E-008

単語間関係を制約として用いた文書クラスタリング
Document Clustering Using Relationships between Words as Constraint

小出 幸典† 栗飯原 俊介‡ 石崎 俊†
Yukinori Koide Shunsuke Aihara Shun Ishizaki

1． はじめに

文書分類タスクでは，単語の出現頻度を元に作成されたベクトルに対し，次元圧縮を施してからクラスタリングを行う手法が一般的である。しかし，文書における単語の出現頻度には偏りがあることが多く，次元圧縮を行う際，標本である分類対象文書集合に対してオーバーフィットを起こしやすいという問題がある。

一般的なクラスタリングにおいて，分類精度を向上させようとした場合，分類対象間に must-link のように制約を付与する半教師あり学習によって，分類精度を上げる手法があることは知られているが，文書分類の際に，これから分類しようとする文書間の関係を推定し，直接的に制約を付与することは難しい上に，これではオーバーフィットが起こりやすいという問題に対しての根本的な解決とはならない。ここで我々は，文書間の関係に比べると比較的容易に手に入れることが可能である，単語間の関係を制約として用いるならば，標本に対するオーバーフィットを回避しつつ，分類精度を向上させることができるのではないかと考えた。

本研究では，確率的潜在意味解析 pLSI[1]に，分類対象文書に依存しない大域的な辞書を用い，外部から単語間関係の制約を付与することによって，オーバーフィットを避けることが可能な文書分類手法を提案する。

2． 潜在的意味解析 pLSI

pLSI は，Aspect model（隠れ変数モデル）に基づいて次元圧縮を行う手法である。Aspect model において，文書と単語の同時生起確率$p(d,w)$は，一般的に式(1)のように求められ，このとき$p(w|d)$は式(2)のように表わされる。

$$p(d,w) = p(d)p(w|d) \quad (1)$$

$$p(w|d) = \sum_{z\in Z} p(w|z)p(z|d) \quad (2)$$

式(2.2)を，ベイズの定理を用いて変形し，同時生起確率$p(d,w)$を求める式(2.1)に代入すると，式(3)のようになる。

$$p(d,w) = \sum_{z\in Z} p(z)p(w|z)p(d|z) \quad (3)$$

このとき，$p(z)$, $p(w|z)$ および $p(d|z)$ は，隠れ変数であるため直接求めることはできない。一般的には，式(4),(5)で表わされる EM アルゴリズムを用いて最尤推定を行う。

・E-step

$$Q_{ijk}^{(t+1)} = \frac{p(z_k)^{(t)}p(d_i|z_k)^{(t)}p(w_j|z_k)^{(t)}}{\sum_k\left\{p(z_k)^{(t)}p(d_i|z_k)^{(t)}p(w_j|z_k)^{(t)}\right\}} \quad (4)$$

・M-step

$$\begin{aligned} p(z_k)^{(t+1)} &= \frac{\sum_i\sum_j n(d_i,w_j)Q_{ijk}^{(t)}}{\sum_k\sum_i\sum_j n(d_i,w_j)Q_{ijk}^{(t)}} \\ p(d_i|z_k)^{(t+1)} &= \frac{\sum_j n(d_i,w_j)Q_{ijk}^{(t)}}{\sum_i\sum_j n(d_i,w_j)Q_{ijk}^{(t)}} \\ p(w_j|z_k)^{(t+1)} &= \frac{\sum_i n(d_i,w_j)Q_{ijk}^{(t)}}{\sum_j\sum_i n(d_i,w_j)Q_{ijk}^{(t)}} \end{aligned} \quad (5)$$

3． 制約付き pLSI

3.1 制約の組み込み

制約付き pLSI のモデルは，基本的に制約の無い場合と同様であるが，関連研究[4]にならい，確率分布$p(Z)$の求め方を，次のように変更する。

$$p(Z) = \prod_k p(z_k) \times \frac{exp\{-\sum_{j,j'}\theta_{jj'}\cdot\delta(z_j\neq z_{j'})\}}{G} \quad (6)$$

