トピックモデルにおけるトピック表現語導出

嘉村　亘太† 　黄　 宏軒†† 　川越　恭二††

† 立命館大学大学院 理工学研究科　〒 525–8577 滋賀県草津市野路東 1-1-1
†† 立命館大学 情報理工学部　〒 525–8577 滋賀県草津市野路東 1-1-1
E-mail: †kamura@coms.ics.ritsumei.ac.jp, ††huang@fc.ritsumei.ac.jp, †††kawagoe@is.ritsumei.ac.jp

あらまし　近年，テキスト形式のデータを取り扱う手法として確率的トピックモデルの有効性が注目されている．トピックとは，単語出現頻度を変動させる (潜在的な) 要因であり，単語生成分布で表現され，特徴や意味については人間が解釈する必要がある．しかし，すべてのトピックを人間が解釈するのは困難かつ主観的である．そこで本稿では，トピックを客観的に解釈可能とするトピック表現語を導出する手法を提案する．まず，単語の組合わせによる総当たり導出手法について述べたのち，計算コスト削減を目的とした，経験則を用いた経験的導出手法について述べ，総当たり導出手法の結果比較によって評価する．

キーワード　トピックモデル，テキストマイニング，潜在的ディリクレ配分法

Determination of Topic Description Terms in Topic Model

Kota KAMURA†, Hung-Hsuan HUANG††, and Kyoji KAWAGOE††

† Graduate School of Science and Engineering, Ritsumeikan University
Nojihigashi 1—1—1, Kusatsu, Shiga, 525—8577 Japan
†† Colledge of Information Science and Engineering, Ritsumeikan University
Nojihigashi 1—1—1, Kusatsu, Shiga, 525—8577 Japan
E-mail: †kamura@coms.ics.ritsumei.ac.jp, ††huang@fc.ritsumei.ac.jp, †††kawagoe@is.ritsumei.ac.jp

Abstract　Recently, the topic model has been getting much attentions and there have been a lot of studies on the model. Although a topic in the model is defined as a factor which fluctuates document word frequencies, it is necessary but difficult for a user to interpret semantic meanings for obtained topics. Therefore, we propose a novel determination method of topic description terms to describe a topic. In this paper, we propose one of the ideal determination methods, where the topic description terms is obtained so as to minimize a total topic description cost for any combination of terms. We also propose several heuristic determination methods to decrease the computational cost drastically.

Key words　topic model, text mining, Latent dirichlet allocation

1. はじめに

社会の高度情報化に伴い，生成される情報量は日々加速度的に増加している．大量の情報から有用な知見の抽出や新たな知識の獲得を行う様々な技術が提案されている．特に，テキスト形式の情報に関する知見抽出や知識獲得等を行う手法として，確率的トピックモデルの有用性が注目されている [2] [3]．確率的トピックモデルは文書が生成される過程を確率的に表現したモデルである [1]．確率的トピックモデルでは，単語出現頻度を変動させる (潜在的な) 要因であるトピックが単語の分布として表現される．確率的トピックモデルは高精度で文書をモデル化できることで幅広く応用することができる．たとえば, 情報検索 [7] や画像認識 [8]，推薦システム [9] などに確率的トピックモデルを利用することができる．以降，確率的トピックモデルを簡単のためにトピックモデルと呼ぶ．また，トピックとは確率的トピックモデルで得られるトピックに限定する．

トピックモデルを用いて得られた個々のトピックは，トピックに関連した単語とその生成確率のペアの集合として表現できる．したがって，トピックモデルによって得られるトピックの意味は，トピックに関連した単語とその生成確率を見ることで適切に解釈する必要がある．表 1 に簡単なトピックモデルの記述例を示す．このトピックモデルでは，２つのトピックと８個の単語に関する生成確率が得られたとしている．

Topic1 では,「ニュース」「メディア」「新聞」などの単語が

あり，おのおの 0.0794，0.0700，0.0657 等の単語生成確率の数値が対応している．この単語分布により Topic1 が記述されている．このとき，Topic1 の意味が何か，あるいは Topic1 の意味を的確に示す単語は何かを決定することは非常に困難である．たとえば，生成確率の大きな単語をトピックの意味を示す単語とすることは容易であるが，それがこのトピックの意味を確かに表現できるかという点では正確性に欠ける．表 1 の例では，ニュースが最も大きな生成確率であるが，これら 8 つの単語の生成確率を考えると，生成確率が 2 番目に大きな値を持つメディアが Topic1 の意味を示す単語としては適切であると考えられる．Topic2 ではバラエティが最も生成確率が大きな単語であるが，これら 8 つの単語の生成確率を考えると，生成確率が 2 番目に大きな値を持つテレビが適切であると考えられる．

