TOSAR—A System for the Structural Formula-Like Representation of Concept Connections in Chemical Publications†

R. FUGMANN, H. NICKELSEN, I. NICKELSEN, and J. H. WINTER*

Hoechst A.G., 6230 Frankfurt (Main), Germany

Received October 8, 1974

A system is described for representing graphically the conceptual contents of a document, such as preparations and processes delineated in a patent or journal article.

Much work has been undertaken in the past to provide the scientist with more efficient methods of information retrieval, and much progress has already been achieved toward this goal. However, strictly speaking it is not sufficient to deliver the literature of (potential) interest to the scientist on his desk. For the human never has available for literature studies more than a very limited reserve of time and attention, and within these limited reserves he must be able to process mentally both the literature submitted to him as a response to his (current awareness or retrospective) search as well as the special journals or patent classes to which he subscribes. Because it was phrased in a language difficult to understand, many a document of considerable interest to him might escape consideration, even though it may be lying on his desk or he may be holding it in his hands.

In this respect the position of the chemist is relatively fortunate compared with that of other scientists. The most important concepts with which he is continually concerned are the substance concepts, and these are represented most unambiguously and lucidly by the structural formula. At first glance he can recognize the constituents of a molecule and the way in which these constituents, the atoms, are typically connected with each other in each individual substance. In this respect, the structural formula is greatly superior to any other kind of mode of expression including scientific nomenclature. It is not an exaggeration to state that, except in cases of extremely simple molecules, the representation of substance concepts other than by structural formulas is of little value to the chemist in the laboratory or in the plant. For instance, only a few chemists are able to recognize the molecular structures implied by the following names within a reasonable expenditure of time of, say, five minutes.

2-Oxabicyclo[3.1.0]hexene (3,4)
2-Oxatricyclo[4.1.0.0^{3,5}]heptane
Pentacyclo[19.3.1.1^{3,7}.1^{15,19}]octacosa-1(25),3,5,7(28),9,11,13(27),-15,17,19(26),21,23-dodecaene

Had the corresponding structural formulas been given instead of nomenclature, no obstacle in conveying the information would have arisen. In other words, the employment of (linear!) language text instead of the multidimensional structural formula may constitute a serious communication barrier in chemistry.

We should now like to deal with another kind of conceptual opacity in chemical literature, which also constitutes a serious communication barrier. As an example, we shall discuss the text of a patent claim which is typical of the polymer field: Thermoplastic moulding compositions from (A) 5–99% by weight of a graft polymer prepared by reaction of 10–95% by weight of a mixture of 50–90% by weight of styrene and 50–10% by weight of acrylonitrile, whereby both components can be replaced entirely or partly by their alkyl derivatives, with 90–5% by weight of a polymer consisting of at least 80% by weight of a conjugated diolefin; (B) 0–94% by weight of a copolymer prepared by polymerizing 50–95% by weight of styrene and 50–5% by weight of acrylonitrile, respectively, the alkyl derivatives of both monomer components, whereby the sum of acrylonitrile and styrene in the components A and B must be at least 50% by weight, characterized by an additional content of 0.1 to 3% by weight, based on the weight of all components, of a stabilizer combination, consisting of (a) 2,6-di-*tert*-butyl-*p*-cresol and (b) thiodipropionyl ester, whereby the alcoholic component must have 9–20 carbon atoms in the hydrocarbon chain and the weight proportion between both components a and b varies between 1 to 6 and 6 to 1.

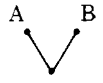
If an expert has to deal with many of these formulations every day, then inevitably he will have to leave many of them unconsidered or at least will have to content himself with an insufficient depth of analysis.

The root of this conceptual opacity and of the resulting communication barrier is quite similar to the nomenclature examples discussed in the foregoing: An entity of knowledge which is multidimensional in nature is forced into a one-dimensional mode of expression, the only difference between both examples being that in the latter case it is the connections of whole conceptual entities instead of the connections of the atoms in the nomenclature examples, which are obscured by the employment of linear text.

If it is intended to improve on this situation, then this can, with any degree of promise, obviously be done only by representing concept relations, also, in a similarly multidimensional manner.

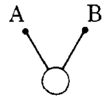
We were concerned with this problem in our work to develop a computerized search system in which not only certain concepts such as (sub-) structures, properties, processes, apparatus, etc., would be admissible as search requirements, but also the connections into which they are put in an individual document in the file. The documentation system TOSAR[1] (topological representation of synthetic and analytical concept relations) which thus came into being requires a multidimensional representation of the typical concept relations of a document before these relations can be stored in the computerized file. Thus, as a highly desirable by-product, the model of a multidimensional, structural formula-like representation of an entire document originated from this approach too.

