E_003

検索に有効な知識の自動獲得 —質問拡張を超えて—

Automatic Term Acquisition Effective for Information Retrieval - Beyond the Query Expansion -

山本 英子† 井佐原 均†
Eiko Yamamoto Hitoshi Isahara

1. はじめに

近年，知りたいことがあれば，インターネットで検索するのが日常的なこととなっている．しかし，検索エンジンに適切なキーワードを入力することは時に非常に困難である．膨大なページが検索された場合に，上位に表示されるページが適切な情報を示しているとは限らないし，キーワードを不用意に追加すれば，意図と違うページが検索されたり，まったく検索結果が得られなかったりする．多くの場合，検索する人(ユーザ)は，対象を熟知しているわけではないので，得られた結果，あるいは必要な結果が得られないという事実が正しいのかどうかを判断できない．

このような状況を克服するため，さまざまな試みがなされてきた．適合フィードバックによる類似文書検索や，検索結果からのキーワード抽出・提示などの有効性が示されているが，これらの手法はユーザが自分の希望に近いページ群に到達してからは有効であるが，膨大なページが表示されている場合には，たとえば適合フィードバックのためにユーザのニーズに近いページを見つけることは困難である．このような場合にユーザの意図に沿った，あるいはユーザにとって有益なページに到達するためには，検索結果を適切に制限するようなキーワードを選択することが必要となる．つまり，ある一まとまりのページだけを検索するキーワード群，あるいは現在の検索結果を適切に限定する追加のキーワードをユーザが知ることが必要である．

このため，入力した検索キーワードに対し、既存のシソーラスに基づく同義語や上位語・下位語を提示することによる検索支援機能を検索エンジンに追加するという試みも行われているが，より的確な検索支援のためには，このような語を提示するだけでは十分ではなく，検索する文書集合において，因果関係や連想関係など，何らかの関連を持って出現する語(関連語)の獲得・提示が必要である．

本研究では，Web 文書集合から共起する語の包含関係に基づいて関連語を抽出する方法を提案し，得られた関連語セットを用いて実際に検索を行い，このような関連語セットがユーザを有益なページに導きうることを検証する．

2. 自動階層構築方法

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[6,7]．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) [2]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．我々は形容詞概念に対応する抽象名詞の階層構造を自動構築し，既存の辞書の階層構造と比較することによって，この手法の有用性を示した．ここで用いたデータは，形容(動)詞と抽象名詞の修飾関係を収集したデータで，抽象名詞は共起する形容詞・形容動詞の概念名(たとえば，「赤い」に対する「色」)を表すものと定義され，抽象名詞の概念を形容(動)詞の概念の上位概念とみなしている[3].

† 独立行政法人 情報通信研究機構, NICT

2.1 出現状況のベクトル表現

この研究では，二つのベクトル間の包含関係を下記の式で定義される補完類似度(CSM)を用いて，数値化する．CSM は非対称性を持つ尺度である．我々は，対象とする文書集合中の共起語の出現状況をベクトルで表し，より多くの共起語が出現する単語はより広い意味を持つといえることから，二語間の CSM の値をその文書集合における二語間の上位下位関係を表す関連度とした．

単語(ここでは抽象名詞) w_i が 1 から n までのどの形容(動)詞と共起するかを 1，0 で表現したベクトルを $Vi = (v_{i1}, ..., v_{in})$，単語 w_j についてのベクトルを $Vj = (v_{j1}, ..., v_{jn})$ としたとき，$CSM(Vi, Vj)$ は次のように定義される．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)詞の数を表す．n は形容(動)詞の総数で，n = a+b+c+d である．

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

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

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

2.2 階層構造の構築

CSM によって上下関係にあると推定された文書集合中の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 値が閾値以上である単語対のみを使う．これは，CSM 値が低ければ，その単語間の関係は信頼性が低いためである．ここでは，例を用いて構築手法を説明する．詳細は文献[6,7]に譲る．

