

動的シソーラスによる情報自己組織化の研究

6E-7

劉 野 陳 漢雄 藤原 譲
筑波大学 電子・情報工学系

1 はじめに

近年、情報高度利用のため意味関係の情報自己組織化の重要性が認識されている。情報の自己組織化とは情報資源に内在する意味関係を構造化情報として抽出し、それを用いて情報資源全体を自動的に組織化することを指している。概念構造に基づく情報の自己組織化は一つの有力な方法である。従来、概念間の関係はシソーラスによって表現するが、情報自己組織化のため、シソーラス構造の柔軟性を十分に考慮しなければならない。このため、本研究では情報自己組織化のための動的シソーラス構築方法を研究している。動的シソーラスの構築は動的複合概念構造に基づいて行なう。動的複合概念構造の生成は概念分解と組合せ、用例チェックなどの手法によって行なわれる。

2 複合概念構造

まず概念構造及び分解と組合せなどの関連概念を説明する。

原概念：原情報から抽出された概念は原概念という。

分割と結合： k_r は一文字を表す。

概念 $c = k_1 k_2 \dots k_i k_{i+1} \dots k_j \dots k_n$ を概念 $c_2 = k_i k_{i+1} \dots k_j$ によって行なう分割を $c \ominus c_2$ と書き、分割の結果が概念の集合

$\{k_1 k_2 \dots k_{i-1}, k_i \dots k_j, k_{j+1} \dots k_n\}$ になる。

概念 $c_1 = k_1 k_2 \dots k_{i-1}$ と概念 $c_3 = k_{j+1} \dots k_n$ の結合は

$c_1 \oplus c_3 = k_1 k_2 \dots k_{i-1} k_{j+1} \dots k_n$ と表す。

アトミック概念：他の概念より分割できない概念をアトミック概念という。

語幹定義のため次のペンイ便宜記法を使う。

$\{c_1, \dots, c_n\} \ominus u = U_i(c_i \ominus u)$

$c \ominus U = (c \ominus u) \ominus (U - \{u\})$

語幹：概念 $c = k_1 \dots k_n$ がアトミック概念による分割が $c \ominus U = \{u_1, u_2, \dots, u_t\}$ 且つ $c = u_1 \oplus u_2 \oplus \dots \oplus u_t$ ならば、 u_t は c の語幹という。また、語幹の集合は $\mathcal{R}$ と書く。

複合概念：複合概念の集合を C とし、新たな複合概念 c は次のように生成される。

$c = u \oplus r, u \in C \cup U, r \in \mathcal{R} \cup C$

また、この時 c が r の下位概念(対称的に、 r が c の上位概念)ともいう。

複合概念構造：概念 c が U による分割が $c \ominus U = \{u_1, u_2, \dots, u_m, r\}$, $u_i \in U, r \in \mathcal{R}$ としたら、 c から分解された複合概念の上下位関係で形成する DAG (Directed Acycle Graph) をこの概念の構造という。

次に概念分解と組合せにより自動的に複合概念を形成するプロシーダを説明する。

ステップ1 概念の分割：原情報から一つのアトミック概念 c を選ぶ。 c がアトミック概念集合 U により分割される。分割できない場合に、この概念はアトミック概念となる。アトミック概念の集合 U は次のように求められる。

原概念 c が与えられた時、

$U \leftarrow \{c' | c' \in (c \ominus u) \cup (u \ominus c), u \in U\}$

分割できる場合には、ステップ2を行なう。

ステップ2 アトミック概念組合せチェック：既存複合概念により、二つずつのアトミック概念の組合せ関係をチェックする。チェックの結果は組合せ行列 A に保存する。既存複合概念に二つのアトミック概念の組合せが存在すれば、 $a_{ij} = 1$ を書く、存在しない場合には $a_{ij} = 0$ を書く。

また、実体の対応しないアトミック概念の大量発生を防ぐため、最長一致、英語単語対応や辞書検索などのチェック機構を用いる。

ステップ3 複合概念の生成：定義により、最後のアトミック概念を語幹として抽出する。

アトミック概念の組合せチェック結果とあわせ、各複合概念を生成する。

ステップ4 複合概念間の上下位関係：生成された各複合概念の間の上下位関係により DAG を形成する。

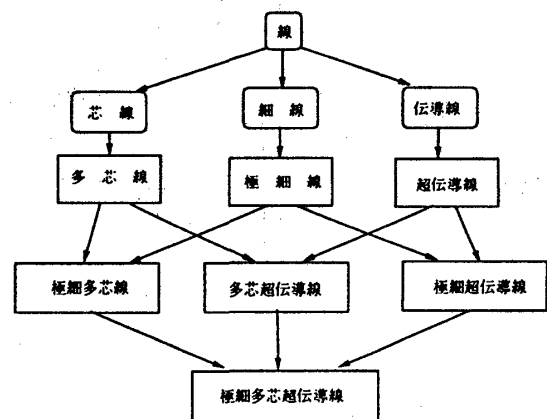


図1 複合概念構造の DAG

Self Organization of Information based on Dynamic Thesaurus

Ye Liu, Hanxiong Chen, Yuzuru Fujiwara
Institute of Information Science and Electronics,
University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305, Japan

