Vol.2013-MPS-92 No.11
2013/2/27

OCRによる文字認識誤りを考慮した重み付きトピックモデルに関する検討

田村 一樹[1,a)]　吉川 大弘[1]　古橋 武[1]

概要：近年，スキャナの普及により，紙媒体の文書の電子化が急速に進んでいる．しかし，大量の電子化文書に対して個々にタグ付けやフォルダ分けを行うことは，時間や労力の面から困難である．そこで，スキャナにより自動で取り込まれた電子化文書に対し，OCR（光学文字認識）から得られるテキスト情報を用いて，分類・検索を行うシステムが有用であると考えられる．本稿では，LDA（潜在的ディリクレ配分法）を用いて，文書間の関係を抽出する手法について検討する．LDA は，OCR の認識率が低い場合にはその性能が低下することが報告されている．本稿では，単語の認識に対する信頼度を定義し，その信頼度に基づいて LDA における各単語の重み付けを行う手法を提案する．信頼度の妥当性を予備実験において確認したのち，実際の OCR 文書分類の実験を行い，提案手法により文書分類性能が向上することを示す．

キーワード：光学文字認識（OCR），潜在的ディリクレ配分法（LDA），文書分類，N-gram，トピックモデル

A Study on Weighting Topic Model Considering False Recognized Characters by OCR

Abstract: Recently, the digitization of paper-based documents is rapidly advanced through the spread of scanners. However, tagging or sorting a huge amount of scanned documents one by one is difficult in terms of time and effort. Therefore, the document retrieval system using the texts in the documents, which is available by OCR (Optical Character Recognition), will be useful. The aim of this study is to extract the relationships between documents using LDA (Latent Dirichlet Allocation). It is reported that the performance of LDA declines along with poor OCR recognition. This paper proposes the method which defines the reliability of the recognized words and weights the words in LDA based on their reliabilities. First, adequacy for the reliability is confirmed through the preliminary experiment. Then, the experiment to classify actual OCR documents are carried out, and it shows the improvement of the performance for the classification of documents by the proposed method.

Keywords: OCR, Latent Dirichlet Allocation, Document classification, N-gram, Topic Model

1. はじめに

近年，スキャナ及びスキャナ機能を持つプリンタの普及により，紙媒体の文書をコンピュータに取り込み，電子データとして扱う機会が増大している．特に企業においては，2005 年に施行された e-文書法により，多くの紙媒体文書が電子データで保存されるようになっている．また，タブレット端末の急速な普及により，気軽に電子的な文書を閲覧できることで，一般の消費者においても，大量の文書データが電子的に保存・蓄積されるようになってきている．さらに，クラウドコンピューティングの普及により，今後様々な種類の文書データを，一括管理する機会も急増していくと考えられる．しかし一方で，蓄積される文書データが多くなるほど，ユーザが目的とする文書を探し出すのに必要な時間と労力も多大なものになると予想される．スキャナによって取り込まれた文書のテキスト情報を検索などに利用するには，光学文字認識 (OCR: Optical Character Recognition) ソフトウェアを用いてテキスト部分を読み取ることが必要となる．しかし一般に，OCR で変換されたテキストは，少なからず読み取り誤りや変換誤

[1] 名古屋大学
Nagoya University
[a)] tamura@cmplx.cse.nagoya-u.ac.jp

りを含むため，文書の持っているテキスト情報を全て正しく電子化することはできない．OCR の性能を高める研究も行われているものの，不鮮明な活字や手書き文字など，未だに困難な課題が多く存在しており，それらを誤りなく認識することは難しい．

