疑似ラベルを用いた潜在的ディリクレ配分法の提案

Pseudo Labled Latent Dirichlet Allocation

鈴木聡子 [1*]　小林一郎 [2]
Satoko Suzuki[1]　Ichiro Kobayashi[2]

[1] お茶の水女子大学　理学部　情報科学科
[1] Department of Information Science, Faculty of Science, Ochanomizu University
[2] 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
[2] Advanced Science, Graduate School of Humanities and Science, Ochanomizu University

Abstract: In recent years, topic models have been widely used for many applications such as document summarization, document clustering etc. Labeled latent Dirichlet allocation (LLDA) was proposed based on latent Dirichlet allocation (LDA), and it regards the tags, i.e., labels, put on documents by humans as the ones expressing the contents of the documents, and uses them as supervised information to estimate latent topics of the documents. Moreover, it is reported that LLDA exceeds the ability of LDA in terms of topic estimation. However, normal documents usually do not have such tags with them, so, the use of LLDA is considerably limited.In this study, therefore, we make pseudo labels from the documents to be estimated their latent topics instead of tags put on documents by humans, and aim to make LLDA available for all documents.

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，人によって文書に予め付けられているタグを，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．

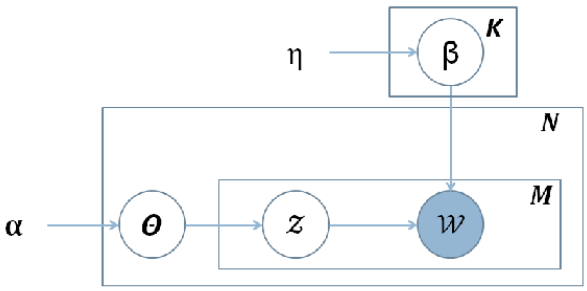

図 1: L-LDA のグラフィカルモデル

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 2 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．

*連絡先：お茶の水女子大学理学部情報科学科小林研究室
〒 112-0012 東京都文京区大塚 2-1-1
E-mail: g0920519@is.ocha.ac.jp

そして，文書ごとに射影行列を生成する．

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

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．

3 疑似ラベル生成

本研究では，文書の 2 つの表層的な情報からラベルの代わりとなる疑似ラベルを生成する．

3.1 単語の共起情報に基づく疑似ラベル生成

ここでは，Newman らの研究 [3] よる，文書の潜在的意味の一貫性は単語の共起関係と関連があるということから，共起関係の強い単語より疑似ラベルを生成することを考える．

まず，疑似ラベルを構成する候補として，各文書から TF-IDF の値が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，その文書特有の単語であると考え，抽出したリストから消去する．残った単語を疑似ラベルを生成する単語の候補とする．

次に，抽出した単語の共起関係を自己相互情報量 (PMI) を用いて求める．ここで，2 つの単語が同じ文書において疑似ラベルの候補に抽出されることを，共起していると捉えることとする．PMI が閾値以上の単語同士をグループ化し，グループごとを疑似ラベルとして設定する．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．

疑似ラベルを構成する単語が抽出されている文書に，同じラベルを与える．ただし，複数の単語で構成される疑似ラベルについては，構成する単語が 2 つ以上抽出されている場合にラベルを与える．

3.2 文書の類似度に基づく疑似ラベル生成

文書 d_i におけるベクトル中の重みを式 (4) とし，生成した文書ベクトルをもとに文書分類を行う．

$$w_{ij} = (\log x_{ij} + 1.0) \log(N/n_j) \quad (4)$$

x_{ij} は文書 d_i における語 t_j の出現回数であり，N は全文書数，n_j は語 t_j の出現する文書数である．その結果，類似する文書に同じラベルを与える．ここでは，Leader-Follower 法と Crouch 法の 2 つの方法 [5][6] において疑似ラベルを生成する．この 2 つの方法は，分類の重複を許すアルゴリズムとなっており，1 文書に対し複数のラベルを生成することが可能である．

