情報処理学会論文誌　数理モデル化と応用　Vol.6　No.3　66–74 (Dec. 2013)

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]

受付日 2013年1月15日，再受付日 2013年3月20日，
採録日 2013年4月16日

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．それらの電子化文書は一般的に，タグ付けやフォルダ分けによって管理されている．しかし，大量の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．したがって，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，自動で文書の特徴を抽出し，分類・検索を行うシステムが有用であると考えられる．代表的なトピックモデルの 1 つである LDA（潜在的ディリクレ配分法）は，各文書の特徴や文書間の関係を高性能に抽出する手法として知られている．LDA を用いて情報検索を行う研究は数多く報告されているものの，それらの多くは正しいテキスト情報を持つ文書を想定しており，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．また，日本語の OCR 文書に対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．本論文では，日本語の OCR 文書に対して LDA を適用する際に，トピックの推定性能を向上させる方法について検討を行う．本論文では，N-gram を用いて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．提案手法において定義した信頼度が，適切に誤認識単語を検出できること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

Kazuki Tamura[1,a)]　Tomohiro Yoshikawa[1]　Takeshi Furuhashi[1]

Received: January 15, 2013, Revised: March 20, 2013,
Accepted: April 16, 2013

Abstract: Recently, the digitization of paper-based documents is rapidly advanced through the spread of scanners. These documents are usually managed by tagging or sorting into folders on a computer. However, tagging or sorting a huge amount of scanned documents one by one is difficult in terms of time and effort. Therefore, the system which extracts features from texts in the documents automatically, which is available by OCR (Optical Character Recognition), and classifies/retrieves documents will be useful. LDA (Latent Dirichlet Allocation), one of the most popular Topic Models, is well known as a method to extract the features of each document and the relationships between documents. Though many studies of information retrieval using LDA have been reported, most of them assume the documents containing correct texts. However, it is reported that the performance of LDA declines along with poor OCR recognition. In addition, no study applying Topic Model to Japanese OCR'ed documents has been shown. This paper assumes the application of LDA to Japanese OCR'ed documents and proposes a method to improve the performance of topic inference. The proposed method defines the reliability of the recognized word using N-gram and weights the words in LDA based on their reliabilities. Adequacy for the reliability of the recognized words is confirmed through the preliminary experiment detecting false recognized words. 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，スキャナおよびスキャナ機能を持つプリンタの普

[1] 名古屋大学
Nagoya University, Nagoya, Aichi 464–8603, Japan
[a)] tamura@cmplx.cse.nagoya-u.ac.jp

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，作成・保存を義務付けられている文書や帳票を，電子データで扱うことが認められたため，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．また一方，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．

電子的に作成された文書は，テキスト情報に加え，様々な属性情報を保持している場合がほとんどであり，それらの情報を利用することで，文書内や対象文書の検索が可能である．しかしスキャナによって取り込まれた文書は，そのままでは画像として扱われるため，テキスト情報を利用することができない．そこで，それらの検索を行うためには，光学文字認識（OCR: Optical Character Recognition）を用いてテキスト部分を読み取り，テキストを埋め込むことが必要となる．OCR は，活字の文書の画像をコンピュータが扱える文字コードに変換するソフトウェアであるが，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りを含むため，文書の持っているテキスト情報をすべ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，不鮮明な活字など，いま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．特に日本語の文書では，アルファベットと比べて文字の種類が多いことで，英語などの文書と比べて多くの誤りを含むのが現状である．しかし，膨大な量の文書の誤り箇所すべてを人手で修正することは，時間やコストの面から困難である．また，文書を取り込むたびに，属性情報などの細かなタグ付けやフォルダ分けを行うことも，ユーザへの大きな負担となると考えられる．そこで本研究では，膨大な文書に対して自動的にスキャナ・OCR をかけることを想定し，誤認識を含み，かつタイトルや文書の種類といった属性情報は付加されていない形で電子化された文書を蓄積し，それら文書中のテキスト情報を用いることで，ユーザが必要とする文書を検索するシステムの構築を目的とする．

