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

2007－DBS－143 (64)
2007／7／3

文書クラスタリングのための潜在的ディリクレ配分法による次元圧縮

正田 備也† 喜安 千弥† 宮原 末治†

† 長崎大学工学部 〒852-8521 長崎県長崎市文教町 1-14
E-mail: †{masada,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

あらまし 本論文では，Blei らによって提案された潜在的ディリクレ配分法 (latent Dirichlet allocation) を，特徴ベクトルの次元圧縮法として利用し，文書クラスタリングにおける有効性を明らかにする．評価実験では，日本語と韓国語の Web ニュース記事のクラスタリングをおこない，記事の属するジャンルをクラスタリング結果の評価に用いる．単語の出現頻度をそのまま入力として，混合多項分布モデルを用いたクラスタリングを行う場合と，潜在的ディリクレ配分法によって次元圧縮された特徴ベクトルを入力として，同じく混合多項分布モデルを用いたクラスタリングを行う場合とで，クラスタリング結果を比較評価する．

キーワード 文書クラスタリング，次元圧縮，潜在的ディリクレ配分法

Dimensionality Reduction via Latent Dirichlet Allocation for Document Clustering

Tomonari MASADA†, Senya KIYASU†, and Sueharu MIYAHARA†

† Faculty of Engineering, Nagasaki University Bunkyo-machi 1-14, Nagasaki, 852-8521 Japan
E-mail: †{masada,kiyasu,miyahara}@cis.nagasaki-u.ac.jp

Abstract In this paper, we employ the latent Dirichlet allocation as a method for the dimensionality reduction of feature vectors and reveal its effectiveness in document clustering. In the evaluation experiment, we perform clustering on the document sets of Japanese and Korean Web news articles. We regard the categories assigned to each article as the ground truth of clustering evaluation. We compare the clustering results obtained by using the feature vectors whose entries are term frequencies with the results obtained by using the feature vectors whose dimensions are reduced by the latent Dirichlet allocation.

Key words document clustering, dimensionality reduction, latent Dirichlet allocation

1. はじめに

文書クラスタリングは，テキスト・マイニングの分野では，古典的な問題である．近年では，例えば Web 検索において，検索結果を整理して提示する目的や，検索結果に含まれる複数のトピックを分離するという目的で，Web 検索結果のような小規模な文書集合への適用が，1 つの応用として注目されている [1] [2] [3] [4] [20]．そこでは，高速な処理が目指される．一方，企業の一部門で蓄積された文書や，特定分野の特許文書など，Web ほど大規模ではなく検索結果ほど小規模でもない，中規模の文書集合に対してテキスト・マイニングを適用し，重要な情報を抽出するという応用がある [9] [12] [15]．本論文では，後者を視野に入れ，中規模文書集合に対して 1CPU で 1 日あたり 2〜4 回の周期で実行できるマイニング技術として，Blei らによって提案された潜在的ディリクレ配分法 (latent Dirichlet allocation) に注目する．潜在的ディリクレ配分法の主な特徴は，1 つの文書に複数のトピックがふくまれることを明示的にモデル化できること，モデルのパラメータ推定にベイズ推定が用いられることにある．今回の実験では，潜在的ディリクレ配分法により文書の特徴ベクトルの次元を圧縮し，圧縮後の特徴ベクトルを使って Web ニュース記事をクラスタリングする．

実験で用いるデータセットは，日本語と韓国語の Web ニュース記事で，いずれも数万件の中規模文書集合である．クラスタリング手法には混合多項分布モデルに基づく確率論的手法を用いる．これは，ナイーヴ・ベイズ分類器の教師なし学習版 [18] で，パラメータ推定に EM アルゴリズム [10] を用いる．評価のベースラインは，各文書での単語の出現頻度をそのまま特徴量として用いる場合である．これと比較するのは，潜在的ディリクレ配分法において，変分ベイズ法を利用して得られるパラメータの値を，文書の特徴量として用いる場合である．そのパ

表 1 JIC データセット
Table 1 JIC dataset

データセット：JIC		
カテゴリ	文書数	文書長和
携帯・ワイヤレス	3,049	499,368
Web ビジネス	9,059	1,214,335
E コマース	2,522	327,264
Web ファイナンス	2,994	398,995
Web テクノロジー	6,109	922,164
Web マーケティング	4,596	746,119
計	28,329	4,108,245

表 2 S2005, S2006 データセット
Table 2 S2005 and S2006 datasets

S2005			S2006		
カテゴリ	文書数	文書長和	カテゴリ	文書数	文書長和
経済	6,172	461,592	行政	1,503	124,657
国際	3,048	216,462	文化	4,870	347,438
政治	3,608	286,375	経済	6,745	549,081
社会	9,221	590,190	芸能	1,710	125,787
			国際	2,498	186,753
			政治	3,806	324,076
			地域	3,923	280,676
			社会	8,946	607,158
			スポーツ	3,016	185,054
計	22,049	1,554,619	計	37,017	2,730,680