Leader-Follower 法

ここで用いるのは，Leader-Follower 法である．本研究で用いたアルゴリズムの概要を以下で説明する．

1. 文書をクラスタに併合するための閾値を設定する．
2. 1 つめの文書を読み，クラスタとして設定する．
3. 1 文書ずつ読む．全ての文書が読み終わったら処理を終了する．
4. 読み込んだ 1 文書と，その時点で存在する全てのクラスタとの類似度を計算する．
5. 閾値以上の類似度をもつクラスタにその文書を併合し，クラスタの語の重みの値を更新する．その文書との類似度が，どのクラスタにおいても閾値を超えない場合は，新しいクラスタとして生成する．
6. 手順 3 に戻る．

ここでクラスタ C_h 中の語 t_j の重みを式 (5) と定義する．

$$\hat{w_{hj}} = \log \sum_{d_i \in C_h} x_{ij} + 1.0 \quad (5)$$

また，類似度にはコサイン類似度を用いる．これによって生成されたクラスタについて，同じクラスタに含まれている文書に同じ疑似ラベルを与える．クラスタを構成する文書数が 1 の場合には，疑似ラベルを与えないこととする．

Crouch 法

この手法は Leader-Follower 法を拡張したものであり，クラスタの設定とクラスタへの文書の割り当てを 2 段階の処理によって行うことが特徴である．
Crouch 法では，Leader-Follower 法と同様に設定したクラスタと，全文書の類似度を計算し，閾値以上の値を持つ文書に同じ疑似ラベルを与える．クラスタと文書の類似度は式 (6) によって求める．

$$s(d_i, C_k) = f \frac{\sum_{j=1}^{M} min(w_{ij}, \hat{w}_{kj})}{min(\sum_{j=1}^{M} w_{ij}, \sum_{j=1}^{M} \hat{w}_{kj})} \quad (6)$$

ここで，d_i，C_k，w_{ij}，$\hat{w}_{kj}$ については，上述した変数であり，M は全語彙数とする．

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各手法と LDA との比較を行う．

4.1 実験仕様

使用するデータは，20 Newsgroups[1]の 20 カテゴリの内，10 個のカテゴリを選び，その中からそれぞれ 100 文書を選んだ，合計 1000 文書を用いる．この合計 1000 文書から成る文書集合を 2 セット用意した．(2 つの文書集合を setA，setB と区別する．)

選んだカテゴリを表 1 に示す．

表 1: 選択カテゴリ

setA	setB
alt.atheism	alt.atheism
comp.graphics	comp.graphics
com.sys.mac.hardware	comp.ibm.pc.hardware
rec.sport.baseball	misc.forsale
sci.med	rec.autos
sci.crypt	rec.motorcycles
sci.electronics	sci.electronics
sci.space	sci.space
talk.politics.guns	talk.politics.guns
soc.religion.christian	talk.politics.misc

２つの文書集合における 10 個のカテゴリは，カテゴリ内で内容が偏らないように選んだ．また，setA と setB でカテゴリが共通する場合は，異なる文書が選ばれている．それらの文書集合は，ストップワードを除いた後，ステミング処理を施す．提案手法の実験は，単語の共起情報により疑似ラベルを生成した場合と，文書の類似度から疑似ラベルを生成した場合の 2 つのパターンにおいて行う．

単語の共起情報により疑似ラベルを生成した場合（パターン 1 とする）では，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．また，1 単語から構成されるラベル数は，出現回数が上位 5 位までの単語を選んだ．ここで，予備実験より PMI の閾値はラベルが複数できる範囲，setA では [4.5,6.2]，setB では [4.8,6.2] において実験を行うものとした．

文書の類似度から疑似ラベルを生成する場合（パターン 2 とする）では，Leader-Follower 法と Crouch 法の 2 つの方法で疑似ラベルを生成した．類似度の閾値は，[0.1,0.9]（パターン 2a とする）と 0.1 以下（パターン 2b とする）において実験を行った．全ての実験で，L-LDA に与えるハイパーパラメータの値は，α=0.1，η=0.1 とした．比較対象である LDA では，まずトピック数を設定するための予備実験を行った．その結果，それぞれの文書集合における最適トピック数を setA では 16，setB では 28 と設定した．LDA において与えるパラメータの値は，提案手法と同じく α=0.1，η=0.1 とした．