Of course, a few rules had to be established on how to represent concept relations graphically analogous to those which chemists had to establish a hundred years ago, when beginning to draw structural formulas. In particular, three fundamental graphical figures were laid down in the TOSAR system:

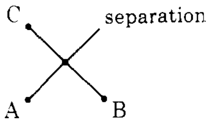
1. 

means A AND B

† Presented at the IUPAC Congress, Hamburg, Germany, Sept 9, 1973.
* Author to whom inquiries should be addressed.

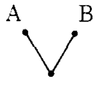
2. 

means EITHER A OR B

3. 

means "separation of C into its components A AND B"

The components of C are, for the sake of lucidity and logic, even in highly complicated cases, represented in a separate, ancillary graph, which in this case looks like:

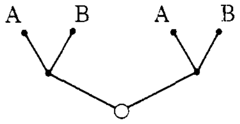


Mixture C

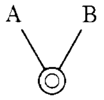
"Concept C is a mixture of the components A AND B"

(In other words a node "•" at the root of a figure in the shape of the letter V is typical of logical conjunction, a node "O" for logical disjunction).

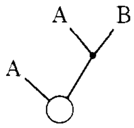
For two frequently occurring kinds of more complex logical concept relations, abbreviated graphs containing the nodes "◎" and "⊗" at the root of an elementary graph were admitted for the sake of brevity:



"EITHER [A AND B] OR [A OR B]":

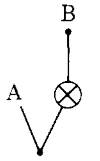


(synonymous with "A AND resp. OR B")



"EITHER A OR [A AND B]":

(synonymous with "A AND OPTIONALLY B)



Let us consider a simple example of the application of these elementary graphs to a patent claim (Figure 1). Acrylonitrile is, in the presence or absence of another comonomer, polymerized in an inert atmosphere and in the presence of a special catalyst to form a polymer of a particularly low degree of discoloration. In this reaction, dimethylformamide or a hydrocarbon or an ether serves as a solvent, each of them also being admitted in mixture with the other one(s) recorded.

The outcome of a process always appears at the root node of a V figure. Logically, such a node bearing result-concept(s) can constitute the "starting" node of another succeeding process. An example of this situation is shown in Figure 2. An acrylic fiber (see center of graph) is stretched and then submitted to mild oxidation in the presence of oxygen and water vapor at 100–160°. The resulting oxidized fiber is carbonized at temperatures below 2600° in an inert atmosphere, *e.g.*, hydrogen, argon, nitrogen, to yield carbon fibers of particularly high tensile strength, useful for materials in aircraft construction.

The broken connective lines in the upper part of the figure indicate that the corresponding processes were not expressly described in the document to be represented but only implied. In this example it is intimated that fibers suitable for undergoing the sequence of processes described in detail can be prepared by spinning an acrylic polymer, which may in itself be prepared by polymerization of acrylonitrile in the presence or absence of one or several comonomers taken from the group consisting of (meth-)acrylic acid (respectively derivatives of the kind indicated in the graph), acrolein, vinylacetate, methylvinylpyridine. In the following figures, which hopefully are self-explanatory, additional sequences of processes are graphically depicted.

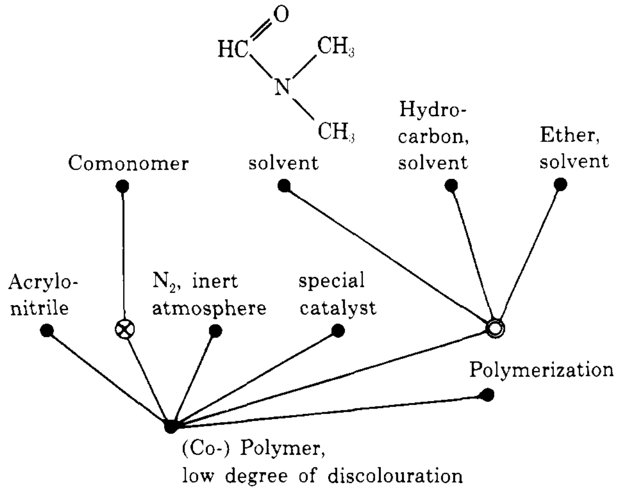


Figure 1.