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．その後，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，得られる階層構造は Z→A→B→D→E であるが，これは Z→A→B→C→D→E に含まれる構造であるため，階層構造のリストから削除する．このように，我々はより多い単語で構成される階層構造のみを利用することにした．

3. 単語セットの抽出

3.1 自動階層構築方法の応用

前節に述べたように，我々はこれまで抽象名詞と係り受け関係にある形容(動)詞を共起語として用い，単語対が持つ CSM 値を階層的な意味関係における関連度として扱ってきた．CSM 値は共起語の出現状況の類似性に基づく関連度であり，さまざまな関係での共起語の出現状況を用いた場合，階層構造だけではなく，因果関係などの，別の意味的関係を持つ単語セットが得られると考えられる．また，この手法は共起語に基づく出現状況だけを情報源とするため，得られる単語セットは対象とした文書集合に依存する．したがって，対象を分野特化した文書集合にすることで，その分野固有の知識を抽出しうる．このような考えの下，本研究では，自動階層構築方法を相互に関連を持つ単語セットの抽出に応用することを試みた．

3.2 医学用語間における単語セットの抽出

本研究で用いた手法は文書から知識を自動獲得する手法であり，文書の分野に依存しない．先で述べるように，検索に有効な関連語の選択の過程で用語の階層構造(シソーラス)が必要であるが，医学用語のシソーラスとして，MeSH シソーラス[5]が利用可能である．また近年，医学分野において，症状から病名を推測するための知識収集や Web 上の医療に関する情報取得支援といった需要に応えるために，医学分野の様々な言語資源を用いて，医学用語間における関係抽出や用語辞書の構築などを行う研究が進められている[1,4].

このような状況から，医学分野を対象に Web 文書集合から，検索支援に役立つ知識となる単語セット(検索キーワード群)の抽出を行い，Web 検索の支援を試みた．

3.3 実験データの作成方法

本研究では，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)からその分野固有の知識を獲得することを試みた．まず，文書集合中の文を KNP により構文解析し，「の，を，が，に，は」の 5 つの助詞による係り受け関係，すなわち，各文から「A<の>B」，「P<を>V」，「Q<が>V」，「R<に>V」，「S<は>V」のパターンにあてはまる関係を収集する．ここで，<X>は助詞，A，B，P，Q，R，S は名詞，V は動詞を表す．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．」という文からは次の 5 つの関係が抽出できる．

- **「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」**

これら以外の助詞による係り受け関係は収集しない．したがって，たとえば「光子<から>聞いた」は収集しない．このような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**：名詞間の共起関係に基づくデータ
  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 を集めたデータである．上の例文に対しては「太郎，次郎，花子，指輪，ダイヤ」というデータが得られる．
- **NV データ**：名詞と動詞の係り受けに基づくデータ
  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データである．例文には「次郎<が>贈った」という関係があるので，この動詞の原形「贈る」についてのガ格データに「次郎」が追加される．同様に，「贈る」のヲ格データには「指輪」が追加される．
- **SO データ**：主語と目的語の関係に基づくデータ
  同一文中で同じ動詞 V に係る主語(主格<が>が続く名詞)と目的語(目的格<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分類したデータである．例文には「次郎<が>贈った」と「指輪<を>贈った」の二つの関係があるので，「指輪」についてのデータに「次郎」が追加される．

これらの実験データにおける名詞の出現状況をそれぞれ図 1 に示すベクトルで表現する．ここでは，NN データと NV データについては，名詞ごとにベクトル化し，SO データについては，主語となる名詞ごとにベクトル化した．各ベクトルに対し，CSM を用いて関連度を測定した．