ラメータは，Blei らの原論文 [8] にある γ_{ik} である．i は文書の添え字であり，k はトピックの添え字である．この γ_{ik} を，個々の文書における各トピックの “頻度” とみなし，混合多項分布モデルに基づくクラスタリングの入力とする．クラスタリング結果の評価には，ニュース記事に人手で付与されているカテゴリを用いる．また，今回の評価実験では，クラスタリング・アルゴリズムに，カテゴリ数を入力として与えることにする．つまり，正しいクラスタ数の推定は，考察の対象外とする．

2. 従来研究

潜在的ディリクレ配分法については，文書と画像のような異種情報統合 [5] [17]，書誌情報における著者と研究トピックとの関連性解析 [13] など，各方面への応用が提案されている．日本語文書への適用については，貞光らによる組織的な実験研究 [23] [22] が，PLSI [14] や混合ディリクレ分布モデル [16] など，他の文書モデルとの比較も行っている点で，新しい文書モデルの日本語文書への適用可能性の調査として重要である．しかし，貞光らは，各種言語モデルについて，文書分類や文書クラスタリングなどの具体的な問題への応用可能性の評価というより，パープレキシティを尺度とすることで，それぞれの言語モデルとしての有効性を評価している．

本論文も日本語文書を扱うが，文書クラスタリングの結果という，ユーザに直接提示される結果を評価の対象とする．本論文では，潜在的ディリクレ配分法を，文書の特徴ベクトルの次元圧縮に利用する．同種の実験は Blei らの原論文 [8] にもあるが，英語の文書情報を対象としており，また，文書クラスタリングではなく文書分類への利用である．Elango らの研究 [11] でも，潜在的ディリクレ配分法により特徴ベクトルの次元圧縮をおこない，次元圧縮後のベクトルをクラスタリングしているが，画像情報が対象である．本論文では，日本語だけでなく，韓国語の文書集合も対象とし，潜在的ディリクレ配分法によって文書の特徴ベクトルの次元を圧縮した結果を，英語とは言語としての特質が異なると予想されるアジアの言語で書かれた文書のクラスタリングに用いた場合，どのようなふるまいを見せるかを明らかにしようとする．

3. 潜在的ディリクレ配分法

潜在的ディリクレ配分法 [8] は，混合多項分布モデル [18] や混合ディリクレ分布モデル [19] とは異なり，文書が複数のトピックを含むことを表現できる，マルチトピック文書モデルである．PLSI [14] も同じくマルチトピック文書モデルだが，未知の文書の生成確率を求めるためにヒューリスティクスが必要なことや，過学習を起こしやすいことなど，欠点がある．

文書の集合を $D = \{d_1, \ldots, d_I\}$，語彙の集合を $W = \{w_1, \ldots, w_J\}$，トピックの集合を $T = \{t_1, \ldots, t_K\}$ とする．以下，タイプとしての単語，つまり単語の種類をいうときには「語彙」という言葉を使い，トークンとしての単語，つまり個々の語彙の出現をいうときには「単語」という言葉を使うようにする．潜在的ディリクレ配分法では，文書集合全体について，ディリクレ分布 $P(\theta;\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k}$ を想定し，これにしたがって，文書ごとにトピック集合 T 上に定義された多項分布が 1 つ選ばれる．そして，各トピックには，語彙集合 W 上に定義された多項分布が 1 つ対応する．文書が違っても，同じトピックには同じ多項分布が対応する．トピック t_k に対応する語彙の多項分布での語彙 w_j の出現確率を β_{kj} とする．もちろん，$\sum_j \beta_{kj} = 1$ がすべての k で成り立つ．潜在的ディリクレ配分法のモデル・パラメータは，この $\beta_{kj}(k = 1, \ldots, K,\ j = 1, \ldots, J)$ と，ディリクレ分布のパラメータ $\alpha_k(k = 1, \ldots, K)$ であり，パラメータの個数は全部で $K + KJ$ 個である．