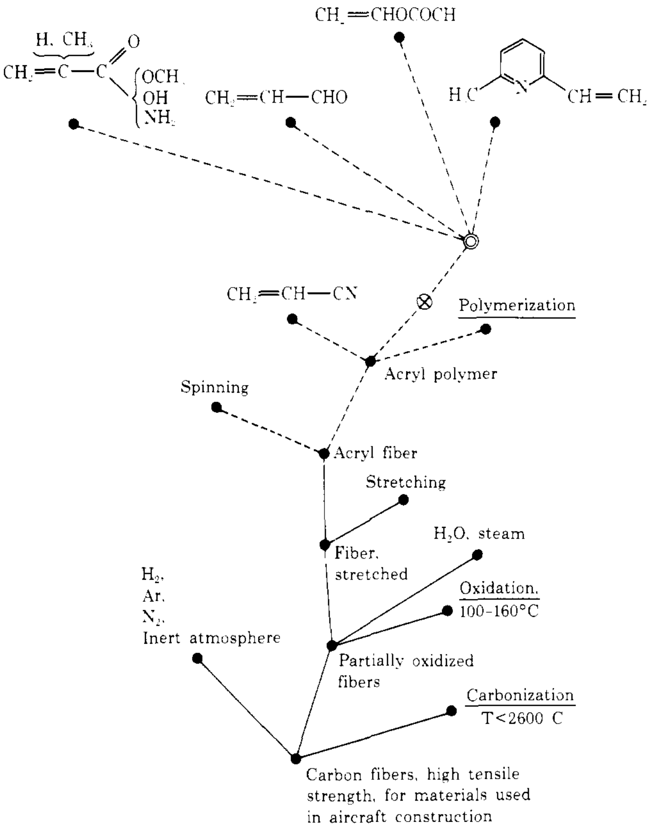


Figure 2.

We are now prepared to reconsider the natural-language patent claim for thermoplastic moulding compositions which was discussed in the beginning (see Figure 3). In the graphical representation of this patent claim, it is apparent at first glance that it consists of at least two components, a graft copolymer A and a mixture of stabilizers, a third one (copolymer B) being only optional. The way in which both polymers are synthesized requires no further explanation, neither does the logical compatibility of the various monomers in the polymerization processes.

If one is also interested in the typical proportions in which the components of the composition and the various monomers are reported in the patent claim, then they too can be introduced into a graph of the TOSAR kind. Above all, it is possible to depict most lucidly which of the particular components add up to 100%. The corresponding per-

