2007－ICS－148 (4)
2007／7／4

英和辞典を用いた単語階層構造の動的構築

Dynamic Construction of Word Hierarchy using English-Japanese Dictionary

下司義寛[1*] 廣川佐千男[2]

[1] 九州大学システム情報科学府情報理学専攻
[1] Department of Informatics, Graduate School of Information Science and Electrical Engineering, Kyushu University
[2] 九州大学情報基盤研究開発センター
[2] Research Institute for Information Technology, Kyushu University

Abstract:
This paper proposes a method which constructs a hierarchy of words from given a set of documents automatically and dynamically. The hierarchy of the words is constructed as a hypernym relation that is defined by the document frequencies and co-occurrence probability of words. The hierarchy is obtained not only to the whole set of documents, but also to any subset of documents. A typical example of such documents is the search results of a keyword. The hierarchy obtained to this set is the hierarchy of related words of the keywords.
Empirical evaluations are conducted for the word hierarchy derived from "Eijiro", an English-Japanese dictionary which contains 1,648,628 descriptions of words. The hierarchies are compared with the proposed method, Niwa's method and Shrinivasan's method with respect to coverage and granularity.

1 はじめに

電子化されたインターネット上の文書の量は爆発的に増え続けている．大量の文書群から必要とするものだけを閲覧することは困難になっている．必要とする情報を効率よく検索する技術が必要とされている．検索結果が多すぎる時には絞り込みが必要であり，結果に必要とする文書が含まれていなければ異なる単語による検索が必要である．いずれの場合もユーザの必要とする情報を逃さないように行わなければならない．検索拡張は新しい検索クエリ候補を提示する方法であり，同義語，類義語，狭義語などの元のクエリと関係のある単語を提示することで次に検索すべき方向を示す．検索拡張が様々なユーザの要求に対応するために，人間が本を読むことで知識を蓄えるように，システムもまた専門用語集やシソーラスを用いて知識を強化することが必要である．しかし，人手による辞書の構築やメンテナンスはコストがかかる．また，WWW 上では日々新しい概念を表す単語が出現することを考慮すると，自動化あるいは効率化は必須である．

*連絡先：九州大学システム情報科学府情報理学専攻
〒 812-0053 福岡県福岡市東区箱崎 6-10-1
E-mail: y-shimo@i.kyushu-u.ac.jp

本稿では，与えられた文書集合からユーザの検索クエリに合わせて動的に特徴語とそれらの上位下位の関係を抽出し，グラフによって可視化する手法を提案する．

2 関連研究

Hearst[2] は辞書からテンプレートを使って上位語を抽出する方法を示した．しかし，テンプレートや自然言語処理の技術を用いると，特定の言語に依存することになる．例えば英語のテンプレートを使った方法は日本語の文書には適用できない．言語独立性をシステムに持たせるために，本研究では単語の関係抽出には統計的手法を用いる．統計的手法を用いる単語階層構築法として丹羽等の研究 [4] や，Srinivasan の研究 [6] がある．

2.1 丹羽等の方法

丹羽等 [4] は単語間の上位下位関係を文書頻度と条件付確率を用いて以下のように定義した．文書集合 D において二つの単語 u が単語 v の上位であることは単語 u, v が以下の二つの条件を満たすことである．

$$df(u,D) > df(v,D). \quad (1)$$

$$P(u|v) = max\{P(w|v)|w \in d, d \in D\}. \quad (2)$$

ここで, $df(w,D)$ は, 文書集合 D において単語 w を含む文書の数, すなわち, $df(w,D) = |\{d \in D|w \in d\}|$ を表す. また, $P(w\ v)$ は, 条件確率, すなわち, $P(w|v) = |\{d \in D|w \in d, v \in d\}|/|\{d \in D|v \in d\}|$ を表す.

(1) は, u が v よりも出現頻度の観点で一般的であることを表す. (2) は, 出現する条件付確率が最大になる u についてだけ, 上位になることを意味する. 単語を節, 上位下位関係を有向枝とする有向グラフとして表すと, 丹羽等の方法では上位語が複数ある場合全ての上位語に枝を引く.

