情報メディア 7－4
(1992. 7. 10)

自動構築シソーラスによる情報の構造化

賴 静娟* 北川 博之** 藤原 譲**

*筑波大学工学研究科 **筑波大学電子情報工学系

要 旨

全文データベース、知識ベース、マルチメディアデータベースなどにおける大量情報の蓄積、管理および検索の問題は大容量の記憶媒体の普及に伴い、ますます重要になってきた。本論文では、大量情報の構造化に関する方法を提案する。具体的には、概念間の意味的関係を自動的に構築するシソーラスを用いて情報の概念構造を組織化される方式と実例について報告する。この方法の信頼性と実用性を示す。

Structuralization of Information by the Automatically Constructed Thesaurus

Jingjuan Lai* Hiroyuki Kitagawa** Yuzuru Fujiwara**

*Program in Engineering Sciences University of Tsukuba
*Institute of Information Sciences and Electronics University of Tsukuba

Abstract

It becomes more important that massive amount of information is stored, managed and retrieved in fulltext databases, knowledge-bases and multi-media databases, as huge capacity of storage media is widely used. This paper describes a method of the structuralization of information. In particular, this shows the method that conceptural structures of information are organized by the automatically constructed thesaurus. Reliability and practicality of the method are shown.

1　はじめに

ハードウェア技術の飛躍的進歩につれ、データベースの取り扱うデータタイプは従来の文字、数値主体から、文章、図形、イメージなど複雑あるいは定め難いデータタイプに変化してきた。その上、CD-ROM, 光ディスク, IC カードなどの大容量の記憶媒体の普及により大量データを計算機によって集積、管理することが可能となっている。今日は、データベース技術の進歩においてこれらの問題が如何に上手に対処できるかがボトルネックとなる。われわれは多様データタイプと大量のデータを扱える、かつ推論、学習と言った高度な機能をもつような情報の資源を情報ベースとよぶ。情報ベースの概略図は図 1 に示している。本研究では、情報ベースにおける情報空間の構造化に重点をおいた手法と成果を取り上げる。タイトルに書いたように、主にシソーラス自動構築法によって情報空間の構造化を実現した。実験の結果により、本研究当初の目的を達成することを示す。しかも情報を自動的に構造化するだけではなく情報構造の保守、更新も自動的に行える。

2　情報の構造化

便宜上、以後は情報空間の構造化を情報の構造化と略称する。対象物の認識により派生する情報の集合体にはそれぞれが対応な実体間の状況を反映し、複雑な関係が存在する。この本来存在する複雑な関連性により、情報群を体系づけることが情報の構造化である。情報ベースにおいて、重要な構造として次の三つを考える：

物理構造
概念構造
論理構造

物理構造とは文章本来もつ内部的構造を指す。科学技術の文献の場合には、タイトル、著者名、抄録、各章、参考文献により構成され、章が節からできている。近年、脚光を浴びているハイパーテキストシステムは主にこういった種類の構造を取り扱っている。概念構造とは専門分野の概念は孤立的に存在せず、まわりの概念と同値関係、上下関係、部分全体関係、類似関係などの関係をもつ構造を指す。概念構造はシソーラスとして記述される。すなわち、シソーラスは情報ベースにおける情報アクセスと概念間の構造化といった二つの役割を果たしているわけである。論理構造とは因果関係、継承関係、順序関係をもつ概念間の構造を指す。論理構造はタキソノミーとして記述される。

多様な情報タイプと大量の情報を扱っている情報ベースににとって、情報を収集し利用に供するためには、情報の構造化は非常に重要であることをいうまでもない。本研究の情報ベースでは、学習、帰納、類推などの機能を概念構造と論理構造によってもたせている[28,29,30]。