テキスト情報から文書の持つ特徴を捉える手法として，確率的潜在意味解析 (pLSA: Probablitistic Latent Semantic Analysis)[1] や潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation)[2] などのトピックモデルが報告されている．これらのトピックモデルは，文書に出現する単語とその出現回数の情報から，それぞれの文書に潜在的に存在するトピックを，精度よく推定することができる手法として知られている．しかし，OCR による誤りを含む文書に対してトピックモデルを適用すると，トピック推定性能が低下することが報告されている [3]．そこで本稿では，OCR によって電子化されたテキストに対して，代表的なトピックモデルの 1 つである LDA を適用する際に，トピックの推定性能を向上させる手法について検討する．

本稿では，OCR によって誤認識された部分が，言葉として不自然な並びになっている場合が多いことに着目する．そこで，文書から得られる単語の認識の信頼度を，N-gram 確率を用いて定義した上で，LDA に対し，信頼度が高い単語の出現を重視する重み付けを行う方法を提案する．

本稿では，視覚的に文書間の類似関係を把握するシステムを想定する．そこで実験では，得られた各文書のトピック分布から，文書間の距離を計算し，それらを 2 次元平面上に配置する問題を設定する．分類ラベルを保持する OCR 文書に対して実験を行い，同一正解ラベルの文書が近くに配置されることを分類精度として評価する．従来の LDA と，提案する重み付けによる LDA を適用した結果を比較し，分類精度の面で提案手法が優れていることを示す．

2. 従来研究

トピックモデルを用いた情報検索の研究は，これまでに数多く報告されている [4][5]．しかし，それらの多くの研究では，正しいテキスト情報を持つ文書を想定しており，誤りを含む OCR 文書に対してトピックモデルを適用した研究はあまり見られない．その中で，OCR 文書にトピックモデルを適用した研究も報告されているものの [7][6]，実際の誤認識に対するアプローチとしては，低頻度語の除去を行う程度であり，十分な対策や工夫がされているとは言い難い．また，OCR の認識率がトピックモデルに与える影響を調べた研究としては，Walker らの報告がある [3]．この研究では，様々な認識率の文書を想定し，各認識率の文書に対して LDA を適用している．実験によって，認識率が下がるほどトピック推定の性能も低下するという結果が示されているが，具体的にその問題を解決する方法については言及されていない．また，日本語の OCR 文書を想定してトピックモデルを適用した研究は，これまでに見当たらない．

一方，LDA に対して重み付けを行う手法は，Wilson らによって提案されている [8]．しかし，これは機能語や高頻度語の影響を抑えることでモデルの性能を向上させることが目的であり，本研究の目的とは異なる．

3. Latent Dirichlet Allocation

LDA は，文書が複数の潜在的なトピックを持ち，それらのトピックを媒介して単語が生成されることを仮定したモデルである．Blei らの LDA[2] ではトピックの出現を多項分布とみなし，その事前分布をディリクレ分布で仮定している．また，Griffiths らはこの LDA を拡張し，単語の分布にもディリクレ分布を導入した LDA を提案しており [9]，広く用いられている．本稿では，後者の Griffiths らによる LDA を採用する．

LDA において，文書の生成過程は以下のようにモデル化される．

(1) 各トピック $t \in \{1, \cdots, T\}$ について，単語分布 ϕ_t をディリクレ分布に従って生成する．
$$\phi_t \sim \mathrm{Dir}(\beta)$$
(2) 各文書 $i \in \{1, \cdots, D\}$ について，トピック分布 θ_i をディリクレ分布に従って生成する．
$$\theta_i \sim \mathrm{Dir}(\alpha)$$
(3) 文書 i に出現する単語 $j \in \{1, \cdots, N_i\}$ について：
(a) トピック $z_{i,j}$ を多項分布に従って生成する．
$$z_{i,j} \sim \mathrm{Mult}(\theta_i)$$
(b) 単語 $w_{i,j}$ を多項分布に従って生成する．
$$w_{i,j} \sim \mathrm{Mult}(\phi_{z_{i,j}})$$