2.2 Srinivasan の方法

Srinivasan[6] は自動的にシソーラスを構築する方法を提案した. まず, 文書頻度を用いて単語の大域的な階層をつくり, 次に異なる階層間で上位下位関係の計算を行う.

まず, 全単語集合 W を重複のない K 個の部分集合に分割する. 分割方法は単語の文書頻度を用いる. 全単語集合での最大の文書頻度 max_{df}, 最小の文書頻度 min_{df} を用いて部分集合 W_i の文書頻度の幅 $df_i - df_{i+1}$ を決定する.

$$max_{df} = max\{df(w,D)|w \in W\} \quad (3)$$

$$min_{df} = min\{df(w,D)|w \in W\} \quad (4)$$

$$R = (max_{df} - min_{df})/K \quad (5)$$

$$df_i = max_{df} - R \times i \quad (6)$$

$$W_i = \{w \in W|df_{i+1} \leq df(w,D) < df_i\} \quad (7)$$

$$W = \cup_{i=1}^{K} W_i \quad (8)$$

筆者等はまず, $K = 10$ として, Srinivasan の方法に従い予備的実験を行った. Srinivasan の方法を実際の自然言語文書に適用した場合, $K = 10$ であるにもかかわらず 2〜3 個の部分集合にしか分割できなかった. このことは自然言語文書における文書頻度が冪分布に従うことに起因すると考えられる. そこで, 本稿では頻度による分割を, 線形ではなく対数尺度とした.

$$R = (\log(max_{df}) - \log(min_{df}))/K \quad (9)$$

$$df_i = e^{max_{df} - R \times i} \quad (10)$$

上位語抽出には次の定義を用いる. 単語 u, v の類似度は次の式 $Cohesion$ で計算する. ここで, $df(u*v,D)$ は D 中の文書で u, v の両方を含む文書の個数である.

$$Cohesion(u,v) = \frac{df(u*v,D)}{\sqrt{df(u,D) \times df(v,D)}} \quad (11)$$

$u \in W_i, v \in W_{i+1}$ が次の条件を満たす時, u は v の上位であるとする.

$$Cohesion(u,v) = \max_{w \in W_{i+1}} Cohesion(u,w) \quad (12)$$

u は隣接する下位階層の要素で類似度が最大になる要素 v の上位語となる. ただし, 下位の単語が隣接する階層になければダミーを下位の階層に作り, ダミーと上位下位関係にある単語を元の単語と上位下位関係にある単語として扱う.

3 概念グラフ

単語を節として上位下位の関係を枝とする有向グラフとして表現することで, 単語の階層構造を直感的に把握することができる. 前節における丹羽や Srinivasan らの手法による階層構造も可視化しシソーラスや検索キーワードの推薦に利用することができる. 本稿では, 筆者等が [5] で導入した概念グラフによる階層構造と, 丹羽や Srinivasan による階層構造の比較を行なう.

例えば, サンテミリオンとは白ワインの銘柄の 1 つである. 白ワインはワインである. これら 3 つの階層構造は' ワイン'-' 白ワイン'-' サンテミリオン' となり左から順により一般的な単語である (図 1).

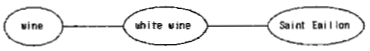

図 1: ワイン–白ワイン–サンテミリオン

上位下位の関係には 2 つの観点が必要と考える. 1 つはサンテミリオンの上位がワインと白ワインであるということである. もう 1 つはサンテミリオンの隣接する上位はワインではなく白ワインであるということである. 本稿では上位下位の関係と隣接する上位の関係二つの観点で捉える. 可視化する際には図 1 のように隣接する上位にのみ枝を引く.

