2009－CVIM－166 (16)
2009／3／13

αダイバージェンスに基づく単語連想と文書分類への適用

別所 克人† 内山 俊郎† 内山 匡†

†日本電信電話株式会社 ＮＴＴサイバーソリューション研究所

〒239-0847 神奈川県横須賀市光の丘1-1

E-mail: †{bessho.katsuji, uchiyama.toshio, uchiyama.tadasu}@lab.ntt.co.jp

あらまし 単語と、単語に付随する意味属性とが、コーパス中で共起する頻度を算出することにより得られる共起ベクトルは、単語間の意味的類似性を反映する性質をもつ。本稿では、共起ベクトル間の距離尺度として、カルバック・ライブラー距離を拡張した距離尺度であるαダイバージェンスを適用することにより、様々なレベルの上位・下位・兄弟概念の単語が連想されることを報告する。また、共起ベクトル間のαダイバージェンスの文書分類への適用について述べる。

キーワード 共起ベクトル，αダイバージェンス，単語連想，文書分類

Word Association based on α-divergence and Application to Document Classifcation

Katsuji BESSHO† Toshio UCHIYAMA† and Tadasu Uchiyama†

† NTT Cyber Solutions Laboratories, NTT Corporation

1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan

E-mail: †{bessho.katsuji, uchiyama.toshio, uchiyama.tadasu}@lab.ntt.co.jp

Abstract The co-occurrence vectors that are the co-occurrence frequencies between words and semantic attributes in a corpus reflect the semantic similarities between words. This paper shows that words of superordinate or subordinate or sibling concept at various levels are associated by applying α-divergence that is the distance measure to which Kullback-Leibler Metric is expanded as the distance measure between co-occurrence vectors. Moreover, it describes the application of α-divergence between co-occurrence vectors to the document classification.

Keyword Co-occurrence Vector, α-Divergence, Word Association, Document Classification

1. はじめに

単語と、単語に付随する意味属性とのコーパス中における共起により、単語の意味表現としての共起ベクトルや概念ベクトルを生成する手法がある[1]。単語間の意味的類似性は、対応するベクトル間の類似度として定量化できるので、様々な言語処理に活用が可能となる。[2]においては、この共起ベクトル間の距離尺度として、カルバック・ライブラー距離をとることにより、一つの固定した単語との距離の小さい単語群が、該単語の上位概念や下位概念に相当する傾向があることを示した。

しかし、単語の上位・下位概念といっても、より詳細なレベルによる違いがある。本稿では、共起ベクトル間の距離尺度として、カルバック・ライブラー距離を拡張した距離尺度であるαダイバージェンスを用い、パラメータであるαの値を変化させることにより、様々なレベルの上位・下位・兄弟概念の単語が連想されることを示す。

また、αダイバージェンスに基づく単語間関連度を用いる文書分類手法としてαダイバージェンス法を提案する。単語とその概念ベクトルの対の集合を概念ベースと呼ぶが、[3]においては、その概念ベースを用いた文書分類手法である概念ベース法を提案している。概念ベース法にαダイバージェンス法を加味することにより、精度が向上することを示す。

以下、2節で共起ベクトル生成について説明し、3節でαダイバージェンスによる単語連想について説明する。4節では、αダイバージェンスを用いた文書分類手法とその評価実験について述べ、5節でまとめを述べる。

2. 共起ベクトル生成

共起ベクトル生成では、コーパスにおける単語と、単語に付随する意味属性との共起頻度をとる。この意

味属性とは、日本語語彙大系[4]における一般名詞意味体系の意味属性のことである。

日本語語彙大系における一般名詞意味体系は、名詞と用言の意味を体系立てたシソーラスであり、各ノードを意味属性と呼ぶ。このシソーラスは12階層あり、2715個のノードからなる。

本手法では、形態素解析プログラムとして JTAG[5]を用いる。JTAGが参照する単語辞書では、各名詞と用言に意味属性が付与されている。形態素解析結果において、各単語には、対応する意味属性の情報が付随している。

共起ベクトル生成では、まず、コーパスを用意し形態素解析する。名詞、用言等の内容語のみを残す（本稿において、用言は終止形を使用する）。任意の単語と、単語に付随する（第1番目の）意味属性とが1文中に共起する頻度をカウントし、各行が単語に対応し、各列が意味属性に対応しているような共起行列を作成する。単語と意味属性との共起のカウントの仕方には、単語と該単語内部に含まれる意味属性との共起をカウントしない自己非共起方式と、カウントする自己共起方式がある。

