社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－176（9）
2006／11／22

非階層関係にある関連語集合の抽出と発想支援への適用

山本英子　　井佐原均

情報通信研究機構

本稿では，上位語下位語や同義語，反義語といった分類的関連を持つ単語集合ではなく，連想関係や因果関係といった主題的関連を持つ単語集合をテキスト集合から抽出することを試みる．後者の単語集合が持つ関連は，前者と違って，シソーラスのような知識ではなく，辞書に載っていない知識である．本研究では，そのような知識が発想支援に利用できるのではないかと考え，得られた関連語集合を用いて実際に Web 検索をすることにより，提案手法によって得られた関連語集合が発想支援に適用できることを示す．

Extraction of Related Word Sets with Non-Hierarchical Relation and its Application to Creativity Support

Eiko Yamamoto　　Hitoshi Isahara

National Institute of Information and Communications Technology

In this paper, we tried to extract sets of related word with thematic relations such as associated relation and causal relation, which are not taxonomical relations such as hypernym-hyponym relation and synonym-acronym relation. The relations between words composing the latter related word set can be regarded as knowledge which is not thesaurus-like knowledge and which there is not in the dictionaries. We think such related word sets can be used to support creativity. Then, we estimate the availability for creativity support that the related word sets we extracted can be had, through verification of their availability to Web retrieval.

1　はじめに

単語間に存在する関係は分類的関連(taxonomical relation)と主題的関連(thematic relation)の二つに大きく分けられる．単語間の関係は単語によって表現される概念間の関係に相当する．分類的関連とは，二つの概念間の属性的な類似性に基づく関係である．たとえば，「牛」と「馬」，「馬」と「動物」といった関係である．一方，主題的関連とは，主題の状況を通した二つの概念間の関係である．たとえば，「牛」と「ミルク」といったある状況を思い出させる関係，この場合は「牛のミルクを絞る」という状況を思い出させる関係である．これまでに，二つの物体を認識する上で，分類的関連だけでなく，主題的関連もまた重要であるということが報告されている(Wisniewski and Bassok, 1999)．また，創造性を評価する実験において，分類的関連にある単語集合を提示した場合より，主題的関連にある単語集合を提示した場合のほうが発想に広がりが観られたと報告されている(Taura and Nagai, 2005; Harakawa et al., 2005)．

このように，最近創造性の観点では，分類的

関連を持つ単語集合よりもむしろ主題的関連を持つ単語集合を提示することによってより幅広い発想が得られることが注目されている．新しいものを創造するには，発想力が必要不可欠である．しかしながら，発想するという行為には簡単な推論や演繹以上のものが要求されるので，非常に難しい．このことから，今日，発想支援に対するニーズが高まっている．簡単な発想支援は，ユーザの想像を展開できる単語をユーザに提示することである．しかし，理論的には，どんな単語であっても想像を広げる手助けとなりえるが，単に単語をランダムに提示したのでは，アイデアを発散させてしまい，実際には発想支援のために有用であるとはいえない．このような背景から，どういった関係にある単語が発想を促し，想像を広げていくか，また，想像がどのような方向に広がっていくかが研究され始めている．

このような背景から，我々は，主題的関連を持つ単語からなる有用な関連語集合を文書集合から抽出することを試みた．このような関連語集合を抽出するために，これまでに提案した文書集合からの自動階層構築方法を拡張し，関連語集合を抽出する．抽出した関連語集合を階層関係にある関連語集合と非階層関係にある関連語集合とに選別し，非階層関係にある関連語集合は主題的関連を持つ単語集合であるとみなして，その有用性を検証する．実際には非階層関係にある関連語集合を入力として Web 上で検索を行い，ユーザを有益なページに導きうるかどうかを調査する．その結果に基づき，提案する関連語集合の抽出手法と得られた関連語集合の発想支援への適用可能性を評価する．