テキスト情報から文書の持つ特徴をとらえる手法として，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，精度良く推定することができる手法として知られている．これらトピックモデルは，これまで主に，正しいテキストの情報が与えられた文書に対して適用されてきた．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]．これに対し，英語の OCR 文書を対象として性能向上を試みた研究はあるものの [4]，そもそも英語は分かち書きされた単位で意味を持つため，ある単語の認識の正誤を一意に定めることができる．一方，日本語は分かち書きがされておらず，何らかの形で意味を持つ単位へと分割する必要がある．しかし，日本語で用いられる漢字は表意文字であり，1 文字でも形態素として機能する．したがって，ある切り出された形態素の認識の正誤について，その形態素の情報のみでは一意に決めることができない．そのような日本語の特性を考慮し，本論文では，周辺の形態素の情報を用いて単語に対する認識の信頼度を計算し，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．

本論文では，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．それら誤認識の部分は，大規模なコーパスから得られる N-gram 確率において，低い値を持つと予想される．そこで本論文では，文書から得られる単語が正しく認識されたものか，誤って認識されたものかを判断する信頼度を，N-gram 確率を用いて定義したうえで，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．

本論文では，Fortuna ら [5] や Iwata ら [6] と同様に，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．これは，蓄積される文書が膨大になるほど，直感的に分類・検索を行うことが有益となるためであり，たとえば検索語を明示的に与えることなく検索を実行するためには，ユーザへの能動的なアプローチである可視化によってユーザに文書の類似関係を提示するという方法が有効であると考えられるためである．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 [6], [7], [8]．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として，Newmann らの研究 [9] や，Blei らの研究 [10] などがある．これらは OCR 文書を実験に用いてはいるものの，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いい難い．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．英語の OCR 文書を対象とした Yang らの研究はあるものの [4]，分かち書きされた英語の文書を想定しており，日本語のように分かち書きされていない文書や，表意文字である漢字を含む文書を想定していない．

また，日本語の文書を対象にトピックモデルを適用した研究も，近年数多く報告されている [11], [12]．しかし，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 [13]．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）

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見なし，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 [14]，広く用いられている．本論文では，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，Dir(·) はディリクレ分布，Mult(·) 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．

文書における単語列を $\mathbf{w}$，それぞれの単語に対応するトピックを $\mathbf{z}$ とする．それらの結合分布は $p(\mathbf{w}, \mathbf{z}|\alpha, \beta)$ で表される．ここで，トピックと単語の独立性により，この結合分布は式 (1) のように表すことができ，式 (1) の各項はそれぞれ式 (2)，式 (3) のように表せる．

$$p(\mathbf{w}, \mathbf{z}|\alpha, \beta) = p(\mathbf{w}|\mathbf{z}, \beta)p(\mathbf{z}|\alpha) \tag{1}$$

$$p(\mathbf{w}|\mathbf{z}, \beta) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(N_{(\cdot)jt} + \beta)}{\Gamma(N_{(\cdot)(\cdot)t} + W\beta)} \tag{2}$$

$$p(\mathbf{z}|\alpha) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{i=1}^{D} \frac{\prod_{t=1}^{T} \Gamma(N_{i(\cdot)t} + \alpha)}{\Gamma(N_{i(\cdot)(\cdot)} + T\alpha)} \tag{3}$$

ただし，N_{ijt} は，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表し，また添え字の (·) はその変数についての総和を表す．W は総語彙数（単語の種類数）である．

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 [14]．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．N_{ijt} のうち，位置 l を除いたものを N_{ijt}^{-l} と表記すると，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．

$$\begin{aligned} p(z_l|z_{\setminus l}, \mathbf{w}) &\propto \frac{p(\mathbf{w}|\mathbf{z})p(\mathbf{z})}{p(w_{\setminus l}|z_{\setminus l})p(z_{\setminus l})} \\ &= \frac{N_{(\cdot)jt}^{-l} + \beta}{N_{(\cdot)(\cdot)t}^{-l} + W\beta} \cdot \frac{N_{i(\cdot)t}^{-l} + \alpha}{N_{i(\cdot)(\cdot)}^{-l} + T\alpha} \end{aligned} \tag{4}$$

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，文書 i のトピック分布 $\boldsymbol{\theta}^i$，トピック t の単語分布 $\boldsymbol{\phi}^t$ 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 (5)，式 (6) でそれぞれ求めることができる．

$$\theta_t^i = \frac{N_{i(\cdot)t} + \alpha}{N_{i(\cdot)(\cdot)} + T\alpha} \tag{5}$$

$$\phi_j^t = \frac{N_{(\cdot)jt} + \beta}{N_{(\cdot)(\cdot)t} + W\beta} \tag{6}$$

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．得られるトピックの情報を用いることで，単語レベルの共起情報では計れなかった文書どうしの関係性などをとらえることが可能となり，文書検索などの分野で応用することができる．しかし，OCR によって得られるテキストには多くの誤りが含まれる．したがって，そのテキスト

