制約知識を用いた潜在トピック抽出に関する基礎的検討と応用

理学専攻・情報科学コース　立川 華代 (指導教員：小林 一郎)

1　はじめに

近年，ウェブ上の大量の文書データの中から有益な情報を抽出する目的の下，テキストのトピック抽出やカテゴリ分類などに関する研究が多く行なわれているトピック抽出手法として，LDA(潜在的ディリクレ配分法)[1] や NMF(非負値行列因子分解)[2] などが存在する．本研究では，この生成モデルに事前知識を与えることでトピックの推定の精度を向上させることを目的とし 3 つの手法を提案した．以下，2 章では，LDA において単語出現の事前確率分布としてディリクレ森分布を用いることで単語に割当てるトピックに制約を付与するとし，その制約を文書中から自動的に抽出しトピック推定のための制約知識として用いる手法を示す．3 章では，ノンパラメトリックベイズ法の基本となるディリクレ過程において採用される中華レストラン過程 (CRP) に制約知識を組み込むことにより，制約知識を反映したトピック推定を行うための基礎的検討を示す．4 章では，NMF において，初期値に制約知識を適用しトピック抽出の精度向上に対する基礎的検討を示す．

2　制約付き LDA によるトピック抽出

2.1　概要

LDA を利用し，制約を組み込んで潜在トピックの抽出を行うために，ディリクレ分布にディリクレ森分布 (DF:Dirichlet Forest Prior) を用いる．DF はディリクレ分布を階層化したものであり通常の LDA で利用する超パラメータ α, β に加え，制約の強さを反映する η を導入する．DF を用いた LDA(LDA-DF) での文書生成過程は以下のようになる．

$$\theta_{d_i} \sim \mathrm{Dirichlet}(\alpha) \quad (1)$$

$$z_i|\theta_{d_i} \sim \mathrm{Multinomial}(\theta_{d_i}) \quad (2)$$

$$q \sim \mathrm{DirichletForest}(\beta, \eta) \quad (3)$$

$$\phi_{z_i} \sim \mathrm{DirichletTree(q)} \quad (4)$$

$$w_i|z_i, \phi_{z_i} \sim \mathrm{Multinomial}(\phi_{z_i}) \quad (5)$$

ここで，θ, ϕ はそれぞれ α, β を超パラメータとする多項分布であり，d_i, z_i を i 番目の単語 w_i が含まれる文書および割り当てられるトピックとして，w_i の生成過程を示している．

2.2　提案手法

Newman ら [3] は，トピックの結束性に関する様々ん評価指標について考察しており，その指標の一つである単語間の自己相互情報量 (PMI:Point-wise Mutual Information) をトピック内の結束性を測る指標として利用する．本研究では与える制約知識の構築方法として二つの手法を提案する．制約知識を構築するために，トピックを代表するとみなせる単語 (以下,「重要単語」と呼ぶ) を選択する必要がある．本研究では重要単語の選択方法は，(i) 対象とする文書群に万遍なく高頻度で現れるもの (頻度情報)，(ii) 他の単語と多く共起関係にあるもの (共起情報)，とする．

(i)(ii) の手法で選択した重要単語について，共起関係に基づきいくつかのグループに分類する．この時共起関係の指標は PMI を用い，予め設定した閾値以上のものをひとつのグループにまとめる．その後，グループ内の単語と共起する単語を PMI により取得し，その値が高いものを追加することにより制約知識を構築する．追加する単語数によって制約が変化するため，本研究では追加する単語数を 1~ 4 個で変化させた．

2.3　実験

2.3.1　実験仕様

トピック抽出実験に用いる文書は，アメリカ ABC News，イギリス BBC NEWS など英語圏各国の主要新聞社や TV 会社のもので，2012 年 1 月 16 日の「イタリア豪華客船座礁事故」に関する英字新聞 (文数 967, 語彙数 2267) など計 4 記事のニュースとした．また LDA-DF で用いるディリクレ分布のパラメータは $\alpha = 0.1, \beta = 0.1, \eta = 100$ とし，イテレーション回数は 50 回，トピック数 $K = 10$ とした．また実験結果は，パープレキシティを指標として比較する．

2.3.2　結果と考察

「イタリア豪華客船座礁」の記事について与える制約の個数を増やしていった際のパープレキシティの変化を結果を図 1 に示す．グラフより共起情報に基づいて構築した制約を与えた場合には，与える制約の個数の増加とともにパープレキシティが定価する様子が見られ，制約を与えていない通常の LDA と比較して安定したモデルとなることがわかった．また，重要単語に対して追加する単語の個数は 1 つまたは 2 つ程度で十分なこと，与える制約の個数は 3 個程度でパープレキシティが最低値となることも分かった．

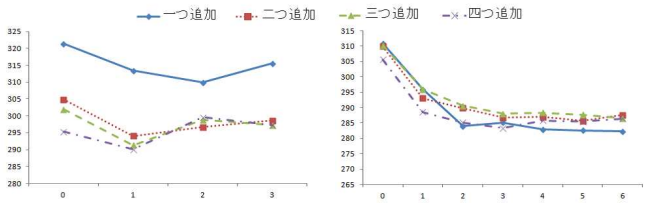

図 1:「豪華客船座礁」(左：頻度情報，右：共起情報)

3　CRP への制約知識の導入

3.1　概要

2 章で述べた LDA では文書内のトピック数について，予め決定しておく必要がある．しかし実際は内容が未知の文書群に関してトピック数は不明である．そこで Teh ら [4] は階層ディリクレ過程を用いたディリクレ配分法 (HDP-LDA) を考察し，トピック数 K を対象に表現した K 項のディリクレ多項分布モデルにおいて $K \to \infty$ としたモデルとして中華レストラン過程 (CRP:Chinese Restaurant Process) を提案した．

