情 報 学 基 礎 35－5
（1994. 11. 9）

概念空間のモデルと専門用語の構造化

賴 静娟　陳 漢雄　藤原 譲

筑波大学 電子・情報工学系

情報の意味処理のために情報組織化の必要性と重要性が徐々に認識されるようになってきた。概念を表現する単位としての用語の組織化は大量かつ多様な情報の組織化に有用である。SS-KWIC 法は用語の造語規則に基づいて用語間の階層関係と関連関係が自動的に抽出される方法である。スーパーグラフ・モデルに基づく SS-KWIC 法の原理および実現方法を示す。SS-KWIC によって多くの階層と関連関係が抽出され、いくつかの実験結果の例も挙げた。またスーパーグラフの物理データ構造およびスーパーグラフに対する問い合わせや更新などのための言語 SML を用いて記述される。

キーワード：専門用語、概念構造、スーパーグラフモデル、構造操作言語

The Model of Conceptual Spaces and The Structuralization of Technical Terms

JingJuan Lai　HanXiong Chen　Yuzuru Fujiwara

Institute of Information Science and Electronics, University of Tsukuba

Importance and necessity of organizing information has been gradually recognized and organization of terms as units for expressing concepts is practical for organizing information. The method of SS-KWIC are useful for extracting hierarchical relationships and associative relationships automatically based on the coining rules of terms. The principle and implementation of the method of SS-KWIC based on the supergraph model are shown, and exam ples of the experimental results are shown. Detailed hierarchies and rich associative relationships were obtained, and the results proved reliability and practicality of the method. A supergraph phisical data structure is presented. A supergraph manipulation language (SML) used for querying and updating labelled supergraphs is introduced.

Key word: technical terms, conceptual structures, supergraph model, supergraph manipulation language

1 はじめ

意味的に自己組織化された情報は類推、反証推論及び帰納推論など情報の高度な利用には欠かせない。概念を表現する単位としての用語の組織化は大量かつ多様な情報の組織化にとって重要な一環である。本研究の目的の一つは用語の表現する概念空間のモデル化およびこのモデルに基づく用語の自己組織化方法の開発である。

一つ新しいデータモデル――スーパーグラフ・モデルが概念の意味空間を記述するために提案された。スーパーグラフはverticesの集合をsupervertices として扱う。スーパーグラフ・モデルに基づく専門用語における意味関係の自動抽出方法が開発された。統計的な方法の中には、同値関係を抽出するためのC-TRAN法、階層関係と関連関係を抽出するためのSS-KWIC法、多様な意味関係を抽出するためのSS-SANS法がある。抽出された意味関係によって自動的に概念構造を構築する方法も考案した。

本論文には、先ず専門用語の意味空間を記述するためのスーパーグラフ・モデルを示す。意味関係を抽出するC-TRAN法とSS-KWIC法のアルゴリズムが記述される。スーパーグラフの物理データ構造とスーパーグラフ操作言語（SML）を示すことによって概念構造の構築の実現方法が説明される。

2 SS-KWIC法と概念構造

SS-KWIC法とは専門用語の造語規則に基づく階層関係および関連関係を自動的に抽出する方法である。この方法には三つのアルゴリズムが含まれる。具体的には基本用語を選択するアルゴリズムと、共通する基本用語をもつ複合語をカプセル化にするアルゴリズムと、品詞性と造語規則を利用して上の用語間における階層関係および関連関係を判定するアルゴリズムである。複合語における造語規則を表1にまとめて示されている。多くの場合用語の接尾によって英語の品詞を区別することができる。例えば、名詞接尾には、 -tion, -sion, -ism, -ness, -ment などがある。動詞接尾には、 -ize, -ise, -ify などがある。形容詞接尾には、 -ible, -ful, -ical, -less などがある。ただし、例外も少なくない。文における用語が多品詞や多義の場合には、品詞判別、前後関係、意味選択を有効に結合して利用しなければ、正確な意味関係の抽出が期待できない。

表1：英語の複合語造語規則

Coining rules	Role
adjective+noun noun+ing noun+past participle adverb+past participle adverb+ing noun+noun adjective+noun adjective+noun-ed noun+adjective adjective+-ing adjective+past participle noun+tight noun+proof noun+free -ing+noun adjective+adjective preposition+noun	attributive (limitation)
compound verb	
other	e.g. output

同値、階層および関連関係によって用語が体系化されたものを概念構造と呼ぶ。概念構造に基づいて、概念間の類似性および類似度を知ることができる。概念構造は情報高度利用に大変重要な役割を果たす。図1にはSS-KWIC法によって得られた階層関係をもつ用語の集合が示されている。図2には、図1の結果から構築された一つの概念構造が示されている。

