BT A'", A' being broader than A.)

1.2 Counting words versus counting terms

Again, this distinction is important. In processing, say, an abstract, a computer program can only recognize words as a string of characters between two blanks. It cannot recognize multi-words terms immediately, unless such terms are identified prior to computer processing by an editor, as for example in free term indexing. To a certain extent, multi-word terms can be identified in a second step through syntactical analysis of a sentence and through statistical analysis (see Section 3.1 (3)).

1.3 Units of text

"Units of text" (broadly defined) can be one of the following:
- search request statements as submitted by the user
- search request formulations in terms of descriptors
- sets of indexing terms contained in document representations
- indexing languages and thesauri (they constitute sets of terms)
- titles of documents
- abstracts of documents
- individual sentences of documents or abstracts
- paragraphs of documents
- full text of documents.

"Corpus of text" is any assembly of units of text of one or more types.

1.4 Methods of counting (total versus unit-wise, weighting)

There are two main methods of counting frequencies and instances of co-occurrence.
(1) *Total count:* a term or concept occurring nine times in one unit is counted nine times to obtain the total count.
(2) *Unit-wise count:* a term or concept is counted only once for each unit, even if it occurs nine times in the unit.

Furthermore, counts can be weighted as follows:
(1) Weighting by source:
A higher weight is assigned to a term or a concept if it occurs in an important source than if it occurs in a marginal one. This method is particularly appropriate if statistics are based on a count of the number of other thesauri and similar sources in which the term or concept occurs.

Remark: In a situation where documents indexed by free terms serve as sources the following modified procedure for weighting by source has been used: it is possible that the term profile of a document contains only terms that, due to low frequency, would not qualify as descriptors. Thus, none of the terms used to index the document would be included in the indexing language, and the document would not be accessible at all in retrieval. In order to avoid this the weight of a document is decreased each time one of its index terms is selected as a descriptor. (In the beginning all documents have the same weight). After each weight modification the frequency count is done all over again. This enhances the chance of documents that are indexed only by seldom-used terms to have at least some of their terms included in the indexing language. Whether or not this method is useful in a fully automated selection procedure can be left open in this paper.

(2) Weighting by importance (position) in the source: For example, a term occurring in an important position is counted 2 or 3 instead of just 1. Or one may simply select a concept as term if it has been used among the four most important terms in indexing any one document. (This example presumes that the terms assigned to a document have been ranked according to their importance for the document.)

(3) If the count is unit-wise, the within-unit frequency can be used as weight. (In many cases this will be equivalent to a total count.)

1.5 Actual collection of data

First of all, the corpus of text must be established. Various types of "units of text" can be collected through search request statements and experimental indexing solicited in the material collection phase of thesaurus development, from the text run, from other operating ISAR (Information Storage and Retrieval) systems and, for purposes of updating, from the operation of the ISAR system for which the system has been built. Titles, abstracts, and full text of documents are available in abundance and the main problem is proper sampling.

The next, and on a practical level often more pressing problem is how to actually obtain a frequency and co-occurrence count. Titles, abstracts, and full-text documents must be available in machine-readable form, except for very small studies. The collection of data on the frequency of descriptor use in search request formulations and document representation, is very easy in mechanized ISAR systems, however, it is difficult in manual ones. In a card catalog one may check to see whether the volume of cards filed under a descriptor has become too large. (This procedure is facilitated if each descriptor has a guide card with a tab.) Still the catalog has to be scanned regularly. With edge-notched cards or peek-a-boo cards it is difficult to obtain any statistics at all. One possibility is monitoring the frequency of descriptors while searching. (If the search results show that a descriptor is used very frequently or very rarely, one may take action on this particular descriptor.) But this is a haphazard kind of procedure. With peek-a-boo cards descriptors that are used very frequently or very seldom can be selected just by going through and having a short glance at every card. With additional effort it is even possible to get association measures for specific pairs of descriptors. (There is an apparatus that counts holes in Termatrex cards (peek-a-boo) or combinations of those cards.

The possibilities of data collection in mechanized ISAR systems are illustrated by the plans formerly developed by ASTIA to produce three listings to provide the thesaurus builder with frequency data and related information:

Example:

Descriptor frequency listing.

