3E1-R-6-7

単語間関係を利用した関連番組検索の検討

A Study of TV Program Recommendation Using Automatic Acquired Word Relations

山田 一郎[*1] 宮崎 勝[*1] 住吉 英樹[*1] 古宮 弘智[*1] 田中 英輝[*1]
Ichiro Yamada Masaru Miyazaki Hideki Sumiyoshi Hironori Furumiya Hideki Tanaka

[*1] NHK 放送技術研究所
Japan Broadcasting Corporation, Science and Technology Research Laboratory

This paper presents a novel method of calculating similarity between two TV programs by using summaries as a part of Electronic Program Guide (EPG). Most previous methods used statistics such as *tf-idf* based cosine measure of word vectors, whose words are appeared in the summaries. However these approaches were not effective for calculating similarity between TV programs because broadcast summaries are too short to obtain reliable statistics. Our method generates a graph structures whose vertexes are TV programs and words. These words are connected by word relations which are extracted from Web automatically. Similarity between two TV programs is calculated based on the relativeness of two TV program's vertexes in the graph structure. Through experiments, our method showed effectiveness of calculating similarities between two TV programs compared with the baseline approaches.

1. はじめに

NHK では、ブロードバンド回線を通じて放送した番組を配信する NHK オンデマンド [1] という動画サービスを行っている。NHK オンデマンドでは、見逃し番組、ニュース番組、そして過去に放送した番組である特選ライブラリーなど、2012 年 1 月現在で 7,000 本近い番組が配信され、今後も配信番組数の増強が予定されている。これらの多くの番組からユーザが嗜好に合った番組を選択できるように、NHK オンデマンドのインターフェースでは、選択した番組に関連する番組が提示される。ユーザは、提示された関連番組の中から一つの番組を選択し、さらには選択した番組に対して提示された関連番組を一つ選択するといった操作を繰り返すことにより、嗜好に合った番組を芋づる式に探し出すことができる。しかし、ユーザが選択する関連番組はその提示順位に大きく依存し、たとえ関連番組として提示されても下位の順位にある番組は選択されにくい(図 1)。ユーザの嗜好をより良く反映した番組検索を実現するためには、どのような関連番組を提示するか、また、提示する関連番組をどのようにランキングするかが重要となる。

これまで我々は、テレビの電子番組表(EPG)中の番組概要文に現れる単語を手掛かりとして、関連番組を提示する手法を提案している[Goto 2010]。しかし、この手法では単語表記の完全一致を手掛かりとしているため、類似性評価の際に問題が生じることがある。例えば、以下の 2 つの文は同じような内容を述べているが、表記が一致する名詞や動詞が出現しないため類似していないと判定されてしまう。

[文 1] 生活習慣病の治療のポイントを伝える。
[文 2] 高血圧を改善するための減塩や薬物療法の進め方など、視聴者からの疑問に答える。

もし、「高血圧」が「生活習慣病」の一つであり、「減塩」や「薬物療法」が「治療」の一つであることが分かれば、この 2 つの文は類似していると判断できるかもしれない。

そこで本稿では、単語間の関係(因果関係や上位下位関係など)を Web から抽出し、この単語間の関係を利用した文書間の類似性評価手法を提案する。提案手法では、2 つの文書に出現する名詞間を自動獲得した関係で結ぶことによりグラフ構造を生成し、ランダムウォークを用いてグラフ上のノード間の関連度スコアを算出することにより、2 つの文書の類似性を評価する。実験では NHK オンデマンドで実際に提示された関連番組を対象とした類似性評価を行い、提案手法の結果は従来手法と比較して人手による類似性評価結果に近いことを示す。

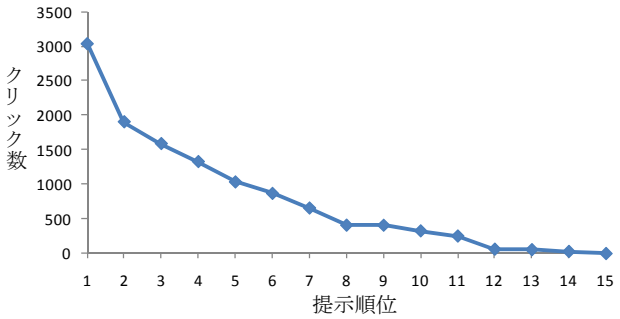

図 1. 関連番組の提示順位に対するユーザの選択数
(2010 年 9 月～2011 年 5 月における
NHK オンデマンドのクリック数)

