2D5-2

# グラフモジュール分解に基づく専門用語辞書からのオントロジー自動構築

# Modular Decomposition based Automatic Construction of Ontology from a Technical Term Dictionary

稲木 誓哉[*1] 邱 シュウレ[*1] 渡部 雅夫[*1] 梅基 宏[*1] 岡本 洋[*1]
Seiya Inagi Xule Qiu Masao Watanabe Hiroshi Umemoto Hiroshi Okamoto

[*1] 富士ゼロックス(株)研究技術開発本部
Research & Technology Group, Fuji Xerox Co., Ltd.

近年, 各専門領域の特異性を反映したオントロジーの構築が求められている. そこで我々は, 専門用語辞書からオントロジーを自動構築する方法を提案する:「見出し語」ー「説明文中の単語」の二部グラフを作る;このグラフをモジュール分解する;分解結果に対してパターンマッチおよび後方文字列一致判定を行う;概念間のクラス-インスタンス関係および Is-a 関係を構築する. 製造業の専門用語辞書にこの方法を適用した結果を報告する.

## 1. はじめに

### 1.1 背景

情報検索あるいはデータ統合を行う知識処理システムの基盤技術として, オントロジーが注目されている. 特に, 専門領域に特化したシステムにおいてはその領域に特有の情報を記述したオントロジーの構築が求められている. しかしながら, 人手によるオントロジーの構築には, コストがかかる, あるいは, メンテナンスが困難であるという問題が伴う. そのため, オントロジーを自動的に構築する方法が近年盛んに研究されている.

現在, オントロジー自動構築においては, コンピュータによる統計処理に基づくライトオントロジーの構築が主流である. ライトウェイトオントロジーは, 概念(クラス)の集合, および, 概念間の上位・下位関係(Is-a 関係)を記述する. ライトウェイトオントロジー構築の情報資源として, 専門用語辞書, Web ページ, あるいは専門用語を含む技術文書が利用される. 特に専門用語辞書は, 個々の見出し語の説明文が限られた専門家により編集されているため, 情報の信頼性や一貫性という点から利用価値が高い.

本稿では, 専門用語辞書から構築した「見出し語」ー「説明文中の単語」の二部グラフのモジュール分解に基づいてライトウェイトオントロジーにおけるクラスを抽出する方法を提案する. さらに, パターンマッチおよび後方文字列一致に基づいてクラスラベルを推定し, 人手でクラス間に Is-a 関係を付与することでライトウェイトオントロジーを構築した.

### 1.2 関連研究

ライトウェイトオントロジー構築においては, k-means, 凝集クラスタリング, その他のクラスタリング方法をベクトル空間モデルに適用し, テキストコーパスからクラスを抽出する方法が示されている[Cimiano 2004]. ベクトル空間モデルでは, クラスタリングの対象, 例えば文書をベクトルで表現する. 文書 $d$ は, $V$ 次元ベクトル $\{n_{di}\}_{i=1}^{V}$ で表される. $V$ は対象となる全ての文書に含まれる単語の語彙数, $n_{di}$ は文書 $d$ における単語 $i$ の出現数である. このベクトル間の距離あるいはコサイン類似度によって文書間の関連度を表す. 例えば, 文書 A, B および C のベクトル表現が, それぞれ, (1, 1, 0), (1, 1, 1)および(0, 0, 1)で与えられたとする. このとき, ベクトルのコサイン類似度が最も高いのは文書 A と文書 B のペアであるため, ベクトル空間モデルでは文書 A と文書 B が最も関連度の高いペアと判断される.

しかしながら, このベクトル空間モデルには次の問題点があることが指摘されている[Jing 2010]. 文書 A, B および C のベクトル表現が, それぞれ, (1, 0, 0), (0, 1, 0)および(0, 0, 1)で与えられた場合を考える. この場合, すべてのベクトルペアに対するコサイン類似度は 0 となる. 従って, どの文書ペアにも関連がないと判断される. しかしながら, 実際の文書ペアにおいては, ベクトル表現によるコサイン類似度が 0 であったとしても, 関連があるとみなすべき場合がある. 例えば, ベクトルの第 1, 第 2 および第 3 成分が, それぞれ, 「自動車」, 「エンジン」および「りんご」を表すとしよう. 「自動車」と「エンジン」との間には実際には意味的に関連がある. 従って, 文書 A と B のペアは, 本来, 文書 A と C のペアよりも関連度が高いと判断されるべきである. このように, ベクトル空間モデルではベクトル要素間の関連性を考慮していないため, 文書ペアの意味的関連性を適切に表現できない場合がある.