Descriptor	*Frequency in indexing*	*Frequency in searching*
Jet planes	*2216*	*37*
Jet sea planes	*22*	*9*

Low-frequency descriptor manual file.

Descriptor	*Document numbers*
Alpha chambers	*AD 204 929*
First aid kits	*AD 219 127*
	AD 222 912

This file can be used to assess the value of the infrequent descriptors by looking at the documents. (In addition this file is very useful for retrieval; in searching for infrequent concepts manual look-up is faster than computer search.)

List of context descriptor sets.

Aircraft	*5325 (total frequency)*
Co-occurring with	
Engine	*2733 (co occurrence frequency)*
Wing	*2201*
Rudder	*2182*
Stabilizer	*2180*
Airframe	*2023*
Fusilage	*1845*
Autopilot	*1673*
Supersonic	*1580*
Rotor	*1512*

Such lists are useful for the methods dealt with in Section 3. From some mechanized ISAR systems frequency counts are available.

2. Selection of preferred terms and identification of descriptor candidates

I would like to reiterate that the results of the methods to be described should be used only as suggestions that have to undergo thorough scrutiny. The final selection decisions should be based mainly in substantive considerations.

2.1 Selection of preferred terms

If one member of a class of synonymous and quasi-synonymous terms has a considerably higher frequency than any other member, it is a strong candidate for selection as the preferred terms representing the class. (But even in this case one should not overlook the possibility of coining a new term.) The most appropriate data for the purpose are:

- a unit-wise count of terms in search request statements representative of the ones to be expected
- a unit-wise weighted count of terms used as preferred terms in other indexing languages and thesauri.

2.2 Selection of descriptor candidates; level 1

First of all, counts of concepts rather than counts of terms should be used for descriptor selection. Appropriate data for this purpose are:

- a unit-wise count of concepts in search request statements (this is the most important one)
- a unit-wise count of concepts in abstracts, full-text documents, or other document representations
- a unit-wise count of concpets used as descriptors in other indexing languages

Frequency data collected from search request statements are straightforward to use: the more frequent the concept, the more important its use as descriptor (unless the concept is used to index, say, 80 % of the documents and is therefore almost useless in retrieval).

Frequency data collected from documents or document representations are more difficult to interpret. The problematical concepts are those that occur very frequently and those that occur very rarely. They must be examined critically to determine whether they should be selected as descriptors or not. The other concepts are strong descriptor candidates.

(1) *Concepts used very frequently*. If a concept occurs very frequently in documents, it does not have much discriminatory power in searching if it is used alone. If it is also used very seldom in searching its usefulness is in doubt. If however the concept is used with reasonable frequency in searching one should investigate to determine which of the following explanations applies:

(1.1) The concept is of general application and mostly used in combination with other concepts. This type of concept can be very useful in searching.

(1.2) The concept pertains to a specific subject field and is often used by itself (as the "thematic" concept) in search requests. In this case further subdivision should be considered.

(2) *Concepts used very seldom*. If a concept occurs very seldom in documents it has very high discriminatory power. If such a concept is used frequently in searching this high discriminatory power is very welcome. For example, a concept used for indexing seven out of a hundred thousand documents (0.007 %) and occurring in 5 % of the search requests is of tremendous usefulness in searching and should be considered as a strong descriptor candidate. In fact, this concept is much more useful than a concept used for indexing five thousand documents (5 %) and occurring in 1 % (or only 0.1 %) of the search requests. On the other hand, if the concept is used seldom in searching it may be too specific, and a USE instruction to a broader concept or to a combination of semantic factors might be appropriate in order to keep the indexing language within reasonable limits. In order to achieve specific indexing it might often be useful to retain as decriptors those low-frequency concepts that belong to the central areas of the thesaurus.

Note: In the case of a concept newly introduced in the subject field no conclusions should be drawn from low frequency.

The above considerations can be formulated more precisely in terms of costbenefit analysis; the inclusion of a concept in the indexing language incurs costs (larger files, indexing more difficult as size of indexing language increases, etc.). These costs have to be distributed over the documents indexed by that concept. If the documents are few, the cost per document is high. This cost can be justified only if there is a corresponding benefit on the searching side, that is, if the concept in question is used often in search requests.