このとき$\delta(\cdot)$はデルタ関数であり，隠れ変数z_jと$z_{j'}$の値が等しい場合は 0，異なる場合は 1 をとる。また，$\theta_{jj'}$は制約項である。

3.2 平均場近似を用いた E-step の計算

本研究において，提案モデルのパラメータ推定は EM アルゴリズムを用いて行う。前章でも述べたように，制約付き pLSI における EM アルゴリズムによる更新手続きは，制約の無い場合の更新手続きと同じである。しかし，本研究で提案する制約付き pLSI では，制約を付与することにより，各隠れ変数がそれぞれ独立であるという Aspect model での仮定が成り立たなくなるため， E-step では以下の確率計算を行う必要が出てくる。

$$p(z_k|D) = \sum_{Z_{-k}} p(Z|D) \propto \sum_{Z_{-k}} p(D|Z)p(Z) \quad (7)$$

平均場近似[3]では，上記の確率分布$p(Z|D)$を式(8)のような形の分布$Q(Z)$で近似する。

$$Q(Z) = \prod_{i,j,k} Q_{ijk} \quad (8)$$

上で示した近似分布$Q(Z)$のパラメータは，真の分布$p(Z|D)$からの KL ダイバージェンスが最少となるものを選ぶ。すなわち，近似分布を算出する作業は，任意のi,jに対し，以下の式(9)を最小化する問題となる。

$$\sum_k Q(Z)\, log\, \frac{Q(Z)}{p(Z|D)} \quad (9)$$

この目的関数に対し，ラグランジュの未定乗数を導入してQ_{ijk}の極値条件を求めると，式(10)を得る。

$$Q_{ijk} \propto p(z_k)p(d_i|z_k)p(w_j|z_k) \times exp\{-\sum_{j'\neq j}(1-Q_{ijk})\theta_{jj'}\} \quad (10)$$

† 慶應義塾大学 政策・メディア研究科
‡ 東京大学 情報理工学系研究科

式(10)より，近似分布のパラメータを次の更新式に基づく反復法で求める。

$$Q_{ijk}^{(t+1)} \propto p(z_k)p(d_i|z_k)p(w_j|z_k) \times \exp\{-\Sigma_{j'\neq j}(1-Q_{ijk}^{(t)})\theta_{jj'}\} \quad (11)$$

上の式について，$\Sigma_z Q_{ijk}=1$ という条件を考えると，最終的に制約を組み込んだ場合の E-step における更新式は，次のようになる。

$$Q_{ijk}^{(t+1)} = \frac{p(z_k)^{(t)}p(d_i|z_k)^{(t)}p(w_j|z_k)^{(t)}\cdot\exp\{-\Sigma_{j'\neq j}(1-Q_{ijk}^{(t)})\theta_{jj'}\}}{\Sigma_k\{p(z_k)^{(t)}p(d_i|z_k)^{(t)}p(w_j|z_k)^{(t)}\cdot\exp\{-\Sigma_{j'\neq j}(1-Q_{ijk}^{(t)})\theta_{jj'}\}\}} \quad (12)$$

3.3 EM アルゴリズムにおける初期値の設定

EM アルゴリズムには局所最適解に陥りやすいという問題があり，それは各パラメータの初期値に依存する。そのため本研究では，初期値の設定についても単語間の制約を反映させながら行う。具体的には，語w_jと語$w_{j'}$の間に制約を科す場合，確率$p(w_j|z_i)$及び$p(w_{j'}|z_i)$が，全ての i について等しくなるように初期値を設定した。

$$p(w_j|z_i) = p(w_{j'}|z_i) \qquad \{i=1,2,\ldots,k\} \quad (13)$$

4．実験

今回の実験では，クラスタリング対象文書データとして，毎日新聞データ集 2004 年度版より，国際・経済・家庭・文化・読書・科学・芸能・スポーツ・社会の 9 トピックについて，それぞれ 1 トピックあたり 100 記事ずつ，合計 900 文書を収集した。用意した文書集合から，文中に出現する名詞，動詞および形容詞と，その出現頻度を素性として取り出し，文書ベクトル$\vec{d_i}$とした。なお，形態素解析には MeCab を用いた。