ここで，$\mathrm{Dir}(\cdot)$ はディリクレ分布，$\mathrm{Mult}(\cdot)$ は多項分布を表し，α と β はそれぞれのディリクレ分布におけるハイパーパラメータである．また，T は総トピック数，D は総文書数，N_i は文書 i の総単語数を表す．

なお，トピックの推定には，容易に高精度な解が得られる手法として知られている，ギブスサンプリングを用いる [9]．ギブスサンプリングでは，ある位置 l のトピック z_l を，位置 l 以外の情報を用いて推定する．文書 i に含まれる単語 j のうち，トピック t に割り当てられたものの数を N_{ijt} と表し，また変数についての総和を添え字の $(\cdot)$ で表す．N_{ijt} のうち，位置 l を除いたものを N^{-l}_{ijt} と表記し，総語彙数（単語の種類数）を V とすると，ギブスサンプリングにおけるトピックの更新式は以下で表される．

$$p(z_l | z_{\setminus l}, \mathbf{w}) \propto \frac{N^{-l}_{(\cdot)jt} + \beta}{N^{-l}_{(\cdot)(\cdot)t} + V\beta} \cdot \frac{N^{-l}_{i(\cdot)t} + \alpha}{N^{-l}_{i(\cdot)(\cdot)} + T\alpha} \qquad (1)$$

トピックを十分な回数更新することで得られたサンプルから，全文書のトピック分布 $\boldsymbol{\theta}$，全トピックの単語分布 $\boldsymbol{\phi}$ についての MAP 推定量を得ることができる．文書 i でト

ピック t が生成される確率を θ_t^i，トピック t から単語 j が生成される確率を ϕ_j^t とすると，それらは式 (2)，式 (3) でそれぞれ求めることができる．

$$\theta_t^i = \frac{N_{i(\cdot)t} + \alpha}{N_{i(\cdot)(\cdot)} + T\alpha} \tag{2}$$

$$\phi_j^t = \frac{N_{(\cdot)jt} + \beta}{N_{(\cdot)(\cdot)t} + V\beta} \tag{3}$$

4. 提案手法

4.1 目的

LDA に代表されるトピックモデルは，文書中に出現する単語の種類とその回数の情報から，トピックを推定している．日本語などの分かち書きがされていない言語においては，形態素分割を行って単語の情報を得る必要があるが，誤りを含むテキストでは，不適切な分割が多く発生するため，得られる単語の情報も誤りを含んだものとなる．

ここで，“コミュニティシステム”という文字列を，“コミュ=デイシステム”と誤認識した例について述べる．形態素解析器 MeCab[10] を用いて，それぞれの文字列に対して形態素解析を行った結果を，**図 1** に示す．図 1(b) のように，誤認識部分が名詞として不適切に切り出されていることが確認できる．そこで本稿では，隣接する名詞や未知語を結合し，1 つの単語として扱った上で，単語 N-gram 確率を用いて求まる，構成する形態素同士の隣接確率を用いて，単語の認識の信頼度を定義し，それをトピックの推定に導入する．以降で単語の信頼度について述べ，続いてその重みを用いたトピックの推定について述べる．

コミュニティ	名詞, 一般, *, *
システム	名詞, 一般, *, *

(a) “コミュニティシステム”の解析結果

コミュ	名詞, 一般, *, *
=	名詞, サ変接続, *, *
デイシステム	名詞, 一般, *, *

(b) “コミュ=デイシステム”の解析結果

図 1 MeCab による形態素解析の結果

4.2 単語の信頼度

単語 N-gram 確率とは，ある N 個の単語（形態素）が隣接して出現する確率である．この確率は大規模なコーパスから得られ，確率が高いものは一般的に多く出現する自然な隣接パターンであり，低いものは不自然な隣接パターンであるということができる．本稿では $N = 2$ とした単語 Bi-gram 確率を，一つの単語内の形態素の隣接確率とし，信頼度計算に用いる．ここで，ある単語 w を構成する形態素が $t_1 t_2 \cdots t_n$ である場合を考えると，単語 w の Bi-gram 確率は，以下で表される．

$$p(w) = p(t_1) \times p(t_2|t_1) \times \cdots \times p(t_n|t_{n-1}) = p(t_1) \prod_{i=2}^{n} p(t_i|t_{i-1}) \tag{4}$$