そこで本稿では，トピックの客観的解釈を可能とする単語 (以降，トピック表現語) を導出する方法を提案する．まず，トピック表現語の定義のひとつとして，トピック表現特徴量を最大にする単語の組み合わせを用いる．この単語組み合わせによるトピック表現語を導出するには，トピックの単語分布と単語組合わせ総当たりにより導出することになる．しかし，総当たりに計算する必要があるため，計算量が増大し，実際のトピック数や単語数では現実にその最大解を求めることが不可能である．これを解決するために，経験則によりトピック表現語を導出するいくつかの導出方法を提案する．

2. トピックとトピック表現語導出方法

2.1 トピックモデルとトピック

まず，文書集合 $D = \{d_l\}$ を考える．ここで，文書とは，ニュース記事や Web ページなどの単語列で表現されるテキストである．本稿では，これら文書集合 D から Latent Dirichlet Allocation(LDA) の手段で得られたトピックを対象とする．トピックモデルは，トピック集合 $P = \{p_k\}$ に関し，各トピック p_k の単語分布 r_k と，文書 d 中のトピック比率 q_d から構成されるトピック表現である．ここで，単語分布 r_k は単語 t_i とその生成確率 w_{ki} のペア (t_i,w_{ki}) の集合，トピック比率 q_d はトピック p_k とその比率 r_{dk} のペア (p_k,r_{dk}) の集合である．すなわち，$r_k = \{(t_i, w_{ki})\}$，$q_d = \{p_k, r_{dk}\}$ である． なお，どのトピックについても文書集合 D 内のすべての単語 t_i の生成確率が求められているとする．文書集合内のすべての単語の集合を T_o とし，以降，単にトピック単語集合と呼ぶ．

Topic1		Topic2	
ニュース	0.0762	バラエティ	0.0728
メディア	0.0754	テレビ	0.0672
新聞	0.0557	ニュース	0.0601
バラエティ	0.0526	ドラマ	0.0563
ラジオ	0.0505	新聞	0.0462
ドラマ	0.0453	シネマ	0.0461
シネマ	0.0428	ラジオ	0.0438
テレビ	0.0367	メディア	0.0363

表 1　トピックの単語生成確率例

たとえば，表 1 の 2 つのトピック Topic1(p_1 とする) と Topic2(p_2 とする)，2 つの文書 d_1, d_2 から構成されるトピックモデルは，

$p_1 = \{(t_1, 0.0762), (t_2, 0.0754), (t_3, 0, 0557), (t_4, 0.0526), (t_5, 0.0505), (t_6, 0.0453), (t_7, 0.0428), (t_8, 0.0367)\}$

$p_2 = \{(t_1, 0.0601), (t_2, 0.0363), (t_3, 0.0462), (t_4, 0.0728), (t_5, 0.0438), (t_6, 0.0563), (t_7, 0.0461), (t_8, 0.0672)\}$

$d_1 = \{(p_1, 0.65), (p_1, 0.35)\}$

$d_2 = \{(p_1, 0.24), (p_1, 0.76)\}$

ここで，

$T_o =< t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8 >$,

t_1 = ニュース，t_2 = メディア，t_3 = 新聞，t_4 = バラエティ，t_5 = ラジオ，t_6 = ハドラマ，t_7 = シネマ，t_8 = テレビである．

2.2 トピック表現語特徴量

トピック表現語を規定するために，下記のトピック表現語特徴量を導入する．

すべてのトピックへトピック表現語を割り当てるとする．そして，あるトピック $p_k \in P$ に割り当てるトピック表現語を次のトピック表現語割り当て部分ビット列 b_k で表現することにする．b_k の s 番目のビットはトピック単語集合 T_o の s 番目の単語がトピック表現語であるとき 1 の値をもち，それ以外のとき 0 の値を持つ．