2. 関連研究

従来、ユーザが選択した商品の関連商品を提示する推薦システムの研究は盛んに取り組まれている。これらの推薦システムで用いられるアプローチは、対象コンテンツの内容に基づいてユーザに商品を推薦する内容ベースフィルタリングと、類似する嗜好を持つユーザの情報を利用して商品を推薦する協調フィルタリングの 2 種類に分類できる。

内容ベースフィルタリングは従来の情報検索技術を利用するもので、コンテンツやユーザプロファイル、さらにはユーザが入力するクエリー間の類似度をもとに、商品をユーザに推薦する。Goto らは、放送番組の概要文に含まれる単語に対して Okapi-BM25[Robertson 1999]によって重み付けを行うことにより複数の番組間の類似性を評価し、ユーザが一つの番組を選択した際

連絡先：山田一郎，NHK 放送技術研究所，東京都世田谷区砧 1-10-11，03-5494-3377，yamada.i-hy@nhk.or.jp

[1] http://www.nhk-ondemand.jp/

に、類似する別の番組を推薦する手法を提案している[Goto 2010]。奥らは、時間帯や天気、同伴者、予算などのユーザの状況に応じて変化するユーザプロファイルをモデル化し、状況依存型の推薦システムを提案している[奥 2007]。

近年、推薦システムとして協調フィルタリングを利用した手法が実用システムとして効果を上げている。Amazon では、ユーザの購入履歴を基にアイテム間の類似性を評価する Item-to-Item Collaborative filtering を行い、「この商品をチェックした人はこんな商品もチェックしています」といった商品の推薦を行っている[Linden 2003]。Koren らは、ユーザと商品を行と列としてユーザの嗜好を数値化した行列を作り、この行列を特異値分解して欠損値(ユーザの嗜好が不明な商品に対する嗜好の評価値)を推定する Matrix Factorization を利用したアルゴリズムを提案し、NetFlix が主催したコンテストで優秀な成績を挙げている[Koren 2009]。

本稿では内容ベースフィルタリングを拡張し、商品に関するテキストに含まれる単語表記が一致しなくても適切に商品間の類似性を評価する手法を提案する。本手法は協調フィルタリングにおいても、ユーザの商品に対する嗜好を数値化した行列を作る処理などで有用である。また、内容ベースフィルタリングと協調フィルタリングの両方を用いるハイブリッドなアプローチも提案されており[Melville 2010]、本手法はこのようなアプローチにおける精度向上にも繋がると考えられる。

3. 単語間関係の獲得

本章では、提案する文書間の類似性評価手法で利用する単語間の関係の獲得手法について説明する。単語間の関係として、番組概要文の類似性評価で効果が期待できる以下の 4 種類の関係を利用する。

- 上位下位　例:「生活習慣病／高血圧」
- 原因結果　例:「かび／ニオイ」
- 名物　例:「六義園／しだれ桜」
- 材料　例:「ビール／麦」

また、「対象(entity)／属性名(attribute)／属性値(value)」という 3 項からなる属性関係を利用して、上記 4 種類以外の関係も獲得する。属性関係は、「対象」と「属性値」に該当する 2 つの単語が「属性名」という関係を持つと解釈できる。例えば「七人の侍／キャスト／三船敏郎」という属性関係では、「七人の侍」と「三船敏郎」の関係は「キャスト」と判断できる。

これらの関係獲得のために ALAGIN Forum[2] で公開されているツールを利用する。以下に、このツールで使われている各関係獲得手法の概略と獲得した関係の精度の調査結果を記す。

3.1 Wikipedia からの関係獲得

単語の上位下位と属性関係の獲得では、Wikipedia を利用した上位下位関係抽出ツール [3] を利用する。日本語 Wikipedia には現在約 80 万本の記事が存在する。上位下位関係抽出ツールは、この記事の階層的なレイアウト構造やカテゴリタグ、さらには記事中の第一文(見出し語の定義文)を利用する[隅田 2009]。例えば、「生活習慣病」を記事タイトルとするページには「高血圧」、「糖尿病」という単語が見出し語として存在する。Wikipedia の見出し語となっている単語(例えば「高血圧」)が、レイアウト構造で上位にある記事タイトルや見出し語(例えば「生活習慣病」)と上位下位関係にあるか否かを教師有りの機械学習で判定することにより、大量の上位下位関係を高精度に獲得することができる。

