

2 項関係抽象化支援システムの基本機能

7S-7

松川智義、中沢正幸、安達久博*、天野真家
(株)日本電子化辞書研究所

1. はじめに

(株)日本電子化辞書研究所(EDR)では、実例文の分析に基づく大規模な自然言語処理用電子化辞書の研究開発を行っている。電子化辞書の開発においても、知識ベース構築における知識の記述¹⁾と同様、カテゴリが設定され、設定されたカテゴリを用いて様々な言語内知識が抽象化されて記述される。

このようなカテゴリの妥当性を検証するためには、カテゴリを用いて記述された知識を実例によって裏付ける、という方法が考えられる。そのためには、カテゴリに実例を対応付けて検証作業者に提示するための手だてが必要である。

そこで、電子化辞書構築のための素データとなる様々な2項関係の実例データをカテゴリによる抽象化された記述に対応付けて可視化することでその記述を検証する作業や、2項関係の実例データを解析して抽象化可能な部分を見つけ出す作業、を支援するシステムを作成した。ここで抽象化とは、要素の集まりをカテゴリ化し、カテゴリのペアとして要素間の2項関係を記述することを指す。

2. 2項関係抽象化支援システムの概要

2.1 用語の説明

本システムの作業の対象となるデータは、図1のような記号のペアの集合として表された、2項関係の実例データである。本稿ではこのようなデータを「元データ」と呼ぶことにする。また、元データ中に現れる記号を「要素」と呼ぶことにする。

元データは、1/0行列として表すこともできる。本システムでは、このような1/0行列状の元データの中から1で埋まった部分行列を抽出することにより、カテゴリによって抽象化された記述を得ることができる。また、元データ中の部分行列を図2のように行列状に可視化して、様々な操作をその部分行列に施すこともできる。本稿では、操作の対象となるこのような部分行列のことを「グループ」と呼ぶことにする。

容	座	+	+	+	1	1	1	1	2	男	/	以	外	の	平	均	的	質				
音	座	+	+	+	1	1	1	1	2	男	/	供	養	・	種	族	・	人	種			
情	報	・	情	報	+	+	+	+	2	男	/	子	供	養	・	種	族	・	人	種		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
機	能	(座	落)	・	機	能	(座	落)	+	+	+	+	+	+	+	+	+	+		
情	報	・	情	報	+	+	+	+	2	男	/	以	外	の	平	均	的	質				
群	集	・	群	集	+	+	+	+	2	女	/	人	間	・	男	女	・	種	族	・	人	種
群	集	・	群	集	+	+	+	+	2	男	/	子	供	養	・	種	族	・	人	種		
群	集	・	群	集	+	+	+	+	2	女	/	子	供	養	・	種	族	・	人	種		
群	集	・	群	集	+	+	+	+	2	女	/	性	の	な	い	そ	の	他	の	集	団	
群	集	・	群	集	+	+	+	+	2	男	/	性	の	な	い	そ	の	他	の	集	団	
群	集	・	群	集	+	+	+	+	2	男	/	性	の	な	い	そ	の	他	の	集	団	
群	集	・	群	集	+	+	+	+	2	女	/	性	の	な	い	そ	の	他	の	集	団	
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		
空	間	関	係	・	空	間	関	係	+	+	+	+	+	+	+	+	+	+	+	+		

図1 2項関係の実例データ

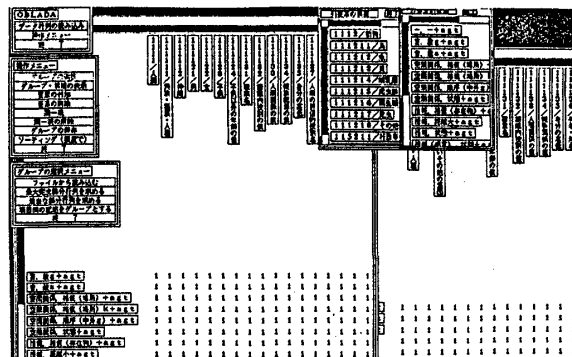


図2 グループの可視化

2.2 本システムの機能

本システムの機能は、a. 1つのグループに対する操作、b. 複数のグループに対する操作、の2つに大きく分けられる。主なものについて述べる。

2.2.1 1つのグループに対する操作

- **グループの編集** : 図2のように、グループを行列状に可視化した上で、グループへの要素の付加や、グループからの要素の削除を画面上で行うことができる。付加される要素は、グループに含まれ得るペアを多く持つもの(候補)を表示させ、その中から選ぶことができる。
- **ソーティング** : 表示されているグループの行側、列側の要素をそれぞれグループ内でのペアの出現頻度について並べ替えることができる。
- **グループの消去** : 元データ上からグループの部分のペアをすべて仮想的に消去することができる。
- **自動グループ化** : 完全にペアが埋まった部分行列や、ある程度ペアが埋まった部分行列を自動的に抽出してきてグループとすることができる。この処理は、文献7)のアルゴリズムを応用して実現している。

2.2.2 複数のグループに対する操作

複数のウィンドウを開くことで、並行していくつかのグループを編集することができる。また、複数のグループをシステムに取り込んで、それらの行側あるいは列側の間の、a. 包含関係を調べたり、b. 集合差や交わりを求めて新たなグループを作ったり、することができる。

3. 電子化辞書構築のための素データ

EDRで作成している電子化辞書構築のための素データのうち、本システムによる抽象化の対象となり得るデータには以下のようなものがある。

- a. **項目間記述データ⁵⁾** : 概念分類項目²⁾間の関係記述データ。2つの概念分類項目が1つの関係子で結ばれたもの(項目間記述タプル)の集合として表

* 現在、(株)東芝 総合研究所
Basic Functions of the Environment for Binary Relation Categorization
Tomoyoshi MATSUKAWA, Masayuki NAKAZAWA, Hisahiro ADACHI, Shin-ya Amano
Japan Electronic Dictionary Research Institute, Ltd.