から得られる単語の情報も，同様に誤りを含んだものとなる．英語などの分かち書きがされる言語では，認識の誤りの有無にかかわらず，空白文字を情報に単語を切り出すことが可能である．一方，日本語などの分かち書きがされない言語においては，事前処理によって文字列から単語などの特徴量を取得する必要がある．一般的に，その処理には形態素解析が多く用いられる．形態素解析では，文字列に対し，意味を持つ最小の単位である形態素に分割を行う．この分割の基準は電子文書・OCR 文書で共通であるため，それらの間で横断的に用いることができる．また，テキストの特徴量として，極大部分文字列を用いる方法も提案されている [15]．この方法では，文書中の情報のみを用いて，文書を特徴づける長い文字列を抽出し，文書分類に役立てている．しかし，OCR 文書では文字列に不規則なノイズが含まれるため，特に電子文書と OCR 文書とで得られる部分文字列の長さなどの傾向が異なり，共通の特徴として用いることが難しくなると考えられる．そのため本研究では，それらから統一した基準で特徴を抽出することのできる，形態素解析を用いる．

ただし，形態素解析で得られた情報をそのまま扱うと，正しく認識された単語も誤って認識された単語も同列に扱うこととなる．誤って認識された単語はコンピュータにとってノイズであり，トピックの推定に悪影響を及ぼすと考えられる．そこで提案手法では，誤認識単語の影響を抑え，正しく認識できたと考えられる単語の影響を重視することを目的とする．

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．形態素解析器として広く用いられている，「MeCab」[16] と「ChaSen」[17] 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**，**図 2** に示す．図 1 (b)，図 2 (b) のように，誤認識部分が名詞や未知語として不適切に切り出されていることが確認できる．また，これはその他の誤認識文字を含む文字列でも同様であることが確認できた．これら OCR の誤りによって得られる形態素は，コンピュータにとってノイズになると考えられるものの，その形態素自身から認識の正誤を判断することは困難である．そこで本論文では，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うえで，単語 N-gram 確率を用いて求まる，構成する形態素どうし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例
Fig. 1　Morphological analysis using MeCab.

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　未知語
=　未知語
デイシステム　未知語

(b) “コミュ=デイシステム” の解析結果

図 2　ChaSen による形態素解析の例
Fig. 2　Morphological analysis using ChaSen.

4.2　単語の信頼度

N-gram 確率は，N 個の文字または単語（形態素）の隣接の確率であり，それぞれ文字 N-gram，単語 N-gram として用いられ，$N = 1$ では Uni-gram，$N = 2$ では Bi-gram と呼ばれている．また，N-gram 確率は大規模なコーパスの統計的な頻度データから得られるため，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，低いものは一般的には登場しない不自然な隣接パターンであるということができる．本論文では形態素 Bi-gram 確率を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 $c_1 c_2 \cdots c_n$ である場合を考えると，単語 w の形態素 Bi-gram 確率は，以下で表される．

$$\begin{aligned} p(w) &= p(c_1) \times p(c_2|c_1) \times \cdots \times p(c_n|c_{n-1}) \\ &= p(c_1) \prod_{i=2}^{n} p(c_i|c_{i-1}) \end{aligned} \tag{7}$$

式 (7) から，単語 w における形態素隣接確率の相乗平均値 $p_{\bar{c}}(w) = p(w)^{\frac{1}{n}}$ により，単語 w_i の信頼度 $m(w_i)$ を式 (8) のように定義する．

$$m(w) = \frac{\log p_{\bar{c}}(w_i) - \arg\min_{w \in W} \log p_{\bar{c}}(w)}{\arg\max_{w \in W} \log p_{\bar{c}}(w) - \arg\min_{w \in W} \log p_{\bar{c}}(w)} \tag{8}$$

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）[13] では，3 章の 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 [13] 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 章にある LDA では，ある位置 l の単語とトピックは，それぞれ W 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 (9) で表すことができる．

$$p(z_l|z_{\backslash l}, \mathbf{w}) = \frac{M^{-l}_{(\cdot)jt} + \beta}{M^{-l}_{(\cdot)(\cdot)t} + W\beta} \cdot \frac{M^{-l}_{i(\cdot)t} + \alpha}{M^{-l}_{i(\cdot)(\cdot)} + T\alpha} \tag{9}$$

