OCR文書における単語の重み付けによるトピックモデルの性能向上に関する検討

[1]田村　一樹，[1]吉川　大弘，[1]古橋　武，[2]鈴木　誠
[1]名古屋大学
[2]ブラザー工業株式会社　ＮＩＤ開発部
tamura@cmplx.cse.nagoya-u.ac.jp

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，潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation)[1] に代表されるトピックモデルがある．LDA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 [2]．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LDAを適用する際に，トピック推定性能を向上させる手法について検討する．ここでは，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に基づく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．文書の類似関係を2次元平面上に配置し，可視化することで分類・検索等を行うことを想定した実験を，OCR文書に対して行い，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　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 [1][3]．LDAでは，文書におけるトピックの出現と，各トピックにおける単語の出現を多項分布で仮定し，各事前分布にディリクレ分布を導入することで，ギブスサンプリングによるトピックの推定を行っている [3]．

ここで，文書 i に含まれる単語 j のうち，トピック k に割り当てられたものの数を N_{ijk} で表し，また添え字の $(\cdot)$ はその変数についての総和とする．そのうち，位置 l を除いたものを N^{-l}_{ijk} と表記すると，ある位置 l のトピック z_l は，位置 l 以外の情報を用いて式 (1) で更新される．

$$p(z_l|z_{\backslash l}, \mathbf{w}) \propto \frac{N^{-l}_{(\cdot)jk} + \beta}{N^{-l}_{(\cdot)(\cdot)k} + V\beta} \cdot \frac{N^{-l}_{i(\cdot)k} + \alpha}{N^{-l}_{i(\cdot)(\cdot)} + K\alpha} \qquad (1)$$

なお，α，β は，それぞれ各文書のトピック分布，各トピックの単語分布を表すディリクレ分布のハイパーパラメータであり，V，K はそれぞれ総語彙数，総トピック数である．トピックを十分数更新して得られたサンプルから，文書 i でトピック j が生成される確率 θ_i^k の MAP 推定量を，以下のように求めることができる．

$$\theta_i^k = \frac{N_{i(\cdot)k} + \alpha}{N_{i(\cdot)(\cdot)} + K\alpha} \tag{2}$$

3 提案手法

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[4] 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

3.1 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．

$$\begin{aligned} p(w) &= p(t_1) \times p(t_2|t_1) \times \cdots \times p(t_n|t_{n-1}) \\ &= p(t_1) \prod_{i=2}^{n} p(t_i|t_{i-1}) \end{aligned} \tag{3}$$

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 (4) のように定義する．

$$m(w_i) = \frac{\log p_{\bar{t}}(w_i)}{\arg\max_{w \in W} \log p_{\bar{t}}(w)} \tag{4}$$

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．

3.2 Weighting LDA

Wilson らの重み付け手法 (WLDA)[5] では，2 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 [5] 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，2 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_{ijk} を，文書 i に含まれる単語 j のうち，トピック k 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 (5) で表すことができる．

$$p(z_l|z_{\setminus l}, \mathbf{w}) \propto \frac{M^{-l}_{(\cdot)jk} + \beta}{M^{-l}_{(\cdot)(\cdot)k} + V\beta} \cdot \frac{M^{-l}_{i(\cdot)k} + \alpha}{M^{-l}_{i(\cdot)(\cdot)} + K\alpha} \tag{5}$$

本稿では，WLDA における重みに，3.1 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を試みる．

4 実験

本稿では，文書の類似関係を可視化する問題を設定し，実験結果を定量的に評価することで，手法間の性能比較を行った．可視化は，各文書のトピック分布の距離を Jensen-Shannon 情報量 [6] とし，多次元尺度構成法 [7] を用いて 2 次元平面に文書を提示することによって行った．また，定量的な評価指標として，可視化空間上における k 近傍法による予測精度を用いた [8]．これは，ある文書のラベルを近傍 k 個の文書のラベルの多数決で予測し，正しく推定された文書の割合を分類精度と定義するもので，類似した文書が近くに配置されるほど高い値となる．本実験では，$k = 5$ とした k 近傍法による予測精度を用いて評価を行った．

4.1 適用文書

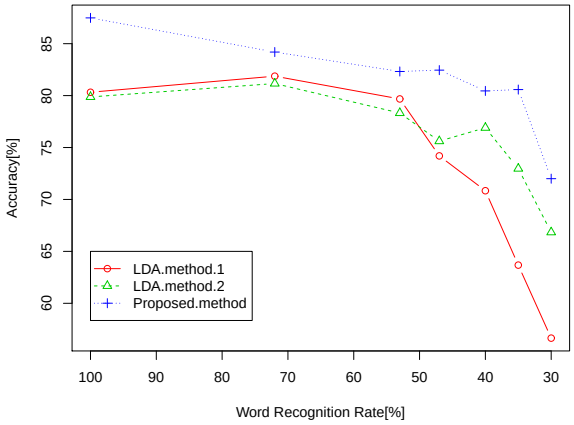
本実験では入力文書として，情報処理学会第 74 回全国大会の講演論文を用いた．そのうち，4 セッション計 31 文書（データ 1），6 セッション計 48 文書（データ 2）からなるデータセットを作成し，それらを用いて評価を行った．データには，電子文書に元々埋め込まれている誤りのないテキストと，文書画像に対して OCR ソフトウェアを用いることで得られる，誤りを含んだテキストを用意した．そのうち誤りを含むテキストには，印刷の不鮮明な文書や，手書き文書など，OCR の認識率が低い文書が含まれることを想定し，文書画像にランダムにノイズを加えて OCR をかけることで，異なる認識率のテキストを作成した．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．