2.2.1 Evaluation of frequency data from operating ISAR systems

This Section relates mainly to frequency of descriptors in indexing, but many of the points are important for other types of counts also. First of all, frequency data from one's own ISAR system are much better than frequency data from another ISAR system. How much better they are, this depends on the similarity of the other ISAR systems in subject matter of coverage and user community to be served. Frequency data from one's own ISAR system should be collected on a continuing basis unless the costs for doing so could not be justified.

Next, the frequency of a concept in an operating ISAR system has to be judged with a view to the following factors:

- relatedness of that system to the system for which the thesaurus is being built
- size of the collection
- age and subject field of the collection
- time elapsed since the first use of the concept within the ISAR system and increase of the collection within that timespan
- rules used in indexing (if generic posting in indexing is used – i. e., with a specific descirptor all the broader descriptors are to be used in indexing as well – the count of the more general descriptors is inflated)
- frequency of the concept at hand as compared with the frequency of other concepts.

If the ISAR system uses very exhaustive indexing, resulting in a large number of descriptors per document, descriptor frequencies in general tend to go up. It might therefore be better to use the rank of a concept in a list arranged by decreasing frequency rather than frequency itself.

2.3 Selection of descriptor candidates; level 2: analysis of frequency patterns

A more sophisticated procedure is as follows: Determine concepts that occur with high within-unit frequency in a few units. These concepts have more discriminatory power than concepts that occur in many units with about the same frequency. (The total count for both concepts may be the same.) Some statistical measure has to be established to determine that the deviation from equal distribution over documents is big enough to make a useful descriptor. We reiterate that this type of analysis is more appropriate for concepts than for terms.

3. Detection of term or concept relationships from co-occurrence patterns

3.0 Nearness measures

We want to determine quantitatively which pairs of terms co-occur more often than others. For this purpose we must define a measure of co-occurrence (association, nearness). A very simple and often-used measure is the following.

Example:

$$r(A,B) = \frac{2c}{a+b},$$

where:

a = frequency of A
b = frequency of B
c = frequency of the co-occurrence of A and B

There are many other nearness measures, many of them more complicated and some of them more appropriate. In some systems a relationship between two terms is introduced in the thesaurus whenever the nearness measure is above a certain threshold. These relationships are then used indiscriminately in retrieval. However, a high nearness measure can mean many different things and it is therefore advisable to make sure first how the relationship between two terms should be interpreted.

3.1 Interpretation of high association between two terms A and B and between two concepts A and B

High association between two terms *A* and *B* can mean any of the following.

(1) Definitional relationship.
(1a) *A* and *B* are synonymous.
(1b) *A* and *B* are quasi-synonymous (designate equivalent concepts).
(1c) *A* and *B* designate concepts that are similar in meaning.
(1d) *A* and *B* designate concepts that are in a class-inclusion or topic-inclusion relationship.
(2) *A* and *B* designate concepts that are in a relationship of contextual contiguity.
(2a) Part-whole relationship.
(2b) Other connected hierarchical relationships.
(2c) Empirically connected.
(3) Two words form a multi-word term; for example, *Information* and *Retrieval* co-occur heavily.

The synonymity interpretation is highly unlikely if the units are sentences because two synonyms are seldom used in one sentence. It is also unlikely if the units are abstracts or sets of indexing terms because an abstractor is unlikely to use Synonymous Terms within an abstract. Synonymity is likely, however, if the units are paragraphs and even more so if the units are full-text documents because people tend to use synonyms in order to achieve variation. Synonymity is also very likely if the terms co-occur in search requests or interest profiles where the users have been instructed to include all synonyms they can think of as OR-combinations. In fact, if two terms occur in an OR-combination in *one* search request one should immediately consider the possibility that they are synonymous or nearly related. Interpretation as equivalence, similarity in meaning, class inclusion or topic inclusion and as contextual contiguity may be appropriate whatever the units are. However, if the units are sentences the relationship "empirically connected" is the most likely unless we have a multi-word term. The interpretation as a multi-word term makes sense only if the units on which the computations are based are sentences.

In dealing with class inclusion, topic inclusion, and contextual contiguity it is more appropriate to use concepts instead of terms and compute co-occurrence data accordingly. Similarity in meaning is in between.

High association between two concepts indicates contextual contiguity. If this is not a hierarchical relationship but

rather the relationship "empirically connected", one should check to see whether the concept should be introduced as a precombined descriptor.