共起行列の各行ベクトルは、対応する単語の共起パターンを表しており、この行ベクトルを対応する単語の共起ベクトルと呼ぶ。意味的に近い2単語は、同じ意味属性と共起する傾向があるので、それらの共起ベクトルは近くなる。

日本語語彙大系は、単語間のあらゆる兄弟関係や上位・下位関係を網羅しているわけではない。しかし、類義語が一つの意味属性に分類されているため、共起ベクトルが、単語・単語間共起をとったときよりも、高品質なものとなる。単語を共起ベクトルで表現することにより、直接、1文や1文書に共起していない単語対を含め、任意の単語対の類似性を定量化することが可能となる。また、導出される単語間の意味的類似性は、使用したコーパスの内容を反映したものとなり、意味的類似性を動的に抽出できる。

3. αダイバージェンスによる単語連想

[2]で、共起ベクトル間の距離尺度としてカルバック・ライブラー距離をとることにより、上位・下位概念の単語が連想されることを示した。一方、[6]においては、確率分布間の距離尺度としてαダイバージェンスが述べられている。これはパラメータαをもち、αの値によって、距離の小さい確率分布間の包含関係が規定される。この性質に基づき、共起ベクトルにαダイバージェンスを適用することにより、より詳細なレベルで、単語間の上位・下位関係を導出できる可能性が考えられる。

このため、共起ベクトルを確率分布ベクトルに変換する。具体的には共起ベクトル $(a_1, a_2, \cdots, a_n)\,(a_i \geq 0 \quad (1 \leq i \leq n))$ を

$(x_1, x_2, \cdots, x_n)$ 但し $x_i = a_i \Big/ \left(\sum_{1 \leq f \leq n} a_f \right) \quad (1 \leq i \leq n)$

に変換する。

単語の対 X, Y に対し、X, Y のベクトル $v(X)$, $v(Y)$ が、

$v(X) := (x_1, x_2, \cdots, x_n)$, $v(Y) := (y_1, y_2, \cdots, y_n)$

のようになっているとき、X から Y へのαダイバージェンス $P_\alpha(X \| Y)$ を、

$$P_\alpha(X \| Y) = \frac{1 - \sum_{1 \leq i \leq n} x_i^{\alpha} y_i^{1-\alpha}}{\alpha(1-\alpha)} \quad \alpha \neq 0, 1$$

として算出する。ここで、$P_\alpha(X \| Y)$ の値を常に有限値にするため、$(1/0) = F \ (>0)$（F：定数）と定義し、

$\alpha < 0$ かつ $x_i = 0$ のとき、$x_i^{\alpha} = F^{-\alpha}$

$1 - \alpha < 0$ かつ $y_i = 0$ のとき、$y_i^{1-\alpha} = F^{-(1-\alpha)}$

として算出する。

特定のαに対し、$P_\alpha(X \| Y)$ は以下のように表される。

$$P_{-1}(X \| Y) = \frac{1}{2} \sum_{1 \leq i \leq n} \frac{(y_i - x_i)^2}{x_i}$$

$$\lim_{\alpha \to 0} P_\alpha(X \| Y) = KL(Y \| X)$$

$$P_{0.5}(X \| Y) = 2 \sum_{1 \leq i \leq n} \left(\sqrt{x_i} - \sqrt{y_i} \right)^2$$

$$\lim_{\alpha \to 1} P_\alpha(X \| Y) = KL(X \| Y)$$

$$P_2(X \| Y) = \frac{1}{2} \sum_{1 \leq i \leq n} \frac{(x_i - y_i)^2}{y_i}$$

上記で、$KL(X \| Y)$ は、X から Y へのカルバック・

ライブラー距離であり、以下の式で表される。

$$KL(X\|Y)=\sum_{1\le i\le n} x_i \log\frac{x_i}{y_i}$$

このように、αダイバージェンスはカルバック・ライブラー距離を拡張したものである。便宜上、$P_0(X\|Y)=KL(Y\|X)$, $P_1(X\|Y)=KL(X\|Y)$とおく。

$v(X)$を固定したときに$P_\alpha(X\|Y)$を小さくする$v(Y)$は、αが小さい場合は$v(X)$に包含される傾向があり、逆にαが大きい場合は$v(X)$を包含する傾向がある。α=0.5 のときは、$v(X)$とほぼ一致する傾向がある。