2　検索支援と発想支援

本研究では，非階層関係にある関連語集合がユーザを有益なページに導きうるかを調査する．

我々の考える有益なページとは，ユーザが思いつかないようなページやユーザが探しているのに，よいキーワード群が思いつかず，見つけられないページを指す．このようなページへの誘導性は，キーワード群が持つ，一つの検索支援能力と捉えられる．

一方，創造することにおいて，ある単語から発想を広げようとしても，アイデアが発散してしまい，想像が広げられない場合がある．このような場合には，発想にある方向性を与えることによって，創造へと導く必要がある．つまり，想像を的確に広げ，発想を進めるための支援である．また，ある単語を見ても発想が広がらない場合，その単語と関連のある語を提示することで，発想を促し，想像を進めることも可能である．これも同様の発想支援である．

実験では，分野を限定して関連語集合を抽出し，その関連語集合の前述した検索支援能力を調査することで，発想支援への応用可能性を推測する．

分野を限定して関連語集合を抽出し，その検索支援能力を評価することによって，関連語集合の発想支援での有用性をある程度検証できる．これは，ある単語集合が有益な Web ページを検索できるのなら，その単語集合は発想を方向付けるからである．つまり，そのような単語集合であれば，人はそれを見て，それが導きうるページに含まれる情報を推測しうる．創造に答えはないが，分野を限定したキーワード群が Web 上の有益なページを検索する過程は，人があるキーワード群からものを発想する過程と類似していると考えた．このような考えから，抽出した関連語集合を用いて，Web 検索を行うことで，その発想支援への適用可能性を推測する．

3　関連語集合の抽出

3.1　自動階層構築方法の応用

我々はこれまでに文書集合から語彙の階層構造を自動構築する手法を提案した(Yamamoto et al., 2005; 山本他，2006)．この手法は対象とする語が文書集合中でどの語と共起するかという状況に基づいて語彙の階層構造を構築する．この手法において状況間の関係を定めている指標は，補完類似度(Complementary Similarity Measure: CSM) (Hagita and Sawaki, 1995)というベクトル間の重なり度合い(包含関係)を測る尺度である．我々は抽象名詞を対象に，それと係り受け関係にある形容(動)詞を共起語として用い，単語(抽象名詞)対が持つ CSM 値を意味的階層関係における関連度として扱ってきた．抽象名詞は共起する形容(動)詞の上位概念と位置づけられる(Kanzaki et al., 2004)ので，形容(動)詞との共起に基づく CSM 値は抽象名詞の意味的階層関係を数値化したものとなる．実際，この手法を用いることにより，適切な抽象名詞の階層構造を抽出することができた．

一方，本研究では，因果関係や連想関係といった主題的関連を持つ関連語集合を得たいと

考え，その第一段階として，非階層関係にある関連語集合を抽出することを試みる. そのために，自動階層構築方法を拡張し，非階層関係にある関連語集合の抽出に応用した. 具体的には，形容(動)詞と抽象名詞の係り受けではなく，動詞に係る名詞句の出現状況などをベクトル化して，自動階層構築方法のアルゴリズムを適用した.

3.2 出現状況のベクトル表現

本研究では，ベクトル間の包含関係を非対称の距離尺度である CSM を用いて，数値化する.

単語 w_i が n 個の単語のどれと共起するかを 1，0 で表現したベクトルを $\boldsymbol{Vi} = (v_{i1}, ..., v_{in})$，単語 w_j についてのベクトルを $\boldsymbol{Vj} = (v_{j1}, ..., v_{jn})$ としたとき，$CSM(\boldsymbol{Vi}, \boldsymbol{Vj})$ は次のように定義される.

$$CSM(V_i,V_j)=\frac{ad-bc}{\sqrt{(a+c)(b+d)}},$$

$$a=\sum_{k=1}^{n} v_{ik}\cdot v_{jk}, \qquad b=\sum_{k=1}^{n} v_{ik}\cdot(1-v_{jk}),$$

$$c=\sum_{k=1}^{n}(1-v_{ik})\cdot v_{jk}, \quad d=\sum_{k=1}^{n}(1-v_{ik})\cdot(1-v_{jk}).$$

単語 w_i と w_j を抽象名詞，ベクトルの次元に対応する共起語を形容(動)詞としたとき，パラメータ a，b，c，d はそれぞれ w_i と w_j の双方と共起する形容(動)詞の数，w_i だけと共起する形容(動)詞の数，w_j だけと共起する形容(動)詞の数，どちらとも共起しない形容(動)詞の数に相当する. n は形容(動)詞の総数 a+b+c+d である.