本論文では，WLDA における重みに，4.2 節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 節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．本論文では，視覚的に文書間の類似関係を把握するシステムを想定し，各文書間の関係性をできるだけ保持したまま，2 次元平面上に文書を配置する問題を設定する．実験では，OCR 文書に対して提案手法を適用し，内容の類似する文書が近くに，類似しない文書が遠くに，それぞれ配置されることの評価を行う．ここでは，正解となる分類ラベルが付与された文書群に対し，個々に文書のラベルを未知としたときの分類精度を算出・比較することによって定量的な評価を行う．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，構成するセッション数・文書数が異なるデータセットを 2 つ作成し，それぞれ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は，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．これは，印刷の不鮮明な文書などの OCR の認識率が低い文書を想定している．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．データ 1 は，4 セッション（災害時通信，演奏支援，P2P ネットワーク，ニューラルネット）の計 31 文書（講演論文）から構成され，データ 2 は，6 セッション（映像・画像の生成と編集，音声・音楽・ゲーム，ロボットインタラクション，クラウド・大規模ネットワーク，プログラミング教育，検索・分類）の計 48 文書からなる．

また，実際にコンピュータ上で文書を扱う際は，正しいテキスト情報を持つ文書（電子文書と呼ぶ）と，OCR によるテキスト情報を持つ文書（OCR 文書と呼ぶ）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電子文書を，残りは OCR 文書を用いた．

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）[18] 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．以降本論文では，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 [19] を，形態素解析器には MeCab [16] を用いた．N-gram 確率は，Web 日本語 N グラム第 1 版 [20] を用いて，最尤値により算出した．

5.2 予備実験

予備実験では，4.2 節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 節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 1,000 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 [10] 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，再現率を求め，F-measure を用いて比較した．F-measure は，適合率と再現率の調和平均で，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，再現率を 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．また，信頼度の統計情報として，各認識率における単語の平均構成形態素数，信頼度が 0（$m(w) = 0$）となる単語の割合を**表 2** 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．再現率については，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．

5.3 文書分類実験

続いて，5.1 節で述べたデータに対する文書分類精度に

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

表 2　信頼度の統計情報
Table 2　Statistical information of reliability.

(a) データ 1

WRR	平均構成形態素数	$m(w)=0$ の割合
72	2.408	0.251
53	2.499	0.352
47	2.653	0.362
40	2.476	0.361
35	2.485	0.411
30	2.748	0.430

(b) データ 2

WRR	平均構成形態素数	$m(w)=0$ の割合
77	2.575	0.265
66	2.428	0.335
57	2.484	0.395
49	2.634	0.427
43	2.571	0.459
34	2.683	0.484
30	2.764	0.524

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 [21] を用いて各文書間の距離を計算する．式 (10) 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 [22]．

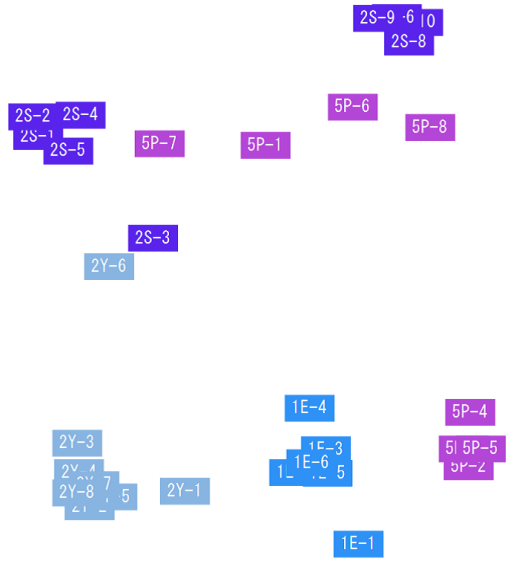


図 3　可視化結果の例
Fig. 3　Example of visualization result.

$$D(d_i,d_j) = \sum_{k=1}^{K}\left[p(z_k|d_i)\log\left(\frac{2p(z_k|d_i)}{p(z_k|d_i)+p(z_k|d_j)}\right) + p(z_k|d_j)\log\left(\frac{2p(z_k|d_j)}{p(z_k|d_i)+p(z_k|d_j)}\right)\right] \quad (10)$$

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）[23] 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 (11) 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．

$$E(\chi)=\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(l^*_{i,j}-l_{i,j})^2 \quad (11)$$

5.1 節のデータ 1 を用いたときの可視化結果の一例を**図 3** 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．