連絡先:稲木 誓哉, 富士ゼロックス(株)研究技術開発本部, 〒220-8668 神奈川県横浜市みなとみらい6丁目 1 番, E-mail:inagi.seiya@fujixerox.co.jp

## 2. 提案方法

### 2.1 二部グラフによる文書の表現

本稿では, 二部グラフのモジュール分解を用いてクラスを抽出する方法を提案する. この方法では, 専門用語辞書から「見出し語」および「説明文中の単語」をノードとする二部グラフを構築することにより, 説明文中の単語の間の関連性を反映した見出し語の表現を得ることができる.

二部グラフとは, 「二種類のノード」および「異種ノード間をつなぐエッジ」からなる形式である. 文書と単語をノードとし, ある単語がある文書中に出現する場合にその単語ノードとその文書ノードをエッジで繋ぐことにより, 「文書」-「単語」の二部グラフが得られる(図 1b). ベクトル空間モデルによる表現(図 1a)と二部グラフによる表現(図 1b)とは, 数学的には等価である. すなわち, 一

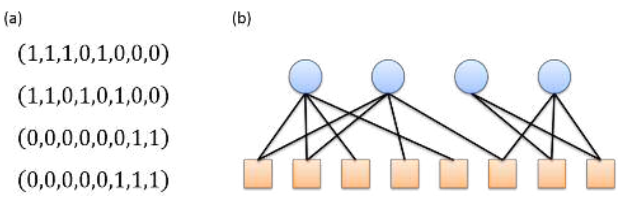


図 1 文書の表現方法. (a)はベクトル空間モデルによる表現, (b)は二部グラフによる表現の模式図を示す. 青丸は文書, 赤四角は文書に含まれる単語を表す.

方から一方を完全に復元することができる. さらに二部グラフ表現は, ベクトル空間モデルでは直接表すことが難しい単語間の関連性を, ノード間のグラフ距離(ノード間をまたぐリンクの数の最小値)として表す. 関連性の高い単語同士は同一文書内で共起する確率が高いため, 関連性の高い単語ノードペアは文書ノードを介して短いグラフ距離で繋がり, 関連性の低い単語ノードペアは長いグラフ距離で繋がる(図 2). 従って, ベクトル表現では内積が 0 となる文書ペアに対しても, 二部グラフ表現では単語ノード間の関連性に応じて関連度を定めることができる.

### 2.2 モジュール分解方法

本稿では, 専門用語辞書から作成した二部グラフから, [岡本 2014]によるモジュール分解方法を用いてクラスを抽出する方法を提案する. グラフの中の, その内部のノード同士の繋がりは密であり, その内部のノードと外部のノードとの繋がりは疎である部分のことを「モジュール」あるいは「コミュニティ」と呼ぶ(本稿では「モジュール」の呼称を用いる). 一般に, グラフにおける個々のモジュールには何らかの意味あるいは機能が対応する. 従って, 二部グラフから抽出されたモジュールに含まれている「見出し語」ノードの集合が, 「クラス」に対応すると考えられることは妥当である. 従って, モジュールとそれに含まれる見出し語の間の関係がクラス-インスタンス関係に対応する.

グラフ内のエッジの密度に基づいてグラフをモジュールに分解する方法が数多く提案されている(図 3). [岡本 2014]により, 確率モデルを用いてグラフ構造をモジュール分解する方法が示されている. 岡本の方法は, グラフ上にエッジに沿ってランダムに動き回るエージェント(ランダムウォーカー)を想定し, 各ノードにおけるランダムウォーカーの存在確率を用いてモジュール分解を実行する. この方法は, エッジが密になっている領域にランダムウォーカーが存在する確率が高く, その存在確率が各ノードのモジュールへの帰属確率に対応する, という仮定に基づいている. 各ノードのモジュールへの帰属確率は最尤法によって推定される. 岡本の方法では, モジュールの数を指定しない. その代わりに, ランダムウォーカーの行動範囲を制限するハイパーパラメータ $\alpha$ を導入しており, この $\alpha$ の大小によって得られるモジュールの大きさや数を変えることができる.

### 2.3 Is-a 関係構築方法

以下の手順でモジュール分解によって得られたクラス間の Is-a 関係を構築する.

手順 1: 各クラスのラベルをパターンマッチ[柴木 2009]および後方文字列一致[玉川 2010]に基づいて決定する.

