Paper

Construction of Semantic Structures in the Self-Organizing Information-Base Systems

Norihiko Uda[†] Won-gyu Lee[†] Jingjuan Lai[†] Yuzuru Fujiwara[†]

Semantic processing of information is required for sophisticated functions such as learning, inductive inference, analogical reasoning. It is useful to organize source information according to semantic structures. Most data models have strong restrictions for storing complicated information and for dealing with structures of information. This paper describes construction of semantic structures in a way of self-organization based on a new infomation model called the semantic structure model. Constructed semantic structures consist of conceptual ones and logical ones. In the information-base, the former is realized as a thesaurus and the latter as a taxonomy. The procedeure of automatic construction of the thesaurus is also explained in this paper. Application to polymer information shows that the model is appropriate for representation of information for advanced research.

Keywords: semantic structure, self-organizing information-base system, learning, analogical reasoning, automatic thesaurus, information model

1. Introduction

Databases and knowledge-bases which can effectively deal with complicated information of the real world are required. Most data models are not flexible for representing wide variety of repationship in information. Therefore, many attempts are reported to extend data models to overcome restrictions. Although basic form of the relational model are extended by Codd and others[1,2,3], they have difficulty in treating complex objects, especially various relationship. Other models such as the object-oriented data model[4,5], and the entity-relationship model[6] are not appropriate for management of very large scale databeses and for handling relativity of information.

This paper explains information structures based on *the semantic structure model*[7] and construction of semantic structures by self-organization.

The semantic structure model is one which has ability to deal with meaning of information by structuralizing information. It is a model derived from IBS/SORITES (Information-Base Systems with Self Organizing Receptor Interconnections[8,9]) which provides sophisticated functions such as inductive inference, analogical reasoning, and learning including semantic processing. The model is applied to polymer information and the system is called PM-IBS(Polymer Materials-Information-Base Systems).

The architecture of the PM-IBS is illustrated in Figure 1. In the PM-IBS, information structures consist of physical ones and semantic ones which include thesauri for conceptual structures and taxonomies for logical ones.

2. The Information Structures

There are many kinds of information in the real world, information in the PM-IBS is

[†] Institute of Information Science and Elecrtonics, University of Tsukuba

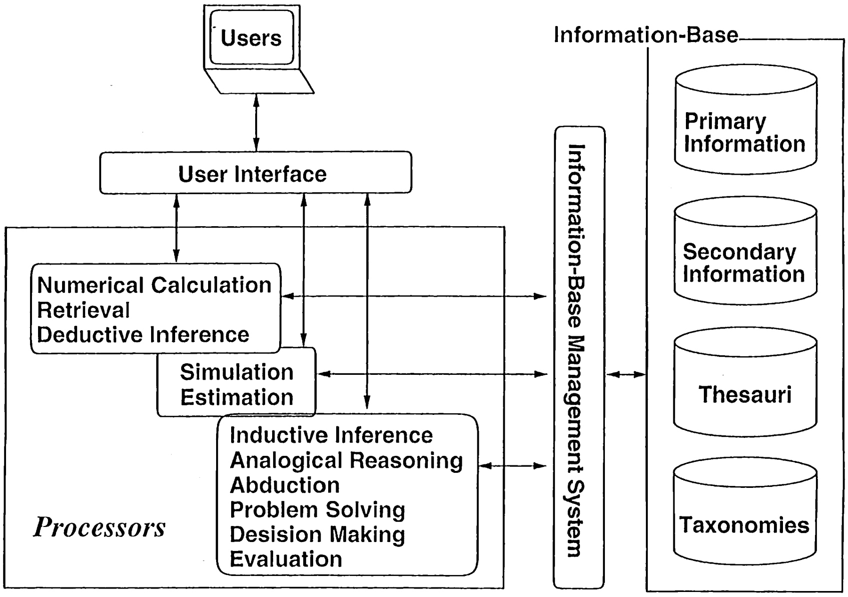

図 1 The Architecture of PM-IBS

organized by three kinds of structures;

- *Physical structures* represent physical relationships such as location of source information.
- *Conceptual structures* represent the relationships such as equivalent ones and hierarchical ones of concepts.
- *Logical structures* represent logical relationship such as 'cause and result'.