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，20Newsgroups 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．

[1]http://qwone.com/ jason/20Newsgroups/

4.2 評価手法

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（7）に示される相互情報量を利用した．

$$MI(L,A) = \sum_{l_i \in L, \alpha_j \in A} P(l_i, \alpha_j) \cdot \log_2 \frac{P(l_i, \alpha_j)}{P(l_i)P(\alpha_j)} \tag{7}$$

$L = \{l_1, l_2, \cdots, l_k\}$ は，k-means 法により分類された文書ラベルの集合であり，$A = \{\alpha_1, \alpha_2, \cdots, \alpha_k\}$ は分類された文書の正解ラベルである．また，$P(l_i)$ は分類により l_j にラベル付けされる確率，$P(\alpha_j)$ は正解データにおいて α_j である確率，$P(l_i, \alpha_j)$ はこれら 2 つが同時に起こる確率である．

ここで，相互情報量を [0,1] の値で得るために式（8）により正規化を行う．

$$\widehat{MI} = \frac{MI(L,A)}{MI(A,A)} \tag{8}$$

4.3 実験結果

k-means 法を用いた分類をそれぞれの手法の各閾値において 10 回ずつ行い，評価値 $\widehat{MI}$ の平均を求めた．setA での実験結果を図 2〜4，setB での実験結果を図 5〜7 に示す．全てのグラフに関して，横軸は閾値，縦軸は評価値 $\widehat{MI}$ を示す．なお，比較のために LDA の $\widehat{MI}$ もグラフに示す．

LDA を含めた全手法に関して，setA の方が setB を用いた実験よりも良い $\widehat{MI}$ を得ている．図 2，5 から分かるように，単語の共起情報を用いた手法では，どちらの文書集合においても LDA よりも $\widehat{MI}$ は低くなった．また，2 つのグラフに類似性は見られない．文書の類似度を用いた手法では，一部の閾値において LDA を上回る結果を得ている．図 3，6 より，[0.1,0.9] では，両手法ともに $\widehat{MI}$ は閾値によって減少増加し，どちら

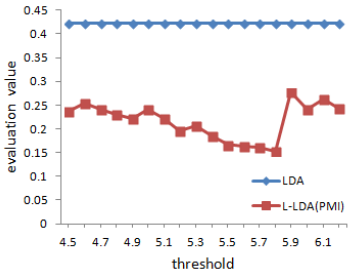

図 2: $\widehat{MI}$ パターン 1 (setA)

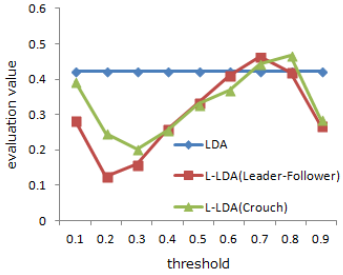

図 3: $\widehat{MI}$ パターン 2a(setA)

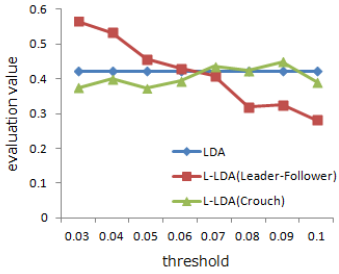

図 4: $\widehat{MI}$ パターン 2b(setA)

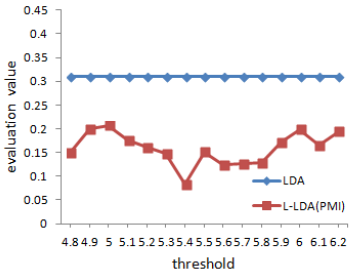

図 5: $\widehat{MI}$ パターン 1 (setB)

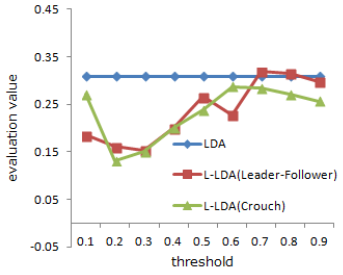

図 6: $\widehat{MI}$ パターン 2a(setB)