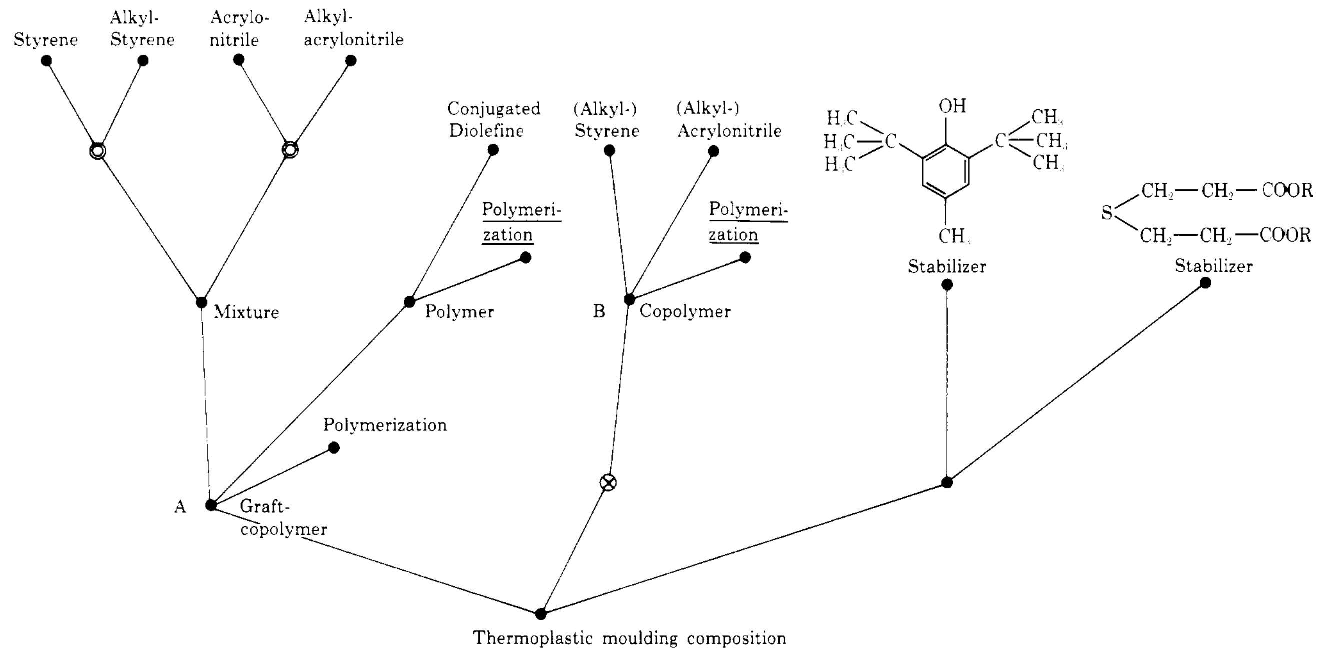


Figure 3.

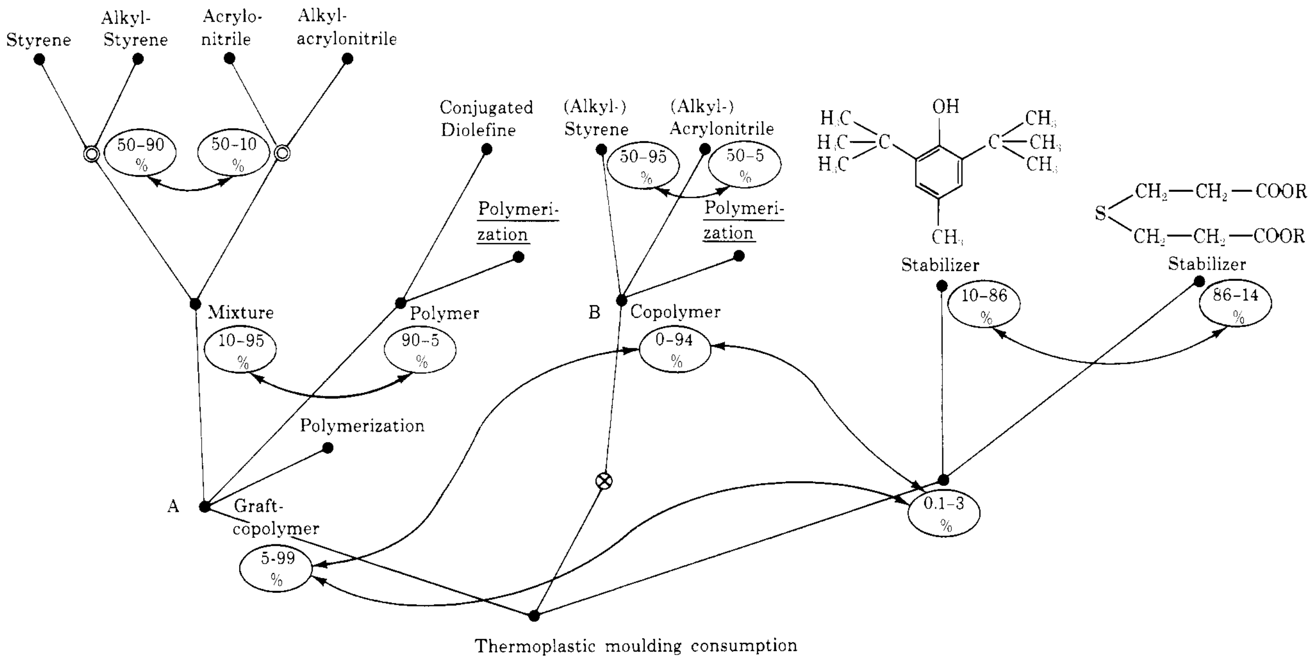


Figure 4.

centage data are circled and interconnected with arrowed, nonlinear lines in Figure 4.