一般に、下位概念の単語と共起する意味属性とは、その上位概念の単語も共起する傾向があると考えられる。したがって、下位概念の共起ベクトルは、その上位概念の共起ベクトルに包含される傾向がある。また、兄弟概念の共起ベクトルは、一致する傾向がある。ゆえに、Xを固定したときに$P_\alpha(X\|Y)$を小さくするYは、αが小さい場合はXの下位概念である傾向があり、逆にαが大きい場合はXの上位概念である傾向がある。α=0.5 のときは、Xの兄弟概念である傾向がある。

このことを検証するために、特定の単語からの連想がどのような様相となるかを見る。共起行列は、2,871,343 個の Q&A 文書を入力コーパスとして、自己非共起方式により生成した。単語数は 382,025 である。$F=10^{24}$, $\log(1/0)=100$とし、Xを「徳川家康」としたとき、α=0, 0.5, 1, 2 のときの、単語Yのランキングは、表 1, 2 のようになった。α=0 のときは、徳川家康の家臣である「鳥居元忠」といった「徳川家康」の下位概念がランクされている。α=0.5 のときは、「織田信長」といった「徳川家康」の兄弟概念がランクされている。α=1 のときは、「江戸時代」といった「徳川家康」の上位概念で、比較的下位レベルの単語がランクされている。α=2 のときは、「強い」といった「徳川家康」の上位概念で、比較的上位レベルの単語がランクされている。このように、αの値に応じたレベルの概念の単語を連想させることができる。αの値を大きくしていくに従い、下位概念から兄弟概念を経て上位概念まで、連想の様相を変化させていくことができる。

表 1 X:「徳川家康」からの連想（α=0, 0.5）

順位	Y	$P_0(X\|Y)$	Y	$P_{0.5}(X\|Y)$
1	徳川家康	0.000000	徳川家康	0.000000
2	稔人	1.525293	織田信長	0.265888
3	光孝	1.549081	家康	0.270975
4	鳥居元忠	1.554235	徳川	0.292918
5	守重	1.586084	秀吉	0.297556
6	龍造寺氏	1.609280	豊臣秀吉	0.300399
7	としひと	1.637266	将軍	0.335583
8	幸隆	1.658294	信長	0.345957
9	前漢	1.670602	大名	0.366035
10	多羅尾	1.678140	戦国時代	0.369774

表 2 X:「徳川家康」からの連想（α=1, 2）

順位	Y	$P_1(X\|Y)$	Y	$P_2(X\|Y)$
1	徳川家康	0.000000	徳川家康	0.000000
2	江戸時代	0.601364	時代	1.577685
3	江戸	0.624708	有名	1.785757
4	歴史	0.672570	日本	2.353522
5	人々	0.693871	存在	2.963021
6	現代	0.725923	与える	3.659257
7	時代	0.728738	もつ	3.907280
8	人物	0.737904	強い	3.984850
9	説	0.739020	いる	4.027017
10	有名	0.743940	知る	4.193812

4. 文書分類への適用

提案するαダイバージェンスを用いた文書分類手法（αダイバージェンス法）は、概念ベースを用いた文書分類手法である概念ベース法と組み合わせて用いるので、まず、概念ベース法について説明した後、αダイバージェンス法について説明する。

4.1. 概念ベース法

[3]で提案した概念ベース法では、概念ベースを用いるが、これは共起行列から以下のようにして生成する。共起行列の各成分を品質向上のため平方根に変換した後、共起行列を特異値分解し、共起行列を、列数が縮約された行列に変換する。変換後の行列の各行ベクトルを長さ 1 に正規化したものが、対応する単語の概念ベクトルである。単語とその概念ベクトルの対の集合が概念ベースである。

