

文書群からの概念グラフの構成

廣川佐千男 † 下司義寛 †† 和多太樹 ††

単語の意味を、その単語が現れる文書の集合により規定するという立場から、単語の上位下位関係の定式化を試みる。特定の文書群と全体の文書群が与えられたとき、ある単語が出現する文書の過半数がその文書群の文書であるとき、その単語はその文書群の特徴的単語と見なす。次に、その特定文書群において、単語 u が出現する文書の過半数において単語 v が出現し、かつ、 u の文書頻度の方が v の文書頻度より大きいとき、 u は v の上位概念とする。こうして得られる上位下位関係から有向グラフを構成する。約 500 人の教員活動概要の文書群に対して概念グラフを構築し、人手による単語の上位下位関係と比較評価を行った。

Construction of Concept Graph from Documents

SACHIO HIROKAWA †, YOSHIHIRO SHIMOJI †† and TAIKI WADA ††

This paper proposes a novel formulation of superordinate relation between words based on the document frequency of words in a given set of documents. A word u is hypernym to v when more than half of documents that contains v contains u and the number of documents that contains u is larger than that of v . An algorithm is shown to generate a directed graph from this superordinate relation. The algorithm is applied to generate concept graphs for documents of activity overview of university researchers.

1. はじめに

インターネットの普及により大量の文書が容易に入手できるようになり、検索結果についてさらに効率よく絞り込んだり、検索をやり直したり、あるいは、検索結果を分かり易くまとめるための技術が必要とされている。検索要求を表現する適当なキーワードを利用者が分からない場合に対応するため、関連するキーワードを提示する検索質問拡張^{2),10)}の研究がある。一件ごとの検索に対応するのではなく、特定の分野を対象とした専門用語集やシソーラスなどを構築することにより、検索システムの側の知識を増やす研究^{3),12),14)}もある。新しい分野のようにそもそも専門用語が揃っていない分野や、Blogなどの評判情報のように毎日大量の新規文書が作り出される領域では、自動化が必須である。確立された分野においても、知識共有のために共通のメタデータ、シソーラス、あるいはオン

トジーが必要とされている⁵⁾。また、検索結果の理解を助けるためクラスタリングや要約や可視化の研究がある。相互の関連を表示するだけでなく、検索結果の文書や文書群の内容を表す適当なキーワードを割り当てることが重要となっている。大量の文書群に対して文書群のクラスタリングと、各クラスを表すキーワードの割り当てができれば、検索システムの向上だけでなく、知識共有のためのシソーラスとしても利用が考えられる。このように、文書群から各文書の関連抽出や特徴的単語の関連抽出は、単に検索システムのためだけでなく知識システムの構築の基礎となる。

シソーラスの構築は古くからある研究テーマである¹²⁾。シソーラスというときに、似た意味の単語をまとめた類語辞典という意味で使われることもある。関連する単語をまとめたものもある。また、さらに上位下位概念を含む場合もある。本稿では、単語の上位下位の概念を、具体的な文書群から自動的に構成する方法を提案する。共通の上位概念、あるいは共通の下位概念を持つ単語として関連語を捉えることもできる。

2. 文書頻度による単語の上位下位関係

どのようなキーワードを選ぶかという問題は、検索システムへの入力という観点から見ても、検索結果の

† 九州大学情報基盤センター
Kyushu University, Computing and Communications Center

†† 九州大学大学院システム情報科学府
Kyushu University, Graduate School of Information Science and Electrical Engineering

分かりやすい表示という観点から見ても重要な問題といえる。検索システムへの入力の場合には、結果として想定する文書群を特徴付けるキーワードを選択することが重要である。検索結果の出力においては求めた文書群を分かりやすく表示するためのランキングやクラスタリングに必要となるだけでなく、クラスタに適切な名前をつけることあるいは、そのクラスタの特徴的な単語を表示することが重要となる。いずれの場合でも基本的となるのは、文書群を規定する単語、あるいは、単語により文書群が規定されるという考え方である。これは、単語を内包、文書集合を外延と見る自然な捉え方といえる。

