# ジャンル別 LSA の結果統合による新聞記事のジャンル推定

○箕浦健太郎，田村哲嗣，速水悟（岐阜大）

## 1. はじめに

昨今では趣味の多様化，ニーズの多様化が唱えられている．そうした要求に対して応えるためには，ユーザーの趣向に見合ったものを提示できると望ましい．それについて，ジャンル推定というタスクは重要である．

ジャンル推定において文章の特徴を抽出するということは重要であり，これまで LSA(Latent Semantic Indexing：潜在的意味索引)[1]や LDA(Latent Dirichlet Allocation：潜在的ディリクレ配分法)[2]が用いられてきた．LSA は文章(若しくは単語)を概念空間上に表現する手法である．また LDA はカテゴリを潜在パラメータとして仮定した上で確率論に基づいて形成される手法である．

本論ではその手がかりの一つとして，ジャンルごとに LSA を行いその結果を統合することによって次元圧縮を実現し，それに対して今までジャンル推定の分野で利用が試みられることのなかった GMM（Gaussian Mixture Model：混合正規分布）を適用するという手法を提案する．

## 2. 提案手法

テキスト情報処理において次元圧縮手法として多く利用されている LSA はテキストの特徴を抽出するものであるが，単にデータ集合に対して LSA を用いて次元圧縮を行うとデータそのものの特徴を抽出することとなり，本来得たいジャンルを分類するに足る要素を取り逃すこととなる．このことはモデルの汎用性にも影響する．

従って，一度ジャンル毎に LSA を行いその結果を統合したものを次元圧縮行列とすることで，ジャンル毎の特徴を抽出し，かつ次元圧縮を実現する．

また，LSA を用いる場合多くは次元圧縮後コサイン尺度によって類似度が計算される．しかし，コサイン尺度は単一の文章との類似度を計算することに向いているものの，クラス分類問題には適していない．そこでサブカテゴリが表現されることを期待して GMM という形でジャンル毎の特徴をモデリングすることとする．

本論では具体的に図 1 の手法を用いて推定を行う．以下ではその主要な項目について詳細に記述する．

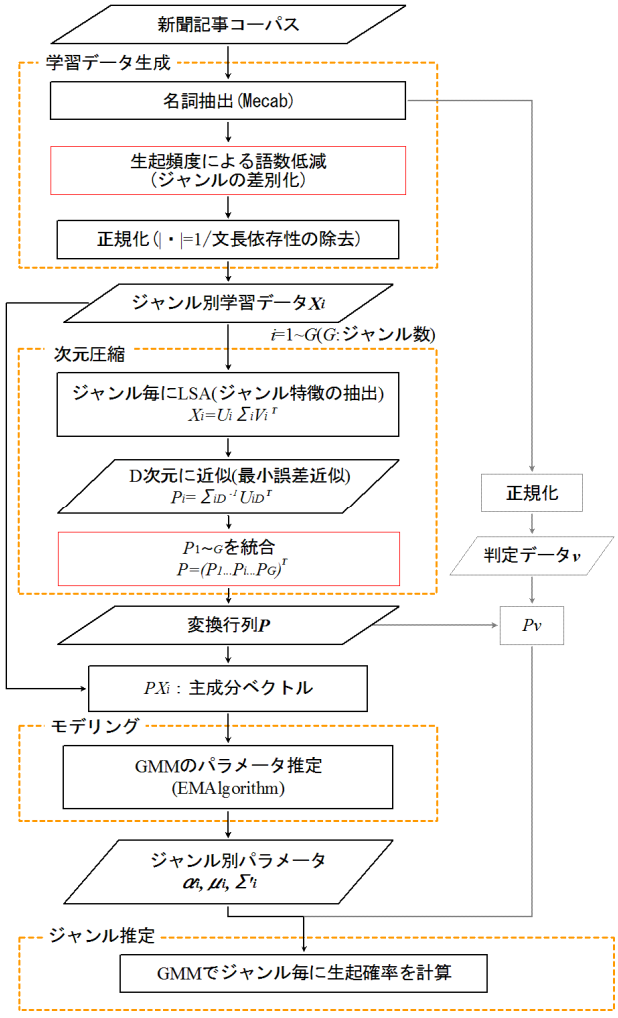


図 1　手法の概略

### 2-1. 語数低減

コーパスから抜き出した単語は出現率がジャンルに依存しないものを含み，これらは頻出する傾向があることから特異値分解において強く作用するような単語となりやすいため，ジャンル推

定の際の精度を下げることとなる場合が多い.