3.1 特徴語

同じ単語でも, それが使われる環境によって意味がかわり,2 つの単語の間の上位下位関係は異なる. そこで本研究では全ての単語の階層構造を扱わず一部の単語の階層構造を構築する. 一部とは特定の文書群の特徴を表す単語である. ユーザの必要とする分野に特徴的

な単語の関係を提示することが有用である．本研究ではユーザの必要とする分野をユーザの入力キーワード q を含む文書集合とする．これによりユーザの入力にあわせた動的な特徴語抽出が可能となる．以下では，単語 w を含む文書集合を $D(w)$ と表記する．

単語 w が

q をユーザの入力キーワード，$D(q)$ を q を含む文書集合とするとき，w が $D(q)$ の特徴語であるとは，$df(w, D(q))/df(w, U) > 0.5$ という条件を満たすこととする．ここで，$df(w, D(q))$ は語 w の $D(q)$ での文書頻度，$df(w, U)$ は w の全体集合 U での文書頻度である．すなわち，単語 w が出現する過半数が文書集合 $D(q)$ に含まれているとき，w は $D(q)$ の特徴語である．

3.2 上位下位関係

本節では単語間の上位下位関係の抽出法を説明する．

単語 u が v の上位であることを，より一般的な単語であること (式 13) と v から見た u の関連が強いこと (式 14) で定義する．

$$df(u) > df(v) \tag{13}$$

$$P(u|v) > 0.5 \tag{14}$$

ただし，本稿では単語間の関係は前節の特徴語抽出と同様に，特定の分野，文書集合に着目して行う．そのため，式 (13), (14) はユーザの入力キーワード q を含む文書集合 $D(q)$ での文書頻度を用いて次の様に書き換える．

$$df(u, D(q)) > df(v, D(q)) \tag{15}$$

$$P(u|v) = df(u * v, D(q))/df(v, D(q)) > 0.5 \tag{16}$$

v よりも u が $D(q)$ でより多く出現し，v が出現する過半数で u も出現しているとき，u は v の上位である．また，u を v の上位語と呼ぶ．

式 (15), (16) を満たす全ての特徴語のペアを上位下位関係にある単語のペアとして抽出する．

3.3 隣接上位関係

前節の上位下位関係の定義により抽出された単語 w の上位語から隣接する上位語を抽出する．これにより，単語の上位下位の階層構造を構築する．

隣接上位関係の定式化のために単語 v の上位語集合 $UP(v)$ と隣接上位語集合 $DUP(v)$ を定義する．

$$UP(v) = \{u \in D(q) | u \text{ は } v \text{ の上位語 }\} \tag{17}$$

$$DUP(v) = \{u | \forall w \in UP(v), u \notin UP(w)\} \tag{18}$$

v の上位語 u が u 以外の上位語の上位ではないとき $u \in DUP(v)$ であり，u は v の隣接上位である．また，u を v の隣接上位語と呼ぶ．

全ての特徴語間の隣接上位関係について下位の単語から上位の単語に向いた枝を引き有向グラフとして可視化する．

4 英和辞典からの単語階層構築

4.1 実験データ

本節では，英和辞典の英辞郎[1]を文書集合として利用した．英辞郎は 1,648,628 件の項目からなり，1 項目は単語や熟語とその説明文から成る．1 項目を 1 文書として日本語と英語の混じったコーパスから単語の階層構造を構築するシステムを実装した．検索エンジンには国立情報学研究所の高野等による汎用連想検索エンジン GETA[2]を利用した．

表 1: 英辞郎のデータ

項目数 (文書数)	1,648,628
単語異なり数	986,410
単語数	8,644,997
全データサイズ (バイト)	112,624,437
1 文書あたりのデータサイズ (バイト)	68.3

1 項目の平均の単語数は 68.3 バイトであり，日本語で 34 文字と短い文書からでも単語の関係を抽出できることである．

4.2 ワインのグラフ

本節では，具体的な概念グラフを用いて，新しい知識や検索キーワード発見が可能かという観点で定性的評価を行う．

検索キーワードとして wine という単語を選択した．wine という単語を英辞郎で引くと以下のような説明がある．

◇wine

[名] ワイン，果実酒 (かじつしゅ)，ブドウ酒 ◇語源はラテン語「vinum」.BC4000-3000 のエジプトが最初に作った．BC1000 頃のギリシャでは水割りで飲んだ．ストレートで飲んだのは野蛮人とされた．

[1] http://www.alc.co.jp/
[2] http://geta.nii.ac.jp/

- My father likes wine. 父はワインが好きだ.
- A glass of wine a day is good for your health. 一日一杯のワインは健康に良い.

