情報知識学会誌 Vol.8, No.1

Research Paper

Semantic Structures of Chemical Data for Problem Solving Systems

Yuzuru Fujiwara
Fucalty of Science, Kanagawa University
Jianghong An
Institute of Electronics and Information Science, University of Tsukuba

There are diversified semantic relationships such as equivalence, inclusion, causality, and relativity in chemical data. Chemical problem solving systems depend on whether the semantic relationships among chemical concepts can be sufficiently represented and effectively processed.

The semantic structures represent the semantic relationships in an easily understandable way for both users and computers. In order to construct this huge and complicated semantic structures, a self organizing approach, i.e., an automatic method is necessary. The information model for semantic representation, method of self organization of conceptual structures of compounds (a kind of semantic structures), and experimental results are described. The functions of the problem solving systems include similarity measurement of compounds, analogical reasoning of reactions, naming of molecular structures and generating molecular structures from names, as well as substructure search of compounds.

1. Introduction

In order to solve problems during research and development, both a large amount of knowledge and mechanisms for manipulating the knowledge to create solutions to the problems are necessary. However, the acquisition and manipulation of knowledge depend on semantic processing. It is well known that computers are very powerful for numerical computing and symbolic handling but are not so good in dealing with meaning of information. Although numerous attempts have been made on the theory and practice of semantic processing of information, the problem remains unsolved.

Since the knowledge necessary for the real world problem solving is usually complicated in forms and huge in amount, both a flexible representation form and an efficient acquisition method are required. However, conventional models of databases and representation languages of knowledge bases such as relational model[1, 2], E-R model[3], functional model[4],SDM[5], object-oriented model[6], semantic network[7, 8], framework[9], and so on, have very strict constraints in representing semantic relationships[10, 11].

In recent years, several researches and attempts have been made to construct large-scale knowledge bases[18, 19, 20]. However, it is apparent that an automatic way should be used because the quantity of information is so huge. Self organization of information, which can be considered as a kind of process of learning, is an approach for the subject.

Furthermore, in many applications, not only

the closed world operations such as retrieval of known facts or deductive inference, but also the mechanisms for open world relevant manipulation of information or knowledge such as generation of lacking information, analogical reasoning, induction and abduction are required to create solutions to problems. Analogical reasoning is one of effective methods for problem solving. However, selection of the analogies and measurement of the similarities are not easy because not only the space of analogies is very large but also the measurement of similarities is related to semantic understanding. Fortunately, it is possible to have a broad view on the information space by using semantic structures which represent concepts and the relationships among them systematically. Moreover, it is possible to generate some lacking information in many cases by taking advantage of the structures.

The self organized semantic structures of chemical data such as the conceptual structure of compounds, the conceptual structure and the logical structure of reactions can be used to select analogies to measure similarities of compounds and reactions, to reason out reactions analogically, and so on. Moreover, names of compounds can also be generated by analogical reasoning based on names of similar compounds.

2. The model of semantic structuresIntroduction

The model used here is based on the homogenized bipartite model which may be considered as extended hypergraphs[11]. A *hypergraphs* can be defined as $HG = (N, E)$, where N is a finite set of nodes and $E \subseteq 2^N$ is a set of edges[21]. *Hypergraphs* are extended by allowing labelling, direction, recursion and nesting structures.

The model of semantic structures is defined as follows:

$$E \subseteq 2^V \quad (1)$$

$$V = V \cup E \quad (2)$$

$$E = E \cup V \quad (3)$$

$$\sigma : L \rightarrow E \cup V \quad (4),$$

where V, E and L are sets of concepts, relationships and labels respectively.

The formula (1) gives out the basic structure of the model, i.e., relationships among concepts are defined by the power sets of concepts (2^V) instead of $V \times V$ in standard graphs. The formula (2) means that E can be treated as V to construct structures recursively. In addition, V should possess internal structures, i.e., should also be treated as relationships. The formula (3) gives the mechanism for it. Finally, the formula (4) maps labels to V and/or E to represent meaning such as names, directions, roles, properties, and so on.

Fig.1 shows an example of a part of semantic structures represented by the mode, where c_i with an ellipse stands for a compound and R_i with a rectangle stands for a reaction.

3. Semantic structures of chemical data

The information of chemical data mainly consists of information on compounds and that on reactions. The semantic structures should be able to represent diversified semantic relationships among compounds, among reactions, between compounds and reactions, and so on.