本稿では、ある単語を含む文書の集合がその単語の意味を規定するという考えに基づき、単語の上位下位関係を定式化する。文書群が与えられたとき、その文章群における単語の上位下位関係を文書頻度を用いて定式化する。具体的には単語 u と v について、 v が現れる文書の過半数の文書には u が現れているときに、 u は v より概念として上位にあるとする。

定義 1 D を文書集合、 w を単語とする。 w が現れる D 中の文書の個数(文書頻度)、すなわち、 $\#\{d \in D \mid w \text{ が } d \text{ 中に現れる}\}$ を $df(w, D)$ で表す。二つの単語 u, v の両方が現れる文書数を $df(u * v, D)$ であらわす。単語 u と v が次の関係を満たすとき、 $v \triangleleft u$ と表し、「文書頻度の観点から u は v の上位である」ということにする。

$$\begin{aligned} df(u * v, D) / df(v, D) &> 0.5 \\ df(u, D) &> df(v, D) \end{aligned}$$

上で定義した上位下位関係は一般に順序関係とはならない。例えば、図 1 において、 x の上位の単語を考える。4 つの円はそれぞれ単語 u, v, w, x を含む文書集合を表すものとする。 x のすぐ上の単語は v であり、 v の上位に u と w がある。 w については、 $x \triangleleft v \triangleleft w$ であり、しかも $x \triangleleft w$ であるので、推移律が成り立っている。しかし、 u については、 $x \triangleleft v, v \triangleleft w$ であっても、 $x \triangleleft u$ ではないので、推移律が成り立っていないことが分かる。このように、 \triangleleft は順序関係には限らない。

3. 局所的な上位下位関係の抽出

前章で定義した上位下位関係を、単語を点として上位下位関係を有向枝とする有向グラフで表すことにする。ところで単純に上位下位関係のある二つの単語の間に線を引くだけでは分かりやすい表示になるとは限らない。例えば、ある単語 u が単語の v の上位となっ

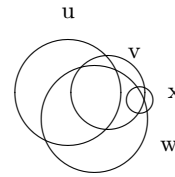


図 1 推移律の反例

ているとき、 v の上位の単語がすべて u にとっても上位の単語となっているとする。もし、 v が 100 個以上の上位単語を持っていたら、 v から出る 100 本の枝の他に、 u から出る枝も 100 本あることになる。 u から v への枝を一つ引くだけとして、 u から 100 個の単語には線を引かないようにすれば、見やすいものとなる。しかも、 v が u より上位にあり、100 個の点は v から上位方向に繋がっているので、 u にとっても上位になることを視覚的に認識できる。本章では、前章で定義した上位下位関係を可視化するために、各点に対しすぐ上の点を求め、「直上」の点との間だけに線を引く。これにより、各単語について局所的な上位下位関係にある単語を求めることができる。

定義 2 集合 S 上の二項関係 \triangleleft について、ある自然数 K が存在し、 K よりも長い列 $p_1 \triangleleft p_2 \triangleleft p_3 \triangleleft \dots$ が存在しないとき \triangleleft は有限列条件を満たすということにする。

二項関係 \triangleleft が有限列条件を満たすときには次のことが成り立つ。

- (1) $p \triangleleft p$ とはならない。
- (2) \triangleleft は閉路をもたない。すなわち $p_0 \triangleleft p_1 \triangleleft p_3 \triangleleft \dots \triangleleft p_n \triangleleft p_0$ となる列は存在しない。

定義 3 p より上に現れる要素の集合を $up(p)$ 、すぐ上に現れる要素の集合を $dup(p)$ と表す。 q が p のすぐ上のときに $p \triangleleft^1 q$ と表現することにする。

- $up(p) = \{q \in S \mid p \triangleleft q\}$
- $dup(p) = \{q \in S \mid p \triangleleft q, p \triangleleft q \text{ かつ } r \triangleleft q \text{ となる } r \text{ は存在しない}\}$
- $p \triangleleft^1 q$ iff $q \in dup(p)$