本論では, ジャンルを推定するにあたって, 文章全体で生起数の高い単語上位$TH$%(以下削除率/小数点以下切り上げ)を取り除くことによってジャンルに依存しない単語を削除し, これにより精度の向上を図ることとする.

これは LSA によるジャンル毎の特徴の抽出を補助する形となる.

### 2-2. データの正規化

データの記事の長さに対する依存をなくすため, データにおける単語の出現回数を正規化(|・|=1)する. これによって, 記事における単語の生起比率だけを特徴量としてクラス分類することができ, その記事がどの程度の長さであるのかについて考慮する必要がなくなる(ただし出現する単語の種類数に対する依存は残る). また, テストデータも同様に正規化を行うこととする.

### 2-3. LSA (Latent Semantic Analysis)

潜在意味分析(LSA)とは, 特異値分解(SVD)を用いて文章-単語行列から文章・単語の出現頻度(TF)等の類似度を考慮した疑似文章ベクトル, 及び疑似単語ベクトルを生成する手法である. また, この手法は同時に次元圧縮も実現する.

特異値分解は次式(1)で表される.

$$X = U\Sigma V^T \tag{1}$$

単語数を$W$, 文章数を$N$ (ただし$W \geqq N$)とおくと, $X \in \mathrm{R}^{W\times N}$, $U \in R^{W\times W}$, $\Sigma \in R^{W\times N}$, $V \in R^{N\times N}$で, $U$,$V$は直交行列, $\Sigma$は(i,i)成分にのみ値を持つ行列である(i:1~N). また, $U$の各行は疑似単語ベクトル, $V$の各行は疑似文章ベクトルであり, これらは$X$の各行, 各列の単語, 文章を別の特徴空間(以下概念空間)で表現したものである.

一方, $\Sigma$の(i,i)成分は特異値と呼ばれ, その値は$U$,$V$の各行に固有な重み付けを行う. これを利用し, $U$,$V$,$\Sigma$を特異値の大きい方から$k$行のみ順に抽出することで$X_k$を次式(2)のように定義する.

$$X_k = U_k\Sigma_k V_k^T \tag{2}$$

ここで, $U_k \in \mathrm{R}^{W\times k}$, $\Sigma_k \in R^{k\times k}$, $V_k \in R^{N\times k}$であり, これによって与えられる$X_k$は元の$X$の近似を最小二乗誤差で与える.

ところで, $V_k$の各行は文章について概念空間によって表された$k$次元のベクトルであり, これは元の$X$の各文章$x$をW次元から$k$次元に圧縮したものと見なせる. これを$x'$とおくと, 次式(3)で表される.

$$x' = \Sigma_k^{-1}U_k^T x = Px \tag{3}$$

ここで$P \in R^{k\times W}$である. この方法によって求めた次元変換行列$P$を用いることによって, データの主成分を抽出することができる.

### 2-4. ジャンル別 LSA の結果統合

ジャンル推定においては文章全体の特徴ではなくジャンルごとの特徴が得られるのが望ましい. 従って, それぞれのジャンルに属するデータ毎の次元変換行列$P_i$を求め, 次式(4)のようにその行列を縦方向へ連結することによって, 次元変換行列$P$とすることを提案する. ただし, このとき全てのジャンルについて圧縮した次元数は同じであるとする.

$$P = (P_1 \ldots P_i \ldots P_G)^T \tag{4}$$

これにより, 単一のベクトル上にジャンル毎の特徴を同時に表現することができるため, コーパス全体の特徴に左右されにくいような特徴を抽出することができると考えられる.