単語 w における隣接確率の相乗平均値 $p_{\bar{t}}(w) = p(w)^{\frac{1}{n}}$ により，単語 w_i の信頼度 $m(w_i)$ を式 (5) のように定義する．

$$m(w_i) = \frac{\log p_{\bar{t}}(w_i)}{\arg\max_{w \in W} \log p_{\bar{t}}(w)} \tag{5}$$

ここで，W は文書集合中の全単語を表す．なお，$p(w) = 0$ のとき，$m(w) = 0$ とする．

4.3 Weighting LDA

Wilson らの重み付け手法 (WLDA)[8] では，3 節の LDA を発展させ，単語に対して重みを付けた形でのギブスサンプリングを行い，トピックを推定している．文献 [8] には明記されていないものの，この重み付けは多項分布を数学的に実数に拡張したものだといえる．具体的には，3 節にある LDA では，ある位置 l の単語とトピックは，それぞれ V 次元，T 次元の 1-of-K ベクトルで表される．つまり，該当の単語やトピックの次元の値のみ 1 で，その他の次元がすべて 0 であるベクトルである．WLDA では，該当の次元に実数値を割り当てることで，単語の重みをトピックの推定に反映させることができる．M_{ijt} を，文書 i に含まれる単語 j のうち，トピック t に割り当てられた重みの合計値とするとき，ギブスサンプリングにおけるトピックの更新式は式 (6) で表すことができる．

$$p(z_l|z_{\setminus l}, \mathbf{w}) \propto \frac{M_{(\cdot)jt}^{-l} + \beta}{M_{(\cdot)(\cdot)t}^{-l} + V\beta} \cdot \frac{M_{i(\cdot)t}^{-l} + \alpha}{M_{i(\cdot)(\cdot)}^{-l} + T\alpha} \tag{6}$$

本稿では，WLDA における重みに，4.2 で定義した単語の信頼度を用いる．認識の信頼度が高い語を重視したトピックの推定を行うことで，OCR 文書におけるトピックの推定性能の向上が期待できる．

5. 実験

初めに，4.2 で定義した信頼度の妥当性を評価するための予備実験を行う．ここでは，信頼度が低い単語に含まれる誤認識単語の割合を，F-measure を用いて検討する．続いて，実際の適用場面を想定した文書分類実験を行う．得られる分類精度を手法に対する定量的な性能評価指標とし，従来手法と提案手法とを比較する．

5.1 適用文書

本実験では入力文書として，情報処理学会第 74 回全国大会の講演論文を用いた．そのうち，4 セッション計 31 文書（データ 1），6 セッション計 48 文書（データ 2）からなるデータセットを作成し，それらを用いて評価を行った．データには，電子文書に元々埋め込まれている誤りのないテキストと，文書画像に対して OCR ソフトウェアを用い

ることで得られる，誤りを含んだテキストを用意した．そのうち誤りを含むテキストには，印刷の不鮮明な文書や，手書き文書など，OCR の認識率が低い文書が含まれることを想定し，文書画像にランダムにノイズを加えて OCR をかけることで，異なる認識率のテキストを作成した．なお，それぞれの文書について，属していたセッションを分類の正解ラベルとした．また，実際にコンピュータ上で文書を扱う際は，電子文書と OCR 文書が混在する場合が多いことを想定し，各セッションのうちランダムに選んだ半数の文書は誤りのないテキスト情報を，残りは OCR で読み取られたテキスト情報を用いた．

