6Q-3

疑似ラベルを用いた潜在的ディリクレ配分法の一考察

鈴木 聡子 † 小林 一郎 ††

† お茶の水女子大学理学部情報科学科 †† お茶の水女子大学大学院人間文化創成科学研究科理学専攻

1 はじめに

近年，文書の潜在情報であるトピックを考慮したトピックモデルが文書要約や文書分類に利用されている．潜在ディリクレ配分法 (LDA)[1] に基づいて提案された Labeled LDA(L-LDA)[2] は，人によって文書に付けられたタグを，その文書の意味内容を表すものと捉え，潜在トピック抽出における教師信号として利用することを考えたモデルであり，複数のタグ付き文書に対しての LDA を上回る性能を示すと知られている．しかし実際は，世の中のほとんどの文書にはタグが付与されておらず，L-LDA の使用される範囲は限られている．そこで本研究では，文書集合からタグの代わりとなる疑似ラベルを作成し，全ての文書に対して L-LDA が有用になることを目的とする．

2 Labeled LDA

L-LDA は，LDA におけるトピック分布を推定する過程で，文書に付与されたタグの情報を考慮したモデルとなっている．図 1 に L-LDA のグラフィカルモデルを示す．L-LDA と LDA との違いは，ラベル（文書に与えられているタグ）の情報が，θ を推定する際に影響を与えているという点である．

まず，文書ごとに付与されているタグの情報から，文書ラベル $\Lambda^{(d)}$ を生成する．

$$\Lambda^{(d)} = (l_1, \cdots, l_K) \quad l_k \in \{0, 1\} \tag{1}$$

K は文書群に含まれる重複の無いラベルの個数であり，文書ごとにラベルの有無の情報を 1 または 0 の二値で与える．次に文書におけるラベルのベクトルを定義する．

$$\lambda^{(d)} = \{k | \Lambda_k^{(d)} = 1\} \tag{2}$$

$\lambda^{(d)}$ は，文書 d に付与されているラベル番号である．そして，文書ごとに射影行列を生成する．

$$L_{ij}^{(d)} = \begin{cases} 1 & if\ \lambda_i^{(d)} = j \\ 0 & otherwise. \end{cases} \tag{3}$$

生成した射影行列と設定したハイパーパラメータ α から，文書ごとに新しいパラメータ $\alpha^{(d)}$ を生成する．ラベルの情報により制限された $\alpha^{(d)}$ から，トピック分布 θ を求める．他の過程は，LDA と同様である．

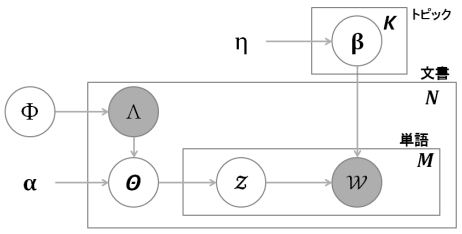

図 1: L-LDA のグラフィカルモデル

3 疑似ラベル生成

本研究では，文書集合の中から単語を抽出し，抽出した単語を共起情報によって分類し，生成したものを疑似ラベルと呼ぶ．本研究における疑似ラベルの生成手法を以下に説明する．

3.1 疑似ラベル候補となる単語抽出

初めに，疑似ラベルを生成するために候補となる単語を抽出する．文書ごとに TF-IDF の高い単語を抽出する．そして，抽出した単語の出現頻度を全ての文書において求める．ここで文書頻度が 1 である単語は，特定の文書においてのみ現れている単語であるため，疑似ラベル候補の中から消去する．

3.2 共起情報による単語のクラスタリング

文書の潜在的意味の一貫性は単語の共起関係に関連があるという Newman らの研究 [3] を参考に，生成するラベルにそのような意味情報が含まれるように，抽出した単語の自己相互情報量（PMI）を求め，共起関係の強い単語群を 1 つのグループとする．そのグループで 1 つの疑似ラベルを表すとする．

また，共起情報によるクラスタリングで作られたラベルの他に，PMI の値は低いが出現頻度は高い単語も，ラベルとして採用する．

4 実験

タグ付けされていない文書集合に疑似ラベルを付与し，文書分類の課題を通じて LDA との比較を行い，提案手法を評価する．

A Study on Pseudo Labeled Latent Dirichlet Allocation

† Satoko SUZUKI(g0920519@is.ocha.ac.jp)
†† Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Department of Informatoin Science, Faculty of Sciences, Ochanomizu University (†)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (††)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