NN データ

	1.....n個の文.....n
名詞	11110010110110101111.....011

NV データ

	1.....n種類の動詞.....n
名詞	11110010110110101111.....011

SO データ

	1.....n種類の目的語.....n
主語	11110010110110101111.....011

図 1：実験データが表す出現状況のベクトル化

4. 得られた単語セットと医学的知識

2003 年版と 2005 年版 MeSH シソーラスに記載されている見出し語(78,194 語)を，医学用語辞書を持つ翻訳ソフトによって和訳し，60,749 語を得た．これらを日本語の医学用語とする．実験データには，このうち 2,557 語が含まれていた．これらの医学用語について，作成した 3 種類の実験データから，CSM に基づく自動階層構築方法を用いて，相互に関連を持つ用語のセットを抽出した．抽出したセットのうち，3 つ以上の用語からなるものを実験では利用した．本節では，抽出した単語セットを MeSH シソーラスと

比較し，CSMに基づく手法の知識獲得への応用可能性を考察する．

図2にMeSHシソーラスに含まれる見出し語の木構造(MeSH Tree)の一部を示す．見出し語は人手により最上位で15のカテゴリに分類される．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．各見出し語は識別番号を持ち，先頭のアルファベットがその語が分類されるカテゴリを示す．たとえば，「親指」は識別番号「A01.378.800.667.430.705」を持ち，カテゴリ"A"「人体・動植物の解剖学的構造 (Anatomy)」に分類される．この識別番号によって「親指」の上位語を辿ることができる．

A01	body region	（身体の部位）
	:	
A01.378	extremity	（四肢）
A01.378	limb	（手足）
	:	
A01.378.800	upper limb	（上肢）
	:	
A01.378.800.667	hand	（手）
A01.378.800.667.430	finger	（指）
A01.378.800.667.430.705	thumb	（親指）
	:	

図2：MeSH Treeの一部 (()内は日本語訳)

表2：単語セットを構成する用語のカテゴリ分布[1]

データの種類	NN	NV			
		ヲ格	ガ格	ニ格	未格[2]
獲得知識数	594	199	62	37	85
用語のカテゴリ分布(割合[%])					
1つ	42	40	14	7	7
	(7)	(20)	(23)	(19)	(8)
2つ	148	55	24	13	28
	(25)	(28)	(39)	(35)	(33)
3つ	120	47	8	3	11
	(20)	(24)	(13)	(8)	(13)
合計(割合[%])	310	142	46	23	46
	(52)	(71)	(74)	(62)	(54)

もし得られた単語セットが医学分野における用語間の階層的な意味関係を表す知識であるならば，その階層を構成する用語は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．表2にNNデータとNVデータから得られた各単語セットを構成する用語がMeSHシソーラスにおける見出し語のいくつのカテゴリに分布したかを示す．

なお，ある単語がMeSHシソーラスにおいて，複数のカテゴリに分類されている場合がある．たとえば，NNデータから「木－森－オラウータン」という単語セットを得るが，「木」は二つのカテゴリに分類されている．「森」は「木」と同じカテゴリに分類され，「オラウータン」は「木」のもう一つのカテゴリと同じカテゴリに分類されている．このような場合，我々は「森」と「オラウータン」は「木」を介して関係があると考え，この単語セットはMeSHシソーラスと一致する知識として扱った．

表2から，構成する用語が3つ以下のカテゴリに分布する単語セットの割合は52%から74%を占めることがわかった．また，「ガ格」のデータから獲得された単語セットがMeSHシソーラスのカテゴリ分類にもっとも合致しているという結果を得た．これは，「ガ格」で表現される主格が他の格と比べて多義性が少ないためと思われる．

5. Web検索での有効性

MeSHシソーラスと類似するものは階層関係にある語であり，自然言語処理には有用でも，検索の場合は単なる検索語の拡張の域を出ない．一方，MeSHシソーラスと一致しないものは，階層関係以外の関係(因果関係など)で関連している語であり，このような単語セットを用いることで，検索結果を適切に制約することが期待できる．