OCR の単語認識率の定量指標には，機械翻訳などの分野において，単語の誤り率として用いられる，PER(Position-independent Word Error Rate)[11] を用いた．PER は，単語の出現位置によらない，正解単語集合からの誤り率である．以降本稿では，単語認識率を $(1 - PER) \times 100[\%]$ で表す．ノイズのない文書画像に対する OCR 文書の単語認識率は約 75% であった．

なお，OCR ソフトウェアには Adobe Acrobat[12] を，形態素解析器には MeCab[10] を用いた．N-gram 確率は，Web 日本語 N グラム第 1 版 [13] を用いて算出した．

5.2 予備実験

予備実験では，4.2 で定義した信頼度の妥当性を評価する．実験データには，5.1 のデータセットのうち，各認識率の文書集合からそれぞれ 1000 語をランダムに抽出し，人手で認識の正誤をラベル付けしたものを用いた．比較は，N-gram に基づく信頼度を用いる方法と，低頻度語の除去 [7] による方法の間で行い，前者は閾値（事前実験により 0.30 と定めた）を下回った単語，後者は出現回数 1 の単語を抽出した．抽出された単語集合に含まれる誤認識単語の適合率，網羅率を求め，F-measure を用いて比較した．F-measure は，適合率と網羅率の調和平均で，高い値ほど優れた性能を表す指標である．単語の認識率を WRR，適合率を P，網羅率を R，F-measure を F と表し，それぞれのデータから得られた結果を**表 1** に示す．

まず適合率について着目すると，低頻度語の除去による方法が，N-gram による信頼度を用いる方法と比較して，著しく低い値となっていた．これは，低頻度語の除去による方法では，正しく認識されている数多くの単語も除去されていることが原因であった．一方，信頼度を用いる方法では，頻度によらず認識の信頼度の値を保持するため，正しく認識されている単語は除かずに，誤認識単語を除くことができていた．網羅率については，低頻度語の除去による方法が高くなった．これは，多くの誤認識単語は 1 度しか出現せず，低頻度語の除去によってほとんど取り除くことができるためであった．これらを F-measure によって評価すると，N-gram による信頼度を用いる方法が，低頻度語の除去による方法を大きく上回る結果となった．この結果から，提案する単語の信頼度において信頼度の低い語は，適切に誤認識単語を表していることが確認できた．

表 1 誤認識単語抽出性能の比較
Table 1 Comparison of Performance to Extract False Recognized Words

(a) 誤認識単語抽出性能（データ 1）

	Low frequency			Reliability		
WRR	P	R	F	P	R	F
72	0.305	0.924	0.459	0.568	0.674	0.616
53	0.452	0.956	0.613	0.756	0.809	0.782
47	0.508	0.936	0.659	0.806	0.754	0.779
40	0.508	0.958	0.664	0.837	0.794	0.815
35	0.568	0.947	0.710	0.879	0.804	0.840
30	0.566	0.957	0.711	0.823	0.807	0.815

(b) 誤認識単語抽出性能（データ 2）

	Low frequency			Reliability		
WRR	P	R	F	P	R	F
77	0.292	0.910	0.442	0.569	0.763	0.652
66	0.443	0.952	0.604	0.728	0.765	0.746
57	0.530	0.954	0.682	0.782	0.803	0.792
49	0.601	0.948	0.736	0.823	0.770	0.795
43	0.626	0.969	0.761	0.858	0.813	0.835
34	0.670	0.937	0.781	0.896	0.808	0.850
30	0.692	0.940	0.797	0.917	0.814	0.863

5.3 文書分類実験

続いて，5.1 で述べたデータに対する文書分類精度に基づき，従来手法と提案手法の性能比較を行った．

5.3.1 可視化システム

実験において想定する可視化システムについて述べる．本システムではまず，入力された各文書に含まれるトピック分布を計算する．トピック分布の類似度を距離指標として，Jensen-Shannon 情報量 [14] を用いて各文書間の距離を計算する．式 (7) で表される Jensen-Shannon 情報量は，Kullback-Leibler 情報量を対称化したもので，確率分布間の距離指標として用いられている [15].