SS-KWIC
network
computer network
data network
circuit switched data network
packet switched data network
packet switched public data network
semantic network
Quillian's semantic network
local area network
wide area network

図 1: SS-KWIC の一例

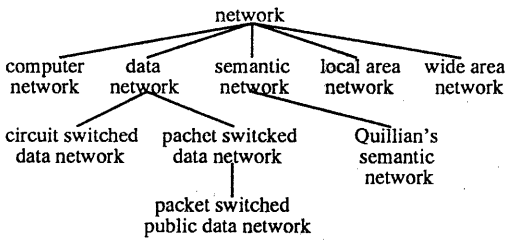

図 2: 概念構造の一例

3 スーパーグラフ・モデルに基づく用語の表現

[定義 (スーパーグラフ)]

$V = \{v_1, v_2, ..., v_n\}$(vertices とよぶ) 上のスーパーグラフ $S = (SV, SE)$ が次のように定義される

$$SV \subseteq 2^V - \{\emptyset\}$$
$$SE \subseteq SV \times SV$$
$$\sigma : V \rightarrow L(V)$$
$$\sigma' : SE \rightarrow L(SE) = \{r_1, r_2, ..., r_m\},$$

ここで、SV の任意一つの要素が supervertex とよばれる。supervertex が vertices の集合である。SE の要素を superedge とよぶ。$r_1, r_2, ..., r_m$ が supervertices のペア間の関係を表す。すなわち、スーパーグラフは集合を supervertices とし、supervertices 間の二項関係を superedges とするラベル付き有向グラフである。$L(V)$ が各 $v_1, v_2, ..., v_n$ に付いているラベル "$L(v_i)$" の集合を表している。

用語 network, semantic network, Quillian's semantic network, QSN を例にスーパーグラフ・モデルを用いて如何に専門用語を表現するかを図 3 で示そう。図中では、実線楕円形で supervertices を表す。ドット線楕円形で vertices を表す。実矢印とドット矢印でそれぞれラベル付き superedges とそのデュアル superedges を表す。例えば、network と semantic network の間に双対の関係 isa と isa^{-1} が存在する。QSN が Quillian's semantic network の省略形なので、両者の間に同値関係がある。

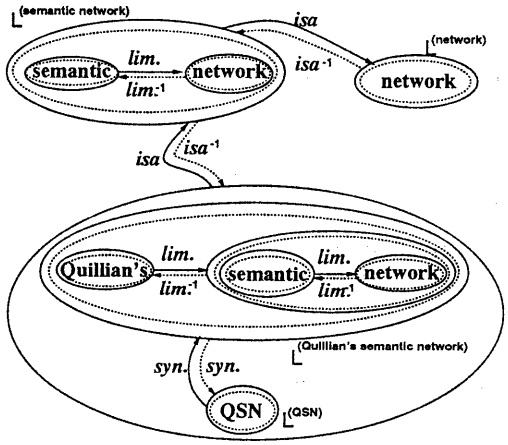

図 3: スーパーグラフによる用語の表現

4 意味関係抽出アルゴリズムおよび実験結果

意味関係を自動的に抽出する方法には C-TRAN や SS-KWIC などがあることを上述した。次に具体的に C-TRAN と SS-KWIC のアルゴリズムを記述する。S が同義語集合である。H が階層関係をもつ用語の集合である。r_e と r_h でそれぞれ同値関係と階層関係を表す。r_l で複合語における限定或いは修飾関係を表す。$BasicTermSet$ が基本用語の集合である。$MatchingString$ がストリング・パターン・マッチング関数である。

T と H' が臨時スペースである。

[Algorithm(C-TRAN)]

```
begin
   S,T = ∅
   input a term a ∈ V
   S = S ∪ {a}
   T = T ∪ {a}
   while T ≠ ∅ do
   begin
      t_i ← T
      T = T − {t_i}
      for all x_j in (V − T − S) do
         if L(< t_i, x_j >) = r_c or L(< x_j, t_i >) = r_c do
         begin
            S = S ∪ {x_j}
            T = T ∪ {x_j}
         end
   end
   output S = {a, x_1, x_2, ..., x_s|
         L(< {a}, {a, x_1, x_2, ..., x_s} >) = r_c}
         as the synonym set of a
end.
```