本研究では 2 章のディリクレ森分布の考えと融合し CRP でトピック数を推定しながらのトピック抽出に，

制約知識を導入する手法を提案する．

3.2 提案手法

CRP とは中華レストランのテーブルをトピック，客を単語とみなし，それぞれの客が着席するテーブルを決定することでトピックを抽出するものであり，レストラン内には無限のテーブルがあると仮定するため，一人以上の客が着席したテーブル数がトピック数ということになる．テーブルを選択する確率は

$$p(z_i = k|z_1, ..., z_{i-1}) = \begin{cases} \dfrac{m_k}{\gamma + i - 1} & (k = 1, \cdots K) \\ \dfrac{\gamma}{\gamma + i - 1} & (k = K + 1) \end{cases} \tag{6}$$

で示される．ここで K は現在までのクラスタ数，k は推定されるトピック，m_k は $i-1$ 番目までの語で既にそのテーブル k に座っている客の人数である．$k = 1, ..., K$ の時，このトピックが選ばれる確率は m_k に比例する．また，$\gamma(\gamma > 0)$ はディリクレ分布のハイパーパラメータであり，この γ に比例する確率でまだ誰も座っていない新しいテーブルに座ることになる．

本研究では中華レストランに制約知識の単語用の個室を導入し，同じトピックに入って欲しい単語群を反映させる．

3.2.1 個室の導入

客は個室または大広間のテーブルに着席し，さらに使用済みまたは未使用のテーブルに着席する．以上より本研究では着席する可能性のあるテーブルの種類について 5 種類に分類し，それぞれ事前確率を求めたところ以下のようになった．

$$P(z_i = (k, j) \mid z_1, ..., z_{i-1}) = \begin{cases} \dfrac{m_k}{\gamma + i - 1} & \cdots \text{(i)} \\ \dfrac{m_k}{\gamma + i - 1} \times \dfrac{m_{kj} + \eta}{m_k + C_k\eta} & \cdots \text{(ii)} \\ \dfrac{m_k + c_k\gamma}{i - 1 + \gamma} \times \dfrac{m_{kj}}{m_k + c_k\eta} & \cdots \text{(iii)} \\ \dfrac{m_k + c_k\gamma}{i - 1 + \gamma} \times \dfrac{c_k\eta}{m_k + c_k\eta} & \cdots \text{(iv)} \\ \dfrac{c_k\gamma}{i - 1 + \gamma} & \cdots \text{(v)} \end{cases}$$

3.2.2 検証

5 つの場合分けについて総和を取ると 1 となることを計算により確かめ，これらの式の正当性が示された．

4 制約付き NMF による文書分類

4.1 概要

非負値で表された行列のデータを二つに分解することにより，データの特徴抽出を可能とする手法として非負値行列因子分解 (NMF:Non-negative Matrix Factorization)[2] がある．NMF では通常，初期値を乱数で与えた行列を更新していくため，結果が初期値に大きく左右されることが知られている．そこで本研究では，トピックを形成するであると考えられる単語群を文書から抽出し，それらを反映するような初期値を与えることによりトピック抽出の精度の改善を検討する．

4.2 提案手法

制約知識となる単語群を抽出するため，まずトピックを代表するような語を選択する．そこで文書群からその文での固有の重要度を示す．tf-idf 値の高いものを重要語とし，次にそれぞれの重要語に対して共起する語を抽出し，それらを制約知識として NMF のトピック行列の初期値として与える．本研究では全語彙から tf-idf 値が高く，かつその語が属する文書の長さが比較的長いものをトピックの数だけ取得した．

4.3 実験

4.3.1 実験仕様

20 Newsgroups の 4 つのトピックの記事を 3 文書ずつ (計 12 文書，語彙数 : 1053)，21 文書ずつ (計 84 文書，語彙数 : 3794) で実験を行った．行列の初期値に関しては 4.2 章の手法で選択された単語の箇所に 10 を代入，他の箇所には 0~ 1 の乱数を与えた．クラスタリングの結果は Purity と Entropy によって評価した．

4.3.2 結果と考察

分類の精度の結果を表 1 に示す．制約を反映させたほうが僅かであるが精度が上がっていることがわかる．

表 1: 分類の精度

3 文書ずつ	Purity	Entropy
制約なし	0.633	0.702
制約あり	**0.658**	**0.648**
21 文書ずつ	**Purity**	**Entropy**
制約なし	0.465	1.136
制約あり	**0.487**	1.136

5 おわりに

本研究では，文書のトピックを抽出する LDA に制約を入れる際に，文書中から抽出した制約知識を組み込んだ．その結果よりトピック推定に精度の向上を確認できた．また，ノンパラメトリックである HDP-LDA について，それを実現する CRP の考え方に，個室の概念を導入することでトピック抽出において制約の反映が可能であるということを示した．さらに NMF においても初期値に制約知識を組み込み精度の向上を示した．今後は，制約知識に対してさらなる工夫を施し，制約を強く反映させる手法について考えていきたい．

参考文献

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

[2] Daniel D. Lee and H. Sebastian Seung:Algorithms for Non-negative Matrix Factorization,In NIPS, 2000.

[3] Newman, David and Lau, Jey Han and Grieser, Karl and Baldwin, Timothy:Automatic evaluation of topic coherence,HLT The 2010 North American Chapter of the ACL,2010.

[4] Y. W. Teh and M. I. Jordan and M. J. Beal and D. M. Blei:Hierarchical Dirichlet Processes,Journal of the American Statistical Association,2006.