Reasoning and inferences are processed on conceptual structures and logical ones. In analogical reasoning, similarity is may be measured as distance in conceptual structures, and reasoning is processed in logical structures as mapping between corresponding substructures.

To represent conceptual structures and logical ones, it is necessary to describe not only basic relasionships among concepts such as equivalent relationship and hierarchical one but also relative, and overlapping ones. Hence the information model which can describe complex information is required.

3. Semantic Structure Model

In order to represent information in the real world, it is necessary to describe hierarchical, relative, and overlapping concepts. The semantic structure model[7] was devised to deal with meanings of information. This model is based on extended hypergraph. A basic hypergraph is defined as follow[10] ;

Let $X = \{x_1, x_2, \cdots, x_n\}$ be a finite set. A hypergraph $H = (E_i, E_2, \cdots, E_m)$ on X is a family of subsets of X such that

$$E_i \neq \phi \ (i = 1, 2, \cdots, m) \qquad (1)$$

$$\bigcup_{i=1}^{m} E_i = X. \qquad (2)$$

The elements $x_1, x_2, \cdots, x_n$ of X are called nodes, and the sets $E_i, E_2, \cdots, E_m$ are called links of the hypergraph.

Since a basic hypergraph can not sufficiently represent relativity, overlap nor hierarchical relationship in information, hypergraphs are extended in the following sense which basic hypergraphs do not have; 1) basic hypergraphs may

```
< Expression > ::= < Set > | < Simple Expression > |
                   < Simple Expression > < Connection Symbol > < Expression >
< Simple Expression > ::= '('< Connection >')'
< Connection > ::= < Set > |
                   < Set > < Connection Symbol > < Simple Expression >
< Set > ::= < String > | < Expression > |
            < String > < Set Symbol > < Set > |
            < Expression > < Set Symbol > < Set >
< Connection Symbol > ::= ' − ' | ' = '
< Set Symbol > ::= ' , ' | ' → ' | ' ← ' | ' ∧ ' | ' ∨ '
```

図 2 The Expression of the Semantic Structure Model

not have a label, whereas each hyperlink in extended hypergraphs always has a label which may have nested structures to describe meanings of information, 2) the extended hypergraph may have a direction, and 3) nodes in basic hypergraphs are primitive elements of a finite set, whereas both of nodes and links in the extended hypergraphs may be extended hypergraphs.

Representation of the extended hypergraph as information model has two types of links i.e. *internal links* and *external* ones. *Internal links* are used for attributes of an object. *External links* are used for explicit relationships between two or more objects.

Figure 2 shows the expression of the extended hypergraph defined in BNF (Backus-Naur Form).

Expression enclosed by '(' and ')' stands for an *internal* or *external link*. '–' means connection of hyperlinks, '=' equivalent of meaning, ' , ' nonordered set, '→', '←' ordered set, '∧' conjunction, and '∨' disjunction respectively. Both types of links have the same expression but are not so in processing them.

Thesauri and taxonomies which describe structures of information are constructed by these expression of semantic relationship.

4. Description of Semantic Structures

An example of conceptual structures is shown in Figure 3. It is transformed from an expression of the semantic structure described in Figure 4.

In Fig. 3, Condensation polymer and polycondensate are synonymous. Condensation copolymer and copolycondensate are synonymous. Condensation polymer is a polymer. Condensation copolymer is a copolymer. The relationship of polymer and condensation polymer are polycondensation. The relationship of copolymer and condensation copolymer are copolycondensation. The relationship of polymer and copolymer are copolymerization. 'Isa' is the internal link. 'Synonym', 'polycondensation', 'copolycondensation', and 'copolymerization are the external links. Each relationship may have direction.

There are different ways to represent conceptual structures in Fig. 3 according to viewpoints. The expressions of Fig. 4 show representation of conceptual structures from several viewpoints.

Equivalent relationship and hierarchical one in conceptual structures are derived from the expressions.