4.1 実験仕様

使用するデータは，20 Newsgroups†の内 10 カテゴリの中からそれぞれ 100 文書をランダムに選んだ，合計 1000 文書を用いる．それら文書は，ストップワードを除いた後ステミング処理を施す．疑似ラベルを生成する際，今回は 2 つの設定において疑似ラベルを生成する単語の共起情報に基づくクラスタリングの閾値を設定し，実験を行った．2 つの設定に共通して，抽出する単語数は各文書ごとに TF-IDF の値が上位 30 単語とした．

閾値は，設定 1 に関しては疑似ラベルの振り分け結果，設定 2 に関しては生成された疑似ラベルの出力結果を考慮して，それぞれの値を決定した．

設定 1. 疑似ラベルの振り分けを重視した場合
出来るだけ多くの文書に疑似ラベルが振り分けられる事を優先させた．クラスタリングする際の PMI の閾値は 5.5 とした．結果，生成された疑似ラベルは 16 個となった．

設定 2. 単語の共起関係の強さを重視した場合
強い共起関係を残すために，PMI の閾値を 6.0 と高く設定した．また，少数の単語で構成されるグループは，その内容を十分に表していないと考え，質の良いラベルのみを残すために，単語数が 3 以下のグループは疑似ラベルとして採用しないとした．結果，生成された疑似ラベルの数は 71 個となった．

L-LDA に与えるハイパーパラメータの値は，α=0.5, η=0.5 とした．比較対象である LDA では，パープレキシティの値よりトピック数を設定した．トピック数ごとに，イテレーションの中での最も低いパープレキシティの平均を求めた．トピック数が 11 で，平均が最小となったため，実験ではトピック数を 11 と設定した．与えるパラメータの値は，提案手法と同じく α=0.5, η=0.5 とした．

4.2 評価

文書のトピック分布 θ から，各文書のトピックで構成されるベクトルを作り，k-means 法により，20 Newsgroups の対象とした 10 カテゴリのグループに文書を分類した際の精度を見ることで提案手法の評価を行う．評価手法には，文献 [4] で用いられている評価手法を採用し，式 (5) で示される相互情報量を利用した．

$$MI(L,A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i, \alpha_j) \cdot \log_2 \frac{P(l_i, \alpha_j)}{P(l_i)P(\alpha_j)} \quad (4)$$

相互情報量を [0,1] の値で得るために正規化を行う．

$$\widehat{MI} = \frac{MI(L,A)}{MI(A,A)} \quad (5)$$

† http://qwone.com/ jason/20Newsgroups/

4.3 実験結果

k-means 法を用いた分類をそれぞれの手法につき 5 回行い，$\widehat{MI}$ の平均を求めた．求めた結果を表 1 に示す．

表 1: 実験結果

実験手法	$\widehat{MI}$
LDA	0.4743
提案手法（設定 1）	0.2155
提案手法（設定 2）	0.1725

4.4 考察

実験結果より，生成された疑似ラベルの内容を重視するよりも疑似ラベルの振り分けを重視した方が結果が良くなることが確認できた．このことから，疑似ラベルは，その内容よりも，多くの文書に振り分けられることを目的として，生成手法を検討していく必要があると考えられる．疑似ラベルの内容を重視した精度が良くなかった原因としては，PMI の閾値を高く設定したことから単語が限定され，多くの文書に共通したラベルを振り分けることが出来なくなったのではないかと考えられる．また，LDA よりも提案手法の精度が良くない原因としては，疑似ラベルの生成方法を単語の共起情報のみに限定してしまっていることから，多くの文書に共通するラベルを生成することができなかったのではないかと考える．

5 おわりに

本研究では，TF-IDF により単語抽出を行い，抽出した単語の共起関係から単語をクラスタリングすることによって，疑似ラベルの生成を行った．生成した疑似ラベルを用いて実験を行い，LDA との精度の比較，評価を行った．結果，LDA より高い精度は得られなかったが，生成される疑似ラベルの精度よりも，疑似ラベルの振り分けを重視した方が良いことが確認できた．今後は，単語のより良いクラスタリング方法，ハイパーパラメータや閾値の値を，実験から検討していく．

参考文献

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

[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning,Labeled LDA: A supervised topic model for credit attribution in multi-label corpora. EMNLP2009, pp. 248-256, 2009.

[3] Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy, Human Language Technologies: NAACL2010, pp. 100–108,Los Angeles, California, 2010.

[4] Gunes Erkan : Language Model-Based Document Clustering Using Random Walks, Association for Computational Linguistics,pp.479-456,2006.