手順 2: Is-a 関係が成り立つラベルの組を人手で抽出し, 階層化する.

ただしここでいうラベルとは, 各クラスに属する見出し語に共通する Is-a 上位語を指す. 例えば, 「ランフラットタイヤ」「スノータイヤ」「スタッドタイヤ」という見出し語からなるクラスのラベルは「タイヤ」となる. 区別のため, 本稿では, 手順 1 で推定する Is-a 上位語を「ラベル候補語」と呼ぶ.

手順1では, あるクラスに属する個々の見出し語に対してラベル候補語を推定し, 最も多くの見出し語に対して推定されたラベル候補語をそのクラスのラベルとする. 以下に, ラベル候補語推定方法の詳細を述べる.

パターンマッチを用いた方法は, 辞書の説明文の末尾の単語あるいは「～の一種」などの特定のパターンの中に現れる単語は, 見出し語の上位概念であることが多いという性質を利用している.

例. 見出し語: 自転車

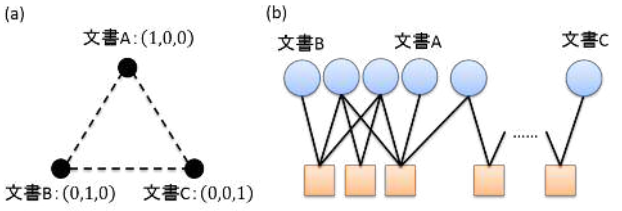


図 2 文書間の距離. (a)はベクトル空間モデル, (b)は二部グラフにおける文書間の距離を示す. 青丸は文書, 赤四角は文書に含まれる単語を表す.

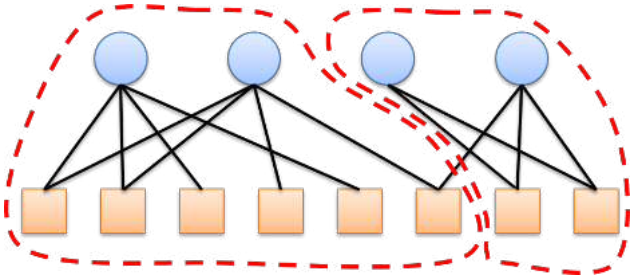

図 3 モジュール分解の概念図. 青丸が見出し語, 赤四角が説明文中の単語を表し, 赤い点線がモジュールを表す. モジュールには見出し語ノードと説明文中の単語ノードの両方が含まれる.

説明文: 二つの車輪を備え, 乗り手がペダルを踏む力によって車輪を回して進む車.

上記の例の場合, 説明文の末尾の「車」が見出し語である「自転車」の Is-a 上位語にあたる. [柴木 2009]は, 独自に定義した文字列パターンにマッチする語を説明文の中から抽出し, それを見出し語に対する上位語として, Is-a 関係を構築している. 本稿では, 説明文の一文目を係り受け解析し, 末尾の文節の主辞をラベル候補語とする. ただし, 抽出された語が「こと」や「一種」などの特定の語であった場合は, 末尾の文節に係る最近接の文節の主辞をラベル候補語とする.

後方文字列一致を用いた方法では, 見出し語のみに注目する. 二つの見出し語間に Is-a 関係がある場合, 相対的に下位に位置する見出し語は, 上位に位置する見出し語を含む複合語で形成される場合が多いという性質を利用している. 例えば, 「エンジン」と「ディーゼルエンジン」は, 「エンジン」が上位, 「ディーゼルエンジン」が下位の Is-a 関係である. [玉川 2010]は, ある見出し語 A が「任意の文字列＋他の見出し語 B」によって表される場合, A を下位, B を上位として Is-a 関係を構築している. 本稿では, 「全ての見出し語およびパターンマッチにより推定されたラベル候補語」を判定対象リストとする. ある見出し語に対して後方文字列一致する語を判定対象リストから抽出し, その中で自身の文字数が最大の語をラベル候補語とする.

手順 2 では, 手順 1 で得られたラベルのみに注目し, Is-a 関係が成り立つラベルの組を人手で抽出する. 得られたラベルの組に対応するクラス間に Is-a 関係を付与し, ライトウェイトオントロジーを得る.