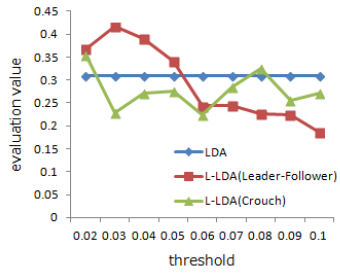

図 7: $\widehat{MI}$ パターン 2b(setB)

の文書集合においても，一部で LDA を上回る結果を得てはいるが，その値にあまり差はない．また，図 4，7 より，[0.03,0.1] では，Leader-Follower 法は左肩上がりのグラフとなっている．一方で，Crouch 法では LDA での $\widehat{MI}$ の付近で増加減少を繰り返すグラフとなっている．

次に，それぞれの手法において生成された疑似ラベルの数を表 2〜4 に示す． 表 2 のラベル数では，複数の単語によって構成されるラベルの数と，括弧の中に 1 つの単語で構成されるラベル数も加えたラベル数を示している．どちらの文書集合においても，閾値 5.2 でラベル数が増大し，5.6 でさらに急増する．そして，5.9 で減少した後，6.3 以上になるとラベルは生成されなくなった．パターン 2 では，Leader-Follower 法と Crouch 法でのクラスタを生成する過程が同じであるため，生成されるラベル数は同じである．どちら文書集合においても，閾値が大きくなるにつれてラベル数が増加し，0.2 で最大となった後に減少している．

表 2: パターン 1

Threshold	Number of labels	
	setA	setB
4.5	2(9)	-
4.6	2(9)	-
4.7	3(10)	-
4.8	3(10)	2(8)
4.9	5(12)	2(8)
5.0	6(13)	1(7)
5.1	5(12)	1(7)
5.2	22(29)	22(28)
5.3	20(27)	27(33)
5.4	20(27)	27(33)
5.5	16(23)	24(30)
5.6	204(211)	193(199)
5.7	204(211)	193(199)
5.8	204(211)	193(199)
5.9	125(132)	124(130)
6.0	125(132)	124(130)
6.1	125(132)	124(130)
6.2	125(132)	124(130)

また，それぞれの文書集合における手法別の生成したラベル数と評価値 $\widehat{MI}$ の関係を散布図で表したものを図 8，9 に示す．グラフの横軸はラベル数，縦軸は評価値 $\widehat{MI}$ を表している．また，LDA は丸，単語の共起情報を用いた手法は菱形，文書の類似度を用いた手法のうち Leader-Follower 法を用いたものは正方形，Crouch 法を用いたものは三角形で表している．各手法ごとに着目してみると，菱形の点は，横軸 0〜50，100〜150，200 付近にまとまって存在し，ラベル数と評価値の相関関係は見られない．正方形では，例外も存在するが，基本的にラベル数が LDA での最適トピック数に近いほど，結果が良く，増加するほど悪くなっている，三角形では，正方形の点と比べて，安定した評価値を得ているが，最適トピックスとの差が大きくなるほど，そのバラツキも大きくなる．

表 3: パターン 2a

Threshold	Number of labels	
	setA	setB
0.1	185	212
0.2	228	220
0.3	202	179
0.4	155	140
0.5	102	93
0.6	59	56
0.7	28	25
0.8	11	9
0.9	2	6

表 4: パターン 2b

Threshold	Number of labels	
	setA	setB
0.02	-	6
0.03	19	21
0.04	42	42
0.05	69	80
0.06	101	119
0.07	118	144
0.08	151	181
0.09	169	194
0.1	185	212