### 2-5. GMM (Gaussian Mixture Model)

混合ガウス分布(GMM)とは, ガウス分布を複数用いることでより汎用的にしたものであり, 次式(5)のように表される.

$$P(x|\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{k=1}^{K}\alpha_k N(x|\mu_k,\Sigma_k) \quad \text{ただし, } \sum_{k=1}^{K}\alpha_k = 1 \tag{5}$$

ここで$K$は用いるガウス分布の個数であり, 混合数(コンポーネント数)と呼ばれる. この分布を用いてジャンル毎のパラメータを推定し, モデルを形成する.

本論でこの GMM を利用するのは, カテゴリ中に存在しているサブカテゴリを GMM の各コンポーネントに表現させることを目的としている.

分布のパラメータ$\alpha_k$, $\mu_k$, $\Sigma_k$は, EM アルゴリズムを用いることによって求めることができる. な

お, 本システムでは$\Sigma$の零割りが発生することを防ぐため, EM アルゴリズムによってパラメータを求める際にベイズ正則化を適応することとする. その際の正則化定数は 1.0e-1 とする.

## 3. 実験・結果・考察

本実験では, コーパスは毎日新聞記事コーパス'94～'07, 使用ジャンルは{社説, 国際, 経済, 家庭, 文化, 読書, 科学, 芸能, スポーツ, 社会}の 10 ジャンルを対象とする.

また検証の条件として, オープンテストで, GMM の混合数$K = 8$, 検証する次元$D = \{20, 30, 40, 50\}$とする.

なお, 表中の*は判別されないジャンルを含んでいることを示す. この場合は目的のタスクに不適であるため, あくまで参考値として扱うこととする.

### *3-1. 次元変換行列 P の生成方法による比較*

#### 3-1-1. 実験

本論での提案手法の効果を示すため, 次元変換行列$P$の生成方法によって結果がどのように変化するかを示す. 一つは学習データ全体による次元変換行列(i), もう一つは学習データのジャンル毎に次元変換行列を求めた後それらを統合することによる次元変換行列(ジャンル別 LSA の結果統合/ii)である. ただし, 学習データ数$N = 1000$, 削除率$TH = \{0, 0.2\}$とする. また, 圧縮後の次元数$D$による変化についても示す.

#### 3-1-2. 結果・考察

結果を表 1 に示す.

**表 1　次元変換行列 P の生成方法による F 値**

| | 全体(i) | | ジャンル毎(ii) | |
|---|---|---|---|---|
| D/ | TH=0 | TH=0.2 | TH=0 | TH=0.2 |
| 20 | 41.57 | 41.42 | 34.05 | 30.69 |
| 30 | 16.48* | 22.56 | 38.72* | 52.28 |
| 40 | 23.86* | 50.24 | 44.79 | 56.31 |
| 50 | 43.36* | 59.09 | 48.34 | 62.41 |

$D = 20$を除いて, 学習データ全体による次元変換行列(i), ジャンル毎に次元変換行列を求めたあとそれらを統合した次元変換行列(ii)共に, 語数の低減を行った方の F 値が高い. また, $D = 20$を除いて, ジャンル毎に次元変換行列を求めた方(ii)の F 値が高いという結果となっている.

これによりジャンル別 LSA の結果統合が実際に有効な手法であることが分かる.

なお$D = 20$において成立しないのは, 語数低減については次元数が少ないため情報の欠損に対し脆弱なのではないか, 圧縮行列についてはジャンル毎に次元変換行列を求めたときに抜け落ちた情報が全体で次元変換行列を求めたときにはより低い寄与率まで参照されるため含まれていたのではないかと考えられる.

### *3-2. 学習データ数による比較*

#### 3-2-1. 実験

学習データ数$N$によって結果がどのように変化するかを示す. ここで$N = \{500, 1000, 1500, 2000\}$, 語数低減の基準$TH = \{0, 0.2\}$とする. また, 圧縮後の次元数$D$による変化についても示す.