3.3 階層構造の構築法

CSM を用いて，文書集合中の二つの単語(抽象名詞)の共起語(抽象名詞を修飾する形容(動)詞)の集合同士が包含関係にあるかどうか，したがって二つの単語が上位下位関係にあるかどうかをと推定することができる. 上位下位関係にあるとされた単語対を順次連結していくことによって階層構造を構築する. 階層構造の構築には，CSM 値があらかじめ定めた閾値以上である単語対のみを使う. 以下で，例を用いて構築手法を説明する.

CSM 値が高い順に並んだ単語対<A, B>, <B, C>，<C, D>，<B, D>，＜Z, A>，<D, E>があるとする．ここで，<X, Y>という表記は，X が Y の上位語，Y が X の下位語と推定された単語対を意味し，この関係を X→Y と表す．階層構造の初期値を<A, B>としたとき，B を上位語として持つ単語対のうちで CSM 値が最も高いものを探し，その下位語を A→B に連結する．ここでは，<B, C>が<B, D>より CSM 値が高いので，A→B→C となる．この工程を繰り返し，A→B→C→D→E まで単語を連結できる．次に，A を下位語として持つ単語対のうちで CSM 値が最も高いものを探し，その上位語をすでに作成した A→B→C→D→E の前に連結する．この例では，<Z, A>があるので，Z を A の前に連結する. この工程も連結できる単語が見つかる間，繰り返す．この結果，この例で得られる階層構造は Z→A→B→C→D→E となる．また，＜B, D>を初期値とした場合，Z→A→B→D→E という階層構造が得られるが，これは Z→A→B→C→D→E に含まれる構造であり，我々はより多い単語で構成される階層構造を採用するため，この構造は結果として得られる階層構造のリストから削除する.

4 実験

4.1 対象とした用語と文書集合

実験では，検証のために，対象とする用語を限定した. 対象とする用語を日本語の医学用語に限定し，関連語集合を抽出する文書集合を医学分野に関連する Web 文書集合(10,144 ページ，225,402 文，37M バイト)とした．日本語の医学用語は，2003 年版と 2005 年版 MeSH シソーラス[1]に記載されている英語見出し語(78,194 語)のうち，医学用語辞書を持つ翻訳ソフトによって和訳できた 60,749 語とした．実際，上記の Web 文書集合から作成した実験データには，このうち 2,557 語が含まれていた.

4.2 実験データの作成方法

3.3 節で述べたように，階層構造を構築する場合には，抽象名詞と形容(動)詞の係りうけを対象としたが，本研究では，以下に示すように文中での共起や，助詞を介しての係りうけを対象とすることにより，階層関係以外の関係を取り出すことを試みた.

まず，文書集合中の文を KNP により構文解析し，「A<の>B」，「P<を>V」，「Q<が>V」，「R<

[1] The U.S. National Library of Medicine created, maintains, and provides the Medical Subject Headings (MeSH®) thesaurus.

に>V」,「S<は>V」のパターンにあてはまる係り受け関係を収集する．ここで，<X>は助詞，A，B，P，Q，R，S は名詞，V は動詞を表し，上記以外の助詞による係り受け関係は収集しない．たとえば，「太郎は光子から次郎が花子にダイヤの指輪を贈ったと聞いた.」という文からは，「太郎 **<は>** 聞いた」,「次郎 **<が>** 贈った」,「花子 **<に>** 贈った」,「指輪 **<を>** 贈った」,「ダイヤ **<の>** 指輪」の 5 つの関係が抽出できる. このように抽出した係り受け関係データから次の 3 つの実験データを作成した.