定理 1 \triangleleft を有限列条件を満たす二項関係とする。このとき、 $p \triangleleft q$ ならば、次の条件を満たす有限個の列 p_0, p_1, \dots, p_n が存在する。

- (1) $p_0 = p, p_n = q,$
- (2) $p_i \triangleleft^1 p_{i+1} (i = 0, 1, \dots, n-1)$

証明 $p \triangleleft q$ となる組 (p, q) に対し距離 $d(p, q)$ を次のように定義する。

$$d(p, q) = \max\{n \mid \exists p_0, p_1, \dots, p_n \text{ s.t.}$$

$$(1) p_0 = p, p_n = q,$$

$$(2) p_i \triangleleft p_{i+1} (i = 0, 1, \dots, n-1)\}$$

\triangleleft は有限列条件を満たすから、 $p = p_0 \triangleleft p_1 \triangleleft \dots \triangleleft p_n = q$ となる列の長さはある定数 K 以下である。従って、そのような列で最長のものがある。これが $d(p, q)$ である。定理は $d(p, q)$ について \emptyset 数学的帰納法で示す。

1. $d(p, q) = 1$ のとき。 $p \triangleleft r \triangleleft q$ となる r は存在しない。故に $q \in \text{dup}(p)$ となり、 $p \triangleleft^1 q$ を得る。よって定理は成り立つ。

2. $d(p, q) \leq k - 1$ まで成り立つと仮定して、 $d(p, q) = k$ のときも成り立つことを示す ($k \geq 2$)。 $d(p, q) = k \geq 2$ より長さ $k (\geq 2)$ の列 $p = p_0 \triangleleft p_1 \triangleleft \dots \triangleleft p_k = q$ が存在する。距離の定義より、 $d(p_0, p_1) < k$ かつ $d(p_1, q) < k$ 。帰納法の仮定より、 $p = s_0 \triangleleft^1 s_1 \triangleleft^1 \dots \triangleleft^1 s_{l-1} \triangleleft^1 s_l = p_1$,

$p_1 = t_0 \triangleleft^1 t_1 \triangleleft^1 \dots \triangleleft^1 t_{m-1} \triangleleft^1 t_m = q$ となる列 $s_0, s_1, \dots, s_l, t_1, \dots, t_m$ が存在する。このとき、

$$s_0 \triangleleft^1 s_1 \triangleleft^1 \dots \triangleleft^1 s_l = t_0 \triangleleft^1 t_1 \triangleleft^1 \dots \triangleleft^1 t_m$$

が求める列である。証明終

定理 1 は、全域的な関係 \triangleleft が局所的な関係 \triangleleft^1 により完全に表わされることを意味する。

4. アルゴリズム

単語 w とそのすぐ上位の単語を結ぶことで、特徴的単語群の関連を表すグラフを構成する。このような「すぐ上の単語」は、具体的には、単語 w より上位にある単語の集合 $up(w)$ の極小元として求めることができる。

特徴語の抽出については様々な手法があるが、本稿では筆者等が¹³⁾で提案した、文書頻度による方法を用いる。対象とする文書集合と全体の文章集合が与えられていると仮定し、ある単語が出現する文書の過半数が対象文書群である場合、その単語は対象文書群の特徴語と見なす。

```

ConceptGraph(D, U) {
# 入力 D:対象文書集合; U:全体文書集合;
# 出力 G = (W, E):有向グラフ;
# W:単語集合; E:単語間の上下位関係;
W = {w | df(w, D)/df(w, U) > 0.5} # 特徴語集合
< = {(u, v) | df(u * v, D)/df(u, D) > 0.5 かつ
      df(u, D) < df(v, D)} #上下位関係
E =  $\emptyset$ ; R = W;
while (R  $\neq$   $\emptyset$ ) {
    foreach w in Minimal(R) {