The expression format of compounds is shown in Fig.2. The structure of a compound is represented as a connection table of atoms. In the Fig.2, *No.* means sequnce number of

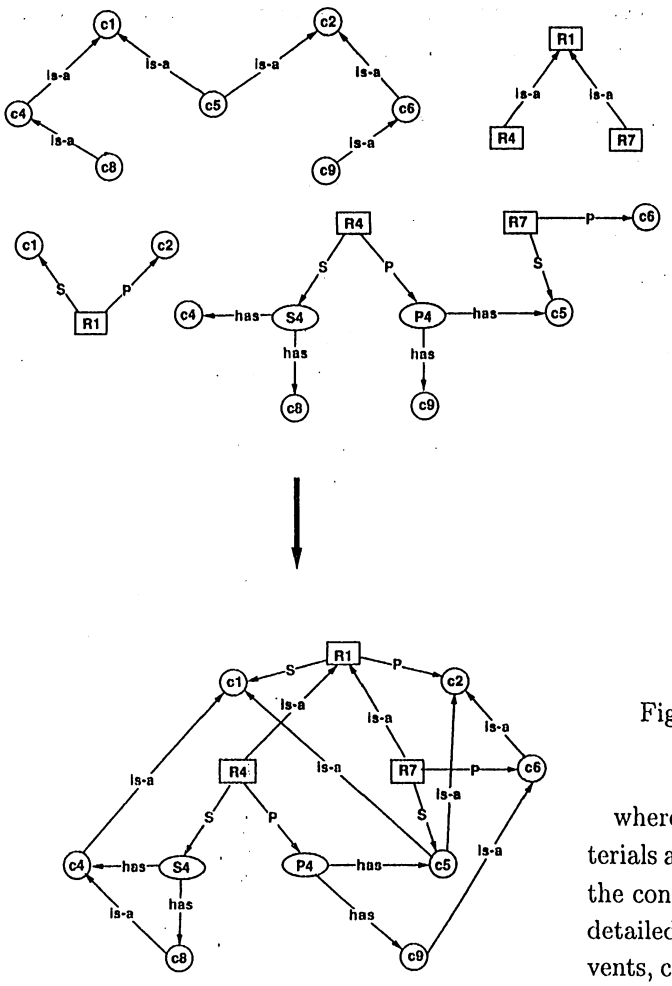
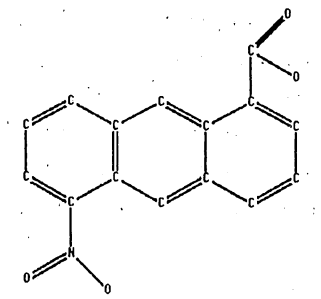


Fig.1: A part of semantic structures represented by the model



No.	Atom	Connectivity	Connection& Bond types	Position (X,Y)
01	C	03	102 104 209 000	368 344
02	C	03	211 107 101 000	411 319
03	C	03	105 206 110 000	282 394
04	C	03	210 101 112 000	368 394
05	C	03	213 103 108 000	239 419
06	C	03	109 203 114 000	282 344
07	C	03	217 119 102 000	411 269
08	N	03	218 120 105 000	239 469
09	C	02	201 106 000 000	325 319
10	C	02	103 204 000 000	325 419
11	C	02	115 202 000 000	454 344
12	C	02	104 215 000 000	411 419
13	C	02	116 205 000 000	196 394
14	C	02	106 216 000 000	239 319
15	C	02	212 111 000 000	454 394
16	C	02	214 113 000 000	196 344
17	O	01	207 000 000 000	448 234
18	O	01	208 000 000 000	196 494
19	O	01	107 000 000 000	454 284
20	O	01	108 000 000 000	274 504

Fig.2: A chemical structure and its connection table

atoms, *atom* means name of atom, *connectivity* means number of atom(s) the current atom bonded to, *connection and bond types* uses 4 colums with the same format "TNN" to describe the detail information of bonds, where T gives the bond type (1: single bond, 2: double bond, and so on) and NN gives the atom number (No.) of partner, *position* means the x-y coordinate of the atom.

On the other hand, the information about reactions is given in the form of:

$$X_1 + X_2 + ... + X_m \xrightarrow{Cond} Y_1 + Y_2 + ... + Y_n,$$

where X_i and Y_i stand for the starting materials and products respectively, *Cond* means the condition of the reaction, which includes detailed information such as reagents, solvents, catalysts, temperature, pressure, yields and so on.

3.1 Construction of the conceptual structure of compounds

The conceptual structure of compounds is a conceptual hierarchy representing *ISA* relationships among compounds.

The conceptual structure can be considered as a kind of semantic structures. *ISA* relationship may be interpreted as *more-general-than* relationships in reverse order. Furthermore, the *more-general-than* relationships among compounds can be represented by *common sub-structure-of* relationships among the chemical structures of the compounds in most cases.

The initial status of the conceptual structure is a set of chemical structures only, i.e., $C = \{c_1, c_2, ..., c_n\}$. The algorithm for constructing the conceptual structure can be described as follows:

For each $c_i \in C$, generate a set of structures C', where $c'_i \in C' \Rightarrow c'_i \prec c_i$. The notation $x \prec y$ means that x is maximal substructure of y, where x is generated by removing bonds, atoms or super-atoms (functional groups or rings) from y. Then, both the generated set C' and *ISA* relationships will be appended to the conceptual structure. The algorithm is run recursively and the hierarchical structures will be organized.