本節では，獲得した単語セットが検索キーワードとして有効であるかどうか，ユーザを有益なページに誘導しうるかどうかを検証するために，得られた単語セットを用いて，実際にWebページを検索した．図3と図4にNNデータとSOデータから得られた単語セットの一部を示す．これらの図に含まれる単語セットを用いた検索結果のいくつかを解説し，有用性を考察する．

NNデータから得られた実験結果に関して，図3の1行目にある「卵巣 - 脾臓 - 触診」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ページが約500ページ(2006年7月現在)得られる．これらの用語のうち，「触診」はMeSHシソーラスでは他の2つとは異なるカテゴリに分類されているため，MeSHシソーラス(を用いた質問拡張)からは，これら3つの用語間の関係を得ることはできない．本手法によって，有益なページを検索できる医学的知識を示す単語セットが獲得できた．また，「卵巣 - 脾臓」だけを使った場合には，5万件近い(同)ページが検索され，その中から，たとえば適合フィードバックに必要な，ユーザのニーズに近いページを選ぶことは困難である．

SOデータについては，図4の1行目にある「潜伏期間 – 赤血球 – 肝細胞」を使って検索すると，「マラリア」に関連するページを得た．「潜伏期間」と「肝細胞」の2つだけを使って検索すると，「肝臓の障害」に関連する多くのページを得る．このように，提案手法によって得られた単語セットを検索キーワードとして使うと，「マラリアの潜伏期間中，患者は肝臓の障害を引き起こす」ということを知っている専門家のように，有益なページを検索できた．このように，単語セットを使って，ユーザはより的確で詳細な情報を検索エンジンに入力でき，有益な結果に到達できる．検索支援として，これらの単語セットをユーザに提示することで，我々が「マラリア」に関連するページに導かれたように，ユーザが興味深い情報を含むページを見つけることに役立てられるだろう．

[1] SOデータについては表に示していない．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，SOデータは「本」について収集され，CSMによって「人間」と「ネズミ」の関係を推定する．ここではMeSHシソーラスと一致しない単語セットを積極的に取り出すため，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．実際，SOデータから得られた単語セットでシソーラスと一致する階層構造は非常に稀であった．

[2] 「未格」は格助詞「<は>」によるデータを表す．

卵巣-脾臓-触診
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3：NNデータから得た結果の一部

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4：SOデータから得た結果の一部

6. まとめ

本稿では，検索支援に有効となる単語セットの抽出を目指した．これまでに提案してきた語彙の自動階層構築方法を応用して医学分野の Web 文書集合から抽出した単語セットが医学的な知識と解釈できるかどうかを考察し，その知識がユーザを適切な，または興味深いページに導きうるかどうかを検証した．その結果，自動階層構築方法は階層関係だけでなく，その他の意味的関係を持つ知識をも獲得でき，検索支援に有効であることがわかった．また，本手法で獲得する知識は対象となる文書集合に依存する知識であり，たとえば本論文のように医学文書を対象に単語セットを収集すれば，医学系の知識を獲得できる．本手法は統計的手法であるため，このように，対象となる文書集合の特徴に柔軟に対応でき，その特徴を捉えた知識を獲得できる．今後の課題のひとつは，知識を獲得するにはどのような文書集合が適切であるかを判定することであり，さまざまな文書集合に本手法を適用することによって，分析を進める予定である．

参考文献

[1] 荒牧英治, 今井健, 梶野正幸, 美代賢吾, 大江和彦. メタ関係を利用したテキストからの人体部位関係の抽出, 言語処理学会第 12 回年次大会発表論文集, pp.508-511, 2006.

[2] 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.

[3] 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.

[4] 木村俊也, 中川晋一, 三角真, 山岡克式, 酒井善則, 島津明. Web 上のがん情報取得のためのがん用語辞書の作成, 言語処理学会第 12 回年次大会発表論文集, pp. 173-176, 2006.

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

[6] 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.

[7] 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.