```

$$up(w, R) = \{u \in R \mid w \triangleleft u\};$$

$$\text{dup}(w, R) = \text{Minimal}(up(w, R));$$

$$E = E \cup \{(w, u) \mid u \in \text{dup}(w, R)\};$$

$$R = R - \text{minimal}(R)$$

}

}

}

Minimal(R) {

$$M = \emptyset;$$

foreach w in R {

$$\text{minimal} = \text{true};$$

foreach u in R {

$$\text{minimal} = \text{false} \text{ if } u \triangleleft w;$$

}

$$M = M \cup \{w\} \text{ if } \text{minimal};$$

}

return M;

}

一般的に、 $u \triangleleft v$ かつ $v \triangleleft w$ であっても $u \triangleleft w$ が成り立つとは限らず、 \triangleleft は順序関係とならない。従って、アルゴリズム ConceptGraph で求まるグラフにおいて、連続する二つの有向枝で繋がっている単語が必ずしも上位下位の関係になっているとは限らない。図 1 において、 $x - v - u$ の方向には同じテーマで文書数が増えているが、 $x - v - w$ の方向については、 $v - w$ のところで二つのテーマに分かれている。 x は $v - u$ の方向のテーマからは外れているような場合 (トピック・ドリフト) に相当する。

しかし、 $v \triangleleft u$ である単語の組については、必ず概念グラフにおいて上下関係の位置に現れることが次の定理により保証される。

定理 2 u が v の上位概念のときには、アルゴリズム ConceptGraph で得られる有向グラフにおいて、 u と v はのどちらとも現れ、かつ、 u は v の上に現れる。

証明 上位下位関係の定義により、 \triangleleft は有限列条件を満たす。 u が v の上位概念とすると、定理 1 により、 v から u にいたる \triangleleft^1 列が存在する。ところが、アルゴリズム ConceptGraph はすぐ上の単語へのエッジを全て求めているので、この \triangleleft^1 列も、ConceptGraph で求まるグラフに現れている。従って、このグラフにおいて、 u は v の上に現れている。証明終

5. 実験と評価

本稿では、九州大学研究者情報として公開されて

<http://hyoka.ofc.kyushu-u.ac.jp/search/>

いる教員の活動概要の文書のうち、表1の5部局について概念グラフの自動生成を行なった。例えば、芸術工学研究院98名の教員の活動概要の文書中の文書頻度3以上の特徴語について提案アルゴリズムで描いたのが図2である。ただし、上位の単語が左側、下位の単語が右側となるように表示している。各単語の横の数は、その単語の文書頻度である。

部局	芸術工学	経済	理学	歯学	システム情報科学
文書数	98	58	178	90	110
平均サイズ	479	1107	980	864	961
名詞総数	2824	3238	7080	3673	4320
文書頻度	6	6	15	10	10
単語数	24	28	65	33	43

表1 教員データ詳細(サイズはハイト)