$$D(d_i, d_j) = \sum_{k=1}^{K} \left[p(z_k|d_i) \log \left(\frac{2p(z_k|d_i)}{p(z_k|d_i) + p(z_k|d_j)} \right) + p(z_k|d_j) \log \left(\frac{2p(z_k|d_j)}{p(z_k|d_i) + p(z_k|d_j)} \right) \right] \quad (7)$$

最後に，文書間の距離関係を 2 次元平面上に可視化する．本システムでは，次元削減の手法として，多次元尺度構成法 (MDS: Multi-dimensional Scaling)[16] を用いる，MDS では，元空間におけるデータ間の距離関係をできるだけ保存しながら，低次元空間の座標にデータを埋め込むことができる．文書 d_i と d_j の元空間での距離を $l_{i,j}$，可視化空間での距離を $l^*_{i,j}$ とすると，各文書の座標

$\chi = \{x_i \in \mathbf{R}^2, i = 1, 2, \ldots, N\}$（$N$：データ数）は，式 (8) の誤差関数を最小化することによって求められる．ただし，可視化空間での距離 $l^*_{i,j}$ は，$l^*_{i,j} = ||x_i - x_j||$ のユークリッド距離で与えられる．

$$E(\chi) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (l^*_{i,j} - l_{i,j})^2 \quad (8)$$

5.1 のデータ 1 を用いたときの可視化結果の一例を**図 2** に示す．図において，“1E-4” などは各文書の講演番号を表し，正解ラベル（セッション）別に色分けを行っている．

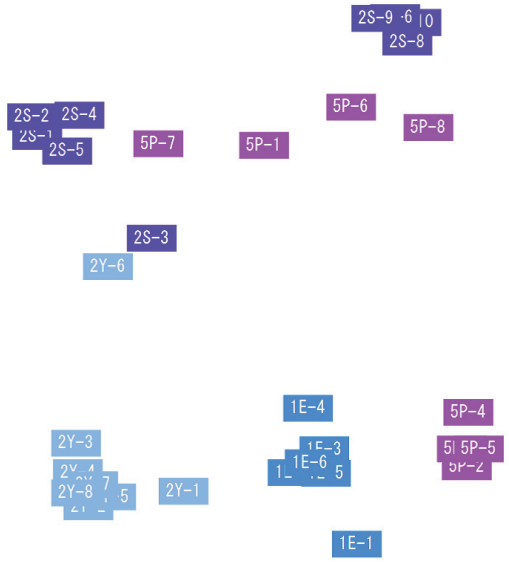


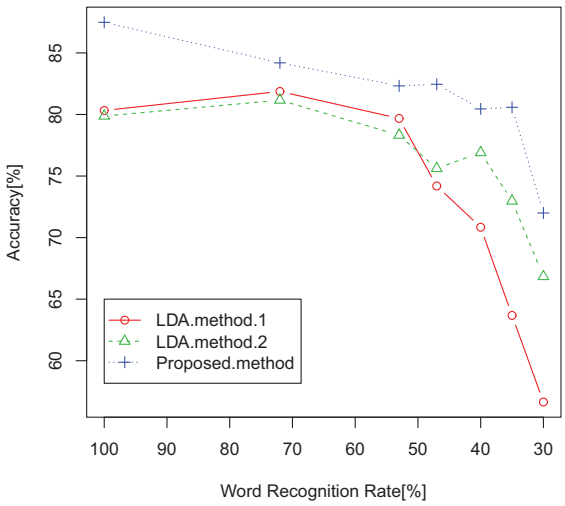
図 2 可視化結果の例
Fig. 2 Example of Visualization Result