## 3. 実験方法

### 3.1 実験データ

本実験では自動車に関する専門用語辞書である*大車林*※1 [飯田 2003]を用いた. *大車林*には自動車に関連する用語(見出し語)とその説明文が記載されており, 見出し語は「エンジン」,

「シャシー」,「ドライブトレイン」,「ボディ」,「試験・性能・能力」,「設計開発」,「生産技術」,「環境・社会・法規」および「その他」の 9 つのカテゴリに分けられている. 本実験に用いたデータは, 紙の辞書である*大車林*に記載された見出し語とそれに対応するカテゴリおよび説明文を手作業で電子化したものである. ただし,「その他」カテゴリに含まれる見出し語は除外した. その結果, 実験データに含まれる見出し語は 6273 語, カテゴリは 8 種類となり, 各見出し語には一つずつカテゴリが付与されている.

まず, mecab※2 を用いて説明文を形態素解析した. 得られた結果に下記の 3 種類の前処理を施し, 7 種類のデータセット(表 1)を作成した.

・前処理 1:品詞選択
  - 条件 1:名詞
    品詞が名詞以外の単語を除外する
  - 条件 2:名詞・動詞・形容詞・副詞
    品詞が名詞・動詞・形容詞・副詞以外の単語を除外する

・前処理 2:単語の重み
  - 条件 1:Binary
    説明文中に出現する単語の重みを 1, それ以外の単語の重みを 0 とする
  - 条件 2:Count
    説明文中に $n$ 回出現する単語の重みを $n$ とする
  - 条件 3:TF-IDF
    単語の TF-IDF 値を重みとする

・前処理 3:フィルタリング
    単語を TF-IDF 値が高い順にランキングし, 下位 10%(条件 1), 20%(条件 2), 30%(条件 3)の単語を除外する

得られた 7 種類のデータセットからそれぞれ二部グラフを作成した.「見出し語」と「説明文中の単語」をノードとし,「単語の重み」をエッジの重みとした.

同様に 7 種類のデータセットから見出し語ベクトルデータを作成した. 説明文中の単語 $i$ の「単語の重み」をベクトルの $i$ 番目の要素の値とした.

### 3.2 クラス抽出

7 種類の二部グラフを, 岡本の方法によりモジュール分解した. 今回の実験では, $\alpha$ の範囲を 10^-3～1 とし, 各データセットに対して 20 回ずつモジュール分解を行った.

また, 比較方法として, クラスタリング方法である k-means, Ward 法および Gaussian Mixture Model(GMM)の 3 種類を見出し語ベクトルデータに適用し, クラスタリングを行った. これらのクラスタリング方法では, 見出し語ベクトルの位置関係に基づいて, 全見出し語ベクトルの集合を複数の部分集合に分割する. このとき得られた見出し語ベクトルの部分集合がクラスとなる. それぞれのクラスタリング方法の実装には scikit-learn※3 を用いた. 分割するクラスタ数は 6, 8, 10, 12, 14 および 21 の 6 種類を指定し, k-means および GMM は初期パラメータをランダムに変えて各データセットに対して 20 回ずつクラスタリングを行った.

### 3.3 評価指標

各方法によって得られたクラスを, データに付与された 8 種類のカテゴリと比較した. 比較指標は後述の Averaged Max F-score を用いた.

※1 *大車林*は株式会社三栄書房の登録商標です.
※2 http://taku910.github.io/mecab/
※3 http://scikit-learn.org/
※4 http://taku910.github.io/cabocha/

表 1 データセット ID と前処理

| | 前処理 | | |
|---|---|---|---|
| データセットID | 1:品詞選択 | 2:単語の重み | 3:フィルタリング |
| 1 | 名詞 | Binary | なし |
| 2 | 名詞 | Count | なし |
| 3 | 名詞 | TF-IDF | なし |
| 4 | 名詞・動詞・形容詞・副詞 | Count | なし |
| 5 | 名詞 | TF-IDF | 10% |
| 6 | 名詞 | TF-IDF | 20% |
| 7 | 名詞 | TF-IDF | 30% |

まず, モジュール抽出方法またはクラスタリング方法によって抽出された $k$ 番目のクラスを $T_k$, カテゴリ $n$ に含まれる見出し語集合を $C_n$ とする.

このとき, Averaged Max F-score は次のように定義される.

再現率: $R_{nk} = \dfrac{|C_n \cap T_k|}{|C_n|}$ , 精度: $P_{nk} = \dfrac{|C_n \cap T_k|}{|T_k|}$,

カテゴリ-クラス対ごとの F-score: $F_{nk} = \dfrac{2R_{nk}P_{nk}}{R_{nk}+P_{nk}}$,

Averaged Max F-score: $F = \sum_{k=1}^{K} \dfrac{|T_k|}{N} \max_n F_{nk}$

ただし$|\cdot|$は集合の要素数を, $N$は全見出し語数を, $K$は抽出されたクラス数を表す.