[自動] ワインを飲む

[他動] ～をブドウ酒でもてなす

システムに wine という単語を入力すると表 2 のような単語が wine を含む文書の特徴語として抽出される. 次に特徴語間の上位下位関係を抽出しグラフを出力する. 図 2 は wine という単語を含む文書群の特徴語の概念グラフの一部である.

単語	文書頻度
wine	391
白ワイン	233
醸造	148
シャトー	134
名産	68
名産地	68
ボルドー	48
ブルゴーニュ	40
ドイツ産白ワイン	19
ローヌ	18
ピノ	17
riesling	17
ドイツ高級ワイン法定地域	14
ロワール	12

表 2: 「wine」の特徴語

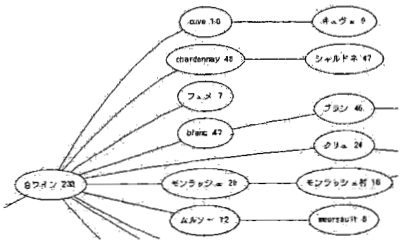

図 2: ワインについての単語階層構造

白ワインの下位の単語に産地や銘柄などの名前が出現している. ユーザがワインに興味を持っているがワインというクエリしか思い浮かばずよい検索結果を得られなかった場合, このグラフを見ることでワインの代わりとして, “ワイン シャルドネ”,” 白ワイン モンラッシュ”, “キュヴェ” などの新しいクエリを見つけることができる.

5 抽出された隣接上位下位関係の評価

5.1 house の関連語についての定性的評価

次節で, 定量的な比較を行う前に, 提案手法と関連研究が抽出する上位下位関係を比較する. 図 3 は house という単語についての単語の階層構造である. 上位下位の関係を表す枝にビット列でラベル付けを行った. 各ビットはどの手法で上位下位関係が抽出されたかを 0,1 で表す. 右から 1 番目は提案手法, 2 番目は丹羽等の方法, 3 番目は Srinivasan の方法によって上位下位関係が抽出されたことを表している.

例えば, “house-貴族院 (House of Lords)” の関係は提案手法と丹羽等の方法で抽出されたが, Srinivasan の方法では抽出されなかった.

全部で 8 種類のラベルが付けられ, それぞれの数は表 3 のようになった. 表から丹羽等の方法はもっとも多くの関係を抽出している.

“上訴委員会-上訴院判事” の関係は丹羽等の方法でのみ抽出されているが, 提案手法中の “上訴委員会-常任判事” と” 常任判事-上訴院判事” の 2 つの関係から自明ではない “上訴委員会-上訴院判事” の関係を理解することができる. つまり, 丹羽等の方法で抽出される関係は, 提案手法で抽出される関係の推移的閉包として表現できている. 言い替えると, 提案手法はより細かい階層を抽出しているといえる.

表 3: 3 手法ごとの隣接上位下位関係の比較

ビット列	Shrinivasan	丹羽	提案手法	枝数
001			√	1
010		√		11
011		√	√	4
100	√			0
101	√		√	4
110	√	√		4
111	√	√	√	0
総数	8	19	9	24

5.2 定量的評価

本節では提案手法と丹羽や Srinivasan の方法との比較を行う. 評価指標は単語階層構造の単語の再現性, 上位下位関係の粒度の 2 つを用いた.

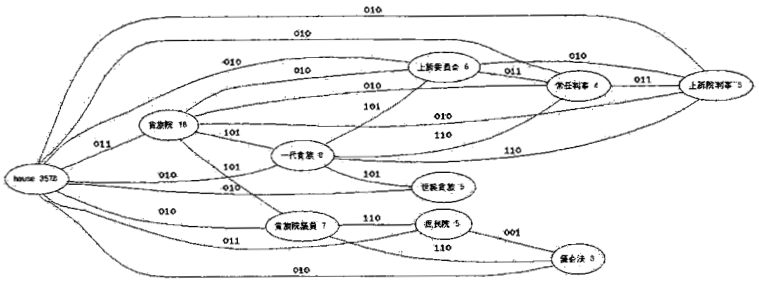

図 3: 上位下位関係の比較

相互被覆率による分類能力の評価

単語階層構造がどれだけ似ているかを単語の集合としてどれだけ共通なものがあるかという観点で比較する方法が知られている [1, 3]. 単語の階層中の各単語より下位の単語集合を考え, 単語の集合の集合を作る. 二つの階層構造 G_1,G_2 からできる二つの「単語の集合の集合」 S_1,S_2 がどれだけ共通な要素を持つかでどちらが階層構造として表現能力あるいは, 分類能力が高いかを比べる.