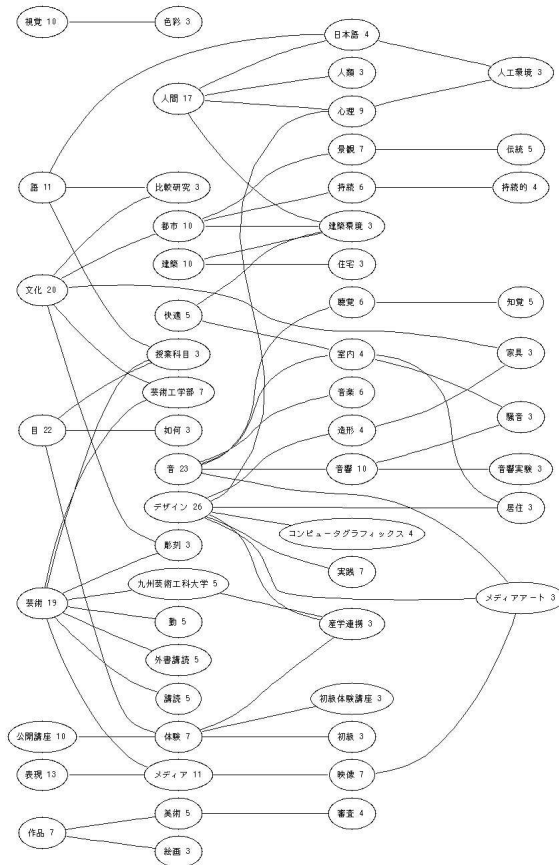


図2 文書頻度3以上の特徴語の概念グラフ

人手による単語の上位下位関係判定と比較するため、各部局の特徴語を50個程度となるように文書頻度の下限を設定し特徴的単語を抽出した。例えば芸術工学

研究院については、特徴語として文書頻度6以上の単語(表2)に限定して評価を行なった。

デザイン(26)	音(23)	文化(20)	芸術(19)	人間(17)	表現(13)	メディア(11)	都市(10)	視覚(10)	建築(10)	音響(10)	心理(9)	コミュニケーション(8)	体験(7)	実践(7)	作品(7)	芸術工学部(7)	景観(7)	近代(7)	映像(7)	聴覚(6)	創造(6)	持続(6)	音楽(6)
----------	-------	--------	--------	--------	--------	----------	--------	--------	--------	--------	-------	--------------	-------	-------	-------	----------	-------	-------	-------	-------	-------	-------	-------

表2 文書頻度6以上の特徴語

出現頻度6以上の特徴語について提案アルゴリズムによる概念グラフは図3のようになった。

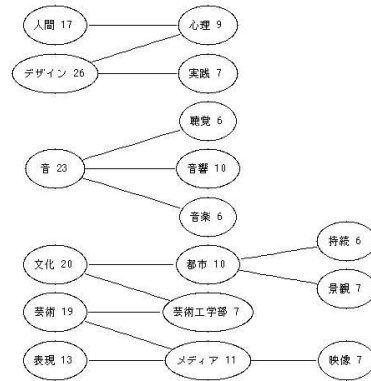


図3 文書頻度6以上の特徴語の概念グラフ

図3と比較・評価するために、12人の被験者に表2の単語のリスト(文書頻度は示さず)から、上位下位関係にあると考えられる単語の組を求めてもらった。被験者の内訳は、教員3名、職員3名、大学院生5名、その他1名である。24個の単語の組について、 w_i が w_j より上位と被験者 p が判定したら $X_p[i, j] = 1$ そうでないとき、 $X_p[i, j] = 0$ となる行列 X_p を作り、アルゴリズム ConceptGraph から得られる同様の行列 X と比較した。 X と X_p から精度 $Prec(X, X_p)$ ならびに再現率 $Recall(X, X_p)$ を次のようにして求めた。

$$Prec(X, X_p) = \frac{\#\{(i, j) \mid X_p[i, j] = 1\}}{\#\{(i, j) \mid X[i, j] = 1\}}$$

$$Recall(X, X_p) = \frac{\#\{(i, j) \mid X[i, j] = 1\}}{\#\{(i, j) \mid X_p[i, j] = 1\}}$$

12人の被験者ごとに $(Recall(X, X_p), Prec(X, X_p))$ をプロットしたのが図4である。精度の平均は0.280、再現率の平均は0.119である。

ConceptGraphにより得られる単語の関連グラフは、各部局ごとの教員の専門性を表しそれなりに意味

のあるものであった。特に、その部局において共通のテーマについて多くの教員がいる場合、そのテーマのより詳細な構造を見出すことができた。しかし、被験者による単語の上位下位関係の抽出結果と較べると、精度再現率とも非常に低い結果となった。これには二つの理由が考えられる。一つ目は、実験対象として文書数が少なく、上位下位関係として誰もが共通に認識できる一般的な関係がごく少数しかなかったことである。これは二人の被験者 X_p, X_q 間の相互評価による正解率の平均が 0.275 であったことから確認できた。また、過半数の人が認める上位下位関係を正解としたとき、被験者の精度、再現率はそれぞれ 0.569, 0.214 であり、ConceptGraph については 0.286, 0.333 であった。つまり、人間同士で相互評価した値と極端な差があるとはいえない。二番目の理由は、上位下位判定の対象とした単語のリストに、例えば、デザイン、メディア、建築などのような技術的専門用語、とその分野の活動を表す一般的用語、例えば、表現、実践、持続など、が混在していたことなどが上げられる。アルゴリズムでは「デザイン-実践」や「都市-持続」などのように異なる観点の単語の間に上位下位の関係を見出していることによるが、人間ではこの関係は推測できなかった。