たとえば，表 1 のトピック p_1 がトピック表現語として t_1 と t_3 を持つとき，

$b_1 = 10100000$

であり，トピック $p2$ がトピック表現語として t_1，t_6，t_7 を持つとき，

$b_2 = 10000110$

である．さらに，トピック表現語割り当てビット列 b は部分ビット列を連接したものとする．すなわち，

$b = b_1 \& b_2 \cdots \& b_{|P|}$

で表現される．なお，& はビット連接演算を示す．たとえば，たとえば、表 1 に関する上記の割り当て b_1 と b_2 より、$b = b_1 \& b_2$ であり，

$b = 1010000010000110$

となる．

トピック表現語割り当てビット列に関するトピック表現語特徴量 $f(b)$ を示すと，

$$f(b) = \sum_k ((\prod_s (g_{ks})) / \sum_s (b_{ks})) \tag{1}$$

ただし，

$$g_{ks} = \begin{cases} w_{ks}/W_s & (if\ b_{ks} = 1) \\ 1 - (w_{ks}/W_s) & (otherwise) \end{cases} \tag{2}$$

$$W_s = \sum_k (w_{ks}) \tag{3}$$

である．ここで，b_{kb} はトピック P_k のトピック表現割り当てビット列の s 番目の部分ビットを示す．もし，トピック表現語特徴量 $f(b)$ の値が同じ場合，トピック表現語割り当てビット列

b において $\sum b_{ks}$ の値が最も小さい b を，トピック表現語割り当てをトピック表現語とする．

たとえば，表 1 に関する上記トピック表現割り当て $b = b1\&b2$ の場合，

$$
\begin{aligned}
f(b) &= \prod_s (g_{1s}) + \prod_s (g_{2s}) \\
&= (w_1/W_1) * (1 - (w_2/W_2)) * (w_3/W_3) * \\
&\quad (1 - (w_4/W_4)) * (1 - (w_5/W_5)) * (1 - (w_6/W_6)) * \\
&\quad (1 - (w_7/W_7)) * (1 - (w_8/W_8)) + \\
&\quad (w_1/W_1) * (1 - (w_2/W_2)) * (1 - (w_3/W_3)) * \\
&\quad (1 - (w_4/W_4)) * (1 - (w_5/W_5)) * (w_6/W_6) * \\
&\quad (w_7/W_7) * (1 - (w_8/W_8)) \\
&= (0.0762/0.1117) * (1 - (0.0754/0.1363)) * \\
&\quad (0.0557/0.1019) * (1 - (0.0526/0.1039)) * \\
&\quad (1 - (0.0505/0.0943)) * (1 - (0.0453/0.1254)) * \\
&\quad (1 - (0.0428/0.1016)) * (1 - (0.0367/0.0889)) + \\
&\quad (0.0601/0.1117) * (1 - (0.0363/0.1363)) * \\
&\quad (0.0462/0.1019) * (1 - (0.0728/0.1039)) * \\
&\quad (1 - (0.0438/0.0943)) * (1 - (0.0563/0.1254)) * \\
&\quad (1 - (0.0461/0.1016)) * (1 - (0.0672/0.0889)) \\
&= 0.005046
\end{aligned}
$$

となる．ここで，$W_1 = 0.1117$，$W_2 = 0.1363$，$W_3 = 0.1019$，$W_4 = 0.1039$，$W_5 = 0.0943$，$W_6 = 0.1254$，$W_7 = 0.1016$，$W_8 = 0.0889$ である．

上記で定義したトピック表現語特徴量は下記の視点から定義したものである．もちろん，上記以外にも様々な特徴量が考えられることはいうまでもない．

(1)　生成確率 w_{ki} が高い単語 t_i は，トピック表現語として適した単語である

あるトピック p_k において，単語 t_i の生成確率 w_{ki} が高いということは，単語 t_i はトピック p_k を含む文書中によく出現し，単語 t_i それだけトピック p_k における重要な単語であると考えられる．よって，生成確率 w_{ki} は，トピック p_k における単語 t_i の重要度を示す重みであると考える．

しかし，相互トピックという観点から，様々なトピックにおける単語 t_i の生成確率 w_{ki} が高い場合，単語 t_i はトピック p_k を表現する特徴的な単語ではなく，一般的な単語である可能性が高い．そこで，式 (2) において正規化を行うことで，生成確率 w_{ki} をトピック p_k における重みではなく，トピック集合 P に対する重要度 w_{ks}/W_s とした．その結果，トピック集合 P に対する重要度 w_{ks}/W_s は，単語 t_i はトピック p_k をどの程度表現できるかの表現度を表す重みとなる．