W を単語の集合,G を概念グラフ,$w \in W$ を単語とする. このとき,G における w の下位語全体を $LOW(w)$ で表わすと, $O(G) = \{LOW(w)|w \in W\}$ として G に対するオントロジーは定式化できる。多数決原理と丹羽らの方式の比較を行なうため 100 通りの単語ごとに得られる 2 通りのグラフ G_1, G_2 を構成し, $O(G_1)$ と $O(G_2)$ を比較する. ただし, G_1, G_2 を概念グラフとするとき, そのオントロジーを定める単語集合 W は 3.1 節で抽出された特徴語集合とする.

一般にふたつのシソーラスやオントロジーの比較において, 使用される単語集合が共通ではない場合を想定して実験がおこなわれる. しかし, 本研究では異なる手法でも, 同じ特徴語集合を関連語候補として使用しているため, 共通の特徴語による下位語集合の共通部分を調べるだけでよい.

具体的には,$S_1 = O(G_1), S_2 = O(G_2)$ として x 座標に $\frac{|S_1 \cap S_2|}{|S_2|}$,y 座標に $\frac{|S_1 \cap S_2|}{|S_1|}$ をプロットすることで, G_1 が G_2 の単語分類能力をどれだけ再現しているか, また逆に G_2 が G_1 の単語分類能力をどれだけ再現しているかを評価する.

G_1 を提案手法によるグラフ,G_2 を丹羽等の方法によるグラフとして, $(\frac{|S_1 \cap S_2|}{|S_2|}, \frac{|S_1 \cap S_2|}{|S_1|})$ をプロットしたものが図 4 である. 全体が $y = x$ の直線よりも下に現れていることがわかる. $\frac{|S_1 \cap S_2|}{|S_2|}, \frac{|S_1 \cap S_1|}{|S_1|}$ それぞれの平均が (0.60,0.48) である. これは, 提案手法は丹羽等の方法の単語分類能力の 6 割を再現し, 丹羽等の方法は提案手法の単語分類能力の約 5 割を再現していることを示している.

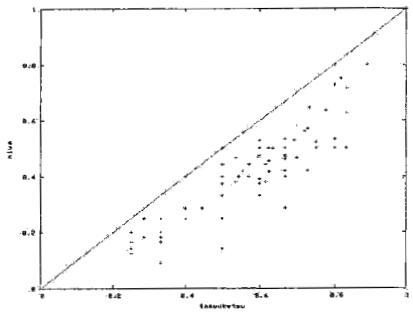

図 4: 提案手法と丹羽等の方法の相互被覆率

このことから, 提案手法と丹羽等の方法の単語分類能力は半分ほど類似し, 再現性は提案手法のほうが若干優れていることが解る.

G_1 を提案手法によるグラフ,G_2 を Srinivasan の方法によるグラフとして, $(\frac{|S_1 \cap S_2|}{|S_2|}, \frac{|S_1 \cap S_2|}{|S_1|})$ をプロットしたものが図 5 である. $y = x$ 付近に分布しているものと, それよりも下方に分布しているものに分けることができるだろうか. $\frac{|S_1 \cap S_2|}{|S_2|}, \frac{|S_1 \cap S_2|}{|S_1|}$ それぞれの平均が (0.35,0.30) である. これは, 提案手法と「Srinivasan の方法」はお互いの単語分類能力を 3 割ほど再現していることを示している.

提案手法と Srinivasan の方法どちらがより優れているかは単語の分類という観点からは判断できない.

5.3 粒度

単語の階層構造における上位下位の関係がどれだけ正確であるかを人手で評価することは困難である. そこで, もっともらしさの指標として粒度という評価値を

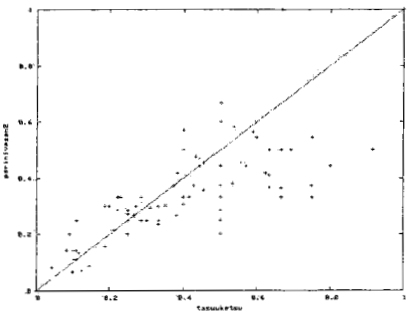

図 5: 提案手法と Srinivasan の方法の相互被覆率

考える.