5.3.2 評価指標

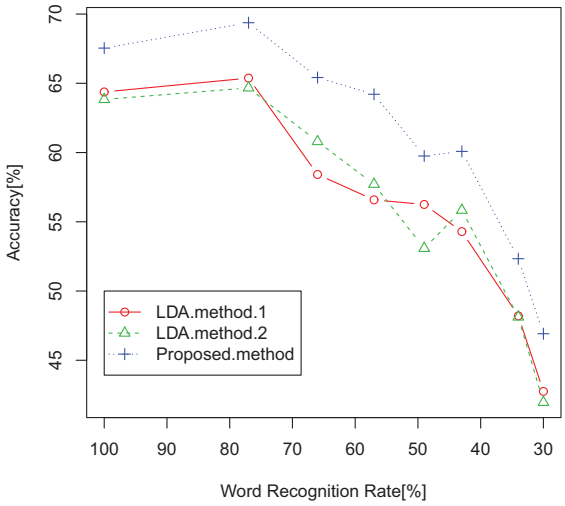
可視化システムにおける目的は，類似した文書を近くに配置して提示することで，ユーザに文書間の関係の直感的な理解を促すことである．これに対する評価指標として，本稿では，可視化空間における k 近傍法による予測精度を用いる [5]．この方法では，ある文書に対するラベルを，その近傍 k 個の文書ラベルの多数決で予測する．そして全文書のうち，正しくラベルが推定された文書の割合を，分類精度と定義する．分類精度は，同じラベルを持つ文書が近くに，異なるラベルを持つ文書が遠くに配置されるほど高い値となる．本実験では，$k = 5$ として，この k 近傍法による分類精度を用いて評価を行った．なお，図 2 の例では，分類精度は約 87% となった．

5.3.3 実験条件

LDA, WLDA のハイパーパラメータは, それぞれ $\alpha = 0.1$, $\beta = 0.1$ とし，サンプリング回数は 1000 回とした．また，これらの結果はランダムに与える初期トピックに依存するため，50 試行の平均を分類精度として用いた．LDA, WLDA におけるトピック数は，電子文書を用いた予備実験において，最も予測精度が高くなったトピック数とし，



(a) 分類精度（データ 1）



(b) 分類精度（データ 2）

図 3 各単語認識率における分類精度の比較
Fig. 3 Comparison of Classification Accuracy in each Word Recognition Rate

データ 1 で $T = 4$，データ 2 で $T = 6$ であった．これは結果として，用いたデータのセッション数（正解ラベル数）と一致した．比較手法には，従来の LDA を用いる方法（以降，LDA 手法 1 と表記する）と，5.2 で検討した，出現回数 1 の単語を除去する前処理を行った上で LDA を用いる方法（以降，LDA 手法 2 と表記する）を用いた．

5.3.4 結果と考察

データ 1,2 それぞれにおける分類精度の結果を**図 3**(a), (b) に示す．まず，LDA 手法 1 について着目すると，データ 1 では 50% 付近，データ 2 は 70% 付近から急激に分類精度が低下していることが確認できた．したがって，文

献 [3] で述べられている，OCR の誤りによる LDA の性能の低下を，分類精度の観点から確認することができた．また，LDA 手法 2 は，データ 1 の認識率が低い部分において，LDA 手法 1 より若干の精度向上が見られたものの，全体的には LDA 手法 1 とあまり変わらない結果となった．それに対し提案手法では，異なる認識率の文書において，総じて LDA 手法 1,2 よりも高い分類精度が得られた．この結果に対して，シダックの統計検定法を用いて多重性を考慮した対応のある t 検定を行ったところ，LDA 手法 1 と提案手法，LDA 手法 2 と提案手法の間でそれぞれ有意差がみられた（$p < 0.01$）．特にデータ 1 について，LDA 手法 1 の性能が大幅に低下する認識率においても，提案手法は依然高い値を保っており，OCR の誤りによるトピックモデルの性能低下を抑える働きをしていることが確認できた．データ 2 においては，データ 1 ほどの効果は見られなかったものの，LDA 手法 1 の性能が低下する認識率付近では，提案手法と LDA 手法 1 の差が大きくなっており，データ 1 と同様の傾向がある結果となっていた．