(2)　複数トピックにおいて生成確率 w_{ki} が高い単語 t_i は，トピック表現語に適さない単語の可能性がある

トピック p_k のトピック表現語割り当てについて考える．表現度が高い単語が，トピック表現語割り当てに含まれなかった場合には，トピック表現語にトピック p_k を特徴的に表現する単語が含まれていないため，トピック表現語特徴量の数値を下げる必要があり，また，表現度が低い単語が，トピック表現語割り当てに含まれた場合，トピック表現語にトピック p_k を特徴的に表現できない単語が含まれており，トピック表現語特徴量の数値を下げる必要がある．そのため，式 (2) の様に，2 つの場合を分けてトピック表現語特徴量を計算することにしている．

(3)　複数単語によるトピック表現語割り当てはトピック表現語に適さない単語の可能性がある

トピック p_k のトピック表現語割り当てについて考える．複数単語がトピック表現語として割り当てられた場合，その単語分だけ人間による客観的解釈の余地を作ることとなる．そのため，式 (2) の様に，単語数で数値を割り，計算することにしている．

2.3　トピック表現語導出方法

トピック表現語はトピック表現語特徴量を最大にするトピック表現割り当てである．すなわち，求めるトピック表現割り当て b^{opt} は，

$$b^{opt} = \arg(\max_b(f(b))) \tag{4}$$

となる．

たとえば，表 1 の 2 つのトピックのトピック表現語を導出するには，トピック表現割り当てのすべての組み合わせ，つまり，トピック表現語割り当てビット列 (16 ビットのビット列) について，0000000000000000 から 111111111111111 までのビット列に対するトピック表現語特徴量 $f(b)$ を計算し，その値が最大となるビット列 b^{opt} が適切なトピック表現語割り当てとなる．もちろん，すべての組み合わせで $f(b)$ を計算する必要はない．たとえば，111111111111111 は 2 つのトピックのトピック表現語が同一であることを意味するため，不要な組み合わせである．

具体的には，つぎのような組み合わせを排除する．

（1）　あるビット列 b_i と b_j が同一の場合

（2）　あるビット列 b_i と b_j に包含関係がある場合

（3）　あるビット列 b_i がすべて 0 の場合

この方法を用いたときのトピック表現語導出の例を以下に示す．$b = 1010000010000110$ の場合の $f(b)$ が 0.005046 であることは先に説明した．今，$\tilde{b} = 1000000000010000$ の場合の $f(\tilde{b})$ は 0.01319 である．したがって，$f(\tilde{b}) > f(b)$ であり，また，他のどんなビット列 $\hat{b}$ について $f(b\tilde{b}) > f(\hat{b})$ となる．したがって，$b^{opt} = \tilde{b}$ でありこのビット列 $\tilde{b}$ が最も適したトピック表現割り当てである．その結果，具体的には Topic1 のトピック表現語はメディアであり．Topic2 のトピック表現語はテレビとなる．これは 1. 章で述べたようにトピックの意味を的確に表現した単語として適切なものである．

3.　経験的トピック表現語導出法

本稿では，以下の 3 方法からトピック表現語を経験的に導出する．

（1）　Naïve 法

（2）　TDDA 法 (Topic Label Determination using Document Analysis)

（3） TDTP 法 (Topic Label Determination using Term Probablistic Distibutions)

以下，3.1 節において Naïve 法について，3.2 節においてに TDDA 法について，3.3 節において TDTP 法ついて述べる．

3.1 Naïve 法

Naïve 方法では，各トピックの単語生成分布の生成確率が高い単語をトピック表現語とする．

あるトピック p_k において，単語 t_i の生成確率 w_{ki} が一番高い場合，単語 t_i をトピック p_k におけるトピック表現語とする．しかし単語 t_i の生成確率 w_{ki} が，他のトピックにおいても一番高い数値を示した場合，それぞれのトピックにおける生成確率の上位 2 つの単語群をそれぞれのトピックにおけるトピック表現語とする．以下，同じように単語群が重複した場合，Top-k の単語の数を 1 つずつ増やしていき，すべてのトピックにおける単語群がユニークになるようにトピック表現語を導出する．

たとえば，表 1 の 2 つのトピックのトピック表現語は，Topic1 の生成確率の最大値はテレビの 0.0794 であり，Topic2 の生成確率の最大値はバラエティの 0.0638 である．すなわち，求めるトピック割り当ては，Topic1：テレビ，Topic2：バラエティとなる．