[Algorithm(SS-KWIC)]

```
begin
   H, H' = ∅
   input a basic term a ∈ BasicTermSet ⊆ V
   H = H ∪ {a}
   for all x_i in V − {a} do
      if MatchingString(a, x_i)=True then
         H' = H' ∪ {x_i}
   proc(a)
end
```

proc is defined as follows.

```
proc(a)
begin
   T = ∅
   T = T ∪ {a}
   while T ≠ ∅ do
   begin
      for all t_i in T do
      T = T − {t_i}
      for all h_j = (x_0 t_i) ∈ H' − H and |x_0| is minimum do
         if L(< x_0, t_i >) = r_l and x_0 ∈ BasicTermSet then
         begin
            H = H ∪ {h_j}
            L(< h_j, t_i >) = r_h
            T = T ∪ {h_j}
            proc(h_j)
         end
   end
   output H as hierarchical relationships of a
   output H ∩ H'(≠ ∅) as associative relationships of a
end.
```

工業用語集を情報源として上記の SS-KWIC アルゴリズムで意味関係抽出実験を行なった。その結果が図4に示される。図5と図6は抽出結果として SS-KWIC の例を二つ示す。

用語数	57936
基本用語数	164
非空 SS-KWIC 集合数	75
最大 SS-KWIC 集合サイズ	71
最長用語の要素数	13 words
非空 SS-KWIC 集合要素の総数	844
一階層の割合	38.98%
二階層の割合	9.12%
三階層の割合	1.42%
四階層の割合	0.12%
関連関係の割合	50.36%

図 4: 実験結果

ADAPTER
MALE ADAPTER
FEMALE ADAPTER
NOZZLE ADAPTER
TAPERED ADAPTER SLEEVE
ADAPTER ASSEMBLY
BEARING WITH ADAPTER ASSEMBLY
ADAPTER PLATE

図 5: SS-KWIC の一例

AFRICA
CENTRAL AFRICA
EAST AFRICA
NORTH AFRICA
NORTH-EAST AFRICA
SOUTH AFRICA (REPUBLIC OF)
SOUTHERN AFRICA
WEST AFRICA

図 6: SS-KWIC の一例

5 概念構造の構築

5.1 スーパーグラフの物理データ構造

図7に示したスーパーグラフの物理データ構造はヘッダ部分、データ部分および superedges 部分によって構成される。ヘッダ部分の "pointer to synonyms" はこの supervertex の vertices に指す。同義語、上位語および下位語がデータ部分に入っる。データ部分と superedges 部分は登録番号、登録名前、接続属性、切り離し属性お

よび行き先へのポインター。図8で用語 "semantic network" を例にしてデータ構造が示される。上位語 "network" と下位語 "Quillian's semantic network","QSN" へのポインターが記憶されるほか、 supervertices"{network}" および "{quillian's semantic network}" へのポインターなども記憶されている。

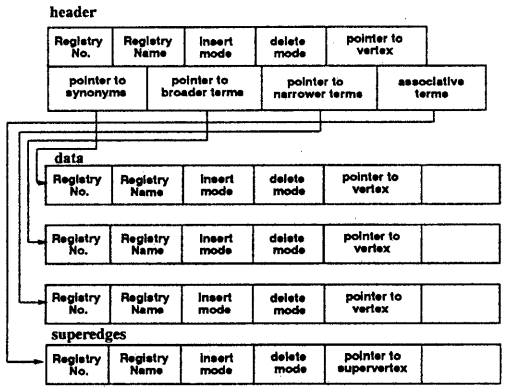

図 7: 物理データ構造

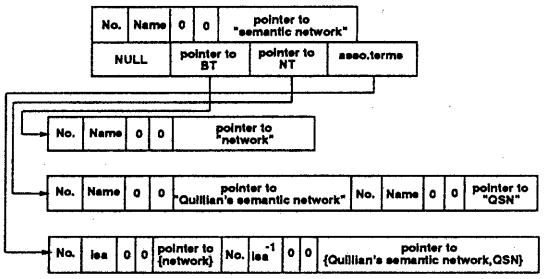

図 8: 物理データ構造の一例

5.2 インデックスと管理ファイル

専門用語のインデックス・ファイルは prefix B^+tree で実現される。葉以外の節点が用語と子ブロックへのポインターで構成する。葉節点は用語と用語 ID （登録番号と登録名前）と属すべき supervertices の物理アドレスによって構成する。 supervertices と superedges を管理するためのファイルは生成や変更や削除などされた supervertices と superedges に対して登録、管理する。