The following notations are introduced to make the description clear:
Degree of a node n, $D(n)$: The number of adjacent nodes of the node n.
Leaf Node, LN: a node n with $D(n) = 1$.
Simple Ring Edge, SRE: an edge which belongs to one and only one ring. $SRE(r)$ is a set of simple ring edges in a ring r.
Simple Ring Node, SRN: a node n which belongs to a ring and $D(n) = 2$. $SRN(r)$ is a set of simple ring nodes in a ring r.
Terminal Ring, TR: a ring r with $|SRE(r)| - |SRN(r)| = 1$.
Fig.3 illustrates an example of molecular structure and notations used in the algorithm.
In order to make the description of the algorithm simple, suppose the structures are numbered from 1 to the total number of the structures. Moreover, some detailed processes are omitted.

LN = 12,13
SRE(r1) = (2,3),(3,4),(4,5),(5,6),(6,7)
SRE(r2) = (1,2),(7,8),(8,9),(9,10),(1,10)
SRN(r1) = 3,4,5,6
SRN(r2) = 8,9
—SRE(r1)— - —SRN(r1)— = 5 - 4 = 1, r1 is a TR.
—SRE(r2)— - —SRN(r2)— = 5 - 2 = 3, r2 is not a TR.

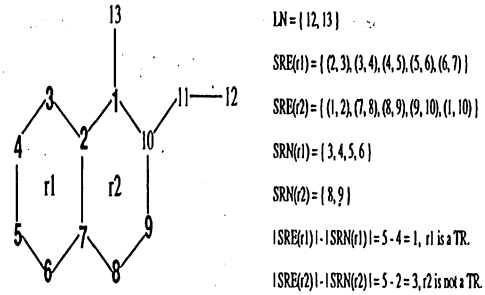


Fig.3: Example of a molecular structure and notations

[Algorithm]

```
Seq_no = N;
FOR each structure S without boarder nodes
DO {
    IF (S is not a primitive structure) THEN
       detect the SSSR(the Smallest
       Set of Smallest Rings) of S;
       FOR(each LN and TR of S) DO {
          remove the LN or TR to
          generate  subgraph SG;
          CALL ISOM subroutine to determine
          if SG exists already;
          IF (not existed) THEN
             Seq_no := Seq_no + 1;
             id of SG := Seq_no;
             register SG;
          END-IF
          c_id :=  id of SG;
          add c_id to the boarder
          concept list of S;
          add id of S to the narrower
          concept list of c_id;
       }
    END-IF
}
```

In the algorithm, the substructures generated are connected graphs because only leaf nodes and terminal rings are removed from structures. The conceptual structure of compounds is self organized systematically.
The subroutine used to detect the rings is based on the SSSR algorithm[22]. It finds the smallest set of smallest rings in a structure. The graph isomorphic determination subroutine *ISOM* is a improved version of the E. H. Sussenguth's algorithm[23].

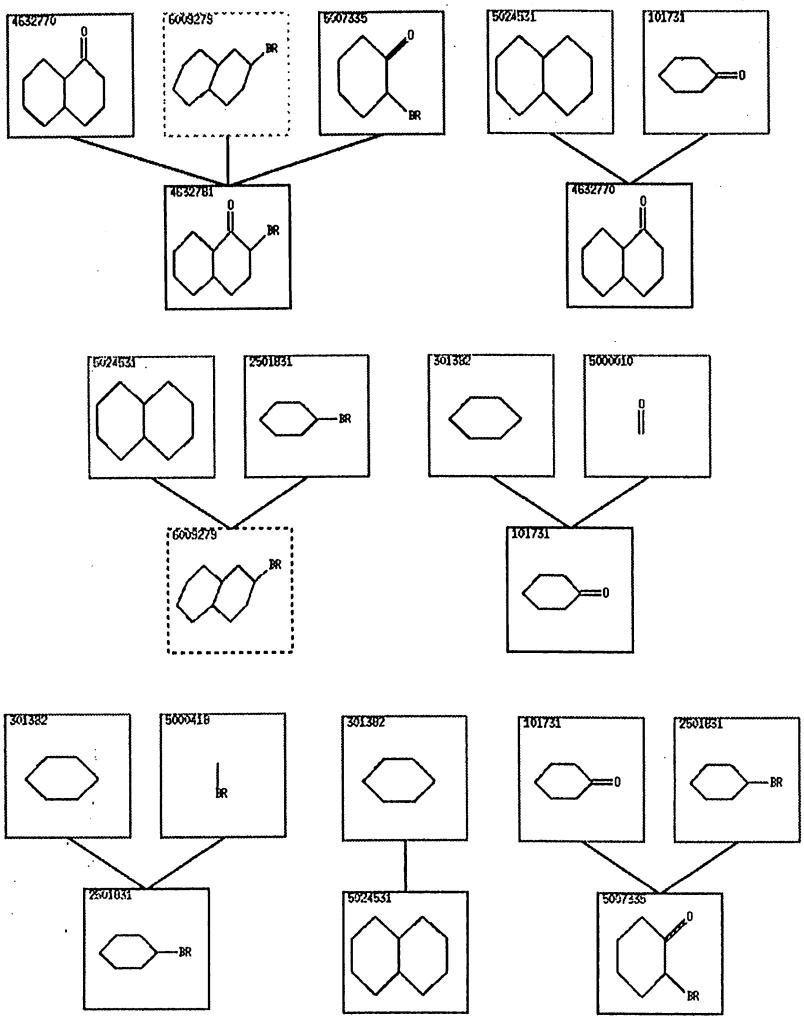


Fig.4: A set of structures and their substructures

Fig.4 and fig.5 show the processing of constructing the conceptual structure of compounds. The substructures of each compound are extracted firstly, then these piece of relationships are integrated together and bigger ones are constructed.