We have developed the TOSAR system for IDC (International Documentation in Chemistry mbH., Frankfurt (Main), Hamburger Allee 26, Germany) with the financial support of the Department of Education and Science of the Federal Republic of Germany. In IDC, this system is at present being used on an experimental scale for the documentation of several thousand polymer documents. This is being undertaken to overcome the well-known deficiencies of the manual and computerized systems for the documentation of polymer literature which are currently available on the market. It has already become apparent that a considerable number of chemists in this field are prepared to learn how to read these graphs when they are represented as responses to their future enquiries. Not only do they appreciate the considerably increased lucidity of the graphical representation but also the much higher degree of order into which the contents of a document are automatically brought when the graph is drawn. This relieves them of much search effort in locating the concepts of interest in documents. It is inherent in any kind of linear text, that closely connected concepts are widely scattered over the sequence of sentences or within one single, giant sentence of a patent claim. For example, any reference to the method of preparation of a certain starting material amid a series of processes will interrupt the linear coherence of this sequence of processes, as will any subsequent indication of the properties and uses of substances already mentioned in the earlier text. This inevitably gives rise to considerable scattering of closely interconnected concepts over the entire natural language text. In order to be certain of having encountered all information of potential interest, the reader will have to scan through the text from the beginning to the end. In a graphical representation, however, it is possible to bring together at a place easy to locate or foresee any information connected with a concept. For example, for a given polymer its method of preparation and further processing, its properties, uses, etc., can be brought together at the node allotted to that polymer in the graph. Thus, inspection of the information allotted to that node is sufficient to ascertain whether or not the information of interest for that polymer is indeed presented in that document. The document can then be very quickly accepted or rejected as a response to an inquiry.

Once a chemist has learned to read the graph for a document, he will advantageously make use of this method for representing and thereby illustrating complicated concept connections he encounters in his profession (and in his daily life!). An example is phrasing a patent claim of one's own or comparing different patent claims with respect to their degree of overlap or the gaps existing between them.

There is every indication that such a guide to the graphical representation of concept connections is sufficiently simple, efficient, and general to be employed on a versatile basis.

ACKNOWLEDGMENT

This publication is dedicated to Professor Dr. Schultheis, the initiator of IDC, on the occasion of his 70th birthday.

LITERATURE CITED

(1) Fugmann, R., Nickelsen, H., Nickelsen, I., and Winter, J. H., "TOSAR—A Topological Method for the Representation of Synthetic and Analytical Relations of Concepts," *Angew. Chem., Int. Ed. Engl.,* **9,** 589–595 (1970).

A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures

GEORGE W. ADAMSON* and JUDITH A. BUSH

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield S10 2TN, England

Received October 14, 1974

A group of 39 structures with local anesthetic activity has been classified automatically by calculating similarity or dissimilarity coefficients between pairs of structure diagrams and applying cluster analysis to the results. The performance of a number of similarity and dissimilarity coefficients has been compared using the relationship between structure and property. Simple coefficients and a distance function give more satisfactory results than functions using probabilistic weighting or standardized distance.

Techniques for the automatic classification of chemical structures could have application in the storage and retrieval of chemical information[1] and in pattern recognition studies on chemical data.[12]

The 20 common naturally occurring amino acids have been classified by Sneath using numerical taxonomic methods, and on the basis of a manual analysis of their structures and some of their physical, chemical, and biological properties.[17] More recently the same compounds were classified using automatic procedures and solely on the basis of their structure diagrams.[1] As the structure diagrams of chemical compounds can be directly related to their properties,[2] then it is possible that structural parameters derived from structure diagrams will be useful in pattern recognition calculations on chemical data.[1,12]

The classification method used in the work described below is broadly similar to that applied to the amino acids;[1] however, it is applied to a group of 39 structures with local anesthetic activity.[6] The local anesthetics are structurally more diverse than the amino acids and thus illustrate the effectiveness of the method when applied to a heterogeneous set of structures.

The classification was carried out using different measures of structural relationship, and their performance was compared by using the measures of relationship and the classifications derived from them to simulate the prediction of the log (MBC), *i.e.,* minimum blocking concentration values of the compounds. The performance of the similarity (SC) and dissimilarity (DC) coefficients is thus estimated in a way which would be useful in situations where the relationship between the structure and the property is important.

METHOD OF CLASSIFICATION

The structure of each anesthetic was described as a redundant connection table, and this was used to obtain a set of augmented atom fragments[5] upon which measures of association were based. The same fragment type was used in the classification of 20 naturally occurring amino acids[1] and consists of an atom, the bonds formed by the atom, and the atoms to which it is bonded, excluding bonds to hydrogen atoms. Single and double bonds in rings and chains were also differentiated in this investigation.[1]

The anesthetics were first analyzed to identify the different augmented atoms occurring, and based on these a set of attributes was chosen to represent each structure. The following two descriptions were used.

(i) For each augmented atom type identified, a suitable set of attributes was selected to cover the different occurrences in each structure. Thus each attribute in the given set was used to indicate whether or not the particular fragment type was present in a structure at the given frequency. Using this qualitative description, multiple occurrences of the same fragment in a structure were then accounted for by additive coding.[19]

(ii) A single attribute was chosen to represent each augmented atom type and it indicated the number of occurrences in a structure of the given fragment type.

* Author to whom inquiries should be addressed.