先ほど示した三つの構造の中で、シソーラスの自動構築法により概念構造の組織化の方法およびその結果を報告する。

3　シソーラスの自動構築法

3.1　シソーラス自動構築の必要性

科学技術のあらゆる分野における概念は専門用語によって表現される。概念の体系を用語の体系によって表現されるのをシソーラスとよぶ。すなわち、用語間の意味関連および階層関係を体系化されるものを指す。一般的に、シソーラスは、意味処理、とくに同義性、多義性をもつ自然言語のアクセスと管理との問題解決には有効であるが, 従来のシソーラス構築はほとんど人手に頼ってばかりで, かつ用語間の意味的関係の判断には専門家の膨大な時間と知的労働を費やしていた。こういうふうにして出来上がったシソーラスは発展迅速の分野において使い始める時点では古くなってしまったケースはしばしばあるし、保守も困難である。ですから、シソーラスを自動的に構築、保

守する方法の実現は重要である。

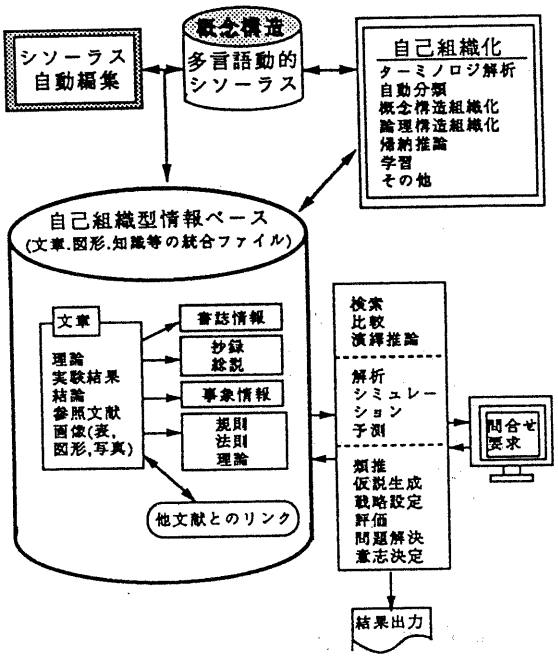

図1　*情報ベース概略図*

3.2　シソーラス自動構築システムの情報源

いくつかのシソーラス自動構築法の提案があったが{12,13,14,15}、情報源として狭い分野範囲に限るものがほとんどであった。研究用データベース、「学術用語集」、「工業標準用語辞典」の機械読み可能形式を情報源にした。工業用語辞典には26分野の64000余りの用語を含まれている。機械読み可能形式は固定長レコードの集まりであり、具体的に、所在、分野、英語、日本語、日本語読みという五つのフィールドから構成されている。広範囲の分野を包含され、各分野で標準化された用語を集め、英日対訳形式を呈するのはこの用語集の特徴である。

この用語集における用語間の関連に着目し，用語集に含まれる専門家の知識を集約して，シソーラスの作成と保守を自動的に実現することを目標として，用語集に基づいたシソーラス自動構成法の開発を行っている{2,3}。この自動構成法の最大な特徴、つまりほかの方法と本質的に異なったところは同時に二つの言語におけるそれぞれのシソーラスができることである。

3.3　シソーラス自動構築システム

開発中のシソーラス自動作成システム（図2に示している）は同値関係抽出，造語規則利用，語義文解析という三つの機能によって構成される。用語集を情報源として，既知の階層関係（上下関係ともいう）を参照して新たな同値関係を取り出すとともに造語規則を利用して新しい階層関係を抽出し，シソーラスを補充，拡張する。また定義式辞書の語義文解析により，より一般的な同値関係，上下関係を抽出する方法の開発につながり，更にシソーラスに対する更新と追加も容易になる{4,5,6}。

3.4　同値関係抽出アルゴリズム

まず本研究で取り扱われている同値関係の定義を与える。同値関係とは複数の用語が同じ概念を表しているとみなされる場合の関係のことである。言語Aの概念から言語Bの概念への翻訳作業の結果はちょうど異なる言語の概念の間に、上述した意味での同値関係の橋をかけるようなものである。同義語 (synonym) とは意味が広範囲の文脈で同一と見なしうるため、実際上、交換可能であるような用語のことを指す。