The algorithm given above is a structure-based approach, that is, the conceptual structure is constructed according to the chemical structures of the compounds. On the other hand, the name-based approach takes names of compounds as terms and analyzes them to extract the relationships among compounds. The SS-KWIC is one of this kind of algorithms[15].

Generally, both the structure-based approach and the name-based one work well for constructing the conceptual structures.
The structure-based approach is suitable to construct more detailed and systematic conceptual structures than the name-based approach.

However, there are some compounds which are not very near in structures but with similar names because they share similar chemical or physical properties. The name-based approach is used to deal with them. More-

over, the generic representation of compounds is also difficult to be processed by the structure-based approach because some generic names cannot be mapped to chemical structures directly.

On the other hand, there are some compounds with similar chemical structures but not be named with similar names for some reason(for example, compounds which had been named before their structures were known). In this case, the structure-based approach is good in constructing the conceptual structures. Therefore, combination of structure-based approach and name-based one is a practical and effective method.

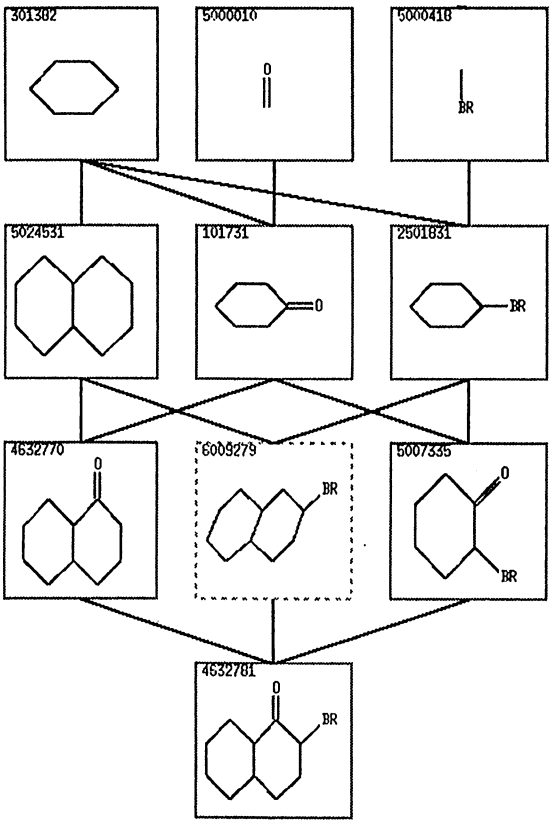


Fig.5: The conceptual structure based on fig.4

3.2 Generation of information

The information space of chemical data should not be closed if it is supposed to be used for problem solving. For example, the target of organic synthesis is to get new compounds, so it is essential that the systems can deal with unknown compounds, that is to say, the systems should be able to generate new compounds and to predict their properties. The properties of generated compounds may be estimated based on the compounds close to them in the conceptual structures. The logical structures should also be augmented to generate the candidates of solutions for new reactions.

The organized information structures can be used to generate lacking information.Generating information means connecting concepts or creating some new concepts and connecting them with other ones.It is a process of extending the constructed information structure to a virtual information space. The virtual space of information consists of concepts which actually exist and those which may exist in the real world or may be generated logically.

The conceptual structure of compounds which is constructed based on the *more-general-than* relationships between compounds is suitable for lacking information generation. The lacking compounds can be generated by augmenting the conceptual structure of compounds. New structures can be assembled based on the atom set and connection rules of molecular structures. Moreover, the position of generated concepts in the conceptual structure can be decided because the relationships between existing concepts and generated ones are computable.

However, some constraints are necessary because the space of chemical structures is infinite. The constraints include starting point (from where the generating process starts), the components be used, the direction of gen-