NN データ：各文について，上記のパターンに含まれる名詞 A，B，P，Q，R，S を集めた名詞間の共起関係に基づくデータである．例文からは，「ダイヤ，指輪，太郎，次郎，花子」が得られる．

NV データ：動詞 V を含むパターンについて，動詞 V と共起している名詞 P，Q，R，S をそれに続く助詞ごとに集めた名詞と動詞の係り受けに基づくデータである．データは助詞ごとに分割され、それぞれ「ヲ格データ」「ガ格データ」「ニ格データ」「未格データ」と呼ぶ．例文からは，動詞「贈る」について，ガ格は「次郎」，ヲ格は「指輪」，ニ格は「花子」が得られ，各々，動詞「贈る」のデータに追加する．

SO データ：同一文中で同じ動詞 V に係る主語(助詞<が>が続く名詞)と目的語(助詞<を>が続く名詞)をすべて集め，目的語となる名詞について主語を集めた, 主語と目的語の関係に基づくデータである．例文からは，「次郎<が>贈った」と「指輪<を>贈った」という関係から，名詞「次郎」のデータに名詞「指輪」を追加する．

NN データについては文数を次元数として名詞の出現状況を，NV データについては動詞の種類数を次元数として, 各名詞について動詞との出現状況をベクトル化した．SO データについては目的語として現れた名詞の種類数を次元数として, 主語として現れた各名詞について目的語として現れた名詞との出現状況をベクトル化した.

5　実験結果

図 1 から 3 に抽出された関連語集合の一部を示す．抽出された関連語集合のうち，3 つ以上の用語からなるものを次の選別の対象とした.

卵巣-脾臓-触診
データ-原因-うつ病-減少-血小板数-骨髄検査
新生児−動脈管開存症−壊死性腸炎
分泌−胃酸−胃粘膜−十二指腸潰瘍
皮膚-アトピー性皮膚炎-ヘルペスウイルス
　-抗ウイルス薬
皮膚-腹部-頚部-口腔-胸部
疲労-子宮筋-妊娠中毒症
水−酸素−水素−水素イオン
疲労-ストレス-十二指腸潰瘍

図 1. NN データから得た関連語集合の一部

アイスクリーム-チョコレート-ワイン（Ni）
出血-発熱-血尿-意識障害-めまい
　-高血圧（Ga）
受胎能力-アクリル樹脂-強心薬-人工血管（Ga）
水疱-鼓腸-腰痛-尺骨神経麻痺-脳内出血
　-閉塞性黄疸（Wo）
心疾患-冠動脈疾患-気管支炎-血栓性静脈炎
　-鼓腸-高尿酸血症-腰痛-尺骨神経麻痺
　-脳内出血-閉塞性黄疸（Wo）
咳-胎児-肺炎（Ha）

図 2. NV データから得た関連語集合の一部

潜伏期間−赤血球−肝細胞
雪−学校−ガス
変化−死−手足
病院−角膜混濁−トリアゾラム
反応−アポトーシス−損傷
研究−調査−味−米
環境−関心−水−肉−下痢
権利−資源−心−教育−森林伐採

図 3. SO データから得た関連語集合の一部

5.1　関連語集合の選別

非階層関係にある関連語集合を得るために，MeSH シソーラスを利用した．このシソーラスでは見出し語(医学用語)は人手により最上位で 15 のカテゴリに分類され，以下順次詳細

に分類されていく．なお，2 つ以上のカテゴリに分類されている見出し語もある．各見出し語は，この分類を示す識別番号を持ち，この番号を辿ることによって，用語間の階層的な意味関係を見ることができる．階層構造構築手法を拡張して得られた関連語集合にはさまざまな関係による関連語の集合が含まれている．集合を構成する用語が意味的階層関係(分類的関連)にあるならば，それらの用語はすべて MeSH シソーラスにおいて同じカテゴリに分類されているであろう．これに対して，非階層関係にある関連語集合を構成する用語は，複数のカテゴリに分布する傾向にあるだろう．これを基準として，関連語集合を選別することにした．表 1 に，SO データを除く各実験データについて，関連語集合を構成する用語がいくつの MeSH カテゴリに分布していたかを示す．