言語Aから言語Bへの対訳付用語集をΩとする。われわれはこのような用語集をもとにし、同値関係Rの推移則（Transitivity Rule）即ちR(ライト,light)とR(light,光)ならば、R(ライト,光)を利用することによって、あるスタートワードの、言語Aに属する推移閉包S^+と言語Bに属する推移閉包T^+を作ることができる。この閉包が同義語集合である。言語Aと言語Bによる対訳付用語集Ωの基で同値関係抽出のアルゴリズムを下に記述する。

言語A、言語Bの同義語集合S,Tを以下のように定義する。まずここでS'とT'がSとTのワーキングスペースとする。任意のスタートワードs ∈ Aを選べ出すと、$S^0 = \{s\}$,$T^0 = \{\}$にする。ここでワーキングスペース

$$T' = \{b_j | (s, b_j) \in \Omega\}$$

をこのように与える。さらに

$$T^{i+1} = T' \cup T^i$$

によって

$$S' = \bigcup_{b_k \in T^{i+1}} \{a_h | (a_h, b_k) \in \Omega\}$$

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

同じくS^{i+1}によって

$$T' = \bigcup_{a_m \in S^{i+1}} \{b_n | (a_m, b_n) \in \Omega\}$$

から新たな$T^{i+1} = T' \cup T^i$が得られるわけである。次の状態

$$T^i = T^{i+1} = T^{i+2} = \cdots\cdots$$

になると、

$$S^i = S^{i+1} = S^{i+2} = \cdots\cdots$$

もなる。アルゴリズムが終了する。

$$T^+ = T^i \subseteq B$$

がB中のsの同義語集合となる。同時にAにおけるsの同義語集合$S^+ \subseteq A$を得ることもできる。

4 結果に関する分析および実例

用語の本質的な特徴は同義性（同義語による）と多義性（多義語による）が存在しているのである。同義語の定義を先ほど与えたが、多義語(polyseme)とは同一の用語が多様な使い方によって複数の概念に対応する用語のことを指す。多義性の発生が二つの面から考えられる。(1)共通語源から違う概念に対応してその意味が変化することによって発生するもの；(2)異なる語源から各自の使い方で偶然な一致によって発生するもの，例えば：

TM＿teacher's manual（教師用手引）

TM＿transcendental meditation（超越瞑想）

TM＿teaching machine（教育学習機器）

TM＿theme music（テーマ音楽）

などたくさん存在する。今回工業標準用語を対象として実験を行ったが，自然語に基礎をおいた工業標準用語はもちろん同義性、多義性の問題も存在する。抽出によるいくつかの結果の実例を挙げる。

例１：
StartWord: （X線）撮影（法）
SynonymSet(Japanese):
（X線）撮影（法）【L】
X線写真【L】
ラジオグラフィー【L】
放射線写真【L】
ラジオグラフ【K】
X線写真【K】

例２：
StartWord: のぞき窓
SynonymSet(Japanese):
のぞき穴【A】
検査穴【A】
のぞき窓【A】
のぞき窓【P】
のぞき窓【H】
のぞき窓【V】
サイトグラス【V】

例３：
StartWord: （インパルス・ホイールスピード・センサの）レゾリューション
SynonymSet(Japanese):
分離度【Y】
決議【Y】
（インパルス・ホイールスピード・センサの）レゾリューション【Y】
解像度（テレビジョンの）【Y】
分解能【Y】
感度限界【Y】
解像力【Y】
分解能（検出器の）【Y】
分解力【Y】
識別【Y】
敷居【H】
感度限界【H】
域値【H】

しきい値【H】
識別【C】
敷居【R】

例4：
StartWord: （クレーン）スパン
SynonymSet(Japanese):
翼幅【C】
範囲【C】
（クレーン）スパン【C】
スパン【C】
翼幅【N】
航続距離【C】
レンジ【C】
飛程【C】
航続距離【Q】
航続時間【F】
航続距離【F】