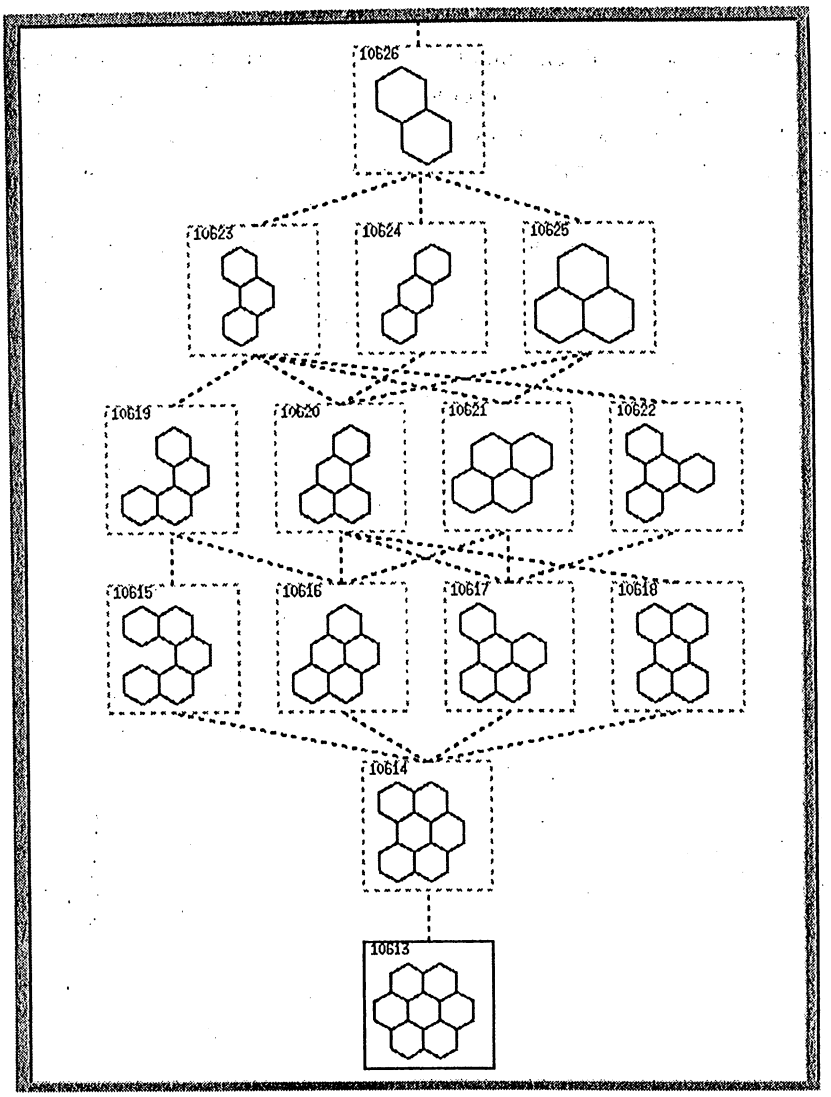


Fig.6: Generation of concepts in a bottom up manner

eration and the distance(levels be generated) or the end point(where the process stops). For example, there are two types of generation according to the direction.

1) Bottom up manner

This means to generate concepts from a narrower one. New concepts are generated by removing components from the base concepts. Therefore, the constraints are not mandatory because the process of generation can stop in the top of conceptual structure. However, the components or distance can be given as options.

Fig. 6 illustrates the generation of rings from a coronene compound. The process stops at the top (single ring, which is not displayed in the figure) automatically.

2) Top down manner

This means to generate the concepts from a broader one. New concepts are generated by adding some components to the base concepts. The components and location should