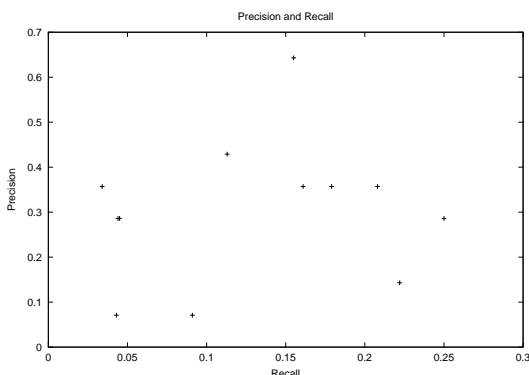


図 4 精度と再現率の評価

6. 応用と今後の課題

単語集合が異なれば構成される概念グラフも異なるので、特徴的単語の抽出に他の方式を適用して比較することが考えられる。特徴的単語の抽出法として、本稿では¹³⁾の文書頻度による方法を用いたが、中川等⁸⁾の接続頻度による方法を適用して得られる単語群を使うことが考えられる。また、同じ単語集合であっても、

文書群が異なれば概念グラフも異なる。今回は筆者等が所属する大学の教員活動概要を対象としたが、異なる大学のデータを使った比較実験や、特定分野の論文アブストラクトについての実験が考えられる。

本手法で得られる上位単語は検索質問拡張に直ちに適用できる。概念グラフのノードには単語とその文書頻度を合わせて表示しているが、実際には、その単語を含む文書集合が分かっている。つまり、概念グラフは、文書群のクラスタリングにもなっている。従って、検索エンジンの検索結果表示方法としての利用も可能である。

今回の実験では、特徴語候補として茶釜⁶⁾による形態素解析で名詞と判定される単語と、連続する漢字、連続する片仮名を考えた。しかし、文書頻度しか使っていないので、原理的には名詞に限定する必要はない。例えば、¹⁵⁾で対象とした形容詞も名詞と一緒に扱えば、評判情報の意味解析が可能となる。また、単語でなく N-gram を考えれば、バイオ関連論文に現れる自然言語の単語とゲノムシーケンスについて、同じ枠組で概念的な上位下位関係を捉えることができる。

本稿では、単語の意味はその単語を含む文書集合で決まるという考え方に基き、二つの単語 u と v の上位下位関係を、その単語を含む文書集合の関係 $df(u * v, D) / df(v, D) > 0.5, df(u, D) > df(v, D)$ として捉えた。 u, v を単語でなく、単語の集合と考えても(計算時間は増えるが)同様な定式化ができる。この場合は、単語の集合とその単語群を含む文書集合との対応を考えることになる。ところで、文書を対象物、単語を属性と見なすと概念束の理論^{1),4)}を適用してグラフを構成する方法が知られている。本稿で提案した概念グラフと概念束の関連の解明は興味深い課題である。また、文書頻度を単語の出現確率と捉えると、本稿で述べたグラフは、閾値を 0.5 としたグラフィカル・モデル^{7),11)}と見なすこともできる。その観点からの分析も必要と考える。

文書を対象物、単語を属性と見なすと、属性間の上位下位関係の定式化を与えたことになっている。例えば、対象物、属性ともに HTML 文書として、対象の HTML 文書が属性としての HTML へのリンクを持つことを、対象と属性の関係と捉えると、HTML 文書間の上位下位関係ならびにそれに伴う HTML 文書のクラスタリングとなる。これは、従来のリンク構造の解析とは異なるグラフ構造を与えることになる。対象物を文献、属性をその文献で引用されている文献とすると、引用関係⁹⁾についての上位下位を表すグラフ構造