5. Thesaurus Construction

There are many ways to transform meanings of concepts from forms written in natural languages to computer-oriented ones. Thesauri for databases are a construct of keywords which contains a semantic relationship of terms including synonyms, polysemes, and hierarchical relationship etc. so that it can be used for organization of information. Although thesauri are useful in many fields e.g. information retrieval and classifi-

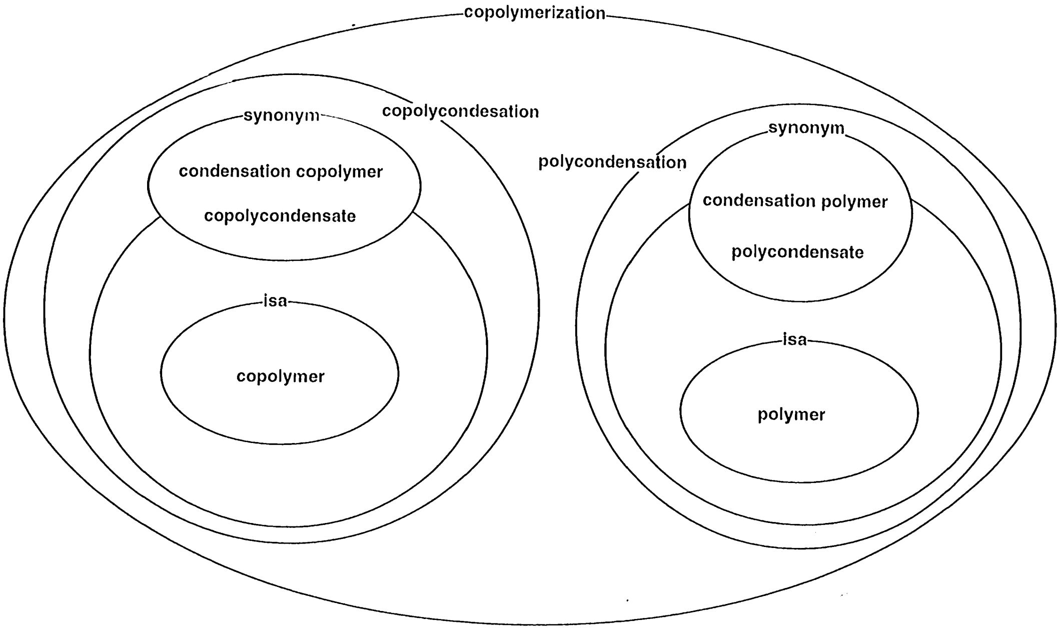

図 3 An Example of Conceptual Structures

((*condensation polymer* = *polycondensate*) − (*isa*) − (*polymer*))

((*condensation polymer* = *polycondensate*) − (*polycondensation*) − (*polymer*))

((*condensation copolymer* = *copolycondensate*) − (*isa*) − (*copolymer*))

((*condensation copolymer* = *copolycondensate*) − (*copolycondensation*) − (*copolymer*))

((*polymer*) − (*copolymerization*) − (*copolymer*))

図 4 An Example of the Expression in Semantic Structure Model

cation, they have difficulties in construction and maintenance, due to necessity of comprehensive and wide expert knowledge, and frequent change of terms.

One of characteristic features of a thesaurus is that the structural relationship of terms are described explicitly. Structures in a thesaurus correspond to conceptual relationships among terms such as equivalent ones and hierarchical ones.

Equivalent relationship is defined as a relationship between two or more terms representing same concept. Thesauri for controlled terms describe these relationships using the following symbols;

UF (use for): the prefix for non-preferred terms,

USE : the prefix for preferred terms,

while thesauri in the PM-IBS has no distinction between preferred terms and non-preferred ones.

A broader term stands for a genus or a whole set, and a narrower term stands for one of individual species or parts. Hierarchical relationship is used to rank broader concepts and narrower concepts in order.

Hierarchical relationship is described by following symbols in thesauri,

BT : the prefix symbol for broader terms,

NT : the prefix symbol for narrower terms.

In hierarchical relationship, there are three kinds of relationships: generic-specific relationship, whole-part one, and type-instance one.

Broader terms happen to appear in equivalent relationship because they cover narrower concepts and are apt to deteriorate synonym set obtained by simple transitive closure due to including excessive polysemes.