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，本論文では，可視化空間における k 近傍法による予測精度を用いる [6]．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

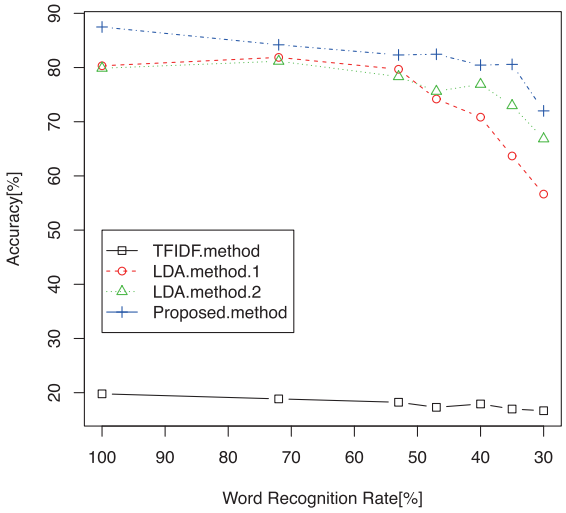
近くに，異なるラベルを持つ文書が遠くに配置されるほど高い値となる．本実験では，$k = 5$ として，この k 近傍法による分類精度を用いて評価を行った．なお，図 3 の例では，分類精度は約 87%となった．

5.3.3　実験条件

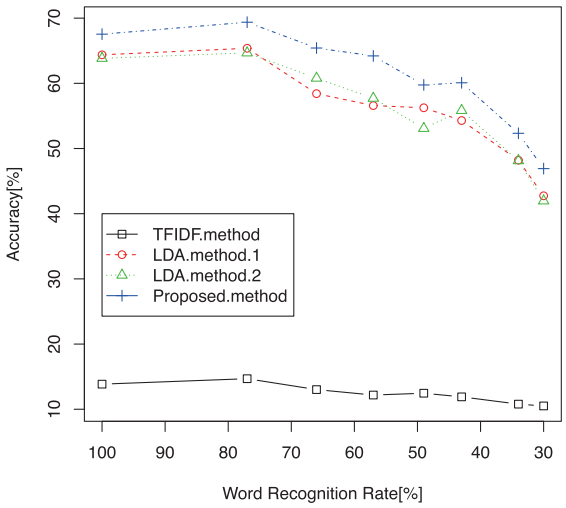
LDA, WLDA のハイパーパラメータは, それぞれ $\alpha = 0.1$, $\beta = 0.1$ とし，サンプリング回数は 1,000 回とした．また，これらの結果はランダムに与える初期トピックに依存するため，50 試行の平均を分類精度として用いた．LDA, WLDA におけるトピック数は，電子文書を用いた予備実験において，最も予測精度が高くなったトピック数とし，データ 1 で $T = 4$，データ 2 で $T = 6$ であった．これは結果として，用いたデータのセッション数（正解ラベル数）と一致した．比較手法には，TFIDF によって重み付けされた単語ベクトルを用いる方法（以降，TFIDF 手法と表記する），従来の LDA から得られるトピック分布を用いる方法（以降，LDA 手法 1 と表記する）と，5.2 節で検討した，出現回数 1 の単語を除去する前処理を行ったうえで LDA を適用し，得られるトピック分布を用いる方法（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 4** (a), (b) に示す．図 4 から，単語ベクトルを用いる TFIDF 手法と比較し，トピック分布を用いる方法は，総じて高い分類精度を得られていることが分かる．この結果から，OCR 文書においても，表層情報のみを用いる TFIDF 手法と比較し，潜在的なトピックの情報を用いる手法が有効であることが確認できた．トピックの情報を用いる方法について，まず，LDA 手法 1 に着目すると，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，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（$p < 0.01$）．なお，多重性はシダックの統計検定法を用いて考慮し，各群の名義水準を $\alpha' = 0.00335$ とした．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



(a) 分類精度（データ 1）



(b) 分類精度（データ 2）

図 4　各単語認識率における分類精度の比較

Fig. 4　Comparison of classification accuracy in each word recognition rate.

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うる，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

本論文では，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 を用いるうえで従来報告されて

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あげられる．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] Hofmann, T.: Probabilistic latent semantic indexing, *Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '99*, pp.50–57 (1999).

[2] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet allocation, *The Journal of Machine Learning Research*, Vol.3, pp.993–1022 (2003).

[3] Walker, D.D., Lund, W.B. and Ringger, E.K.: Evaluating models of latent document semantics in the presence of OCR errors, *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, *EMNLP '10*, pp.240–250 (2010).