潜在的ディリクレ配分法では，各文書は次のように生成される．まず，個々の文書について，ディリクレ分布 $P(\theta;\alpha)$ にしたがって，トピック上に定義された多項分布を 1 つ選ぶ．そして，各単語ごとに，この多項分布にしたがってトピックを 1 つ選ぶ．そして，そのトピックに対応する語彙の多項分布にしたがって，語彙を 1 つ選ぶ．以下同様に，文書を構成する単語の個数だけ語彙を選ぶ．つまり，語彙を選ぶたびに，トピックを選び直している．その結果，1 つの文書が複数のトピックを含むことをモデル化できるようになっている．文書 d_i の l 番目の単語に割り当てられるトピックを表す確率変数を $\mathbf{z}_{il}$ とする．これをまとめて n_i 次元ベクトル $\mathbf{z}_i$ で表す．n_i は文書 d_i の文書長，つまり，d_i における各語彙の出現回数の総和である．文書 d_i の l 番目に現れる単語を表す確率変数を $\mathbf{x}_{il}$ とする．これをまとめて n_i 次元ベクトル $\mathbf{x}_i$ で表す．文書 d_i が生成される確率 $P(\mathbf{x}_i)$ を書き下すと，以下のようになる．

$$P(\mathbf{x}_i;\alpha,\beta)=\int\sum_{\mathbf{z}_i}P(\theta;\alpha)P(\mathbf{z}_i|\theta)P(\mathbf{x}_i|\mathbf{z}_i,\beta)d\theta \qquad (1)$$

文書集合全体 $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_I\}$ が生成される確率は $P(\mathbf{X};\alpha,\beta)=\prod_i P(\mathbf{x}_i;\alpha,\beta)$ となり，その対数 $\log P(\mathbf{X};\alpha,\beta)=\sum_i \log P(\mathbf{x}_i;\alpha,\beta)$ を最大化することで，パラメータの推定を行う．しかし，この推定は解析的に解けない．そこで，本論文では，Blei ら [8] にならって変分ベイズ法 [21] によるパラメータ推定を行う．すると，下の式 (2) のように，各文書ごとに $Q(\theta;\gamma_i)$ と $Q(\mathbf{z}_i;\phi_i)$ という 2 種類の確率分布が新たに導入され，$\log P(\mathbf{X};\alpha,\beta)$ の近似としての下限値が得られる．

$$\begin{aligned}
&\log P(\mathbf{x}_i;\alpha,\beta)\\
&=\log\int\sum_{\mathbf{z}_i}P(\theta;\alpha)P(\mathbf{z}_i|\theta)P(\mathbf{x}_i|\mathbf{z}_i,\beta)d\theta\\
&=\log\int\sum_{\mathbf{z}_i}Q(\theta;\gamma_i)Q(\mathbf{z}_i;\phi_i)\frac{P(\theta;\alpha)P(\mathbf{z}_i|\theta)P(\mathbf{x}_i|\mathbf{z}_i,\beta)}{Q(\theta;\gamma_i)Q(\mathbf{z}_i;\phi_i)}d\theta\\
&\geqq\int\sum_{\mathbf{z}_i}Q(\theta;\gamma_i)Q(\mathbf{z}_i;\phi_i)\log\frac{P(\theta;\alpha)P(\mathbf{z}_i|\theta)P(\mathbf{x}_i|\mathbf{z}_i,\beta)}{Q(\theta;\gamma_i)Q(\mathbf{z}_i;\phi_i)}d\theta
\end{aligned} \qquad (2)$$

上式で，$Q(\mathbf{z}_i;\phi_i)=\prod_{l=1}^{n_i}Q(\mathbf{z}_{il};\phi_{il})$ は，文書を構成する個々の単語に特定のトピックが割り当てられる確率を示す，トピック集合 T 上に定義された多項分布である．ϕ_{ilk} が，文書 d_i の l 番目に出てくる単語にトピック t_k が割り当てられる確率であり，$\sum_{k=1}^{K}\phi_{ilk}=1$ がすべての i,l について成立する．また，$Q(\theta;\gamma_i)$ は，トピックの多項分布上に定義されたディリクレ分布である．その意味は $P(\theta;\alpha)$ と似ているが，変分ベイズ法による近似の結果，各文書ごとに別々のディリクレ分布 $Q(\theta;\gamma_i)$ が想定されている．変分ベイズ法では，$\log P(\mathbf{X};\alpha,\beta)$ の代わりに，式 (2) で得た下限値を最大化するが，パラメータ推定の詳細は，Blei らの原論文に譲る [8]．ここでは，結果として得られる更新式だけを示す．

$$\phi_{ilk}\propto\beta_{kj_{il}}\exp\{\Psi(\gamma_{ik})-\Psi(\sum_{k'}\gamma_{ik'})\} \qquad (3)$$

$$\gamma_{ik}=\alpha_k+\sum_{l=1}^{n_i}\phi_{ilk} \qquad (4)$$

$$\beta_{kj}\propto\sum_i\sum_l\delta_{ilj}\phi_{ilk} \qquad (5)$$

$$\alpha_k=\hat{\alpha}_k+f_k(\hat{\alpha})+\sum_{k'}f_{k'}(\hat{\alpha})\Big/\Big\{\frac{\Psi_1(\hat{\alpha}_k)}{\Psi_1(\hat{\alpha}_\Sigma)}-\sum_{k'}\frac{\Psi_1(\hat{\alpha}_k)}{\Psi_1(\hat{\alpha}_{k'})}\Big\}$$