3.2 Second-order associations between terms for the detection of definitional relationships

The associations, as measured by the nearness measure in the previous paragraph, are called first-order associations. A second-order association can be defined as follows (compare Figure 1): The list of terms associated in first order with *Airplanes* is called the association profile of *Airplanes.* In the same way, we have an association profile for *Aircraft.* The degree of similarity between the two association profiles is called the second-order association between the two terms. If we have a situation where the first-order associations mostly correspond to contextual contiguity then similarity in the association profiles of *A* and *B* means that the concepts designated by the terms *A* and *B* tend to occur in the same empirical context. This we would expect if *A* and *B* were synonymous or in a class inclusion relationship. We therefore conclude the other way around: if two terms *A* and *B* have similar association profiles, then they are expected to be in a definitional relationship. That means they are either synonymous or quasi-synonymous or they designate concepts that are similar in meaning or in a class inclusion or topic inclusion relationship.

Figure 1: Example of second-order association (3.2).

Airplanes	*Aircraft*
Associated terms:	Associated terms:
Wing	Engine
Engine	Wing
Rudder	Rudder
Airframe	Stabilizer
Stabilizer	Airframe
Autopilot	Fusilage
Jet	Autopilot
Supersonic	Supersonic
	Rotor

One may also obtain association profiles directly by asking individuals to name terms associated with a similar term. Terms related by definition as well as terms related by contextual contiguity are obtained by this method.

3.3 The use of inconsistent association profiles for the detection of homonyms

"Consistent association of a given term with two groups of terms, each representing an entirely different discipline, may indicate that a homograph exists and that separate terms should be established. For example, if the term *Precipitation* were frequently associated with such terms as *Climate, Clouds, Temperature, Humidity, Solutions, Chemical reactions,* and *Solubility,* it is evident that two separate concepts are being indexed as one; therefore, two terms should be established or the one term redefined."

3.4 Detection of hierarchical relationships

Earlier it was mentioned that a high second-order association between two terms *A* and *B* may mean that they are synonymous or that the concepts designated by the two terms are in a class-inclusion relationship. A hierarchical relationship may be surmised especially if there is a one-sided overlap. This means if term *B* is considerably less frequent than term *A* and if almost all units containing term *B* also contain term *A* we may suspect that *B* designates a concept that is narrower than the concept designated by *A*. However, *B* may just as well be a rarely used synonym of *A*.

The one-sided overlap criterion can be applied also to concepts, and this application is even more useful. We can suspect that a concept *B* is narrower than concept *A* if almost all units dealing with concept *B* also deal with concept *A* and if the number of units dealing with *A* is much larger than the number of units dealing with *B* (compare the definition of hierarchical relationship in Section C 3.2). (Note that it is quite likely that a specific concept is used more often than a more general concept. For example, there may be articles dealing with a certain biological species, few of which bother to mention the genus to which the species belongs. However, in this case the same genus will usually co-occur with a number of other species so that we do not have the situation of one-sided overlap between the genus and one species.)

For this kind of analysis it is useful to use as units either search requests (where the searches have been instructed to include Narrower Terms in their requests if they want to retrieve material on Narrower Terms also) or sets of indexing terms (if the indexers have been instructed to use Broader Terms for which the document may be of interest, too).

3.5 Combined application of different methods

If full-text documents are used as units, high first-order association may mean any type of relationship. We could now proceed to compute second-order associations. Two terms that have a high second-order association are likely to be in a definitional relationship. We could subtract from the list of pairs with high first-order association those pairs with a high second-order association. The remaining pairs should be in a relationship of contextual contiguity.

Since the detection of contiguity relationships should be based on a count of concepts rather than a count of terms the following procedure might be useful: Detect synonymity relationships by second-order associations and form classes of synonymous and quasi-synonymous terms accordingly. Each class corresponds to a concept. It is now possible to obtain a frequency count on concepts and determine contiguity relationships between them.

4. Automatic derivation of classification schemes ("global" structures)

The previous Sections were concerned (a) with the identification of terms and their pairwise interrelationships and (b) with the identification of concepts and their pairwise interrelationship. The latter sould be called "local" classificatory information. A next step is the automatic derivation of a "global" structure (a classification scheme) to obtain the overall picture. There are two interrelated tasks:

(a) Find useful groupings of concepts.
(b) Find a pattern of subdivision of the set of documents into non-overlapping classes.