[4] Yang, T. and Lee, D.: Towards noise-resilient document modeling, *Proc. 20th ACM International Conference on Information and Knowledge Management*, pp.2345–2348 (2011).

[5] Fortuna, B., Grobelnik, M. and Mladenic, D.: Visualization of text document corpus, *Informatica*, Vol.29, No.4, pp.497–502 (2005).

[6] Iwata, T., Yamada, T. and Ueda, N.: Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents, *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD '08*, pp.363–371 (2008).

[7] Wei, X. and Croft, W.B.: LDA-based document models for ad-hoc retrieval, *Proc. 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '06*, pp.178–185 (2006).

[8] Yao, L., Mimno, D. and McCallum, A.: Efficient methods for topic model inference on streaming document collections, *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *KDD '09*, pp.237–946 (2009).

[9] Newmann, D.J. and Block, S.: 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).

[10] Blei, D.M. and Lafferty, J.D.: Dynamic topic models, *Proc. 23rd International Conference on Machine Learning*, *ICML '06*, pp.113–120 (2006).

[11] 横山正太朗，江口浩二，大川剛直：潜在トピックを用いたブログ空間からの情報伝搬ネットワーク抽出，電子情報通信学会論文誌 D，情報・システム，Vol.93, No.3, pp.180–188 (2010).

[12] 北島理沙，小林一郎：文書上の潜在トピックを捉える事象の検討とその応用，情報処理学会研究報告．自然言語処理研究会報告，Vol.2011, No.3, pp.1–8 (2011).

[13] Wilson, A.T. and Chew, P.A.: Term Weighting Schemes for Latent Dirichlet Allocation, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, *HLT '10*, pp.465–473 (2010).

[14] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, *Proc. National Academy of Sciences of the United States of America*, *National Acad. Sciences*, Vol.101, No.1, pp.5228–5235 (2004).

[15] Okanohara, D. and Tsujii, J.: Text categorization with all substring features, *Proc. 9th SIAM International Conference on Data Mining* (*SDM*), pp.838–846 (2009).

[16] MeCab, available from ⟨http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html⟩.

[17] ChaSen, available from ⟨http://chasen.naist.jp/hiki/ChaSen/⟩.

[18] Och, F.J., Tillmann, C., Ney, H., et el.: Improved alignment models for statistical machine translation, *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp.20–28 (1999).

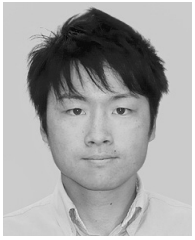
[19] アドビシステムズ株式会社：Adobe Acrobat 9.46, available from ⟨http://www.adobe.com/jp/⟩.

[20] 工藤　拓，賀沢秀人：Web 日本語 N グラム第 1 版．

[21] Lin, J.: Divergence measures based on the Shannon entropy, *IEEE Trans. Information Theory*, Vol.37, No.1, pp.145–151 (1991).

[22] Heinrich, G.: Parameter estimation for text analysis, Technical Note, Ver.2.4 (2008).

[23] Torgerson, W.S.: Multidimensional scaling: I. Theory and method, *Psychometrika*, Vol.17, No.4, pp.401–419 (1952).



田村 一樹　（学生会員）

2012 年 3 月名古屋大学工学部電気電子・情報工学科卒業．同年 4 月同大学大学院工学研究科博士課程前期課程計算理工学専攻に入学，現在に至る．主として自然言語処理，情報推薦に関する研究に従事．

吉川 大弘　（正会員）

1997 年名古屋大学大学院博士課程修了．同年カリフォルニア大学バークレー校ソフトコンピューティング研究所客員研究員．1998 年三重大学工学部助手．2005 年名古屋大学大学院工学研究科 COE 特任准教授．2006 年 10 月同研究科准教授，現在に至る．主としてソフトコンピューティングとその応用に関する研究に従事．博士（工学）．IEEE，電子情報通信学会，日本知能情報ファジィ学会，進化計算学会，計測自動制御学会各会員．

古橋 武

1985 年名古屋大学大学院工学研究科博士後期課程電気系専攻修了．工学博士．2004 年名古屋大学大学院工学研究科計算理工学専攻教授，現在に至る．ソフトコンピューティング，感性工学に関する研究に従事．1995 年日本ファジィ学会論文賞受賞．IEEE，日本知能情報ファジィ学会，電気学会等各会員．