表 1.構成する用語のカテゴリ分布

データ		関連語集合数	カテゴリ分布 (カッコ内は%)	
			1 カテゴリ	2 カテゴリ
NN		594	24 (4)	169 (28)
NV	ヲ格	199	35 (18)	42 (21)
	ガ格	62	12 (19)	19 (31)
	ニ格	37	3 (8)	14 (38)
	未格	85	6 (7)	26 (31)

表 1 から，ガ格データから得られた関連語集合が他に比べて，MeSH シソーラスのカテゴリ分類と合致する率が高いことがわかる．これは，ガ格で表現される主格が他の格と比べ，動詞の特徴をより直接的に反映しているためと思われる．また，NN データと NV データについて比較すると，NV データから得た関連語集合のほうが NN データより MeSH シソーラスのカテゴリ分類と合致することがわかる．言い換えれば，助詞を介した係り受けによる NV データを用いるほうが，文中の共起を集めた NN データを用いるよりも，分類的関連を持つ関連語集合が多く得られたということである．これは，共起関係を制限するほど，分類的関連を持つ関連語集合を正確に抽出できることを示している．

なお，NV データから得られた関連語集合の中には，一つのカテゴリに用語すべてが分類されるべきであるにもかかわらず，2 つ以上のカテゴリに分類されているものが見受けられる．たとえば，図 2 にある「アイスクリーム － チョコレート – ワイン」は，「チョコレート」が「原料」として分類され，他二つは「食物」として分類されていた．しかし，「食べられるもの」としてみれば，チョコレートも「食物」のカテゴリにも分類されるべきである．本研究は，分類的関連を持つ関連語集合を得ることを目的とはしていないが，この事実は NV データを作成し，CSM に基づく手法を用いることにより，文書集合から分類的関連を持つ関連語集合を得られることを示している．

また，SO データについて，本研究では，主語と目的語が係る動詞を制限せずに，データを収集した．たとえば「人間<が>本<を>読む」と「ネズミ<が>本<を>齧る」があるとき，「読む」と「齧る」の差は考慮せず，共に「本を」と共起するものとして，「人間」と「ネズミ」の関係を推定した．同じ動詞に係る同じ目的語を持つ主語間の関係を推定することも考えられるが，そのような制限を設けなかったため，SO データは，NN データよりもさらに制限のないデータである．このことから，SO データからは非階層関係にある関連語集合が他と比べ，多く得られることが予測できた．実際，このデータから得られた関連語集合はシソーラスと合致しなかった．これは，発想支援の観点から見ると，作成した 3 つのデータのうち，SO データから最も発想を広げてくれる関連語集合を得られることを示唆している．

6　検索による検討

本研究では，医学分野の Web 文書集合から作成した実験データから，医学用語からなる関連語集合を抽出した．実際，実験で得た関連語集合は有益であっても医学に関するものであるため，発想支援には直接は利用できないかもしれない．これは，医学は定量的実験と経験に基づかなければならないからである．しかし，初めに述べたように，得られた関連語集合が Web 検索に役立ちうるかどうかを調査することにより，発想支援への適応性を検討することができる．もし関連語集合が有益なページに導いてくれるなら，人はその関連語集合を観て，想像を広げ，そのページに含まれる情報を推測できるかもしれない．

以上により，実際に Web 検索を行って，その結果を調査することにより，発想から創造までに至る過程を模倣することで，関連語集合が持つ発想支援での有効性を検証する．具体的には，選別して得た非階層関係にある関連語集合

を入力として，Google で Web 検索を行い，導かれたページから知識を得られるか，もしくは関連語集合そのものを知識と解釈できるかどうかを調査した．ここでは，関連語集合を用いた検索実験のうち，いくつかを解説する．