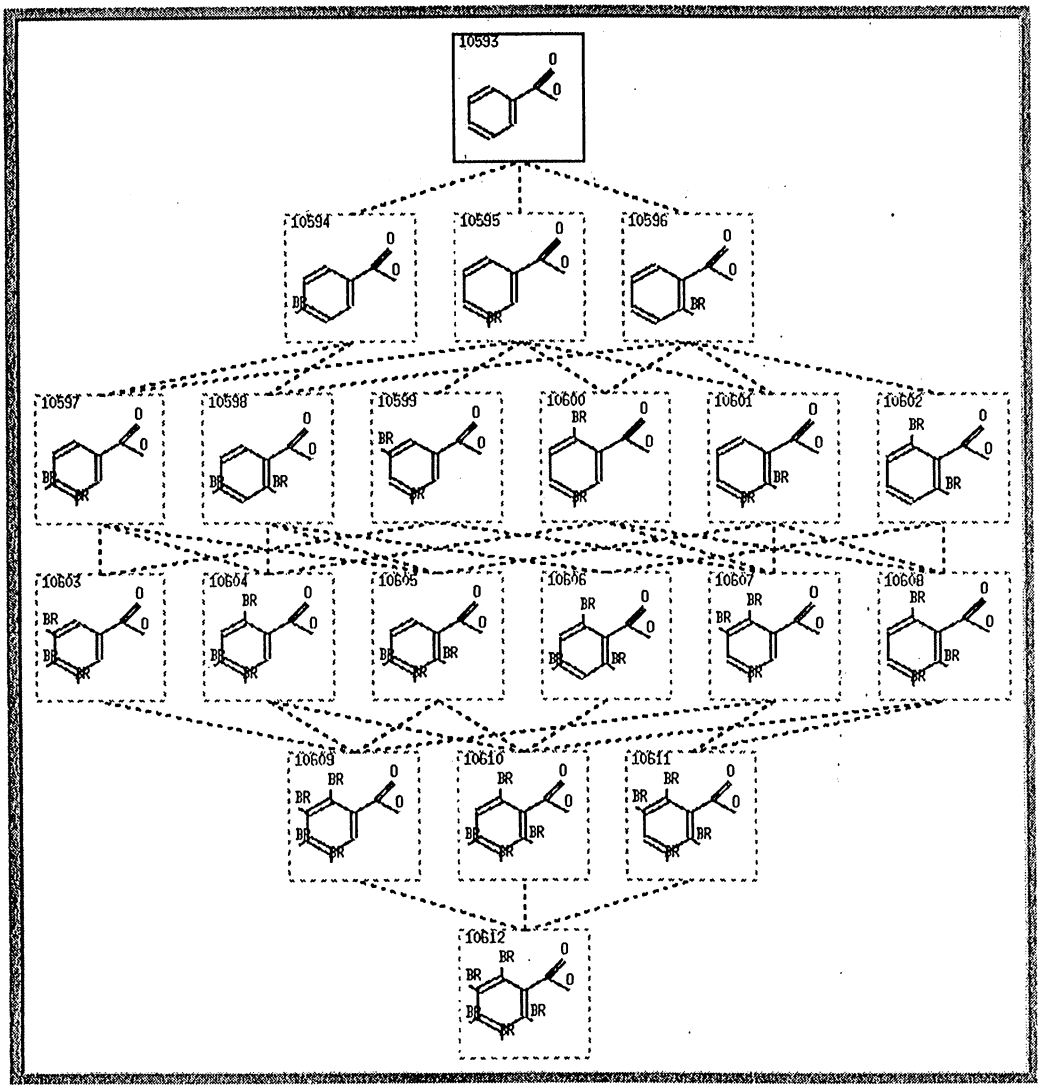


Fig.7: Generation of concepts in a top down manner

be given because the process can not stop automatically.

Fig.7 illustrates the generation of concepts from the top compound by adding bromo *(-Br)* to the ring. The process stops in the bottom because there is no unoccupied position in the ring.

3.3 Other semantic structures of chemical data

Besides the conceptual structure of compounds, the logical structures of compounds which represents the network of reactions can be constructed based on the collection of known reactions. This network can be generated automatically by piling up common compounds for reactions. A graph isomorphism algorithm is necessary to decide if two connection tables represent a same compound. Furthermore, the conceptual structure of reactions which represents the *ISA* relationships among reactions can be constructed also. It is constructed according to generality of change of chemical structures from the starting com-

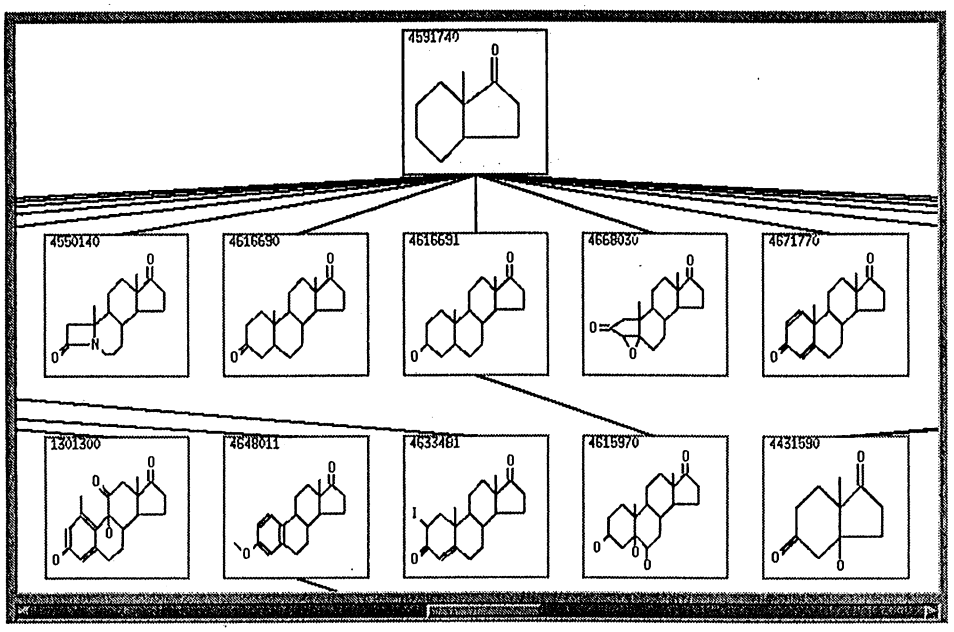


Fig.8: An example of substructure search

pounds to the products together with the reaction conditions. The approach of constructing the structure is a bottom-up manner, that is, using more generic changes on chemical structures to organize reactions, then taking advantage of the knowledge on reaction patterns such as oxidation, reduction, replacement, and so on[14].

4. Substructure search of compound

Conventional substructure search systems take the screening approach which uses inverted file of some fragmental structural features to get the candidate structures of the query. However, they still have to use the time consuming operation of subgraph isomorphism for the final decision because that the relationships or connections among the fragments are not be considered in the stage of screening. Taking advantage of the conceptual structure of compounds, substructure search can be implemented efficiently in two steps: find the entrance node standing for the query substructure by graph isomorphism (not subgraph isomorphism) operation, and then retrieve all candidate compounds by navigating links directly. Fig.8 shows an example of substructure search. The structures which include the top structure are given systematically.