### 3.4 Is-a 関係構築

$\alpha$ の範囲を 10^-6～1 とした岡本の方法によって得られたクラス集合のうち, 最も下層(クラス数が最多)のクラス集合に対し, 2.3 の方法によって Is-a 関係を構築した. その後, ランダムサンプリングにより抽出した見出し語とラベルの関係を評価した. なお, 手順 1 におけるパターンマッチでは, 係り受け解析器として Cabocha※4 を用いた.

## 4. 結果

### 4.1 クラス抽出方法の評価結果

得られたクラスと, *大車林*で定義されている見出し語のカテゴリを Averaged Max F-score により比較した.

岡本の方法(二部グラフを用いた方法)と, k-means, Ward 法および GMM(ベクトル空間モデルを用いた方法)を ID:2 のデータセットに適用して得られたクラスの Averaged Max F-score を図 4 に示す. 二部グラフを用いた方法では, クラス数 6 の場合に最も高い値 0.68 を得た. 一方, ベクトル空間モデルを用いた方法ではクラス数 8 の場合に最も高い値 0.52 を得た. また, 同じクラス数での結果を比較すると, すべてのクラス数において二部グラフを用いた方法の方が高い値を得た.

また, 二部グラフを用いた方法を ID:1~7 のデータセットに適用し, 得られたクラス数が 6 の場合の Averaged Max F-score を図 5 に示す. ID:2 のデータセットで最も高い値が得られた. ただし, データセットの違いによる Averaged Max F-score の変化幅は大きくても 0.02 程度である.

### 4.2 ラベル推定方法の評価結果

3.4 で述べた最も下層のクラス集合として, 130 個のクラスを得た. 個々のクラスに対して 2.3 の手順 1 の方法でラベルを推定し,

105 種のラベルを得た. 推定された 105 種のラベルに人手で Is-a 関係を付与し, ライトウェイトオントロジーを構築した. 提案方法を用いることで, ベクトル空間モデルを用いた方法よりも高精度に, 6273 語の見出し語からなるオントロジーを構築できた.

また, ランダムサンプリングによって, 見出し語とラベルの関係を人手で評価したところ, 正しく Is-a 関係となっている見出し語は 53%であった. なお, このランダムサンプリングにおける標本の大きさは 538 語であり, 信頼度 95%かつ標本誤差 5%以下である.

## 5. 議論

図 4 の結果において二部グラフが有効に働いた実例として, 下記の二つの見出し語に注目する.

見出し語 1:スクエアエンジン

説明文 1:ピストンストロークとシリンダー内径が等しいエンジンをスクエアエンジンといい, 実用回転域の性能と高速回転域の性能の両立を狙ったエンジンに採用される場合が多い.

見出し語 2:ランキンサイクルエンジン

説明文 2:ランキンサイクルを作動原理とする外燃機関で, 蒸気ボイラー, 蒸気機関または蒸気タービン, 復水器, 給水ポンプからなる蒸気原動機.

この二つの見出し語はどちらも「エンジン」カテゴリに属するが, GMM の結果では異なるクラスに, 二部グラフを用いた結果では同じクラスに属していた. この結果は, 二つの説明文のベクトル表現ではベクトル同士の内積が 0 となり, 見出し語間の関連性を表現できなかったが, 二部グラフ表現では「エンジン」を表すノードと「蒸気機関」や「原動機」を表すノードのグラフ距離の近さにより, 関連性を表現できたことによるものと考えられる.

また, サンプリングした見出し語のうち約半数が, 正しく Is-a 関係となるラベルを付与されなかった. この要因の一つは, 二部グラフを用いた方法によって得られたクラスには, Is-a 関係に限らず, 様々な観点で関連のある見出し語が集まっていることである. 本実験で用いた二部グラフは, 見出し語とその説明文中に出現する単語をすべてエッジでつないでいるため, Is-a 関係を持つ見出し語だけでなく Part-of 関係あるいはその他の関係を持つ見出し語とのグラフ距離も近くなっている. この課題への対策として, 単語間の Is-a 関係に重み付けした二部グラフを作成し, モジュール分解することが考えられる. 例えば, 今回利用したパターンマッチの結果に基づいて, マッチした単語と見出し語間のエッジの重みを大きくした二部グラフを作成する.