図1に原概念「極細多芯超伝導線」から構造化した結果DAGを示す。ここではアトミック概念集合 $U = \{\text{極, 細, 多, 芯, 超, 伝導, 線}\}$ 、語幹 $c_i = \{\text{線}\}$ 、複合概念集合 $C = \{\text{細線, 超伝導線, } \dots\}$ などの例も示された。

3 動的シソーラス

複合概念構造は概念間の階層関係を含んでいるため、各複合概念構造を統合して全体の情報概念空間が構築できる。統合する結果はシソーラスで表現する。図2にシソーラス、複合概念構造及び原情報の関係を示す。

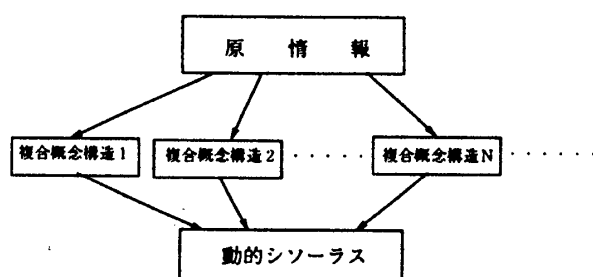


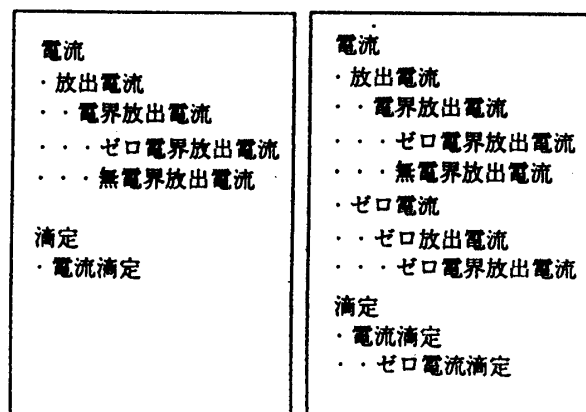
図2 システムの構造

複合概念構造は原情報に基づいて作成されたので、原情報の変化に伴って変化する。さらに、複合概念の動的な変化を随時反映して、動的シソーラスの構造も更新される。

次にこのようなシステムの動的特徴を説明する。

(1) アトミック概念の更新

もし原概念が既存アトミック概念より分割できなければ、原概念はアトミック概念となる。そして、既存のアトミック概念がこの新アトミック概念により分割できる場合に、既存アトミック概念は複合概念となり、アトミック概念ではなくなる。



(a)

(b)

図3 動的的概念構造

(2) 概念構造の更新

原概念が既存アトミックにより分割可能の場合にはまず、この原概念を構造化して、既存概念体系に追加する。次にこの新構造と既存複合概念構造統合して概念体系を更新する。例えば、「電流」、「放出電流」、「電界」、「放出」、「電界放出電流」、「無電界放出電流」、「無」、「ゼロ」、「ゼロ電界放出電流」、「測定」、「電流測定」、の原概念を構造化した結果は図3のaに示す。その後、「ゼロ電流測定」という原概念を処理すると、概念構造の変化は図3のbに示す。「ゼロ電流測定」は「測定」のDAGに追加し、「電流」のDAGに関する構造も変わる。

4 実現の結果

本研究では超伝導分野を例として動的シソーラス構築システムを実現している。処理された用語は78737項目（工業用語集63879、ルートシソーラス13408、超伝導用語集260、論文のキーワード385、定義文、要旨から抽出されたキーワード805）がある。この中に辞書から選択された既存アトミック概念は8852個、概念解析過程中に抽出された新アトミック概念は10829個あり、そして43097個の複合概念構造が生成されたが、同語幹により統合された結果が2165個になった。

処理の適正度

原情報が不十分な場合、複合概念の分割とアトミック概念組合せのチェックができないこともある。間違い分割や要素組合せチェックノイズが存在する。それにもかかわらず、本システムの複合概念分割の適正度と組合せチェックの適正度がそれぞれ93%と96%に達している。

5 おわりに

動的シソーラス構築システムは原情報の増加、発展につれてシステムの概念構造が進化することができるので、情報の自己組織化が実現する。

これからの課題として、動的シソーラスの論理構造や類推などへの応用があげられる。

参考文献

- [1] Z.Q.Wang, Y.Fujiwara et al: Learning and Reasoning in the Information-Base System for Organic Synthesis Research, Journal of Japan Society of Information and Knowledge, Vol.2, No.1, 1991, 71-82.
- [2] J.J.Lai, H.X.Chen, Y.Fujiwara. Extraction of Semantic Relationships Among Terms—SS-KWIC. Proc. of 47th FID Conference and Congress, pp.155-159(1994).
- [3] 高橋 延匡：日本語情報処理。近代科学社。