5. Naming of structures

The conventional nomenclatures are rule-based and are designed to give unique names for all of the chemical structures[17]. The system may be too complex to be maintained and may lose the flexibility of naming a structure from various viewpoints.

Chemical structures can be named by analogical reasoning based on the names of similar structures. Fig.9 shows an example of naming structures taking advantage of the conceptual structure of compounds. Names of compounds around named compounds can be given in an analogical way. The maintenance of this kind of case-based system may

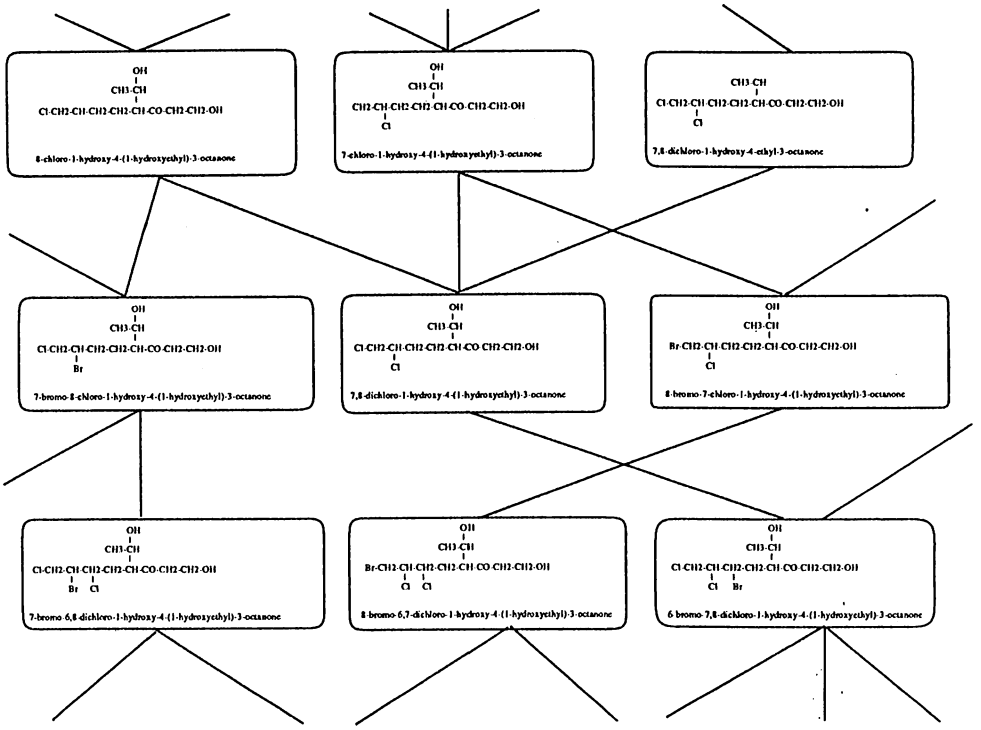


Fig. 9: Naming the similar compounds by analogical reasoning

be easier than the rule-based one. Another example is given in Fig.10, it shows the case of naming a structure from different viewpoints. Structures can be given different names based on their different substructures in the semantic structures, and vice versa.

6. Conclusion

Semantic structures are essential components for problem solving systems because the capability of the systems mainly depends on semantic processing of comprehensive information. The semantic structures serve as knowledge bases of the system.

The homogenized bipartite model for flexible representation of meaning is applied to chemical information. Relationships among concepts such as overlap, nesting, recursion and relativity, which are difficult to be dealt with by conventional set or graph based models can be represented by the model. The algorithm for self organizing the semantic structures is described. The approaches are practical in the processing of chemical data. The chemical information is organized as a conceptual structure of compounds which represents the *ISA* relationship between compounds, a logical structure of compounds which represents the reactive relationships between compounds, and a conceptual structure of reactions which represents the *ISA* relationships between reaction rules.

Taking advantage of the semantic structures, many functions of problem solving systems, especially the functions for open world relevant manipulation of information can be developed. Generation of lacking information, substructure search of compounds and naming of compounds by analogical reasoning are illustrated as examples.