を求めることができる。対象物をiTMSのPlaylist、対象を曲名あるいはアーティストとすると、流行している曲の分類が可能となる。

7. ま と め

単語の文書頻度に基づき、単語間の上位下位関係の定式化を考案し、与えられた文書群からそこに現れる特徴的単語の上位下位関係を表すグラフを自動的に構成するアルゴリズムを提案した。具体的には単語 v を含む文書の過半数に単語 u が現れるとき u は v より上位にあるとする。これは、単語の意味をその単語を含む文書集合で捉えるという観点から見て妥当な定式化といえる。

大学教員の活動概要の文書群に提案アルゴリズムを適用し、意味のある概念グラフが構成されたことを確認できた。人手で決めた上位下位関係との比較では、精度が0.3、再現率は0.1程度しかなく有効とはいえない結果となった。しかし、被験者同士で相互に評価した場合でも、正解率が0.3であり、提案手法の欠点というより、適用した文書の総数が少なかつたためと考えられる。逆に、このように人手で上位下位関係を抽出することが困難な場合にも、結果的に意味のある概念グラフを構成できたことは、本手法の有効性と考えられる。安定した概念集合を対象とした精度と再現率の評価のためには、大規模な文書集合についての実験が必要である。

謝 辞

困難な評価実験を手伝っていただいた方々に感謝します。本研究の一部は、平成17年度科学研究費16650030、16016267による。

参 考 文 献

- 1) Claudio Carpineto, Giovanni Romano, Concept Data Analysis: Theory and Applications John Wiley & Sons, 2004
- 2) 土肥広典, 青野雅樹, 双クラスタリングに基づく検索質問拡張法, 電子情報通信学会第2回Webインテリジェンス研究会, IEICE SIG Notes WI2-2005-18, 2005.
- 3) 藤井敦, 石川徹也, World Wide Webを用いた事典知識情報の抽出と組織化, 電子情報通信学会論文誌, Vol.J85-D2, No.2, pp.300-307, 2002
- 4) Bernhard Ganter, Rudolf Wille, C. Franzke, Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, 1999

- 5) 木村徳信, 溝口理一郎, オントロジー工学に基づく機能的知識体系化の枠組, 人工知能学会論文誌, Vol.17, No.1, pp.61-72, 2002
- 6) 松本裕治, 形態素解析システム「茶筌」, 情報処理 Vol.41, No.11, pp.1208-1214, 2000
- 7) 宮川雅己, グラフィカル・モデリング, 朝倉書店, 1997
- 8) 中川裕志, 森辰則, 湯本紘彰, 出現頻度と接続頻度に基づく専門用語抽出, 自然言語処理, Vol.10, No.1, pp.27-45, 2003
- 9) 難波英嗣, 論文間の引用情報を利用した関連用語の自動収集, 言語処理学会第11回年次大会, 2005
- 10) 新里圭司, 鳥澤健太郎, HTML文書からの単語間の上位下位関係の自動獲得自然言語処理, Vol.12, No.1, pp.125-150, 2005
- 11) Rohini Srihari, Miguel E. Ruiz, Munirathnam Srikanth, Concept Chain Graphs: A Hybrid IR Framework for Biomedical Text Mining, Proceedings of the SIGIR 2003 Workshop on Text Analysis and Search for Bioinformatics, 2003
- 12) Padmini Srinivassan, Thesaurus Construction, in Information Retrieval: Data Structures and Algorithms, Edited by William B. Frakes and Ricardo Baeza-yates, Prentice-Hall, 1992
- 13) 下司義寛, 和多太樹, 安元裕司, 関隆宏, 廣川佐千男, 文書群の局所性と大域性の差を利用したキーワード粒度評価, 情報処理学会研究会報告2005-NL-168, pp.7-12, 2005
- 14) 梅村恭司, 未踏テキスト用シソーラスの自動構築システムの開発, 情報処理振興協会平成13年報告集, 2001
- 15) 安元裕司, 和多太樹, 関隆宏, 廣川佐千男, 病院評判情報の多面的解析, 人工知能学会研究会資料SIG-KBS-A501, pp.1-4, 2005

<http://www.apple.com/itunes/store/>