また、上位下位関係抽出ツールは記事の階層的なレイアウト構造を利用して属性関係を獲得することもできる。例えば、「七人の侍」を記事タイトルとするページには、その見出しに「キャスト」と「三船敏郎」が存在する。記事タイトル(例えば「七人の侍」)を「対象」、上位下位関係があると判定された単語対(例えば「キャスト」と「三船敏郎」)を「属性名」と「属性値」とすることにより、大量に属性関係を獲得することができる[山田 2012]。

3.2 Web テキストからの関係獲得

原因結果、名物、材料の関係の獲得では、ALAGIN フォーラムで公開されている意味的関係抽出サービス[De Saegar 2009]を利用する。このサービスは、少数のシードとなる単語ペアを入力として与えることにより、シードの単語ペアに類似した関係を持つ大量の単語ペアを約 6 億の Web ページに含まれるテキストから獲得できる。この処理では、単語の意味クラスと文脈パターンを利用する。例えば、「X が Y の原因となる」という文脈パターンに出現する単語 X と Y には原因結果の関係がある可能性が高い。さらに、X と Y が属する意味クラスを制限し、例えば、シードの単語ペアに「細菌／悪臭」がある場合は、「細菌」と同じ意味クラスに属する単語と「悪臭」と同じ意味クラスに属する単語のペア(例えば「カビ／におい」)にも同様の関係がある可能性が高い。この意味的関係抽出サービスでは、分布類似度を用いた手法[Kazama 2008]によって単語の意味クラスを自動獲得している。

意味的関係抽出サービスにより獲得される関係には明らかな誤りや曖昧な関係が含まれる。例えば原因結果を対象とした結果には「カビ／症状」という関係が含まれる。この関係は、カビが何の症状を引き起こすのか分からないため有用ではない。そこで除外する単語リストを手作業で作成し、意味的関係抽出サービスの出力に対して、このような不要な関係を除外する処理を行った。

3.3 獲得した単語間関係の精度・カバー率調査

上位下位関係抽出ツールを利用して、2007 年～2011 年の 5 年分の Wikipedia のダンプデータから上位下位関係と属性関係を獲得する処理を行った。また意味的関係抽出サービスでは、各関係数個程度の単語ペアのシードを入力とし、得られた結果の上位ペア(信頼性の高い処理結果)を、再度入力として意味的関係抽出サービスに与えることによって、原因結果、名物、材料の関係を獲得した。ツールにより獲得した単語間関係数とその精度を表 1 に示す。

表 1. 獲得した単語間関係数と精度

関係名	獲得関係数	精度
上位下位	8,591,469	90.0%*
属性	5,213,455	94.0%*
原因結果	77,636	75.0%
名物	183,093	49.0%
材料	49,711	73.0%

* 上位下位と属性に対する精度は文献からの引用

原因結果、名物、そして材料に対する関係獲得結果の精度は、獲得される関係からそれぞれ 200 ペアをランダムサンプルし、一人のアノテータによる判定で算出している。

獲得した全関係に出現する異なり単語数は 3,431,430 語であった。NHK オンデマンドでこれまでに公開された 25,769 個の番組概要文に出現する名詞(異なり数 94,456 語)を取り出し、こ

[2] http://alaginrc.nict.go.jp/

[3] http://alaginrc.nict.go.jp/hyponymy/index.html

れらの名詞に対して、獲得した全関係に出現する名詞のカバー率を調査した。結果を以下に記す。

獲得関係の単語カバー率　　47.7% (45,042/94,456)

自動獲得できる単語間関係は 6%～51.0%の誤りが含まれ、さらに実際に使われるテキストの半分程度の単語カバー率であった。しかし、異なり単語数が約 340 万語、獲得した全関係数が約 1,400 万と大規模なものであり、テキスト間の類似性評価における効果が期待できる。