される。

b. 概念間記述データ²¹ : テキストを構文的・意味的に分析作業した結果得られる、個々の概念間の関係記述データ。2つの概念が1つの関係子で結ばれたもの(概念間記述タプル)の集合として表される。

c. 単語共起データ²¹ : 単語間の共起関係を表すデータ。概念間記述データと同様、テキストの分析作業の結果得られる。2つの単語が1つの共起関係子で結ばれたものの集合として表される。

4. システムの使用例 - データの検証と抽象化

本システムの使用方法は次の2つの側面に分けて整理することができる:a. グループの可視化によるデータの検証、b. 自動グループ化による2項関係の抽象化。

4.1 グループの可視化によるデータの検証

概念間記述データを元データとし、項目間記述タプルをグループとする以下の作業により、項目間記述データ、概念分類データ²¹の検証・修正を行うことが可能である。

○ 作業1(項目間記述に対する事例の可視化):
グループを行列状に表示して、値1の埋まり具合などを基に、タプルの妥当性を検討する。

○ 作業2(項目間記述に基づく概念分類の検証):
1. グループ内のペアの頻度の低い要素を一覧し、不適切な概念が分類されていないかどうか調べる。
2. グループに含まれ得るペアの数の多い概念(候補)を表示させ、それらの項目への分類追加を検討する。

4.2 自動グループ化による2項関係の抽象化

3. で述べた3種類の元データを(半)自動的にグループ化することにより、それぞれ、a. 項目間記述に基づく概念体系の構築、b. 概念間記述データに基づく項目間記述の追加、c. 単語共起データの記述量圧縮、という作業を支援することができる。

4.2.1 項目間記述に基づく概念体系の構築

本システムによって、項目間記述に基づく概念体系の構築作業を支援することができる。この作業は、1. 項目間記述のグループ化処理、2. グループの階層化処理、3. ノードの交わり関係に基づく下位ノードの付加、の3つの工程より成る^{61, 81}。これらは2.2で述べた、1. 自動グループ化機能、複数のグループの、2. 包含関係と集合差、3. 交わり、を求める機能をそれぞれ用いて実現できる。これらの機能を総称して、「階層関係抽出機能」と呼ぶ。

機能実行の形態としては、a. 対話処理、b. 一括処理、の2種類の方法から選べるようにした。a. を用いた場合、1. ~ 3. の各工程ごとに処理を中断し、本システムの編集機能を用いて途中結果としてのグループやノードを適切なものに変更してから、次の工程に移る。b. を用いた場合、1. 2. を一括して処理した後、3. の作業を行う。

4.2.2 概念間記述に基づく項目間記述の追加

概念間記述データを元データとしてグループ化し

た結果は、概念をカテゴリに分類して、カテゴリ間で関係記述を行ったものとみなすことができる。すなわち、この結果は項目間記述タプルと同じレベルの記述であると考えられる。したがって、これを項目間記述データの検証に用いることができる。

概念間記述データを自動グループ化した結果、a. 意味的に共通性のあるペア集合がグループとして得られ、b. 項目間記述データ中にグループに対応する項目間記述タプルがない場合は、新たな項目間記述タプルの追加を検討する。さらにその際、c. グループの行側、列側に対応する概念分類項目がない場合は、新たな概念分類項目の設定も併せて検討する。

4.2.3 単語共起データの記述量圧縮

単語共起データをグループ化することにより、単語共起データの記述量の圧縮を計ることができる。具体的には、次のようにすればよい。1. 得られたグループに1つずつ記号を与える、2. 各単語には、それが属するすべてのグループの記号をリストで持たせる。3. 2つの単語が共起するか否かは、互いのリストの交わりが在るかどうかで判断する。

5. おわりに

事例データに基づいて2項関係を抽象化する作業を支援するシステムを作成した。主な機能は、a. データの可視化、b. 自動グループ化、c. グループの編集とグループ間の関係の解析、である。本稿では、これらの機能を用いて概念体系の構築における種々の作業を支援することが可能であることを示した。

今後の課題としては、以下のことが挙げられる:
a. 使用例として想定している様々な作業に用いることで機能の有効性を確かめる。b. 作業効率を向上させるためにマンマシンインターフェースの充実を図る。

【謝辞】

本研究の機会を与えて頂いた横井所長、並びに有益な意見・コメントを頂いたEDRの研究員の皆様に感謝いたします。

【参考文献】

- 1) Lenat, D. B., et al.: Building Large Knowledge-Based Systems., Addison-Wesley Publishing Company, Inc., p. p. 160-162 (1989).
- 2) 日本電子化辞書研究所: 概念辞書(第2版), IR-012, p. p. 55-73 (1989).
- 3) 日本電子化辞書研究所: 共起辞書(第1版), IR-013, p. p. 4-8 (1989).
- 4) 岸本行生他: 概念分類項目の設定, 情処第41回全国大会(1990).
- 5) 横田英司他: 概念分類項目間の関係記述, 情処第41回全国大会(1990).
- 6) 三池誠司他: 概念分類項目間の関係記述に基づく概念体系の試作, 情処第41回全国大会(1990).
- 7) 松川智義他: 共起関係に注目したDM分解と確率的推定による単語のクラスタリング, 情処 NL-72-9 (1989).
- 8) 松川智義他: 記述量圧縮の観点から見た概念体系の構築, 情処 NL-78 (1990).