SynonymSet(Japanese)における各用語の最後に付いてあるアルファベットがその用語の属する分野を意味する。用語集にはAからZまでの26分野を含まれている。ここでほんの二、三例を示した。例1、例2はわれわれの求める本当の同義語集合である。反して、例3、例4は明らかに同義語集合ではない。いくつかの同義語集合の和集合と見なされたほうが適切であろう。ゆえに、例1、例2のような集合のみに対して同義語集合という呼び名を延用するが、例3、例4のような集合を新たな呼び名ーー多義語集合と名付けよう。便宜上、混同集合という概念を定義する。集合のサイズ（集合が含む用語の数）が二つ以上かつ集合における用語らが少なくとも二つ以上の分野に属するような集合を混同集合とよぶ。下に示したのは全分野において同値関係抽出に関するデータである。

用語集サイズ＝64314
SWリストサイズ＝53348
抽出された集合総数＝95298
混同集合数＝11002
サイズ1の同義語集合数＝80404
サイズ2の混同集合数＝11308
サイズ3の混同集合数＝3586
混同集合の最大サイズ＝112

5 組織化される情報空間へのアプローチ

5.1 多義語集合の分解

上の例でわかるように、同値関係自動抽出アルゴリズムによって同義語集合を抽出されたと同時に多義語集合も抽出されてしまった。多義語集合はいくつかの同義語集合から構成されている（例を参照）ことを注目しながら、多義語集合の分解というアプローチを考案した。

S_i(i=1,⋯,j) が同義語集合とする。多義語集合Pが次の条件

$$P = S_1 \cup S_2 \cup \cdots \cup S_j$$

かつ

$$S_1 \cup S_2 \cup \cdots \cup S_j = \phi$$

を満たされれば、S_i(i=1,⋯,j) をPの分解とよぶ。多義語集合分解の大間かな手順は次の通り

（1）用語集をいくつかのブロック $D_1, D_2, \cdots, D_n$ に分割する。各 D_i(i=1,⋯,n) は接近分野の集まり、例えば、化学、化学技術、環境工学と安全工学、材料、金属工学、木材と紙および繊維という六つ分野の集まりとか。

（2）各 D_i に対して抽出実験を行なう。得られた結果に関して同義語集合であるか否かを判断する（現段階では、この判断作業が部分的に人に頼る）。そして同義語集合を同義語集合ファイル（これから単にファイルとよぶ）に記憶する。判断により多義語集合であれば、いくつかの”ヒント”（キーワードなど）を与えることによって、学習的に分解して、その結果もファイルに記憶する。

（3）全用語集に対して抽出実験を行なう。ファイルを参照しながら多義語集合を分解していく（同形、同音判断による）。ここでの多義語集合についての判断は完全に自動的に行う。新たに得られた同義語集合をファイルに記憶する。

次には一つの例を示している。ファイルにつぎの二つの同義語集合

可視放射【B】
可視放射【I】
可視放射【D】
光【D】
と
窓【R】
ウィンドウ【R】
が含まれているとすると、多義語集合
可視放射【B】
窓【R】
可視放射【I】
ウィンドウ【R】
可視放射【D】
光【C】
光【D】
窓【C】
を、ファイルを参照することによって、新たな同義語集合
可視放射【B】
可視放射【I】
可視放射【D】
光【D】
光【C】
と
窓【R】
ウィンドウ【R】
窓【C】
に分解することができる。

5.2 上位概念による分解

既知の階層関係を利用して，混同集合における用語間の同値関係を判別する方法である。つまり共通の上位概念で同義語同士を結び付ける方法である。三つの同義語集合 $\{\alpha,\beta\},\{\beta,\gamma,\delta\},\{\delta,\epsilon\}$ の中で，δ と δ は多義性を持つ語と仮定する。

抽出実験の結果，単純な推移閉包では意味の異なる三つの同義語集合が同一の集合 $\{\alpha,\beta,\gamma,\delta,\epsilon\}$ にはいることになる。

B_1、B_2、B_3 はそれぞれの同義語集合に共通の上位概念とすると，この共通の上位概念を利用して確実な同義語同士しか結び付かないように分解することができる。このような方法によって，多義語集合も分解され，同値関係も失われない。