5.3 SML

ラベル付きスーパーグラフに対する問い合わせおよび更新などにスーパーグラフ操作言語 (SML) を用いて記述される。

- _create_supervertex(term)
  用語 term の supervertex を作成する。 supervertex がすでに存在するならば、 ERROR を返す。
- _add_vertex(term1, term2)
  term2 の属する supervertex に term1 を挿入する。 supervertex が存在しないならば、 ERROR を返す。 term1 がすでに supervertex の中に存在するならば、 ERROR を返す。
- _delete_vertex(term)
  term を supervertex から削除する。 term は存在しない或は supervertex は存在しなければ、 ERROR を返す。
- _delete_supervertex(term)
  term の属する supervertex を削除する。 supervertex は存在しないならば、 ERROR を返す。
- _add_superedge(term1, term2, r)
  term1 の supervertex の接続属性値がデフォルト値或は r ならば、 term1 の supervertex と term2 の supervertex との間に r と r^{-1} というリンクを張る。
- _delete_superedge(term1, term2, r)
  term1 の supervertex の切り離し属性値がデフォルト値或は r ならば、 term1 の supervertex と term2 の supervertex との間のリンクを削除する。
- _modify_vertex(term1, term2)
  term1 が term2 に代替される。 term1 が存在しない場合、 ERROR を返す。
- _modify_superedge(term1, term2, r1, r2)
  term1 の supervertex と term2 の supervertex との間のリンク r1 と $r1^{-1}$ を指定された r2 と $r2^{-1}$ で代替する。 r1 が存在しない時、 ERROR を返す。
- _retrieval_vertex(term)
  term を検索し、必要に応じ term が対応する同義語、上位語あるいは下位語を返す。 term がないと、 ERROR を返す。
- _traversal_supervertex(term)
  term の属する supervertex とリンクで連結されているすべての supervertices の pointer を返す。

6 結論

スーパーグラフ・モデルに基づくC-TRAN法とSS-KWIC法の原理およびアルゴリズムが示した。スーパーグラフへの問い合わせや更新などのためにSMLが導入された。実験の結果によって上述の統計的な手法で有用な意味関係が自動的に抽出できることが実証された。今後の課題としては、スーパーグラフ・モデルの拡張およびスーパーグラフに対する操作などの完備にある。より複雑な意味関係の抽出は意味解析が必要なので、その研究も力入れる予定である。概念構造はデータベースシステム検索およびエキスパートシステムにおける演繹推論にだけでなく、各科学技術分野の先進的な情報ベースにおける類推、帰納推論および反証推論など高度な機能にも利用できる。

References

(1) E.F.Codd. "A Relational Model of Data for Large Shared Data Banks", communications of the ACM, Vol.13, No.6, pp.377-387, 1970.

(2) J.Banerjee, W.Kim, H.J.Kim, and Henry F. Korth. "Semantics and Implementation of Schema Evolution in Object-Oriented Database", proceedings of ACM SIGMOD, pp.311-322, 1987.

(3) C.Berge. "Hypergraphs", North-Holland, 1989.

(4) H.Boley. "directed Recursive Labelnode Hypergraphs: A New Representation-Language", Artificial Intelligence, Vol. 9, No.1, pp.49-85, 1977.

(5) Y.Fujiwara. "The Model for Self Structured Semantic Relationships of Information and Its Advanced Utilization", Proc. of 47th FID Conference and Congress, Japan, (A72-01-7), 1994.

(6) Y.Fujiwara, J.Lai, and T.Makino. "Management and Advanced Utilization of Semantically Organized Terminology and Knowledge", Proceedings of TKE'93, 1993.

(7) Y.Fujiwara, Z.Wang, S.Zheng and Y.Luan. "The Multicategorical Structures of Information for Inferences and Reasoning in the Self-organizing Information-Base System", Proc. of CAMSE, 2, pp.27-32, 1992.

(8) Donald E.Knuth, James H.Morris.Jr. and Vaughan R. Pratt. "Fast Pattern Matching in Strings", SIAM J.Comput., Vol.6, No.2, pp.323-350, June 1977.

(9) J.J.Lai, H.Kitagawa and Y.Fujiwara. "Structuralization of Information by the Automatically Constructed Thesaurus", Information Media, 7(4), pp.25-32, 1992.

(10) J.Lai, H.Chen and Y.Fujiwara. "Extraction of Semantic Relationships Among Terms-SS-KWIC", Proc. of The 47th FID Conference and Congress, Japan, pp.155-159, 1994.

(11) H.Sano and Y.Fujiwara. "Syntactic and Semantic Structure Analysis of Article Titles in Analytical Chemistry", Journal of Information Science, 19, pp.119-124, 1993.