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

2005－ FI － 80（12）
2005－ NL － 169（12）
2005／9／30

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 比較評価を行な 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 新規文書がつくり出される領域で 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 おうに 、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 Univeristy, 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 求まっ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 適当な 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、$\sharp\{d \in D \mid w$ が d 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、$v \lhd u$ 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。

$$df(u * v, D)/df(v, D) > 0.5$$
$$df(u, D) > df(v, D)$$

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 、$x \lhd v \lhd w$ で あり、しかも $x \lhd w$ 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 、$x \lhd v$, $v \lhd w$ で あって も、$x \lhd u$ 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 、$\lhd$ は 順序関係に な ると 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。

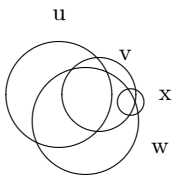

図 1 推移律の 反例

定義 2 集合 S 上の 二項関係 $\lhd$ に ついて 、ある自然数 K が存在し、K よりも長い列 $p_1 \lhd p_2 \lhd p_3 \lhd \cdots$ が存在しな いと き $\lhd$ は 有限列条件 を満たすと いうこと に する。

二項関係 $\lhd$ が有限列条件を満たすと きに は 次の こと が成り立つ。

(1) $p \lhd p$ と は な らな い。

(2) $\lhd$ は 閉路をもたな い。すな わち $p_0 \lhd p_1 \lhd p_3 \lhd \cdots \lhd p_n \lhd p_0$ 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 $p \lhd^1 q$ と 表現する こと に する。

- $up(p) = \{q \in S \mid p \lhd q\}$
- $dup(p) = \{q \in S \mid p \lhd q$, $p \lhd q$ かつ $r \lhd q$ と な る r は 存在しな い $\}$
- $p \lhd^1 q$ iff $q \in dup(p)$

定理 1 $\lhd$ を有限列条件を満たす二項関係と する。この と き、$p \lhd q$ な らば 、次の 条件を満たす有限個の 列 $p_0, p_1, \cdots, p_n$ が存在する。

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

(2) $p_i \lhd^1 p_{i+1} (i = 0, 1, \cdots, n-1)$

証明 $p \lhd q$ と な る組 (p, q) に 対し距離 $d(p, q)$ を次の ように 定義する。

$d(p,q) = max\{n \mid \exists p_0, p_1, \cdots, p_n$ s.t.
(1) $p_0 = p, p_n = q$,
(2) $p_i \lhd p_{i+1} (i = 0, 1, \cdots, n-1)\}$

$\lhd$ は 有限列条件を満たすから、$p = p_0 \lhd p_1 \lhd \cdots \lhd p_n = q$ 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 の 数学的帰納法で 示す。

1. $d(p,q) = 1$ の と き。$p \lhd r \lhd q$ と な る r は 存在し な い。故に $q \in dup(p)$ と な り、$p \lhd^1 q$ 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 $p = p_0 \lhd p_1 \lhd \cdots \lhd p_k = q$ 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、$p = s_0 \lhd^1 s_1 \lhd^1 \cdots \lhd^1 s_{l-1} \lhd^1 s_l = p_1$, $p_1 = t_0 \lhd^1 t_1 \lhd^1 \cdots \lhd^1 t_{m-1} \lhd^1 t_m = q$ と な る列 $s_0, s_1, \cdots, s_l, t_1, \cdots, t_m$ が存在する。この と き、

$$s_0 \lhd^1 s_1 \lhd^1 \cdots \lhd^1 s_l = t_0 \lhd^1 t_1 \lhd^1 \cdots \lhd^1 t_m$$