4. 番組間の類似性評価

本章では、前章で獲得した単語間の関係を利用して NHK オンデマンドにおける番組概要文間の類似性を評価する提案手法を説明する。

4.1 番組概要文

NHK オンデマンドに登録されている番組概要文の平均文字数は 170 文字、含まれる平均名詞数は 26 語であった。本稿では、この番組概要文に含まれる名詞を手掛かりとして、番組間の類似性を評価する。

4.2 関連度スコア計算

2 つの番組の類似性を評価するため以下の手順で 2 つの番組を結ぶグラフを生成する。

1. 番組タイトルと、番組概要文に含まれる名詞をノードとし、番組タイトルのノードと各名詞のノードをエッジで結合。
2. 類似性比較対象の番組に対しても、1.と同様に番組タイトルのノードと各名詞のノードをエッジで結合。
3. 1. と 2.の名詞のノードを、前章で生成した単語間関係により連結。例えば、番組概要文に「生活習慣病」と「たばこ」という単語があり、「生活習慣病／高血圧」、「高血圧／喫煙」、「喫煙／たばこ」といった関係が獲得されている場合は、4 つの名詞ノードを順に結ぶエッジを生成。

生成したグラフに対して 2 つの番組がどの程度強く連結されているかを評価することによって、2 つの番組の類似性を評価する。この処理では、Web 検索などで使われているランダムウォークのアルゴリズムの一つの Green Measures[Yann 2007]を利用する。この手法では、一つのノードから別のノードへ遷移する確率を行列 M で表現し、(1)式で Green matrix を定義する。

$$G := \sum_{t=0}^{\infty}(M^t - M^{\infty}) \qquad (1)$$

ここで、M^t は t 回目のランダムウォークのステップにおける遷移行列を示す。行列 G の i 行 j 列の要素は、ノード i とノード j がどの程度関連するかを示す値と解釈できる。最終的な関連度スコアは、Green matrix を利用した(2)式により定義される。

$$S^i(j) := G_{ij}\log(1/\nu_j) \qquad (2)$$

ここで ν は equilibrium measure と呼ばれ、任意のベクトルμに対し無限に遷移を繰り返した後に収束するベクトル($\mu M^{\infty} = \nu$)を示す。(2)式の対数関数は、遷移を繰り返した後の最終状態としてノード j に収束する可能性が高い場合に対する補正を与え、情報検索などで用いられる *idf* 値と同じような役割を果たす。

式(2)を利用して 2 種類の類似性評価手法を提案する。1 つ目の手法では、$S^{p_1}(p_2)$ を直接用いることにより 2 つの番組 p_1, p_2 の類似度 $S_{dierct}(p_1, p_2)$を定義する。

$$S_{direct}(p_1, p_2) = S^{p_1}(p_2) \qquad (3)$$

ノードを結ぶエッジに与える重みは(4)式、(5)式の値とする。この値は、遷移確率を表す行列 M の各要素となる。

[番組から名詞へのエッジ]

$$e(p_i, n_j) = tf(n_j)idf(n_j)/Z_{p_i} \qquad (4)$$

[名詞間のエッジ]

$$e(n_i, n_j) = 1/Z_{n_i} \qquad (5)$$

ここで、*tf*(n)は名詞 n の番組 p における出現頻度、*idf*(n)は名詞 n の全番組における逆文書頻度、そして、Z_p、Z_n は、それぞれ、p、n から他のノードへのエッジの重みの合計を示す。

2 つ目の手法では、p_1 から p_2 へのパス上にある全ノードに与えられた $S^i(j)$の値の合計を利用し、2 つの番組 p_1, p_2 の類似度 $S_{related}(p_1, p_2)$ を(6)式で定義する。

$$S_{related}(p_1, p_2) = \sum_{v \in vertex(p_1, p_2)} S^{p_1}(v) \qquad (6)$$

ここで *vertex*(p_1, p_2)は、p_1から p_2 へのパス上にある全ノードを示す。この手法においても、ノードを結ぶエッジに与える重みは(4)式、(5)式の値とする。図 2 に 2 つの手法における関連度スコア計算の概略を示す。