ただし

$$f_k(\alpha)=\frac{\Psi(\alpha_\Sigma)}{\Psi_1(\alpha_k)}-\frac{\Psi(\alpha_k)}{\Psi_1(\alpha_k)}+\frac{\sum_i\{\Psi(\gamma_{ik})-\Psi(\gamma_{i\Sigma})\}}{N\Psi_1(\alpha_k)} \qquad (6)$$

j_{il} は，文書 d_i の l 番目に出てくる単語の添え字，δ_{ilk} は，文書 d_i の l 番目の単語が w_j のときにかぎり 1 となり，それ以外の場合 0 となる値である．また，$\alpha_\Sigma=\sum_k\alpha_k$，$\gamma_{i\Sigma}=\sum_k\gamma_{ik}$ と略記した．Ψ はダイガンマ関数，Ψ_1 はトリガンマ関数で，ともにベルヌイ数を利用すれば，容易に実装できる(注1)．ダイガンマ関数の実装については，Beal の学位論文 [6] の補遺が参考になる．トリガンマ関数は，ダイガンマ関数を展開したときの各項の微分を考えれば，やはり容易に実装できる．式 (6) は，α_k だけの更新式なので，全体としては，式 (3)，式 (4)，式 (5) を実行した後，式 (6) を十分収束するまで反復し，式 (3)，式 (4)，式 (5) の実行に戻るという反復計算を行う．今回の C 言語による実装では，モデルのパラメータ α_k の和 $\sum_k\alpha_k$ の変動率が 0.000001 を切ったところで，全体の反復計算を停止させた．

本論文では，変分ベイズ法によって，各文書 d_i について導入されるディリクレ分布 $Q(\theta;\gamma_i)$ のパラメータ γ_{ik} を，文書 d_i におけるトピック t_k の "頻度" と解釈する．そして，K 次元ベクトル $(\gamma_{i1},\ldots,\gamma_{iK})$ を，各文書の次元削減後の特徴ベクトルとする．つまり，潜在的ディリクレ配分法を用いることで，すべての文書について K 次元の特徴ベクトルが計算される．ベクトル $(\gamma_{i1},\ldots,\gamma_{iK})$ を特徴量として用いて構わない理由は，次の通りである．γ_{ik} の更新式 (4) の両辺について，すべての k にわたって和をとると $\sum_k\gamma_{ik}=\sum_k\alpha_k+\sum_k\sum_{l=1}^{n_i}\phi_{ilk}$ となる．右辺第 2 項の $\sum_k\sum_l\phi_{ilk}$ は，$\sum_{k=1}^{K}\phi_{ilk}=1$ より n_i に等しい．つまり，$\sum_k\gamma_{ik}$ は，各文書の文書長と同じディメンジョンを持つ量である．

4. 実験の手続き

4.1 実験データ

本論文では，Japan.internet.com(注2)の 2001 年から 2006 年のニュース記事のうち，携帯・ワイヤレス，Web ビジネス，E コマース，Web ファイナンス，Web テクノロジー，Web マーケティングの 6 つのカテゴリに分類された記事を，日本語のデータセットとして用いた．形態素解析には MeCab(注3)を用いた．このニュース記事集合を，2001 年と 2002 年，2003 年と 2004 年，2005 年と 2006 年の 3 グループに分け，それぞれで，出現回数が 5 回以上，かつ，文書長の総和の 0.1%以下の単語だけを残し，他の単語を削除した．このような仕方で単語を制限したのは，どの単語が特徴量として残すに値しないかは，記事の年次に依存して決めるほうがよいと考えたからである．このデータセットを JIC と呼ぶ．JIC の文書数は計 28,329 件であり，文書長の総和は 4,108,245，平均文書長は 145.02 である．また，含まれる語彙数は 12,376 である．カテゴリが 6 つあるので，クラスタ数を 6 とするクラスタリングを実行する．カテゴリごとの文書数，文書長和は，表 1 のとおりである．

韓国語の文書集合としては，次の 2 つを用いた．1 つは，Web 版ソウル新聞(注4)の 2005 年の記事のうち，経済，国際，政治，社会の 4 つのカテゴリに分類されている記事の集合，もう 1 つは，同新聞の 2006 年の記事のうち，行政，文化，経済，エンターテインメント，国際，政治，地域，社会，スポーツの 9 つのカテゴリに分類されている記事の集合である．形態素解析に