To generate less noisy synonym sets, broader terms should be extracted from equivalent sets of terms, otherwise, synonym sets may contain noise connected by polysemes. Synonym sets are connected to the broader terms after obtaining transitive closures[11].

A set of broader terms is identified and extracted from equivalent sets of terms or from definition of terms in a dictionaries by several ways as explained bellow.

1) The Coinage Rule

Technical terms are often generated by combining words to give new scientific meanings. *Scientific & Technical Terms*, published by the Japanese Ministry of Education and Culture, shows that composite terms of more than one words amount to 70 %. These terms are structuralized by equivalent relationships. Common terms for distinct concepts are broader terms[12].

2) The Definition of Terms

Definition of a term, like in a dictionary, consists of many terms related to it. That is, a new term is defined by means of other well-defined terms. Definitions are converted to expressions of the information model. Since relationships among terms can be extracted automatically, it becomes easier to construct and maintain a thesaurus.

3) The Available Thesauri

The automatic thesauri can be improved by combining it with available thesauri in which relationships among basic terms are well-organized.

The utilization of multi lingual dictionaries to extract equivalent relationship was shown by an example of CD-WORD which is twelve language dictionaries[11,12]. The algorithm to extract synonym sets automatically[13] is shown as follows.

Let $\mathcal{M}$, $\mathcal{N}$ be languages translated to each other, and $\mathcal{T}$ be a set of translational relationships. Let synonym sets in $\mathcal{M}$, $\mathcal{N}$ be S_m, S_n respectively. Let us use S'_m and S'_n as the working sets corresponding to S_m and S_n. If $m \in \mathcal{M}$, $n \in \mathcal{N}$, initial state of the synonym sets are written

$$S^0_m = \{m_0\} \text{ and } S^0_n = \{\},$$

where $m_0 \in \mathcal{M}$ is a starting term. A set of translational relationships is expressed as

$$S'_n = \{n_k | (m_0, n_k) \in \mathcal{T}\}$$

the synonym set is merged with the previous synonym set obtained by translational relationship.

$$S'_n = \bigcup_{m_p \in S^i_m} \{n_q | (m_p, n_q) \in \mathcal{T}\}$$

$$S^{i+1}_n = S^i_n \cup S'_n$$

$$S'_m = \bigcup_{n_s \in S^{i+1}_n} \{m_r | (m_r, n_s) \in \mathcal{T}\}$$

$$S^{i+1}_m = S^i_m \cup S'_m.$$

$S^+_n \subseteq \mathcal{M}, S^+_m \subseteq \mathcal{N}$ are transitive closures of synonym sets in $\mathcal{M}$, $\mathcal{N}$ respectively, if

$$S^i_n = S^{i+1}_n = S^{i+2}_n = \ldots = S^+_n$$

$$S^i_m = S^{i+1}_m = S^{i+2}_m = \ldots = S^+_m.$$

6. Discussions

It is not easy to extract desired information from complicated information. Easily available fragmental information must be organized to built the whole structures.

One of ways to organize information is to construct a thesaurus automatically. Most thesauri describe equivalent relationship and hierarchical one. It is necessary to describe not only basic relationships among concepts but also recursive, relative, and overlapping concepts. Thesauri must be constructed automatically because it is difficult to organize large amount of information.

The semantic structure model has abilty to describe meanings of complex information flexibly and in detail as far as desired, so that thesauri and taxonomies may be constructed by self-organization in the PM-IBS.

7. Conclusion

Information structures based on the semantic structure model and construction of the structures by the self-organization are described.

The semantic structure model for the PM-IBS is used for flexible representation of concepts and relationships among concepts by labeled, directed, recursive links in the structures. Hyperlinks are a set of objects with abstract data type or internal structures as attributes, and it carries meanings as specified relationships among hypernodes which may be hyperlinks at the same time in the model used.

Thesauri with conceptual structures are constructed using synonymous relationship as equivalent one and extraction of broader terms automatically from coinage rules, definitions, and available thesauri. Taxonomies also constructed automatically in the same way for representing logical relationship.