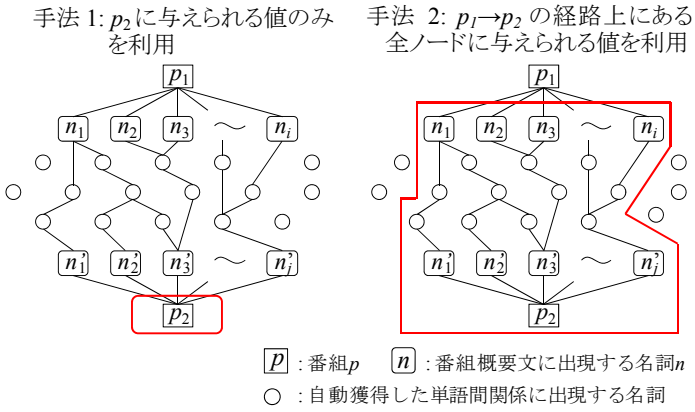

図 2. 2 つの手法における関連性スコア計算の概略

5. 実験

提案手法の有効性を示すために、NHK オンデマンドに登録されている番組を対象とした関連番組のリランキング実験を行った。まず、2010 年 9 月から 2011 年 5 月までに登録されていた 25,769 番組から、以下の制約のもとで 352 番組をランダムにサンプルした。

- 番組タイトルが同じ番組は取り出さない(例えば「NHK スペシャル」は 1 番組のみサンプル)
- 関連番組を 2 番組以上持つ

次に、NHK オンラインで提示された 352 番組の関連番組を対象として、筆者を含まない 3 名のアノテータにより、サンプルした番組とその関連番組間の類似性をランキングする作業を行った。各番組に関する関連番組は Okapi-BM25 を利用した Goto らの手法で抽出され、一つの番組に対して平均 10.4 個の関連番組が提示されていた。3 名のアノテータが付与したランキング結果は、その順位相関(Spearman's rank correlation)の平均が 0.565 であった。これは、一定の一致度であったと解釈できる。最終的に 3 名のアノテータが付けた類似性のランクを平均し、平均ランクの昇順に類似すると判断したデータを基準として、このデータと各手法によるランクを比較することにより評価する。

5.1 ベースライン手法

本節では、提案手法と比較する 3 つのベースライン手法を紹介する。

(1) Okapi-BM25 を利用した手法

Goto らの手法では文書 p_1 に対する文書 p_2 の類似性を、Okapi-BM25 の指標を利用した(7)式で評価する[Goto 2010]。

$$S_{BM}(p_1,p_2)=\sum_{n\in p_1} idf(n)\cdot\frac{tf_{p_2}(n)\cdot(k+1)}{tf_{p_2}(n)+k\cdot(1-b+\frac{|p_2|}{avgdl})}\cdot\frac{(k'+1)tf_{p_1}(n)}{k'+tf_{p_1}(n)} \quad (7)$$

ここで、$idf(n)$は単語 n の逆文書頻度、$|p_2|$は p_2 の文書長、$avgdl$ は平均文書長、k、k'、b はパラメータであり、k=3.0、k'=100.0、d=0.75 を使用している。

(2) *tf-idf* による手法

tf-idf による手法では、文書 p に出現する単語 n に対して(8)式の重みを与えて文書を単語のベクトルで表現する。

$$w_{TFIDF}(n)=tf_p(n)\cdot idf(n) \quad (8)$$

文書 p_1 に対する文書 p_2 の類似性は、2 つの文書のベクトル間のコサイン類似度により評価する。

(3) 単語間関係を利用した手法

自動獲得した単語間関係を用いて文書 p に出現する単語 n を拡張し、文書に出現する単語 $n\in p$ と、n と関係を持つ単語 n_{rel} を要素とするベクトルで文書 p を表現する。n に与える重みは(8)式、n_{rel} に与える重みは(9)式を用いる。

$$w_{rel}(n_{rel})=w_{TFIDF}(n)/N_{rel}(n) \quad (9)$$

ここで、$N_{rel}(n)$は n と関係を持つ単語数を示す。文書 p_1 に対する文書 p_2 の類似性は、2 つの文書のベクトル間のコサイン類似度により評価する。

5.2 評価実験

ランダムサンプルした 352 番組とその関連番組に対して、前節で説明した 3 つのベースライン手法と 4.2 節で説明した 2 つの提案手法を適用して関連番組のリランキング処理を行い、これらの結果とアノテータにより生成した基準データとの相関を、Spearman's rank correlation により評価した。結果を表 2 に示す。