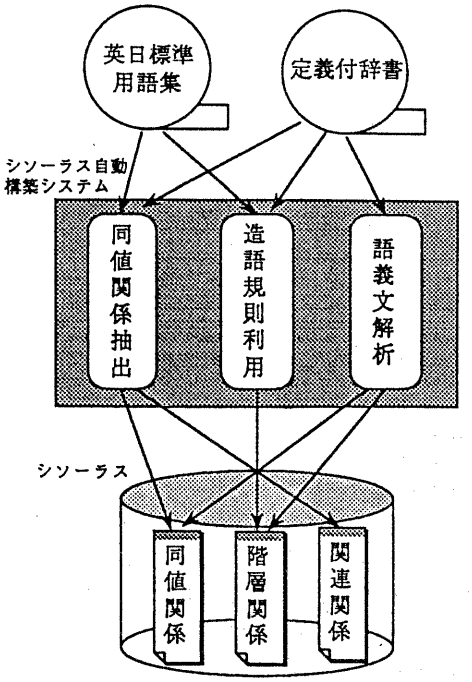

図2　シソーラス自動構築システム

5.3 既存シソーラスの利用

シソーラスを構築するとき、既存のシソーラスを利用することも有効である。ここでいう利用は既存のシソーラスから有用な要素を取り出して新しいシソーラスの構築に役立つ、あるいは、既存シソーラスに対して追加、更新、保守する。シソーラス自動構築システムはこの利用問題を重要な機能として取り扱っている。現段階では、われわれの自動構築システムは階層関係の抽出が不十分であって、多数の専門家による豊富な階層関係を備えるシスーラスの利用を検討した。ISO の ROOT thesaurus は同値関係が少ないが階層関係が豊である。しかしその階層関係がわれわれにとって細かすぎるため、、集約化をしながら、利用する。このアプローチによって自動的に構築されるシソーラスは尚一層強力的であろう。

6 むすび

シソーラス自動構築システムは情報空間の構造化に非常に有力であることはいろいろな実験結果から示した。しかもシソーラス自動構築システムはすでに実用段階にまで達したことも明瞭となった。今後の課題としては、意味解析によりもっと信頼度の高い同義語集合の自動抽出の実現、多言語におけるシソーラス自動構築法への拡張により情報ベースにおける多言語の概念構造の実現などである。

参考文献

[1] F.W. ランカスター著, 松村多美子, 鈴木祐滋訳：情報システムのためのシソーラスの構築と利用. (社) 情報科学技術協会, (1989)

[2] Y. Fujiwara, W. G.Lee, Y. Ishikawa, T. Yamagishi, A. Nishioka, K. Hatada, N. Ohbo and S. Fujiwara: A Dynamic Thesaurus for Intelligent Access to Research Databases. (44)FID Congres, Aug. 1988, Helsinki

[3] Y. Fujiwara, N. Ohbo, T. Itoh, M. Morita, K. Sawai, T. Kawasaki and S. Fujiwara: MULTILINGUAL THESAURI FOR INTERNATIONALLY DISTRIBUTED INFORMATION SYSTEMS. Information, Communication,and Technology Transfer. 1987, pp.47-54

[4] A.Ghose;A.S.Dhawle : Problems of Thesaurus Construction. Journal of American Society for Imformation Science - July 1977,p p.211-217

[5] 藤原譲, 李元揆, 張暁冬, 北川博之, 大保信夫：科学技術用語集に基づくシソーラスの自動作成. 情報学シンポジュム, pp.63-69

[6] Y. Fujiwara, J. He, G. Chang, N. Ohbo, H. Kitagawa and K. Yamaguchi : Self Organizing Information Systems for Material Design. Proceedings of - CAMSE '90, Aug. 1990, Tokyo

[7] 笹森勝之助：シソーラス作成の自動化. 情報管理, Vol. 15 No. 4, pp.260-268

[8] 原田隆史, 細野公男, 田村俊作, 高柳俊子, 後藤智範, 岸田和明, 坂田亮子：複合語の解析による語の上位ー下位関係の自動構築. 自然言語処理, 70-5

[9] 高野文雄, 佐藤誠：JICST ファイルの用語統計分析とシソーラス作成への利用. 情報管理, Vol. 29 No. 12 Mar. 1987, pp.1035-1052