3.2 TDDA 法

TDDA 法では，tf-idf によるトピック表現語導出法 (TDDA-1) と，単語共起によるトピック表現語導出法 (TDDA-2) を提案する．以下，3.2.1 節において TDDA-1 について，3.2.2 節において TDDA-2 について述べる．

3.2.1 tf-idf によるトピック表現語導出法 (TDDA-1)

tf-idf によるトピック表現語導出法では，文書に対して tf-idf 法を使用して単語ごとの重みを利用してトピック表現語を導出する．

あるトピック p_k のトピック比率がある閾値以上の文書 d をトピック p_k を含む文書として，文書 d に出現する単語について tf-idf 値を計算する．トピック p_k を含む文書 d において計算した tf-idf 値を単語ごとに和を計算し，その数値が高い単語をトピック表現語とする．この方法においても，Naïve 方式と同様にトピック表現語が重複しないよう Top-k の単語を使用する．

3.2.2 単語共起によるトピック表現語導出法 (TDDA-2)

単語共起によるトピック表現語導出法では，2 つの単語の共起文書数を利用してトピック表現語を導出する．

ある 2 つの単語 t_{ki_1} と単語 t_{ki_n} がともに存在する (共起する) 文書数を数える．同様に，すべての単語についてある 2 単語の共起文書数を数え，その数値が高い単語をトピック表現語とする．この方法においても，Naïve 方式と同様にトピック表現語が重複しないように Top-k の単語を使用する．

たとえば，表 1 の Topic1 では,「テレビ」と「メディア」の共起文書数が 43,「新聞」,「ラジオ」,「ニュース」との共起文書数がそれぞれ，21，19，36 のとき,「テレビ」の総共起文書数は 119 となる．以下,「メディア」,「新聞」,「ラジオ」,「ニュース」の総共起文書数がそれぞれ，134，68，72，98，であったとき，最も総共起文書数が高いのは,「メディア」の 134 である．すなわち，求めるトピック割り当ては，Topic1：メディアとなる．

3.3 TDTP 法

TDTP 法では，上位単語の生成確率の正規化によるトピック表現語導出法 (TDTP-1) と，単語ごとの生成確率によるトピック表現語導出法 (TDTP-2) を提案する．以下，3.3.1 節において TDTP-1 について，3.3.2 節において TDTP-2 について述べる．

3.3.1 上位単語の生成確率の正規化によるトピック表現語導出法 (TDPD-1)

上位単語の生成確率の正規化によるトピック表現語導出法 (TDPD-1) では，トピック表現語は単語の生成確率が上位 k 単語のうちから表現できるとして，それぞれの生成確率の正規化された数値を利用してトピック表現語を導出する．

トピックモデルでの各トピックの単語分布に関係する単語数は表 1 に示したような 5 個のような少数ではなく，実際には通常数百や数千個の可能性がある．しかし，上位単語以外の出現確率はほぼ 0 に近しいため，トピックの特徴は上位単語から表現できると考える．そこでまず，単語の生成確率の正規化を，トピック p_k における生成確率が上位 k 単語中の割合の計算によって行う．単語の生成確率の正規化を行った後，正規化された数値が最上位である単語 t_i の数値がある閾値 (δ) 以上の場合，単語 t_i を p_k におけるトピック表現語とする．正規化された数値が最上位である単語 t_i の数値が閾値 (δ) 未満の場合，閾値以上となるまで，第 2 位から順に単語 t_i の数値を加えて，トピック表現語を構成する．なお，閾値 (δ) 以上であっても，単語群が他のトピックと重複する場合，さらに単語を 1 つ以上増やし，ユニークにトピック表現語が決まるように，トピック表現語を導出する．

たとえば，表 1 の Topic1 では，5 単語の生成確率の正規化を行うと，結果は表 2 となる．閾値を 0.4 とした場合，上位 2 つの数値の和が 0.5832 となり閾値以上になる．すなわち，Topic1 における求めるトピック割り当ては「メディア，ニュース」となる．

3.3.2 単語ごとの生成確率によるトピック表現語導出法 (TDPD-2)

単語ごとの生成確率によるトピック表現語導出法 (TDPD-2) は，単語の生成確率を全トピック中の割合を計算することで正規化された数値を利用してトピック表現語を導出する．