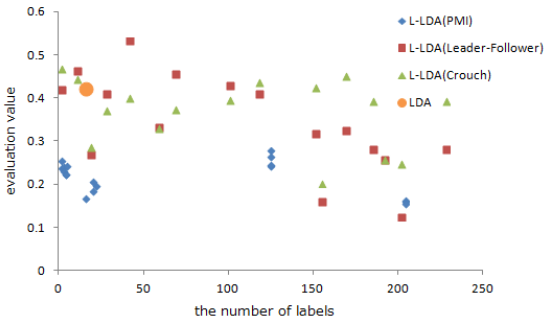

図 8: ラベル数と $\widehat{MI}$（setA）

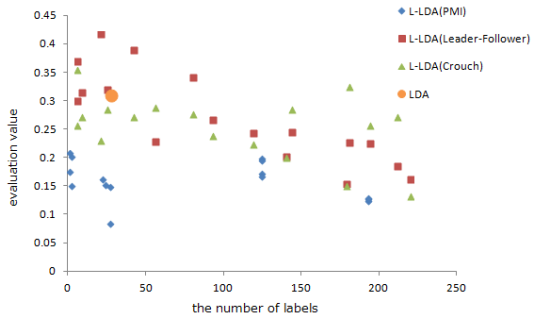

図 9: ラベル数と $\widehat{MI}$（setB）

5 考察

実験結果より，単語の共起情報から疑似ラベルを生成した場合では，2 つの文書集合において，全ての閾値で LDA と比べ精度が上がらなかった．また，ラベル数と評価値との相関関係が見られなかった．これらは，この手法で生成された疑似ラベルは，トピックの情報が反映できていないためであると考える．原因としては，抽出する単語数が少ないこと，もしくは，単語の共起情報のみでは文書集合の全てのトピックについて反映できないということが考えられる．閾値を低く設定した場合，トピックと関連の無い単語の共起情報も認識してしまい，トピックと関係の強い疑似ラベルを生成することが困難になり，一方で，閾値を高く設定した場合には，共起関係が非常に強い単語のみで疑似ラベルを作るため，生成された疑似ラベルはトピックとの関係は強いが，限られた文書にしかラベルを振り分けることができないとことが精度を悪くすると考える．また，今回はカテゴリごとに同じ数の文書を用意したが，文書数に偏りがあった場合には，トピックの情報を反映した疑似ラベルを生成することが，さらに困難になるのではないかと考えられる．

文書の類似度から疑似ラベルを生成した場合，Crouch 法を用いた方が Leader-Follower 法を用いるよりも，全体的に安定した評価値を得ている．また，2 つの文書集合での実験結果に共通して，閾値 0.03 で Leader-Follower 法を用いた場合に，LDA を上回り最も良い結果を得ている．閾値 0.03 では，生成された疑似ラベルの数が最適トピック数と近い値となっている．また setB においては，閾値 0.03 よりも 0.7 の方が，最適トピック数に近い数の疑似ラベルを生成しているが，この 2 つの点における評価値は 0.03 の方が良い．このことから，Leader-Follower 法を用いて，閾値をできるだけ低く設定し，生成される疑似ラベルの数が最適トピック数と近い場合に良い精度が得られることが分かった．

6 おわりに

本研究では，単語の共起情報と文書の類似度の 2 つの表層的な情報から疑似ラベルを生成した．それぞれの手法によって生成した疑似ラベルを用いて実験を行い，文書分類の課題を通じて LDA との精度の比較，評価を行った．その結果，単語の共起情報を用いた手法では，LDA を上回る結果を得ることはできなかったが，文書の類似度を用いた手法では，閾値を変えることによって，一部で LDA よりも良い精度を得ることができた．2 つの文書集合における評価結果から，それぞれの手法における閾値と評価値の関係や生成された疑似ラベル数と評価値の関係を確認した．また，2 つの文書集合における，全ての手法について比較したところ，Leader-Follower 法を用いた文書の類似度を利用した手法で，閾値を小さく設定し，生成される疑似ラベルの

数が最適トピック数に近かった場合に LDA を上回り，最も良い精度が得られることが分かった．

今後の課題としては，単語の共起情報を用いた手法において，生成したラベルを振り分けることのできる文書数が限られる原因として，TF-IDF を用いて抽出する単語の数が少ないことが考えられることから，設定を変えた実験が必要であると考えられる．また全ての提案手法に共通して，分類課題における別の評価方法や，他の課題を通した精度の確認をしていきたい．

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.

[5] 岸田和明: 大規模文献集合に対して階層的クラスタ分析法を適用するための単連結法アルゴリズム, Library and Information Science,No.47 ,2002.

[6] 岸田和明:文書クラスタリングの技法, Library and Information Science,No.49, 2003.