4.2 実験条件

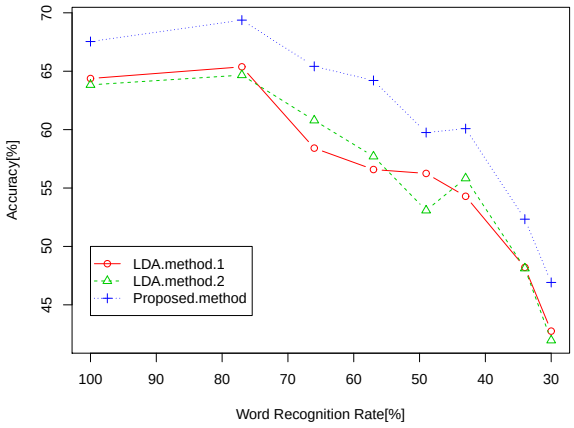
OCR ソフトウェアには Adobe Acrobat 9.46[9] を，形態素解析器には MeCab[4] を用いた．なお，ノイズのない文書画像に対する OCR の形態素単位での認識率は約 75% であった．また 3.1 で示した N-gram 確率は，Web 日本語 N グラム第 1 版 [10] を用いて算出した．LDA，WLDA のハイパーパラメータは $\alpha = 0.1$，$\beta = 0.1$ とし，サンプリング回数は 1000 回とした．また，50 試行の平均を予測精度として用いた．トピック数は，電子文書に対して LDA を用いた予備実験において，最も予測精度が高くなったトピック数とし，データ 1 で $T = 4$，データ 2 で $T = 6$（ともに正解となるセッションの数）であった．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，低頻度語の除去 [11] に対応する方法として，出現回数 1 の単語を除く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．

4.3 結果

それぞれのデータセットに適用し，得れらた結果を図 2(a)，図 2(b) にそれぞれ示す．まず，LDA 手法 1



(a) 分類精度（データ 1）



(b) 分類精度（データ 2）

図 2: 各単語認識率における分類精度の比較

に着目すると，データ 1 では OCR の単語認識率 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 [2] 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の付近では，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となった．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，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し，トピックの推定に反映させる方法などについて検討していく必要があると考えられる．

5 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 を用いる際に，従来報告されている，OCR の誤認識によるトピック推定性能の低下を確認するとともに，その性能低下を抑える方法を提案した．提案手法では，N-gram 確率を用いて単語の認識の信頼度を定義し，LDA に対して信頼度が高い単語の出現を重視する重み付けを行った．実際の講演論文文書の類似性を，2 次元に可視化する実験を行い，従来の LDA と提案手法を比較して，分類精度の面で提案手法が優れていることを確認した．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対する検討や，文書を特徴づける重要語の抽出，それらへの重み付けなどを行うことなどが挙げられる．

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] David M. Blei, Andrew Y. Ng, and Michael I. Jordan: Latent dirichlet allocation, Machine Learning Research, Vol. 3, pp. 993-1022, 2003

[2] Daniel D. Walker, William B. Lund, and Eric K. Ringger: Evaluating models of latent document semantics in the presence of OCR errors, Proc. of EMNLP ’10, pp. 240-250, 2010

[3] Griffiths, T.L. and Steyvers, M.: Finding scientific topics, Proc. of NAS ’04, Vol. 101, No. 1, pp. 5228-5235, 2004

[4] MeCab: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[5] Wilson, A.T. and Chew, P.A.: Term Weighting Schemes for Latent Dirichlet Allocation, Proc. of HLT-NAACL ’10, pp. 465-473, 2010

[6] Lin, J.: Divergence measures based on the Shannon entropy, IEEE Transactions on Information Theory, Vol. 37, No. 1, pp. 145-15, 1991

[7] Warren S. Torgerson: Multidimensional scaling: I. Theory and method, Psychometrika, Vol. 17, No. 4, pp. 401-419, 1952

[8] Tomoharu Iwata, Takeshi Yamada, Naonori Ueda.: Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents, Proc. of KDD ’08, pp. 363-371, 2008

[9] アドビ システムズ株式会社: Adobe Acrobat, http://www.adobe.com/jp/

[10] 工藤拓, 賀沢秀人.:Web 日本語 N グラム第 1 版

[11] David M. Blei and John D. Lafferty: Dynamic topic models, Proc. of ICML ’06, pp. 113-120, 2006