が求める列で ある。証明終

定理 1 は 、全域的な 関係 $\lhd$ が局所的な 関係 $\lhd^1$ 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 上位下位関係;
   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)} #上位下位関係
   E = ∅; R = W;
   while (R ≠ ∅) {
      foreach w in Minimal(R) {
         up(w,R) = {u ∈ R | w ◁ u};
         dup(w,R) = Minimal(up(w,R));
         E = E ∪ {(w,u) | u ∈ dup(w,R)};
         R = R − minimal(R)
      }
   }
}
Minimal(R) {
   M = ∅;
   foreach w in R {
      minimal = true;
      foreach u in R {
         minimal = false  if u ◁ w;
      }
      M = M ∪ {w}  if minimal;
   }
   return M;
}
```

一般的に 、$u \lhd v$ かつ $v \lhd w$ で あって も $u \lhd w$ が成り立つと は 限らず、$\lhd$ 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 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、$v \lhd u$ 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、$\lhd$ 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 $\lhd^1$ 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 $\lhd^1$ 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 バイト)

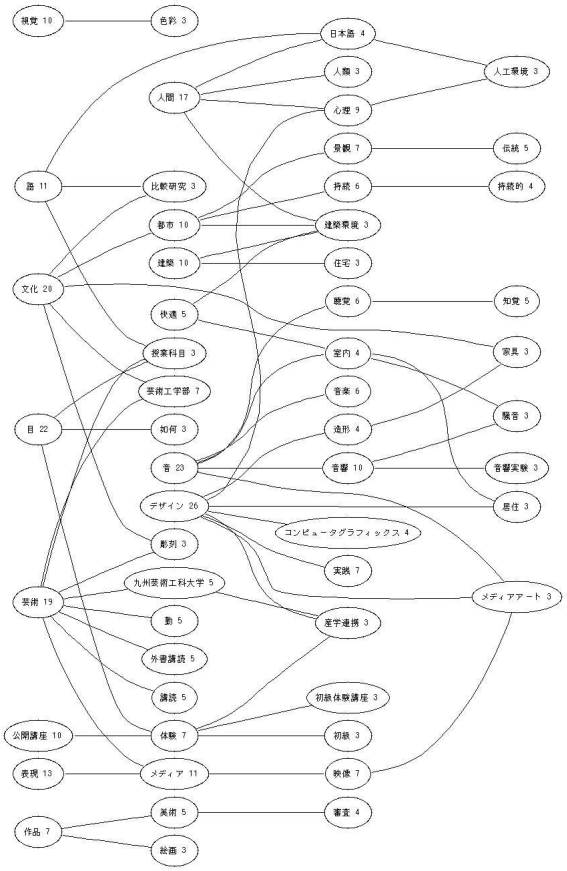

図 2 文書頻度 3 以上の 特徴語の 概念グラフ

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。

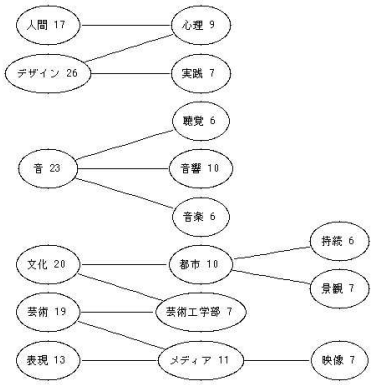

図 3 文書頻度 6 以上の 特徴語の 概念グラフ

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 示さずに) 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。

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

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

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 構造を見い出すこと 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 からも確認で 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 関係を見い出して 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。

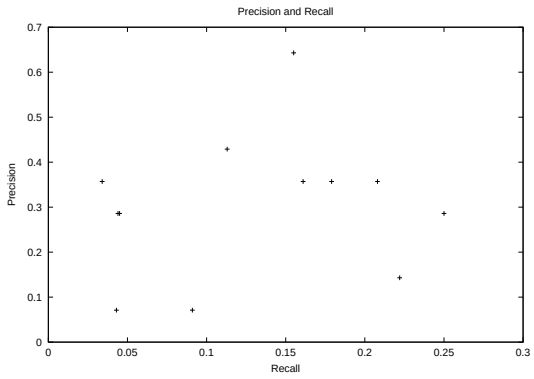

図 4　精度と 再現率の 評価

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 連接頻度に 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 応用で 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 茶筌 [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 基づ 、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 かったためと 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 連接頻度に 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/