(注1)：http://mathworld.wolfram.com/DigammaFunction.html
(注2)：http://japan.internet.com/
(注3)：http://mecab.sourceforge.net/
(注4)：http://www.seoul.co.kr/

は，クンミン大学校言語工学情報検索研究室が公開しているKLT version 2.10b(注5)を用いた．2005年分の記事集合について，出現回数が10回以上かつすべての文書長の和の0.5%以下，の単語だけを残したデータセットを作成し，これをS2005と呼ぶ．S2005の文書数は，計22,049件であり，文書長の総和は1,554,619，平均文書長は70.51，語彙数は14,563である．カテゴリが4つなので，クラスタ数4のクラスタリングを実行する．2006年分の記事集合については，出現回数が10回以上かつ文書長の総和の0.5%以下，の単語だけを残したデータセットを作成し，これをS2006と呼ぶ．S2006の文書数は，計37,017件であり，文書長の総和が2,730,680，平均文書長が73.77，語彙数は25,584である．カテゴリが9つなので，クラスタ数9のクラスタリングを実行する．それぞれ，カテゴリごとの文書数，文書長和は，表2のとおりである．

4.2 クラスタリング手法

本研究は，文書クラスタリングの手法として，混合多項分布モデルに基づくクラスタリングを用いる．パラメータ推定のためのEMアルゴリズムは，論文[18]に詳しい．混合多項分布モデルは，1つの隠れ変数を含み，この変数がとる値の各々に，語彙集合上で定義された別々の多項分布が対応している．同じ多項分布によって生成される文書が，同じクラスタに属する文書である．つまり，隠れ変数のとる値の各々が，別々のクラスタに対応している．混合多項分布モデルのパラメータは，隠れ変数が特定の値をとる確率，そして，隠れ変数の特定の値に対応する多項分布における，各単語の出現確率である．これらについて推定されたパラメータ値を用いると，ある文書を所与とする，隠れ変数の個々の値の，条件付き確率が求まる．この確率を最大とする隠れ変数の値により，各文書が属するクラスタを決定する．

混合多項分布モデルは，通常，スムージングを併用することで，より良い性能を示す．本論文の実験では，各クラスタごとの単語の出現確率に，データセット全体での単語の出現確率を線形混合することで，スムージングをおこなった．具体的には，後者の出現確率の混合比を0.0，0.01，0.1，0.3とする4通りのスムージングを試した．

4.3 クラスタリング結果の評価

クラスタリング結果は，precisionとrecallの調和平均であるF-measureにより評価する．まず，各クラスタに含まれる文書について，そのカテゴリを調べる．そして，各クラスタで最も多数を占めるカテゴリを，そのクラスタの正解ラベルと呼ぶ．クラスタのprecisionとは，(正解ラベルをカテゴリとするクラスタ内の文書数) ÷ (クラスタのサイズ) である．クラスタのrecallとは，(正解ラベルをカテゴリとするクラスタ内の文書数) ÷ (正解ラベルをカテゴリとするデータセット中の全文書数) である．

クラスタリング結果全体を評価するためには，次の計算を行う．各クラスタのprecisionの計算に使う分母と分子について，すべてのクラスタで和をとる．そして，分子の和を分母の和で割ったものを，microaveraged precisionと呼ぶ．recallについても，同様に計算したものを，microaveraged recallと呼ぶ．例えば，クラスタリング結果を構成するクラスタが三つあり，それぞれprecisionが2/3，5/8，3/7とすると，microaveraged precisionは(2+5+3)/(3+8+7)となる．定義より，どのクラスタの正解ラベルにもならないカテゴリがある場合，microaveraged precisionとmicroaveraged recallは，一般には異なる値をとる．そして，microaveraged precisionとmicroaveraged recallの調和平均を，本論文ではmicroaveraged F-measureと呼び，これをクラスタリング結果全体の評価に用いる．本論文の実験では，同じ入力データについて，混合多項分布モデルに基づくクラスタリングを20回実行した．なぜなら，混合多項分布モデルのパラメータはランダムに初期化されており，EMアルゴリズムの実行に初期値への依存性があるためである．そして，ランダムに初期化された20通りの初期値から得られたmicroaveraged precisionとmicroaveraged recallの20個のペアについて，microaveraged F-measureを求め，これら20個のmicroaveraged F-measureの平均と標準偏差を，当該入力データに対するクラスタリング結果の評価値とする．