経路上の全ノードに与えられる関連度スコアを利用した提案手法 2 の rank correlation が 0.425 と最も高く、この手法が他に比べて人手によるランキング結果に近いことが分かる。一方、番組を表すノード p_1 から p_2 への直接の関連度スコアを利用した提案手法 1 は良い結果が得られていない。Green matrix を利用した関連度スコアの値は、直接エッジで繋がれているノード間は大きい値となるが、間接的に繋がれているノード間では極端に小さな値となっていた。そのため、直接繋がれているノード(2 つの番組に共通する単語)のみに影響を受け、従来手法とほぼ変わらない結果となってしまったと考えられる。

表 2. 各手法の評価結果

手法	rank correlation
ベースライン 1 (Okapi-BM25)	0.370
ベースライン 2 (*tf-idf*)	0.350
ベースライン 3 (単語間関係)	0.371
提案手法 1 ($S^{p_1}(p_2)$ を直接利用)	0.351
提案手法 2 (経路上の全ノード利用)	<u>0.425</u>

6. おわりに

本稿では、Wikipedia や Web テキストから自動獲得した単語間の関係を用いることにより、2 つの文書間の類似性を評価する手法を提案した。提案手法では、2 つの文書に出現する名詞間を自動獲得した関係で結んだグラフ構造を生成し、ランダムウォークによりグラフ上のノード間の関連度スコアを算出することにより、2 つの文書の類似性を評価した。評価実験では、2 つの文書を繋ぐ経路上の全ノードの関連度スコアを利用した手法による結果が、従来手法と比較して人手によるランキング結果に近いことを示した。

実験で利用した単語間の関係には誤りが含まれる。また現状では、番組概要文に出現する名詞の半分弱しかカバーできていない。今後、関係数を増加させ、さらには人手により関係のチェックを行い、文書間類似性評価処理の検証を進める予定である。

参考文献

[De Saegar 2009] Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda and Masaki Murata: Large Scale Relation Acquisition using Class Dependent Patterns, In Proceedings of the IEEE International Conference on Data Mining (ICDM'09), pp.764-769, 2009.

[Goto 2010] J. Goto, H. Sumiyoshi, M. Miyazaki, H. Tanaka, M. Shibata, and A. Aizawa: Relevant TV Program Retrieval using Broadcast Summaries, Proceedings of ACM on Intelligent User Interfaces(IUI), pp.411-412, 2010

[Kazama 2008] Jun'ichi Kazama and Kentaro Torisawa: Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations, In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), pp. 407-415, 2008.

[Koren 2009] Y. Koren, R. Bell and C. Volinsky: Matrix Factorization Techiques for Recommender Systems, IEEE Computer, pp.42-49,2009.

[Linden 2003] Greg Linden, Brent Smith, and Jeremy York: Amazon.com recommendations, IEEE Internet Comput., vol7, no.1, pp. 76-80, 2003.

[Melville 2010] P. Melville, V. Sindhwani: Recommender Systems, Encyclopedia of Machine Learning, Springer, 2010.

[Robertson 1999] S. Robertson and S. Walker: Okapi/ Keenbow at TREC-8, In Proceedings of TREC-8, pp151-162, 1999.

[Yann 2007] Ollivier Yann, Senellart Pierre: Finding Related Pages Using Green Measures: An Illustration with Wikipedia, Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, pp.1427-1433, 2007.

[奥 2007] 奥健太,中島伸介,宮崎純,植村俊亮: 状況依存型ユーザ嗜好モデリングに基づく Context-Aware 情報推薦システム, 情報処理学会論文誌. データベース, Vol. 48, SIG11 (TOD_34), pp. 162-176, 2007.

[隅田 2009] 隅田飛鳥, 吉永直樹, 鳥澤健太郎: Wikipedia の記事構造からの上位下位関係抽出, 自然言語処理, vol.16(3), pp. 3-24, 2009.

[山田 2012] 山田一郎, 橋本力, 呉鍾勲, 鳥澤健太郎, 黒田航, Stijn De Saeger, 土田正明, 風間淳一: Wikipedia を利用した上位下位関係の詳細化, 自然言語処理, Vol.19, No.1, pp. 3-23, 2012.