TDPD-2 では，Naïve 方式と同様にトピック表現語が重複しないように正規化された数値から，Top-k の単語をトピック表

Topic1	
ニュース	0.2931···
メディア	0.2901···
新聞	0.2143···
バラエティ	0.2023···

表 2 上位単語の正規化結果

Topic1	
ニュース	0.5590···
メディア	0.6750···
新聞	0.5466···
バラエティ	0.4194···
ラジオ	0.5355···
ドラマ	0.4458···
シネマ	0.4814···
テレビ	0.3532···

表 3 単語ごとの正規化結果

現語として導出する．

たとえば，表 1 の Topic1 では，単語の生成確率を全トピック中の割合を計算すると，結果は表 3 となる．ここで，正規化された数値の最大値はバラエティの 0.6075 である．すなわち，求めるトピック割り当ては，メディアとなる．ただし，他のトピックにおいても「メディア」の数値が最も高かった場合，Topic1 における求めるトピック割り当ては「メディア，ニュース」となる．

4. 評価実験

4.1 評価条件

本稿で提案したトピック表現語の有効性を確認するため，実データを用いて評価実験を行った．実データは 20 Newsgroups data set [16] を用いた．この実データに対して以下に示すフィルタリングを行い，実験で使用するテストデータを構築した．まず，トピックモデルの性質上，「その他」というトピックを生成できない．そのため，20 Newsgroups data set のうち，「misc.forsale」，「comp.os.ms-windows.misc」，「talk.religion.misc」，「talk.politics.misc」を除外した．次に，「comp.sys.mac.hardware」と「comp.sys.ibm.pc.hardware」では，ハードウェアに関連する単語が多く含まれており，「ハードウェア」という 1 つのトピックとしてまとめられやすいことと，本稿では潜在的ディリクレ過程の性能を評価するのが目的ではないことから，「comp.sys.ibm.pc.hardware」を除外した．「rec.autos」も，同様の理由から除外した．構築したテストデータから，ストップワードを省いた，14 カテゴリ，5,383 文書，61,188 単語に対して潜在的ディリクレ過程を用い 14 個のトピックを抽出した．なお、本稿では，潜在的ディリクレ過程を Java 言語を用いて実装した．

4.2 評価指標

本稿では 2.3 節において述べたトピック表現語導出法から得られた単語集合を正解集合として，経験的導出法から得られた単語集合を Jaccard 係数，Simpson 係数，純度の 3 つの指標から評価する．Jaccard 係数 $J(C_i, A_h)$，Simpson 係数 $S(C_i, A_h)$，純度 P_i の 3 つの指標の算出式をそれぞれ (5) 式，(6) 式，(7) 式に示す．

$$J(C_i, A_h) = \frac{|C_i \cap A_h|}{|C_i \cup A_h|} \tag{5}$$

$$S(C_i, A_h) = \frac{|C_i \cap A_h|}{\min\{|C_i|, |A_h|\}} \tag{6}$$

$$P_i = \frac{1}{|C_i|} \max_h |C_i \cap A_h| \tag{7}$$

ここで，C_i はカテゴリ i において，経験的導出法から得られた単語集合，A_h はカテゴリ h において，トピック表現語導出から得られた単語集合である．

4.3 評価結果ならびに考察

潜在的ディリクレ過程から得られた各トピックにおける生成確率が上位 5 単語ならびに，各手法によって得られた単語集合を表 7，指標を用いた各手法の評価結果を表 4，表 5，表 6 に示す．表 7 では潜在的ディリクレ配分法を用いて得られた各トピックにおける生成確率の上位 5 単語と，トピック表現特徴量を用いて導出したトピック表現語，3. 節で述べた経験的導出法を用いて得られた単語群をそれぞれ，上位 5 単語，表現語，Naïve，TDDA-1，TDDA-2，TDPD-1，TDPD-2 として示す．

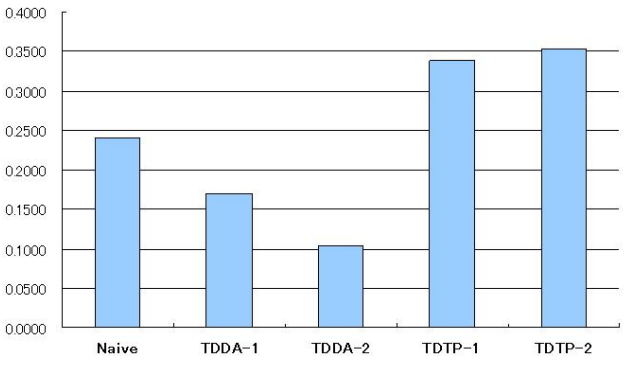

表 4　Jaccard 係数評価結果

表 7 より，20Newsgroups のカテゴリと比較してトピック表現語特徴量をもちいたトピック表現語が適切に割り当てられていることがわかる．例えば，「comp.graphics」に対応するトピックでは，”image ”，”bit ”，”graphics ”，”jpeg ”と，画像系に関する単語が導出できた．また，「sci.med」に対応するトピックでは，”medical ”，”cancer ”，”disease ”と，医療，病気に関する単語が導出できた．