しかし，全体的な性能の向上は見られたものの，提案手法においても，分類精度は認識率の低下とともに低下する結果であった．これは，提案手法は誤認識単語のトピック推定への影響を抑えるアプローチであり，誤認識された単語を正しく修正するものではないため，正しく認識されていればトピック推定に有用であったはずの語の情報を使えていないことが原因であると考えられる．今後は，OCR の誤りによって生じ得る，表記が似ている単語の情報などを用いて正しい単語を推定・修正し，トピックの推定に反映させる方法などについて検討する必要があると考えられる．

6. おわりに

本稿では，OCR で文字認識された文書から特徴を抽出する手法として，LDA を用いる上で従来報告されていた，OCR の誤認識によるトピック推定性能の低下を抑える方法を提案した．提案手法では，OCR によって誤認識された部分は，言葉として不自然な並びになっている場合が多いことに着目し，N-gram 確率を用いて単語の認識の信頼度を定義した．また，LDA において，信頼度が高い単語の出現を重視する重み付けを行い，OCR 文書における LDA の性能の向上を試みた．

初めに，予備実験により，単語の認識信頼度の妥当性を評価し，低頻度語の除去を行う手法と比較して適切に誤認識単語を抽出できていることを確認した．続いて，文書の類似性を 2 次元平面上で可視化するシステムを想定し，従来の LDA と比較して，分類精度の面で提案手法が優れていることを示した．

今後の課題として，OCR の誤認識単語の情報も用いて，トピックの推定性能を向上させる方法についての検討や，文書を特徴づける重要語の抽出，それらへの重み付けを行うことなどが挙げられる．また，実際に分類・検索するシステムを構築し，提案手法に対する有用性の検証を進めていきたい．

謝辞 本研究は，文部科学省科学研究費（基盤研究 (C)，No.22500088）の補助を得て遂行された．

参考文献

[1] T. Hofmann: Probabilistic latent semantic indexing, SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50-57, 1999
[2] D.M. Blei et al.: Latent dirichlet allocation, The Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003
[3] D.D. Walker et al.: Evaluating models of latent document semantics in the presence of OCR errors, EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 240-250, 2010
[4] X. Wei et al.: LDA-based document models for ad-hoc retrieval, SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 178-185, 2006
[5] T. Iwata et al.: Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents, KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 363-371, 2008
[6] D.J. Newmann et al.: Probabilistic topic decomposition of an eighteenth-century American newspaper, Journal of the American Society for Information Science and Technology, Vol. 57, No. 6, pp. 753-767, 2006
[7] D.M. Blei et al.: Dynamic topic models, ICML '06 Proceedings of the 23rd international conference on Machine learning, pp. 113-120, 2006
[8] A.T. Wilson et al.: Term Weighting Schemes for Latent Dirichlet Allocation, HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 465-473, 2010
[9] T.L. Griffiths et al.: Finding scientific topics, Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences, Vol. 101, No. 1, pp. 5228-5235, 2004
[10] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[11] F.J. Och et al.: Improved alignment models for statistical machine translation, Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, pp. 20-28, 1999
[12] アドビ システムズ株式会社: Adobe Acrobat 9.46, http://www.adobe.com/jp/
[13] 工藤拓, 賀沢秀人:Web 日本語 N グラム第 1 版
[14] J. Lin : Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory, Vol. 37, No. 1, pp. 145-15, 1991
[15] G. Heinrich: Parameter estimation for text analysis, Technical Note, Ver.2.4, 2008
[16] W.S. Torgerson: Multidimensional scaling: I. Theory and method, Psychometrika, Vol. 17, No. 4, pp. 401-419, 1952