次に，その文書ベクトル$\vec{d_i}$に対し，3 章で説明した式(12),(5)を用いた EM アルゴリズムによって，以下の式(14)で計算される対数尤度が収束するまで，繰り返しパラメータ推定を行った。本実験では，$\sqrt{(L^{(t)}-L^{(t-1)})^2}$が 0.001 未満になった際，対数尤度が収束したものとして見なし，パラメータ推定を終了した。なお，本実験では，過学習を避けるため，Tempered EM[2]を適用した。

$$L = \Sigma_{w\in W}\Sigma_{d\in D} n(d,w)\log p(d,w) \quad (14)$$

対数尤度が収束したのち，推定されたパラメータ$p(d_i|z_k)$および$p(z_k)$を利用し，式(15)により$p(d_i,z_k)$を計算し，得られた$\{p(d_i,z_1),p(d_i,z_2),\ldots,p(d_i,z_k)\}$のうち，最大となるものをクラスタ番号として割り当てた。

$$p(d_i,z_k) = p(d_i|z_k)p(z_k) \quad (15)$$

また，制約付き pLSI において使用する単語間の制約$\theta_{jj'}$については，Web 文書から集めた 100 万文より，文中で共起する各単語ペアの$\chi 2$乗値を求め，それに定数を掛けたものを重みづけとして用いた。本実験では，提案モデルである制約付き pLSI の有効性を検証するため，制約を用いない場合の pLSI との比較実験を行った。

5．評価

クラスタリング結果の精度・再現率から F 値および精度，再現率を算出し，提案手法の評価を行った。なお，EM アルゴリズムは初期値依存性を持つため，制約付き，制約なし，ともに 5 回試行を行って，平均をとった。制約を用いた場合の F 値は 0.523，制約を用いない場合の F 値は 0.496 となった。この結果から，単語間関係を制約として用いた提案手法のほうが，制約を用いなかった場合に比べてクラスタリング結果が良くなったことがわかる。

表 1. クラスタリング結果の比較

	f-measure	precision	recall
制約あり	0.523	0.530	0.516
制約なし	0.496	0.503	0.489

6．まとめ

評価のところで詳しく説明しなかったが，制約を用いた場合は制約なしの場合に比べて，特に初期値依存が大きくなり，良い初期値から始まった場合はもちろん制約を用いなかった場合に比べて結果も良くなるが，悪い初期値から始まった場合は，制約を用いなかった場合に比べて結果が悪くなる場合もあり，結果の分散が大きくなった。これは，$p(w_j|z_i)$の初期値設定において，違う潜在トピックから派生しているにも関わらず，似た初期値が割り振られたためだと考えられる。これを解決するためには，今回行った must-link のような同じ初期値を設定する手法の逆，つまり，cannot-link のように，大きく違う初期値を割り振ることができたら，結果はさらに安定し，精度が向上すると考えられる。

また，本研究における提案手法について，今回の実験では Web 文書から抽出した単語間の関係を制約として用いたが，距離や重みといった単語間の関係の強さを数値として含む辞書であれば，他の辞書を利用することによっても同様の成果を期待することができる。

7．参考文献

[1] T. Hofmann, “Probabilistic Latent Semantic Indexing”, In Proceeding of the 22nd annual international ACM SIGIR conference, pp.289-269 (1999).

[2] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, Machine Leaning, vol.42, pp.177-196 (2001).

[3] H. Nishimori, “Statistical Physics of Spin Glasses and Information Processing”, Oxford Uni- versity Press (2001).

[4] 高村大也, 乾孝司, 奥村学. “複数語からなる評価表現のモデル化”, 言語処理学会 第 12 回年次大会 (2006).

[5] 鍛冶伸裕, 喜連川優. ”単語の半教師ありクラスタリング”, 情報処理学会 第 70 回年次大会 (2008).