表 4，表 5，表 6 からわかるように，提案した 5 つの経験的導出法のうち，TDPD-2 が最も良い結果を得た．これは，あるトピックにおける生成確率が高く，その他のトピックにおける生成確率が低い単語はトピック表現語に適していたと考えられる．表 5 で TDPD-2 の純度が 1.0 を示しており，正解単語以外の単語を導出しておらず，トピック表現語として適さない単語を導出していないことがわかる．以上のことは表 7 より，「talk.politics.guns」においては ”gun ”が，「rec.sport.baseball」では”baseball”が，「comp.graphics」では，graphics に近しい ”image ”がトピック表現語として選ばれていることからも良い結果が得られたと考えることができる．

5. 関連研究

5.1 トピック検出

トピック検出手法には，LSI（Latent Semantic Indexing）[4]，pLSI（probabilistic LSI）[5] などが存在するが，その中でも，潜在的ディリクレ配分法 (Latent Dirichlet Allocation) [6] (以下，LDA) がよく機能することが知られている. [6]

LDA は，pLSI モデルを文書に関するトピック空間の多項分布にディリクレ事前分布を導入し，拡張したモデルである．LDA は，多くの分野に応用可能であり，ユーザプロファイリング [13] や，次元圧縮 [14]，可視化 [14] など，広く利用されて

表 5　純度評価結果

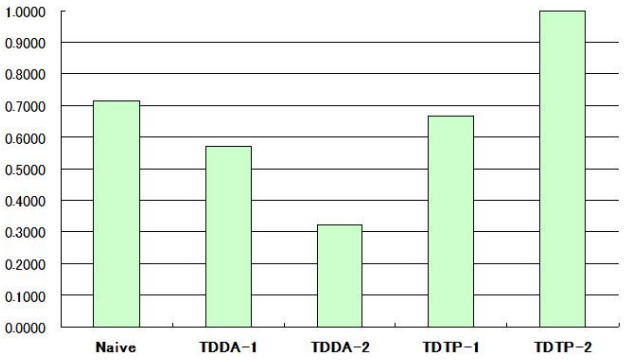

表 6　Simpson 係数評価結果

いる．

5.2　クラスタラベル導出

インターネットの階層構造のカテゴリーは，人間によって注意深く定義される．文書を処理することによってラベル付けを行う手法が提案されてきたが，多くの場合，よいラベルが文書中に出現しないという問題ある．そのため，Wikipedia や WordNet といった外部知識を利用したクラスターラベリング手法が提案されている [11] [12]．しかし，それらの方法は単語の重要度を決定するためにトレーニング・データを必要する．したがって，同意語，下位語，上位語は不明瞭なままである．そこで Fukumoto らは機械が読むことが可能な辞書中の意味関係に着目した，Cluster Labeling based on Concepts in a Machine-Readable Dictionaly [10] を提案した．

Fukumoto らの手法では，クラスタリングに Graph-based unsupervised clustering，単語の重みに TFIDF，PMI，χ statistics，information gain の 4 手法，単語間距離による上位語抽出を用いてラベルのスコア付けと決定を行う．

6. おわりに

本稿では，トピックモデルにおけるトピック表現語導出を提案した．トピック表現語によって，トピックの客観的解釈を可能となる．また，トピック表現語は様々な分野に応用が可能であると考えられる．たとえば，tagging や，トレンドの把握，ユーザープロファイリングなどが挙げられる．

今後は，トピック表現特徴量を用いた表現語導出法の客観的評価法の考察が必要である．また．オントロジーに着目した表現語の導出を考察する予定である．

文　献

[1] Blei D. , Carin L. , Dunson D.:“ Probabilistic Topic Models ”, Signal Processing Magazine IEEE , 27, pp.55—65 (2010)