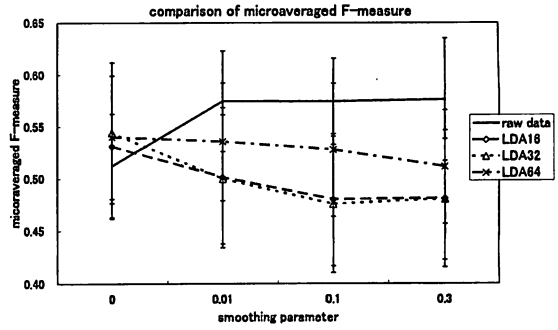

図1 S2005データセットのmicroaveraged F-measure
Fig. 1 microaveraged F-measure for S2005 dataset

5. 実験結果

5.1 Web版ソウル新聞の場合

データセットS2005，S2006についての実験結果を，それぞれ図1と図2に示す．横軸は，混合多項分布モデルにおけるスムージング・パラメータである．この値が大きいほど，文書集合全体での単語の出現頻度が，クラスタ別の単語の出現頻度に，より多く混合される．縦軸は，microaveraged F-measureであり，グラフは，20通りのmicroaveraged F-measureの平均を示している．マーカは，20通りのmicroaveraged F-measureの標準偏差の±1倍の幅を示している．S2005よりS2006が全体として評価値が低いのは，S2005がカテゴリ数4，S2006がカテゴリ数9だからである．図1と図2で，"raw data"が次元圧縮なしの場合，つまり，各文書における各単語の出現頻度をそのままクラスタリングに用いた場合である．この場合は，スムージング・パラメータを非ゼロにし，スムージングを適用するほうが，クラスタリング性能は良くなる．LDA16，LDA32，

(注5)：http://nlp.kookmin.ac.kr/HAM/kor/

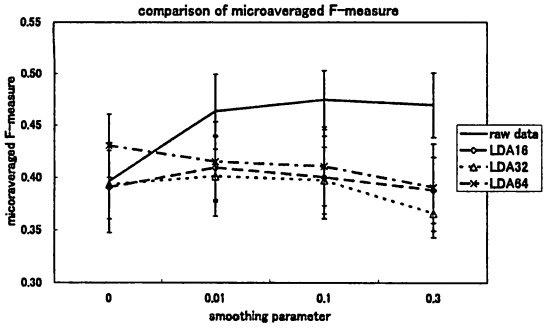

図 2 S2006 データセットの microaveraged F-measure
Fig. 2 microaveraged F-measure for S2006 dataset

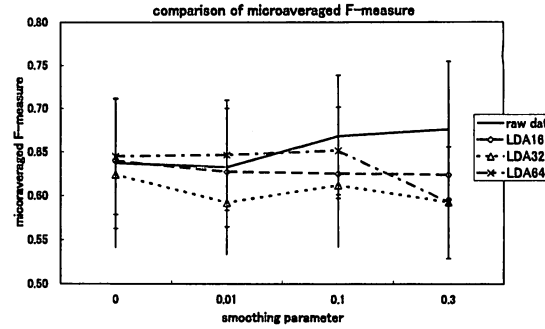

図 3 JIC データセットの microaveraged F-measure
Fig. 3 microaveraged F-measure for JIC dataset

LDA64 は，潜在的ディリクレ配分法により，文書の特徴ベクトルをそれぞれ 16 次元，32 次元，64 次元に圧縮したものをクラスタリングした結果である．いずれも，スムージングによって性能は良くならない．これは，次元圧縮自体がスムージングの効果を含むためだろう．また，標準偏差の ±1 倍の幅を示すマーカが，"raw data" の場合と重なり合っていることから，次元圧縮前後でクラスタリング性能に有意な差がないこともわかる．つまり，次元圧縮により，クラスタリング性能を劣化させることなくデータ量の縮小が実現できた．

ところで，潜在的ディリクレ配分法には，時間計算量，空間計算量が大きいという問題がある．実験に使用した PC は Intel Core2 6600 2.40GHz，メモリ 2G バイトを搭載している．文書数が 37,017，文書長の総和が 2,730,680 のデータセット S2006 の場合，トピック数 16 の場合はパラメータ推定終了まで約 40 分，トピック数 32 の場合は約 90 分だった．それぞれ，パラメータ推定のための反復計算の回数は 623 回，408 回だった．また，トピック数 64 の場合は 5 時間強を要したが，これはスワッピングが起こってしまったためであろう．この場合の反復計算の回数は 663 回だった．今回の実験では，データセット S2005 について，トピック数 64 の場合に他より少し良い F-measure を得たが，大きくクラスタリングの性能が改良されたわけではなかった．そのため，計算量の削減という観点からは，トピック数は 16 で十分と言える．