さらに, 今後は Is-a 関係以外の関係を抽出する方法を提案していきたい. 例えば Part-of 関係をオントロジーに加えることができれば, 装置の物理的な構成を表現することができるようになる. それによって, ある部品を変更した際に影響が及ぶ範囲を示す, といった応用範囲が広がることが期待される. Is-a 関係以外の関係の抽出方法を確立することで, より実用に即したオントロジーの構築を目指す.

## 6. 結論

本稿では, 専門用語辞書から構築した「見出し語」ー「説明文中の単語」の二部グラフのモジュール分解に基づいてクラスを抽出する方法を提案した. さらに, パターンマッチおよび後方文字列一致に基づいてクラスのラベルを推定し, 人手でクラス間に Is-a 関係を付与することでライトウェイトオントロジーを構築した.

提案方法で得られたクラスは, *大車林*のカテゴリに対して 0.68 の Averaged Max F-score を得た. この値はベクトル空間モデルを用いた方法よりも高い. また, ラベルと見出し語の関係をランダムサンプリングによって評価した結果, 正しい Is-a 関係となっているものは 53%となった.

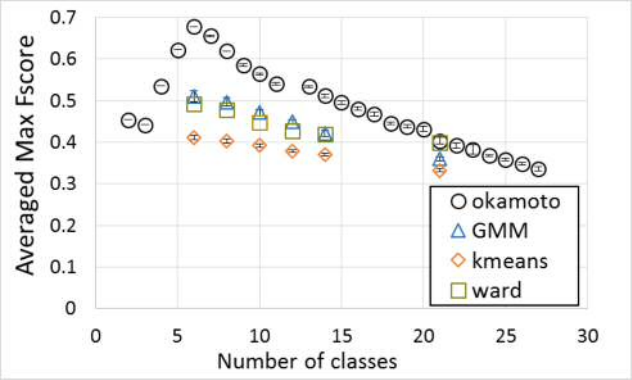


図 4 クラス抽出結果. okamoto は岡本の方法を表す. okamoto, GMM, k-means に付されたエラーバーは標準誤差を表す. (okamoto:N≦20, GMM, k-means:N=20)

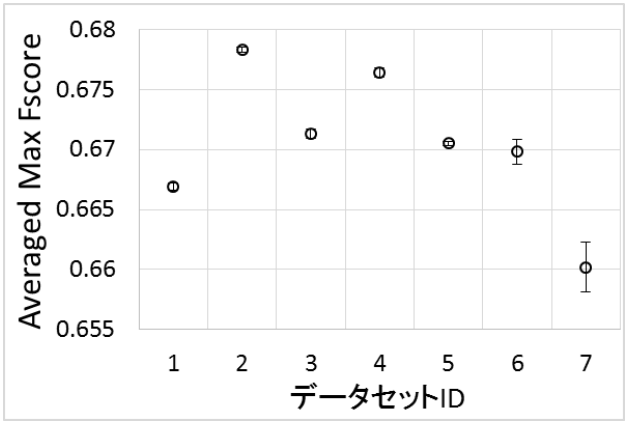


図 5 岡本の方法を各データセットに適用した結果. エラーバーは標準誤差を表す. (N=20)

今後は, 単語間の Is-a 関係に重み付けした二部グラフを利用することでラベルと見出し語が Is-a 関係となる精度を上げるとともに, Is-a 関係以外の抽出にも取り組んでいきたい.

## 参考文献


[Cimiano 2004] Cimiano P., Hotho A., and Staab S.: Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text, *Prestigious Applications of Intelligent Systems (PAIS 2004): Proceedings*, IOS Press, 110, 435, 2004.

[Jing 2010] Jing L., Ng M. K., and Huang J. Z.: Knowledge-based vector space model for text clustering, *Knowledge and information systems*, 25(1), 35-55, 2010.

[飯田 2003] 飯田 一: 大車林, 三栄書房, 2003.

[岡本 2014] 岡本 洋: マルコフ連鎖のモジュール分解:ネットワークからの重なりと階層構造を持つコミュニティの検出, JWEIN2014.

[柴木 2009] 柴木 優美, 永田 昌明, 山本和英: 日本語語彙大系を用いた Wikipedia からの汎用オントロジー構築, 情報処理学会研究報告, 自然言語処理研究会報告, NL-194(4), 1-8, 2009.

[玉川 2010] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, 25(5), 623-636, 2010.