[2] T. Iwata, T. Yamada, N. Ueda: ”Modeling Social Annotation Data with Content Relevance using a Topic Model ”, Advances in Neural Information Processing Systems, pp.835-843 (2009)

[3] D. Blei, J. McAuliffe:”Supervised topic models”, Neural Information Processing Systems 21, (2007)

[4] DEERWESTER S.:“ Indexing by latent semantic analysis ”, Journal of the American Society For Information Science, 41, pp.391—407 (1990)

[5] HOFMANN T.:“ Probabilistic latent semantic indexing ”, SIG-IR, pp.50—57 (1999)

[6] BLEI D.:“ Latent dirichlet allocation ”, Journal of Machine Learning Research, 3, pp.993—1022 (2003)

[7] Koji EGUCHI, Hitohiro SHIOZAKI:“ Wikipedia Retrieval using Multitype Topic Models ”, SIG-SWO, pp.73-80 (2008)

[8] Liangliang Cao, Li Fei-Fei:“ Spatially Coherent Latent Topic Model for Concurrent Segmentation and Classification of Objects and Scenes ”, IEEE 11th International Conference, pp.1-8 (2007)

[9] Tomoharu IWATA, Shinji WATANABE, Takeshi YAMADA, Naonori UEDA:“ Topic Tracking Model for Analyzing Consumer Purchase Behavior ”, International Joint Conferences on Artificial Intelligence, vol.9, pp.2-4 (2009)

[10] Fumiyo Fukumoto, Yoshimi Suzuki: ”Cluster Labelling based on Concepts in a Machine-Readable Dictionary ”, In Proceedings of the 5th International Joint Conference on Natural Language Processing, pp.1371-1375 (2011)

[11] O. S. Chin, N. Kulathuramaiyer, and A. W. Yeo:”Automatic Discovery of Concepts from Text”, In Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pages 1046—1049 (2006)

[12] Gabrilovich and S. Markovitch:”Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis”, In Proc. of the 20th International Joint Conference on Artificial Intelligence, pp.1606—1611 (2007)

[13] Hiroshi FUJIMOTO, Minoru ETOH, Akira KINNO, Yoshikazu AKINAGA:”Personalized Web Content Recommendation based on LDA Profile”, Data Engineering and Information Management (2011)

[14] Tomonari MASADA, Senya KIYASU, Sueharu MIYAHARA:”Dimensionality Reduction via Latent Dirichlet Allocation for Document Clustering”, IPSJ SIG Technical Reports, pp.381-386 (2007)

[15] T. IWATA, T. YAMADA, N. UEDA:”Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents”, Knowledge-Discovery in Data (2008)

[16] 20 Newsgroups Data Set:http://people.csail.mit.edu/ jrennie/20Newsgroups/

	カテゴリー名	上位5単語	表現語	naïve	TDDA-1	TDDA-2	TDPD-1	TDPD-2
1	talk.politics.guns	fbi fire gun guns government	fbi fire gun guns	fbi	fbi	government	fbi fire	gun
2	rec.sport.baseball	game hit baseball team won	baseball hit	game	mattingly	game	game hit baseball	baseball
3	comp.graphics	image bit graphics imfomation jpeg	image bit graphics jpeg	image	graphics	imfomation	image bit	image
4	rec.sport.hockey	game hockey team games pit	hockey game	game	game	game	game hockey team	hockey
5	sci.electronics	ploblem read find post hard	hard post	ploblem	battery	read	priblem read find	hard
6	sci.space	space earth nasa energy shuttle	space earth nasa energy shuttle	space	space	space	space earth	shuttle
7	rec.motorcycles	university bike car front road	bike road car	university	bike	road university	university bike	bike
8	talk.politics.mideast	israel armenian war jews muslims	israel armenian jews muslims	israel	israel	war	israel armenian war	israel
9	sci.med	medical cancer information research disease	medical cancer disease	medical	doctor	research	medical cancer information	cancer
10	soc.religion.christian	god church jesus christ man	church jesus christ man	god	god	christ	god	christ
11	omp.sys.mac.hardwar	apple drive mac computer monitor	apple drive mac monitor	apple	mac	mac	apple drive	monitor
12	comp.windows.x	windows set run file server	windows server	windows	window	set	windows set run file	windows
13	sci.crypt	key government public chip encryption	key public encryption	key	key	information	key government	key
14	alt.atheism	religion point evidence wrong atheism	religion atheism	religion	god	point	religion point evidence	religion

表 7　各手法による単語集合