概念ベース法では、任意の文書Dを以下のように概念ベクトルで表現する。Dを形態素解析し、名詞、用言等の内容語のみを残す。残った異なり単語の集合を$\{L_b | 1\le b\le c_D\}$とする。L_bのDにおける頻度をTF_bとし、L_bの概念ベクトルをv_{L_b}としたとき、

$$\sum_{1\le b\le c_D} TF_b \bullet v_{L_b}$$

の長さを1に正規化したものを、Dの概念ベクトルv_Dとする。

概念ベース法では、分類先のカテゴリの集合を$\{C_j \mid 1 \le j \le p\}$としたとき、各$C_j$の正例文書群$\{D_{jk} \mid 1 \le k \le q_j\}$を用意し、これから以下のようにして学習を行う。任意の1カテゴリC_jの正例文書概念ベクトル$v_{D_{jk}}$（kが異なれば別物とする）の集合をウォード法によりクラスタリングする。その結果得られたクラスタ集合を$\{S_{jm} \mid 1 \le m \le r_j\}$とする。また、クラスタ$S_{jm}$に属する文書集合を$\{D_{jmu} \mid 1 \le u \le t_{jm}\}$とする。クラスタ$S_{jm}$の特徴ベクトル$v_{S_{jm}}$を、$\sum_{1 \le u \le t_{jm}} v_{D_{jmu}}$を長さ1に正規化したものとして算出する。

分類時は、分類対象文書DとC_jとの関連度$CSIM(D, C_j)$を、

$$CSIM(D, C_j) := \max_{1 \le m \le r_j} CSIM(D, S_{jm}) := \max_{1 \le m \le r_j} (0.5 \times (v_D \bullet v_{S_{jm}}) + 0.5)$$

として算出する。$CSIM(D, C_j)$が最も大きいC_jを分類結果とする。

4.2. αダイバージェンス法

概念ベース法は、分類対象文書全体の意味を考慮して分類を行う。これに対し、αダイバージェンスは単語間の関係性を捉えるのに有効な尺度であるので、αダイバージェンス法では、分類対象文書中に、あるカテゴリに相当する単語が出現すれば、該カテゴリが反応するような方式をとる。例えば「歴史」というカテゴリがあった場合、分類対象文書中に、「歴史」を上位概念とする単語「徳川家康」が1回でも出現すれば、カテゴリ「歴史」が強く反応するようにする。概念ベース法は分類対象文書全体の意味を考慮するがために、本来対応するカテゴリと無関係な単語の存在により、該カテゴリの関連度がやや低く出ることがあり、それを、特定の単語だけに感応して分類を行うαダイバージェンス法で補正することを狙いとする。

この方式を実現するため、αダイバージェンス法では、まずαダイバージェンスに基づく単語間関連度を算出する。次に、学習データの学習として、概念ベース法で得られた各カテゴリのクラスタごとに、単語間関連度を用いて該クラスタの代表単語を認定する。分類時は、単語間関連度を用いて、分類対象文書と代表単語との関連度を求める。以下、各処理について説明する。

4.2.1. 単語間関連度算出

単語Xを固定し、$P_\alpha(X \| Y)$の昇順に単語Yをランキングしたとき、単語Xごとに、ランキング上位における$P_\alpha(X \| Y)$のスケールは異なる。例えば表3は、単語Xを「精神病」とし、αを1としたときの、単語Yのランキングである。同じαでも、表2の単語Xを「徳川家康」としたときの単語Yのランキングと比べ、ランキング上位におけるαダイバージェンスのスケールが異なっている。

表3 X：「精神病」からの連想（α=1）

順位	Y	$P_1(X \| Y)$
1	精神病	0.000000
2	精神	0.205408
3	病気	0.214300
4	精神的	0.293497
5	鬱病	0.304062
6	苦しむ	0.324298
7	障害	0.329706
8	患者	0.341999
9	医師	0.360097
10	医者	0.375174

しかし、ランキング上位は、Xごとのαダイバージェンスのスケールの違いに関わらず、αに応じた概念レベルの単語が常に占める傾向がある。このため、ランキングにおける順位に基づいて、XからYへの関連度を表すこととする。

単語Xを固定し、$P_\alpha(X \| Y)$の昇順に単語Yをランキングしたとき、単語Yの順位がeであれば、XからYへの関連度$E_\alpha(X \| Y)$を、

$$E_\alpha(X \| Y) = 1/e$$

と定義する。

パラメータαを固定したとき、単語Xに対し$E_\alpha(X \| Y)$が小さくなるのは、特定レベルの概念の単

語である。様々なレベルの概念の単語が導出されるようにするため、複数のパラメータのもとでの関連度を以下のように合成する。

パラメータ $\alpha_1,\alpha_2,\cdots,\alpha_h$ に対し、任意の単語 X から任意の単語 Y への関連度 $E_{\alpha_g}(X\|Y)\ \ (1\le g\le h)$ を算出する。X から Y への合成後の関連度 $E(X\|Y)$ を、以下のように、$E_{\alpha_g}(X\|Y)$ の最大値として算出する。

$$E(X\|Y)=\max_{1\le g\le h}E_{\alpha_g}(X\|Y)$$

$E(X\|Y)$ は、単語 X を固定したときに、あるパラメータのもとでは関連度が低くても、他のパラメータのもとでは関連度が高い単語 Y について高くなる。$E(X\|Y)$ は、各パラメータ α_g の連想の傾向を兼ね備えた関連度となり、様々なレベルの概念の単語を導出できる。