5.2 Japan.internet.com の場合

データセット JIC についての実験結果を，図 3 に示す．図の見方は S2005，S2006 の場合と同様である．やはり，標準偏差の ±1 倍を示すマーカが互いに重なり，次元圧縮によって，クラスタリング性能が本質的に落ちていない．次元圧縮前との性能の差は S2005，S2006 の場合よりもむしろ小さい．また，潜在的ディリクレ配分法のトピック数を 16，32，64 と変化させても，クラスタリング性能に大きな違いが見られない．

なお，このデータセットについては，次元圧縮にランダム投影法を用いた場合との比較の結果を，図 4 示しておく．ランダム投影法については，論文 [7] でテキスト情報における有効性が主張されているが，データセット JIC については，LDA による次元圧縮には遠く及ばなかった．ランダムに投影する空間の次元を 128，256，512 次元と増やしても，同様の結果しか得られなかった．また，S2005，S2006 についても，LDA に比べ，ランダム投影法は有効ではなかった．

6. おわりに

本論文では，潜在的ディリクレ配分法を，文書の特徴ベクトルの次元圧縮法として利用した．そして，次元を圧縮された特徴ベクトルを文書クラスタリングに用い，その性能を調べた．実験の結果，日本語と韓国語のニュース記事について，各文書における単語の出現頻度をそのまま特徴量として用いた場合に得られるクラスタリング性能をほとんど損なわずに，次元を圧縮できることが分かった．また，ランダム投影法との比較では，潜在的ディリクレ配分法による次元圧縮が，大きく有利であることが分かった．次の課題として，同じマルチトピック文書モデルではあるが，潜在的ディリクレ配分法よりもシンプルで，パラメータ推定に必要な計算量も小さいモデルである PLSI [14] を，次元圧縮法として用い，比較を行いたい．

さらに，今後は，潜在的ディリクレ配分法により得られる他の情報の有効利用について考えたい．なぜなら，今回の実験では，各文書におけるトピックの "頻度" と解釈できるパラメータしか使っていないが，変分ベイズ法をパラメータ推定に用いると，他にも，各文書に現れる各単語に各トピックが割り当てられる確率，また，各トピックにおける各単語の出現確率（表 3），さらに，文書集合全体での各トピックの "重要度" と解釈できる値が得られる．これらの値を，一文書中の重要な部分文書の発見など，具体的な応用に活用する方法を考えていきたい．

文　献

[1] http://www.grokker.com/
[2] http://www.mooter.com/
[3] http://www.quintura.com/
[4] http://vivisimo.com/
[5] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, Vol. 3, pp. 1107-1135, 2003.

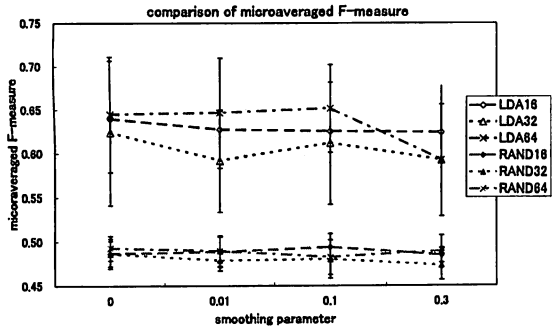

図 4 JIC データセットの microaveraged F-measure（ランダム投影法との比較）

Fig. 4 microaveraged F-measure for JIC dataset (comparison with random projection)

[6] M. J. Beal. Variational Algorithms for Approximate Bayesian Inference. PhD. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[7] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. *Proc. of KDD'01*, pp. 245-250, 2001.

[8] D. M. Blei, A. Y. Ng and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003.

[9] J. G. Conrad, K. Al-Kofahi, Y. Zhao and G. Karypis. Effective document clustering for large heterogeneous law firm collections. *Proc. of the 10th International Conference on Artificial Intelligence and Law*, pp. 177-187, 2005.

[10] A. P. Dempster, N. M. Laird and D. B. Rubin. Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *Journal of the Royal Statistical Society*, Series B (Methodological), Vol. 39, No. 1, pp. 1-38, 1977.

[11] P. K. Elango and K. Jayaraman. Clustering Images Using the Latent Dirichlet Allocation Model. available at `http://www.cs.wisc.edu/~pradheep/Clust-LDA.pdf`

[12] M. Fattori, Giorgio Pedrazzi and Roberta Turra. Text mining applied to patent mapping: a practical business case. *World Patent Information*, Vol. 25, pp. 335-342, 2003.

[13] T. Griffiths and M. Steyvers. Finding Scientific Topics. *Proc. of the National Academy of Sciences*, 2004.

[14] T. Hofmann. Probabilistic Latent Semantic Indexing. *Proc. of SIGIR'99*, pp. 50-57.