There are two main methods to perform these tasks:

(1) Clustering methods.
(2) Graph theoretical methods.

To some extent these methods overlap; one might even say that the clustering method is a special case of methods based on graph theory.

4.1 Automatic derivation of classification schemes by clustering methods

The basic idea is to define a nearness measure in the set of concepts (such as the nearness measure defined in Section 3.0) or in the set of documents, respectively, and then derive clusters of near concepts or near documents, respectively. A cluster of concepts can be defined roughly as a set of concepts that tend to be nearer to each other than to concepts outside the cluster. Various cluster definitions and various clustering procedures are used.

4.2 Automatic derivation of classification schemes by graph-theoretical methods

These methods are probably more appropriate to the relational nature of thesaurus data. In this approach one starts with a set of concepts (the nodes of the graph), and relationships between the concepts (the connections between the nodes called the arcs of the graph). Thus, the input consists of "local" information, namely, terms and pairwise relationships between them (the terms and the interrelationships may have been derived by the automatic methods discussed in Sections 2 and 3). Algorithms derived from graph theory make it possible to put together the over-all structure, the total graph, as in a jigsaw puzzle. One example of this is computer-assisted hierarchy construction by "chaining" BT-NT cross-references. Application of graph theory might lead to more efficient procedures for this purpose. In this case the graph is based on hierarchical relationships. It is also possible to use RT relationships as the base for the graph. In either case one might look for relatively close (strongly connected) subgraphs; these would then correspond to subdivisions of the classification scheme to be developed.

A simple-minded method, based on RT relationships, is as follows: pick any term A. R $(1, A)$ is the set of all terms related to A. R $(2, A)$ is the set of all terms that are related to any term in R $(1, A)$. If R (n, A) = R $(n + 1, A)$, then clearly R $(n + 1, A)$ = R $(n + 2, A)$ and R (n, A) is a closed subset. If the difference between the number of elements in R $(n + 1, A)$ and the number of elements in R (n, A) reaches a minimum then R (n, A) is relatively closed, the closure being sharper as the minimum is smaller.

REPORTS AND COMMUNICATIONS

Classification Research and Development in India, 1968–1974

FID/CR Report 14, to be published in the first half of 1974, will highlight some of the researches in the field of classification carried out in India, since 1968. The present note mentions some of the topics covered in the report.

1. Interdisciplinary Subjects

Interdisciplinary subjects, resulting from multidisciplinary and interdisciplinary research, are emerging at an accelerated pace, particularly during the past two decades. The coextensive representation, either as a subject heading or as a class number, of subjects falling in such interdisciplinary fields has posed problems to designers of schemes for classification and systems of subject headings. Subject specialists have, from time to time, examined and commented upon the pattern of organization of research which produce such interdisciplinary associations and on the types of hybrid subjects generated. A study of these observations and analysis and classification of a number of interdisciplinary subjects have led to a typology of the modes of combination of ideas and of subjects and formation of interdisciplinary subjects. The modes of formation recognized are: Fission, fusion, distillation, lamination 1, lamination 2, clustering, and agglomeration. This typology, together with a few guiding principles for recognition of the core entity of study in a subject-field, has facilitated the formulation of some guiding principles for the representation, classification, and helpful arrangement of inter-disciplinary subjects and subject-fields.

2. Absolute Syntax

The use of a natural language for representing a subject raises, among other things, the problem of linguistic syntax which varies from one language to another. However, at a deeper level – close to or at the plane of formation of ideas and combination of ideas, that is, at the level of thinking – it may be possible to discern a more stable and consistent structure of subjects less constrained by language and culture. The sequence in which component ideas of subjects usually arrange themselves in the minds of the majority of normal intellectuals while thinking about or formulating a subject is called the Absolute Syntax of Ideas among intellectuals. It is conjectured that such a syntax of ideas exists. It may not coincide with linguistic syntax. Findings in the field of linguistics and psycholinguistics, developmental psychology, neurophysiology, biocybernetics, and general systems theory appear to lend support to the postulate of absolute syntax. It is proposed that the sequence of component ideas in a subject – that is facet syntax – should parrallel the absolute syntax such that it would be of maximal acceptability to a wide range of users. The helpfulness of this