4.2.2. 代表単語認定

α ダイバージェンス法では、概念ベース法で得られた各カテゴリのクラスタ集合を学習データとして用いる。各クラスタの代表単語を以下のようにして求める。学習データの各種記号は、4.1 節と同じものを用いる。

正例文書 D_{jmu} の内容語の異なりの集合を $\{X_{jmuv}\,|\,1\le v\le z_{jmu}\}$ とする。

i)正例文書 D_{jmu} と単語 Y の関連度 $E(D_{jmu},Y)$ を、以下の式により算出する。

$$E(D_{jmu},Y)=\max_{1\le v\le z_{jmu}}E(X_{jmuv},Y)$$

ii)クラスタ S_{jm} と単語 Y の関連度 $E(S_{jm},Y)$ を、以下の式により算出する。

$$E(S_{jm},Y)=\left[\sum_{1\le u\le t_{jm}}E(D_{jmu},Y)\right]\Big/t_{jm}$$

iii)クラスタ S_{jm} との関連度から、C_j 以外のカテゴリのクラスタとの関連度の最大値を引いた値

$$E(S_{jm},Y)-\max_{j'\ne j,1\le m'\le r_{j'}}E(S_{j'm'},Y)$$

が最も大きくなる単語 Y を、クラスタ S_{jm} の代表単語とする。

i)により、正例文書 D_{jmu} 中に一つでも、単語 Y への関連度が高い単語があれば、D_{jmu} から Y への関連度が高くなるようになる。ii)によりクラスタ S_{jm} 中の正例文書と平均的に高い単語 Y が特定できる。iii)により、クラスタ S_{jm} に特徴的で、なおかつ、他のカテゴリのクラスタと差異化できる単語 Y を特定できる。C_j 内の他のクラスタ $S_{jm'}$ を考慮しないのは、S_{jm} と $S_{jm'}$ にとって関連度の高い単語が同一であったときに、そのような単語が排除されるのを避けるためである。

カテゴリ C_j に対して、その各クラスタの代表単語の異なりの集合を $\{Y_{jw}\,|\,1\le w\le a_j\}$ とする。

4.2.3. 文書分類

分類対象文書 D の内容語の異なりの集合を $\{L_b\,|\,1\le b\le c_D\}$ とする。D とカテゴリ C_j との関連度 $ASIM(D,C_j)$ を、

$$\begin{aligned}ASIM(D,C_j)&:=\max_{1\le w\le a_j}E(D,Y_{jw})\\&:=\max_{1\le w\le a_j}\max_{1\le b\le c_D}E(L_b,Y_{jw})\end{aligned}$$

として算出する。

クラスタ S_{jm} の代表単語を Y_{jw} としたとき、$E(S_{jm},Y_{jw})$ は大きく、従って $E(D_{jmu},Y_{jw})$ も大きい。一方、別カテゴリ $C_{j'}$ のクラスタ $S_{j'm'}$ の代表単語を $Y_{j'w'}$ としたとき、$E(S_{jm},Y_{j'w'})$ は小さく、従って $E(D_{jmu},Y_{j'w'})$ も小さい。

このことから、分類対象文書 D の内容がクラスタ S_{jm} に相当するならば、$E(D,Y_{jw})$ は大きくなり、$ASIM(D,C_j)$ も大きくなる。一方、別カテゴリ $C_{j'}$ の任意のクラスタ $S_{j'm'}$ に対し $E(D,Y_{j'w'})$ は小さくなるので、$ASIM(D,C_{j'})$ も小さくなる。