[10] 高野文雄, 岡野弘行：シソーラスの利用と今後の問題 (I). 情報管理, Vol. 20 No. 11 Feb. 1978, pp.875-893

[11] 高野文雄, 岡野弘行：シソーラスの利用と今後の問題 (II). 情報管理, Vol. 20 No. 12 Mar. 1978, pp.945-965

[12] 鶴丸弘昭, 日高達, 吉田将：単語間の上位ー下位関係の自動抽出. 情報処理学会研究報告, 86-FI-3

[13] 富浦洋一, 日高達, 吉田将：国語辞典の語義文からの動詞の上位ー下位関係の抽出, 自然言語処理, 73-3, pp.17-24

[14] 荻野孝野, 横山晶一, 荻野網男：シソーラス作成のための辞書関係語の抽出, 第 39 回 (平成元年後期) 全国大会講演論文集 (I), pp.670-671

[15] 横山晶一, 荻野孝野, 荻野網男：国語辞書に基づくシソーラスの計算機処理, 第 39 回 (平成元年後期) 全国大会講演論文集 (I), pp.672-673

[16] 宮地泰造, 吉武淳：シソーラスの構成に関する一考察. 第 39 回 (平成元年後期) 全国大会講演論文集 (I), pp.674-675

[17] 永井秀利, 中村貞吾, 日高達：造語モデルに基づく単語表記の扱い. 第 39 回 (平成元年後期) 全国大会講演論文集 (I), pp.58788

[18] U.Guntzer et al.:Automatic Thesaurus Construction By Machine Le arning From Retrieval Sessions. Information Processing and Managmen t Vol. 25 No. 3,pp.265-273, 1989

[19] 水谷静夫編集：朝倉日本語新講座 1 文学・表記と語構成. 朝倉書店, 1987 年

[20] 岩波講座日本語 9 語彙と意味. 岩波書店, 1977 年

[21] 荒木啓介：単言語シソーラスの設定と発展のための指針「その１」．情報管理，Vol.19 No.5 Aug. 1976, pp.34349

[22] 荒木啓介：単言語シソーラスの設定と発展のための指針「その２」．情報管理，Vol.19 No.6 Sep. 1976, pp.41725

[23] 斉藤和男：「ＪＩＣＳＴ科学技術用語シソーラス」の編成経緯．情報管理，Vol. 18 No.5 Aug. 1975, pp.390-399

[24] 安倍浩二：「ＪＩＣＳＴ科学技術用語シソーラス」１９７５年版の作成Ｉ科学技術用語の選定と関係．情報管理，Vol. 18 No. 6 Sep. 1975, pp 463-472

[25] 小野脩一，佐原卓：「ＪＩＣＳＴ科学技術用語シソーラス」１９７５年版の作成電算機によるシソーラスの編成．情報管理，Vol. 18 No. 9 D ec. 1975, pp.729-739

[26] 緒方良彦：自動分かち書きの手法．情報管理，Vol 9 No. 12 1966, p p.650-656

[27] 鶴丸弘昭，兵頭竜二，松崎功，日高達，吉田将：語義を考慮した単語間の階層構造の抽出について．自然言語処理，64,pp.9-16

[28] Brachman, R.J. and Schmolze, J.G.: An Overview of the KL-ONE Know ledge Representation System. Cognitive Science, Aug. 1985, pp.171-216

[29] Wang, Z.Q., Zheng, S.Q., Yu, X., Yamaguchi, K., Kitagawa, H., Ohbo, N., and Fujiwara, Y.: Learning and Analogical Reasoning in the Information - Base Systems for Organic Synthesis Research. Journal of Japan Society of Information and Knowledge, 2(1), pp.72-79, (1991)

[30] Wang, Z.Q., Zheng, S.Q., Yu, X., Yamaguchi, K., Kitagawa, H., Ohbo, N., and Fujiwara, Y.: Learning and Rule Control in the infor mation Base System for Organic Synthesis Research. 14th Symposium on Chemical Information and Computer Science, Nov.,27-29, (1991), pp.162-165