Automatic construction of thesauri and taxonomies for semantic processing improves their precision and maintainability.

Acknowledgement

A part of this study is being supported by the Grant for Advancement of Science by Ministry of Education and Culture (No. 02402054), and by Special Coodination Funds of the Science and Technology Agency of the Japanese Government as a part of the project "Self Organizing Information-Base Systems Assisting Creative Research and Development".

文 献

1) E. F. Codd: "Extending the Database Relational Model to Capture More Meaning ", *ACM Transactions on Database Systems*, 4(4):397–434, 1979.

2) M. Stonebraker, B. Rubenstein, and A. Guttman: "Application of Abstract Data Types and Abstract Indices to CAD Data", In *Proc. of Ann. Meeting Database Week*, pages 107–115, San Jose, 1983.

3) J. M. Smith and D. C. P. Smith: "Database Abstractions: Aggregation and Generalization", In *ACM TODS 2(2)*, pages 105–133, 1977.

4) W. Kim, J. Banerjee, H. T. Chou, J. F. Garza, and D. Woelk: "Composite Object Support in an Object-Oriented Database System", In *OOPSLA'87 Proceedings*, pages 118–125, October 1987.

5) J. Banerjee, W. K., H. J. Kim, and Henry F. Korth: "Semantics and Implementation of Schema Evolution in Object-Oriented Databases", *ACM SIGMOD*, pages 311–322, 1987.

6) P. P. S. Chen: "The Entity-Relationship Model: Toward a Unified View of Data", *ACM Transactions on Database System*, 1(1):9–36, 1986.

7) Y. Fujiwara, N. Uda, and X. Zhang: "Analogical Reasoning in Polymer Information-Base Systems", the 13th CODATA, October 1992.

8) Y. Fujiwara, J. He, G. Chang, N. Ohbo, H. Kitagawa, and K. Yamaguchi: "Self Organizing Information Systems for Material Design", In *Proceedings of CAMSE'90*, Tokyo, August 1990.

9) Y. Fujiwara: "Self Organizing Information-Base Systems with Learning and Analogical Reasoning", In *Proc. of the 3rd Beijing Int. Symp. on Computer Information Management*, October 14-18 1991.

10) C. Berge: *"Hypergraphs"*. North-Holland, 1989.

11) Y. Fujiwara, N. Ohbo, T. Itoh, M. Morita, K. Sawai, T. Kawasaki, and S. Fujiwara: "Multilingual Thesauri for Internationally Distributed Information Systems", In *Information, Communication, and Technology Transfer*, pages 47–54, 1987.

12) Y. Fujiwara, W. G. Lee, T. Itoh, N. Ohbo, and S. Fujiwara: "Analysis of Scientific and Technical Terms for Multilingual DB Access

System", In *ICIK*, pages 3–4, November 1987.

13) J. Lai, H. Kitagawa, and Y. Fujiwara: "Structuralization of Information by the Automatically Constructed Thesaurus", *IPS Japan Information Media*, 7(4):25–32, July 1992.

14) H. Boley: "Directed Recursive Labelnode Hypergraphs and their use as Representations for Knowledge", In *free session contribution to the Fourth Int. Joint Conf. Artificial Intelligence*, 1975. partially reprinted in: G. Veenker Ed., Zweites Treffen der GI-Fachgruppe Kuenstliche Intelligenz, Univ. Dortmund, Abt. Informatik, Bericht No. 13 (1975).

15) H. Boley: "A Theory of Representation (-Language, -Constructions, and -Relations)", Technical Report IFI-HH-M-38, Inst. fuer Informatik, 1976.

16) H. Boley: "Directed Recursive Labelnode Hypergraphs: A New Representation-Language", *Artificial Intelligence*, 9(1):49–85, 1977.

（1993年5月15日受付）
（1993年6月30日採録）

著 者 紹 介

宇陀則彦

1965年生．1989年図書館情報大学図書館情報学部卒業．1991年同大学院修士課程修了．現在，筑波大学大学院工学研究科博士課程に在学中．情報モデル，意味処理などに興味を持つ．ACM，IEEE，情報処理学会各会員．