6Q-2

OCR文書検索を想定した重み付きトピックモデルの検討

田村　一樹[†]，吉川　大弘[†]，古橋　武[†]，鈴木　誠[‡]
[†]名古屋大学 工学研究科　[‡]ブラザー工業株式会社 ＮＩＤ開発部

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．しかし一方で，蓄積される文書データが多くなるほど，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．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，トピックの推定性能を向上させる手法について検討する．ここでは，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．従来のLDAと，提案する重み付けによるLDAとを比較する可視化実験を行い，文書分類予測精度の上で提案手法が優れていることを示す．

A Study on Weighting Topic Model for OCR Documents Retrieval
Kazuki Tamura[†], Yoshikawa Tomohiro[†], Takeshi Furuhashi[†], and Makoto Suzuki[‡]
[†]Graduated School of Engineering Nagoya University
[‡]NID Research & Development Dept. Brother Industries, LTD

2　提案手法

2.1　単語の信頼度

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 Bi-gram 確率を用いて単語の信頼度を定義する．ある単語間の隣接確率について，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．ここで，単語 w を構成する形態素が $t_1 t_2 \cdots t_n$ の場合の Bi-gram 確率は，以下で表される．

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

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 (2) のように定義する．

$$m(w_i) = \frac{\log p_{\bar{t}}(w_i)}{\arg\max_{w \in W} \log p_{\bar{t}}(w)} \tag{2}$$

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．

2.2　Weighting LDA

Wilsonらの重み付け手法 (WLDA)[4] では，LDA[3] 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．WLDAでは，多項分布を数学的に実数に拡張し，単語やトピックの出現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 (3) で表すことができる．$(\cdot)$ はその変数についての総和である．

$$p(z_l | z_{\setminus l}, \mathbf{w}) \propto \frac{M^{-l}_{(\cdot)jk} + \beta}{M^{-l}_{(\cdot)(\cdot)k} + V\beta} \cdot \frac{M^{-l}_{i(\cdot)k} + \alpha}{M^{-l}_{i(\cdot)(\cdot)} + K\alpha} \tag{3}$$

本稿では，WLDA における重みに 2.1 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．

3 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 情報量とし，多次元尺度構成法を用いて 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 ($k = 5$) による予測精度を用い [5]，50 試行の平均値を用いて評価を行った．

3.1 実験条件

本稿では，情報処理学会第 74 回全国大会の講演論文 31 稿について，属していたセッションを分類の正解ラベルとしてデータセットを作成し，実験に用いた．なお，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．また，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．比較手法は，従来の LDA による方法（以降，LDA 手法 1 と表記する）と，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）とした．詳細な実験条件は，紙面の都合上割愛する．

3.2 結果

3.1 のデータセットに適用し，得れらた結果の一例を図 1 に示す．まず，LDA 手法 1 に着目すると，OCR の単語認識率 50% 付近から予測精度が急激に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，予測精度の観点から確認することができた．また，LDA 手法 2 は，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 よりも高い予測精度が得られた．この結果に対して，多重性を考慮した対応のある t 検定（シダックの検定手法）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確認できた（$p < 0.01$）．

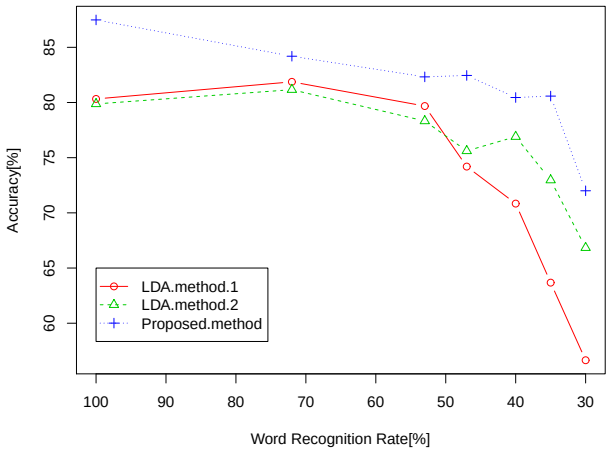

図 1: 各単語認識率における予測精度の比較

4 おわりに

本稿では，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手法を提案した．実験において，従来の LDA と提案手法を比較し，予測精度の面で提案手法が優れていることを確認した．なお本研究は，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] Blei, D.M. et al.: Latent dirichlet allocation, Machine Learning Research, Vol. 3, pp. 993-1022, 2003

[2] Walker, D.D. et al.: Evaluating models of latent document semantics in the presence of OCR errors, Proc. of EMNLP '10, pp. 240-250, 2010

[3] Griffiths, T.L. et al.: Finding scientific topics, Proc. of NAS '04, Vol. 101, No. 1, pp. 5228-5235, 2004

[4] Wilson, A.T. et al.: Term Weighting Schemes for Latent Dirichlet Allocation, Proc. of HLT-NAACL '10, pp. 465-473, 2010

[5] Iwata, T. et al.: Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents, Proc. of KDD '08, pp. 363-371, 2008