隣接上位下位の単語 (u, v) に対する枝の粒度を $df(v)/df(u)$ で定義する. 各単語について全ての隣接する単語間の粒度を測定し, その平均を単語数の頻度順位を横軸としてプロットした. 具体的な 100 個の単語についてこの値をプロットしたのが図 6 である.

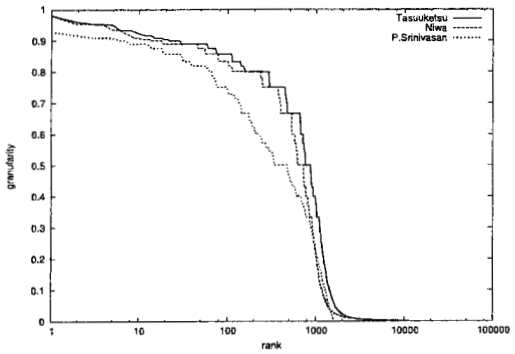

図 6: 粒度

この曲線が x 軸に平行な直線に近いほど各階層の頻度の比が均等であり, y 座標が大きいほど各階層の頻度の差が小さい. すなわち, x 軸に平行な直線に近く y が大きいほど上位下位関係が細やかな階層化といえる. 図 6 から, 提案手法は他の 2 手法よりも 平坦であり, 隣接階層間のギャップが最も小さい. 丹羽や Shrinivasan の方法では, 隣接単語の選択基準として単語の類似度を用いている. 一方, 提案手法では, 順序についての隣接単語としているので, 結果として最も細かい階層ができたと考えられる.

6 まとめと今後の課題

本稿では, 具体的な文書群から, ユーザの入力にあわせて特徴語とそれらの上位下位関係を抽出しグラフによって可視化するシステムを提案した.

ランダムに選んだ,100 種類の単語について, 提案手法と既存の手法との比較を行なった. 提案手法と既存の手法と比較した場合, グラフの表現する単語間上位下位関連が概念を再現できているのかという観点では丹羽等の方法よりも若干優れているが,Srinivasan の方法と比較した場合はどちらが優れているかは判断できなかった.

また, 階層のきめ細かさを, 隣接する単語の頻度の比として定式化し比較した結果, 提案手法が最も高い評価となった.

コーパス作成の単語切り出し作業において, 漢字とカタカナ以外を不要語として除去したが, ひらがな, 記号, 数字などを含む意味のある単語は存在する. また茶筅を使用しているため, 日本語, 英語以外の言語には本システムをそのまま使用することはできない. そこで, N-グラムの頻度に対する統計的方法を用い単語切り出しを行なうことで, 漢字とカタカナ以外の特徴語を抽出することや言語独立性の向上が可能になると考えられる. また, 人手によって評価されたデータを用いた適合率や再現率による評価が必要である. 本稿では扱わなかったが, 検索キーワード拡張への応用が今後の課題である.

参考文献

[1] P. Cimiano, A. Hotho, S. Staab, Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, *Journal of Artcial Intelligence Research*, Vol. 24, pp. 305-339 (2005)

[2] Hearst, M.A.: Automatic Acqusition of Hypernyms from Large Text Corpora, *Proceeding of the fourteenth International Conference on Computational Lingustics* (1992)

[3] A. Maedche, S. Staab, Measuring similarity between ontologies, In Proceedings of the European Conference on Knowledge Engineering and Knowledge Management (EKAW), LNCS 2473,pp. 251~263 (2002)

[4] Y. Niwa et al.: Topic Graph Generation for Query Navigation, *NLPRS'97*, pp. 95–100 (1997).

[5] 下司義寛, 和多大樹, 廣川佐千男, 英和辞典からの知識抽出, 第 68 回情報処理学会全国大会講演論文集 3, pp. 19–20, 2006.

[6] P. Srinivasan: *Tesaurus Construction*, Data Structures and Algorithms, Prentice-Hall (1992).