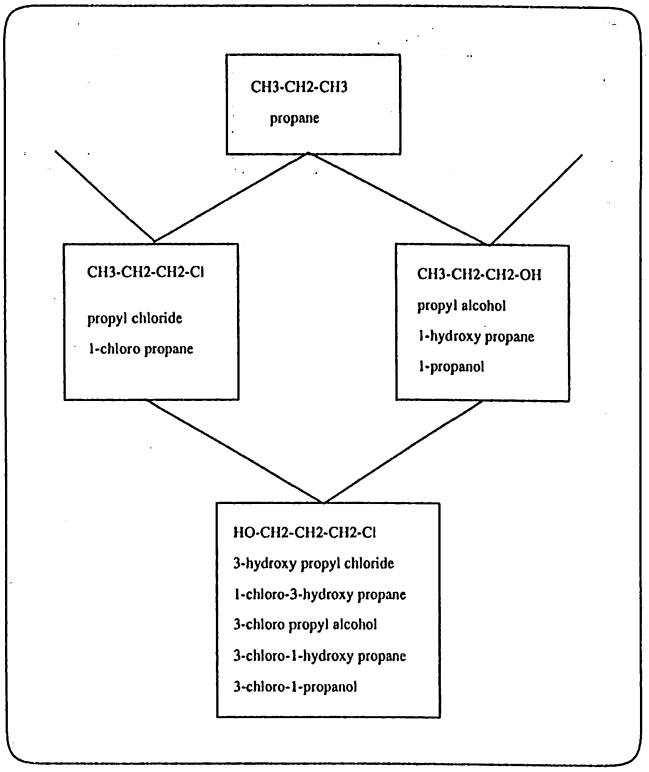


Fig.10: Naming compounds from different viewpoints

References

[1] E. F. Codd: A Relational Model of Data for Large Shared Data Banks, Communications of ACM, Vol.13, No.6, 1970, 377-387.

[2] E. F. Codd: Extending the Database Relational Model to Capture More Meaning, ACM transactions on Database Systems, Vol.4, No.4, 1979, 397-434.

[3] P. P. Chen: The Entity-Relationship Model: Toward a Unified View of Data, ACM transactions on Database Systems, Vol.1, No.1, 1977, 9-36.

[4] D. W. Shipman: The Functional Data Model and the Data Language DAPLEX, ACM transactions on Database Systems, Vol.6, No.1, 1981, 140-173.

[5] M. Hammer and D. McLeod: Database Description with SDM: A Semantic Database Model, ACM transactions on Database Systems, Vol.6, No.3, 1981, 351-386.

[6] J. Banerjee, W. Kim, H. J. Kim and H. F. Korth: Semantics and Implementation of Schema Evolution in Object-Oriented Databases, Proceedings of ACM SIGMOD, 1987, 311-322.

[7] R. M. Quilian: Semantic Memory, Semantic Information Processing, Minsky,M.A. (ed.) The MIT Press, Cambridge, MA, 1968.

[8] J. F. Sowa: Conceptual Structures, Addison-Wesley, 1984.

[9] M. A. Minsky: A Framework for Representing Knowledge, The Psychology of Vision, Winston, P. H. (ed.) McGraw-Hill, New york, 1975.

[10] Y. Fujiwara, Z.Q. Wang et al: The Multicategorical Structures of Information for Inferences and Reasoning in the Self-organizing Information-Base System, Computer Aided Innovation of New Materials II, Elsevier Science Publishers B.V. (M. Doyama ed.), 1993, 27-32.

[11] Y. Fujiwara : The Model for Self-structured Semantic Relationships of Information and Its Advanced Utilization, International Forum on Information and Documentation, Vol.19, No.2, 1994, 8-10.

[12] Y. Fujiwara and H. Gotoda : Representation Model for Relativity of Concepts, International Forum on Information and Documentation, Vol.20, No.1, 1995, 22-30.

[13] Y. Fujiwara, G. Chang and Y. Ishikawa: A Dynamic Thesaurus for Intelligent Access to Research Databases, Proceedings of 43rd FID Conference, 1988, 173-181.

[14] J. An and Y. Fujiwara: Similarity of Compounds and Reactions Based on Self Organized Conceptual Structures of Organic Synthesis Information, Journal of Japan Society of Information and Knowledge, Vol.6, No.1, 1996, 21-35.

[15] J. Lai, H. Chen and Y. Fujiwara: Extraction of Semantic Relationships among Terms – SS-KWIC, Proceedings of 47th FID Conference, 1994, 155-159.

[16] H. Sano and Y. Fujiwara: Syntactic and Semantic Structure Analysis of Article Titles, Journal of Information Sciences Principles of Practice, Vol.19, 1993, 119-124.

[17] International Union of Pure and Applied Chemistry: Nomenclature of Organic Chemistry, Section A, B, C, D, E, F, and H, Pergamon Press, Oxford, 1979.

[18] D. B. Lenat: CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM, Vol.38, No.11, 1995, 33-38.

[19] G. A. Miller: WordNet: A lexical database for English, Communications of the ACM, Vol.38, No.11, 1995, 39-41.

[20] T. Yokoi: The EDR electronic dictionary, Communications of the ACM, Vol.38, No.11, 1995, 42-44.

[21] C. Berge: Hypergraphs, North-Holland, 1989.

[22] J. Gasteiger and C. Jochum: An algorithm for the perception of synthetically important rings, J. Chem. Inf. Comput. Sci., Vol.19, 1979, 43-48.

[23] Edward H. Sussenguth, Jr: A graph-theoretic algorithm for matching chemical structures, J. Chem. Doc. , Vol.5, 1965, 36-43.

(1997年12月15日受付)
(1998年4月7日採録)

著者紹介

Yuzuru Fujiwara(藤原譲) (正会員)
理学博士
神奈川大学理学部情報科学科 教授
E-mail: yfuji@info.kanagawa-u.ac.jp

Jianghong An(安江虹) (正会員)
工学博士
理化学研究所ジーンバンク室
E-mail: ajh@rtc.riken.go.jp