分類対象文書 D とカテゴリ C_j との最終的な関連度 $SIM(D,C_j)$ を、以下のように、概念ベース法での関連度とαダイバージェンス法での関連度との線形結合により算出する。$SIM(D,C_j)$ が最も大きい C_j を分類結果とする。

$$SIM(D,C_j) := \beta \bullet CSIM(D,C_j) + \gamma \bullet ASIM(D,C_j)$$

4.3. 評価実験

提案手法の有効性の検証は、3 節で述べたのと同じ共起行列を用いた。この共起行列から単語数 382,025、次元数 2,000 の概念ベースを生成した。単語間関連度算出では、連想先の単語 Y を、コーパスにおける出現頻度上位 10,000 個に限定し、α=0.5, 1.0, 1.5 に対し $F=10^{24}$, $\log(1/0)=100$ とした上で $E_\alpha(X\|Y)$ を算出した。これらから $E(X\|Y)$ を導出した。

分類の評価用データとして、14 カテゴリにわたる 8,211 個のブログ文書を用いた。カテゴリごとに文書群を 4 分割し、3/4 を学習データとし、1/4 をテストデータとする 4 つのデータを作成し、交差検定を行った。

概念ベース法におけるカテゴリごとの正例文書群のクラスタリングでは、クラスタ間距離の最小値が 2.5 以上になった時点で、クラスタリングを停止した。この結果、1 カテゴリあたりのクラスタ数の 4 データの平均は 13.2 個となった。

αダイバージェンス法では、コーパスにおける出現頻度上位 10,000 個の単語の中から、各クラスタの代表単語を認定した。

概念ベース法による分類と、線形結合の係数をβ=1.0, γ=0.02 としたときの提案手法による分類を実行した。1 カテゴリあたりの再現率、適合率の 4 データの平均と、1 分類対象文書あたりの分類処理時間の 4 データの平均は、表 4 のようになった。ここで分類処理時間は、分類対象文書の形態素解析をし、内容語を抽出した後の時間である。

表 4　分類精度，分類処理時間／文書

分類手法	再現率	適合率	処理時間
概念ベース法	81.37%	82.41%	6.54msec
提案手法	81.42%	83.62%	9.75msec

処理時間が 3.21msec 増えたものの、再現率が 0.05 ポイント、適合率が 1.21 ポイント上昇した。概念ベース法で誤り、提案手法で正しく分類された、カテゴリ「フード・グルメ」が正解のある分類対象文書の、両手法での関連度上位 3 位までのカテゴリとその関連度は、表 5 のとおりであった。

表 5　関連度上位 3 位までのカテゴリとその関連度

	概念ベース法		提案手法	
	カテゴリ	関連度	カテゴリ	関連度
1	旅行・イベント	0.775150	フード・グルメ	0.780778
2	フード・グルメ	0.774111	旅行・イベント	0.775489
3	ビューティー	0.650834	ビューティー	0.651078

概念ベース法では、僅差で正解カテゴリの関連度が不正解カテゴリである「旅行・イベント」の関連度より低くなっている。一方、αダイバージェンス法では、分類対象文書中の単語と、各カテゴリの代表単語との関連度が最大となった組合せは、分類対象文書中の単語「ビール」と、カテゴリ「フード・グルメ」の代表単語「酒」であった。提案手法では、分類対象文書中に、正解カテゴリの代表単語が強く反応する単語が存在したために、正解カテゴリの関連度が上回るように補正がなされた。

5. おわりに

本稿では、共起ベクトル間の距離尺度としてαダイバージェンスを用いることにより、様々なレベルの概念の単語が連想され、また、αダイバージェンスを文書分類に適用することにより、精度が向上することを示した。今後、提案方式を用いて、大量データによる学習と分類の検証を進めていきたい。

文　献

[1] 別所克人，内山俊郎，内山匡，片岡良治，奥雅博，“単語・意味属性間共起に基づくコーパス概念ベースの生成方式，”情報処理学会論文誌，Vol.49, No.12, pp.3997-4006, Dec.2008.

[2] 別所克人，内山俊郎，片岡良治，“単語・意味属性間共起に基づく単語間の階層関係の抽出，”信学技報，Vol.NLC2006-92, pp.31-36, Jan.2007.

[3] 別所克人，内山俊郎，内山匡，“学習データのクラスタリングを用いた文書分類，”信学技報，Vol.OIS2008, Mar.2009.

[4] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦，“日本語語彙大系，”岩波書店，1997.

[5] T. Fuchi, and S. Takagi, “Japanese Morphological Analyzer using Word Co-occurrence-JTAG,” COLING-ACL, pp.409-413, 1998.

[6] T. Minka, “Divergence measures and message passing, ”Technical Report MSR-TR-2005-173, Microsoft Research, 2005.