[15] F.-C. Hsu, A. J.C. Trappey, C. V. Trappey, J.-L. Hou and S.-J. Liu. Technology and knowledge document cluster analysis for enterprise R&D strategic planning. *International Journal of Technology Management*, Vol. 36, No.4 pp. 336-353, 2006.

[16] R. E. Madsen, D. Kauchak and C. Elkan. Modeling Word Burstiness Using the Dirichlet Distribution. *Proc. of ICML'05*, pp. 545-552, 2005.

[17] T. J. Malisiewicz, J. C. Huang and A. A. Efros. Detecting Objects via Multiple Segmentations and Latent Topic Models. available at `http://www.cs.cmu.edu/~tmalisie/sword_allocation/sword_allocation.pdf`

[18] K. Nigam, A. McCallum, S. Thrun and T. Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103-134, 2000.

[19] M. Yamamoto and K. Sadamitsu. Dirichlet Mixtures in Text Modeling. CS Technical report CS-TR-05-1, University of Tsukuba, 2005.

表 3 各トピックの最頻出語 8 個（JIC データセット，トピック数 32）

Table 3 Eight most frequent words for each topic when we use JIC dataset by setting the number of topics to be 32.

売上 億 四半期 前年 予測 増 同期 決算 比 利益
それ 自分 これ 何 場合 キーワード とき 点 問題 今
バイヤー カード 小売 ジャパン 業 マイ 新た 直接 必要 入会
社内 イン 受付 ソフトバンク トラ 業務 円 株 日本 システム
位 掲示板 メディア 新聞 テレビ 書 個人 地方 参考 家
脆弱 ウイルス 攻撃 確認 緊急 ファイル ウェア 長期 休暇 コンピュータ
オプション テープ 突破 問題 顧問 提訴 法務 準拠 ストレージ ソフトウェア
表示 ページ コンテンツ 結果 キーワード テーマ 型 ニュース 連動 クリック
調査 割 回答 結果 代 ポイント 購入 女性 テレビ 増加
プロセッサ 製 サン 業界 デザイン カスタム 価格 ジャケット 監査 デル
契約 業界 オリジナル 担当 大手 部門 顧客 次 明らか 型
拡張 同日 実験 ラボ 有料 以上 拡大 追加 時 運営
個人 保護 機関 団体 会 教育 政府 詐欺 米 米国
接続 通信 ネットワーク 無線 バンド ブロード 実験 回線 高速 機器
料金 無料 マイクロソフト モード 月額 円 通信 端末 料 会員
操作 音楽 搭載 フォン 通話 位置 採用 番号 着信 曲
ドメイン 名 所 レンタル 家 相談 兼 ゴルフ 用品 結婚
パソコン デジタル 写真 カメラ 画像 機器 価格 ノート プレーヤー ビデオ
コンテンツ 電子 デジタル タグ 書 映像 認証 発行 証明 暗号
デスク 定額 協業 マイクロソフト 制 トップ 米国 端末 大学 音楽
提携 共同 事業 構築 協業 両社 顧客 設立 分野 中小
アプリケーション プロセッサ ストレージ プラットフォーム オープンソース 標準 サポート 出荷 環境
ジャパン 集中 誰 業 小売 自由 マイ 直接 自身 明確
買収 株式 売却 従業 員 部門 億 保険 合併 株
プリンタ メモリ 重視 自宅 フロント 購入 画質 キヤノン エプソン ロー
位 キーワード 表 ランク 期間 集計 県 総合 時事 地震
中国 メーカー 市場 産業 世界 製造 通信 アジア 生産 地域
兼 体系 環境 トレーニング 富士通 製 住商情報システム サーバ 所 印刷
契約 コンテンツ テレビ 映画 州 メディア バンド ビデオ 放送 大手
特許 訴訟 侵害 和解 ライセンス 提訴 違法 ファイル 法 問題
ツール 認証 アプリケーション 価格 統合 バージョン 自動 日本語 パッケージ 連携
インター プラス 保存 カスタマイズ フェイス グリー 顔写真 無制限 枚数 ジュエル

[20] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma. Learning to cluster Web search results. *Proc. of SIGIR'04*, pp. 210-217, 2004.

[21] 上田修功. ベイズ学習（全 4 回）. 電子情報通信学会誌, Vol. 85, No. 4,6,7,8, 2002.

[22] 貞光九月, 三品拓也, 山本幹雄. 混合ディリクレ分布を用いたトピックに基づく言語モデル. 電子情報通信学会 D-II, Vol.J88-D-II, No. 9, pp. 1771-1779, 2005.

[23] 山本幹雄, 貞光九月, 三品拓也. 混合ディレクレ分布を用いた文脈のモデル化と言語モデルへの応用. 情報処理学会研究報告, SLP48, pp. 29-34, 2003.