NN データから得られた実験結果に関して，図 1 の 1 行目にある「卵巣 - 脾臓 - 触診」は複数のカテゴリに分布しており，非階層的関係にあると判断されるが，これらを使って Google で検索すると，「卵巣や脾臓の病気は触診で診断される．」という情報を含む Web ページを上位で得る．このページの情報から，この関連語集合中の単語が因果関係にある単語群であることがわかり，また，これらの単語を見ることによって，病気と触診との関係へと発想が広がることが期待できる．同様に，非階層的関係である，図 2 の 2 行目にあるガ格のデータから得た「出血 - 発熱 - 血尿 - 意識障害 - めまい - 高血圧」を用いると，「これらは薬の副作用である．」という情報を得られた．

他の例として，図 1 の 2 行目にある「データ - 原因 - うつ病 - 減少 - 血小板数 - 骨髄検査」という関連語集合に含まれる単語は複数のカテゴリに分布するので，分類的関連ではない，つまり，非階層関係にある関連語集合である．実際，この集合を用いて，検索すると，「骨髄の疾患はうつ病や血小板の減少を引き起こす原因となるので，骨髄検査は必要である．」という文を含む Web ページが得られ，この関連語集合は知識として解釈できる．

SO データについては，図 3 の 1 行目にある非階層的関係「潜伏期間 – 赤血球 – 肝細胞」を使って検索すると，「マラリア」に関連する，「マラリアの潜伏期間中，患者は肝臓の障害を引き起こす」という専門家の知識を得た．

7　まとめ

本稿では，非階層関係にある関連語集合を抽出することを試みた．構文解析した文書集合から，共起関係に基づいた実験データを作成した．手法としては，自動階層構築方法を応用した．実験では，医学分野の Web 文書集合から医学用語からなる関連語集合を抽出し，得られた集合から，MeSH シソーラスのカテゴリ分類を利用して，非階層関係にある関連語集合を選別した．そして，非階層関係にある関連語集合が，ユーザを有益な情報を含むページに導きうるかどうか，また，集合中の単語のセットが，医学的な知識と解釈できるかどうかを実際に Web 検索することで検証した．実験の結果，自動階層構築方法を拡張することにより，階層関係だけでなく，因果関係や連想関係を含む非階層関係にある関連語集合をも抽出できることがわかった．その調査結果から，本手法が発想支援に応用可能であると考えている．

今回の実験では，抽出した関連語集合を用いて検索した結果を分析することで，発想支援への応用可能性を間接的に推測した．この結果を踏まえ，発想支援での有用性を実験的に検証することが今後の課題である．

参考文献

Hagita, N. and Sawaki, M. Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning, In *Proceedings of the SPIE – The International Society for Optical Engineering*, 2442: pp. 236-244, 1995.

Harakawa, J., Nagai, Y. and Taura, T. Study on Conceptual Synthesis in Design Creation -Role of Thematic Relation in Creativity-. *2005IDC International Design Congress IASDR,* on CD-ROM, 2005.

Kanzaki, K., Yamamoto, E., Ma, Q. and Isahara, H. Construction of an objective hierarchy of abstract concepts via directional similarity, In *Proceedings of the 20th Coling*, Vol.2, pp. 1147-1153, 2004.

Taura, T., and Nagai, Y. Primitives and principles of synthetic process for creative design -Taxonomical relation and thematic relation. *Computational and Cognitive Models of Creative Design VI*, Gero, S. J., and Maher, M. L. (eds.), pp.177-194, 2005.

Wisniewski, E. J., and Bassok. M. What makes a man similar to a tie? *Cognitive Psychology*, 39: 208-238, 1999.

Yamamoto, E., Kanzaki, K. and Isahara, H. Extraction of hierarchies based on inclusion of co-occurring words with frequency information, In *Proceedings of the 19th IJCAI,* pp. 1166-1172, 2005.

山本英子，神崎享子，井佐原均．出現状況の包含関係による語彙の階層構造の構築，情報処理学会論文誌，Vol.47，No.6，pp. 1872-1883, 2006.