#### 3-2-2. 結果・考察

結果を表 2, 表 3 に示す. なお, 最下段$W$は語数である.

**表 2　学習データ数による F 値(TH=0)**

| $D/N$ | 500 | 1000 | 1500 | 2000 |
|---|---|---|---|---|
| 20 | 13.04 | 34.05 | 38.74 | 10.14* |
| 30 | 30.27 | 38.72* | 32.57* | 44.71 |
| 40 | 43.50 | 44.79 | 41.10* | 49.63 |
| 50 | - | 48.34 | 49.45 | 51.04 |
| $W$ | 14322 | 20683 | 25166 | 29505 |

**表 3　学習データ数による F 値(TH=0.2)**

| $D/N$ | 500 | 1000 | 1500 | 2000 |
|---|---|---|---|---|
| 20 | 23.98 | 30.69 | 36.09 | 35.60 |
| 30 | 37.09 | 52.28 | 62.02 | 44.59 |
| 40 | 57.55 | 56.31 | 61.92 | 62.20 |
| 50 | - | 62.41 | 63.45 | 63.32 |
| $W$ | 14294 | 20642 | 25116 | 29446 |

ただし$D = 50$, $N = 500$は EM アルゴリズムが収束しないため未掲載である.

$TH = 0$の場合, 判例されないジャンルが存在したときを除いて, おおよそデータ数, 次元数に比例する結果となっている.

一方, $TH = 0.2$の場合, データ数, 次元数に比

例する傾向はあるものの，$TH = 0$ほどそれは強くはなく，不安定な結果となっている．

また，双方を比較すると，おおよそ語数低減を行った場合の方のF値が高いため，語数削減は(次元数が高いときはとりわけ)データ数，次元数の増加に対する補助的な役割を担えるだろうと考えられる．

### *3-3. 従来法との比較*

3-3-1. 実験

従来法との比較実験として LSA 後の文章-単語行列に対してコサイン尺度を用いた場合の結果を示す．ここでいずれにおいても語数低減の基準 $TH = \{0, 0.2\}$，学習データ数$N = 1000$とする．また，圧縮後の次元数$D$による変化についても示す．

3-3-2. 結果・考察

結果を表 4 に示す．

**表 4　従来法(コサイン尺度)による F 値**

| | コサイン尺度 | | 提案手法(GMM) | |
|---|---|---|---|---|
| D/ | TH=0 | TH=0.2 | TH=0 | TH=0.2 |
| 20 | 10.61* | 45.97 | 34.05 | 30.69 |
| 30 | 13.87* | 47.68 | 38.72* | 52.28 |
| 40 | 15.44 | 48.40 | 44.79 | 56.31 |
| 50 | 17.65 | 43.66 | 48.34 | 62.41 |

$D = 20$の場合を除いて，提案手法の F 値の方が上回っている．またその次元に対する精度の改善幅も提案手法の方が大きいことが分かる．

### 3-4. *GMM について*

いずれの場合も，有効な混合数が 1~2 程度となる．これは圧縮行列が写像と同じ作用をするため，圧縮後の要素がそれぞれ線形独立なサブカテゴリを表現することとなり，これにより圧縮後のデータが複数のコンポーネントを持たなくなったことによるものと考えられる．しかしながら正規分布を用いた場合逆行列を持つことができなくなることがあるため，GMM により近似することによってこれを回避することを目的として使用する．

## 4. 今後の課題

### *4-1. 精度について*

実用について本論での精度は遠く及ばない．これについては，テキスト情報処理におけるデータ数，単語数としては本論で用いているデータ数は少ないため，より規模の大きいデータにおいてどのようになるのかを検証することで，考察を行えるのではないかと考えられる．これらが実現できなかったのは SVD においてのメモリ容量上の問題があったためであり，SVD の逐次アルゴリズム等を利用することによって解決を図りたい．

### *4-2. ジャンルの差別化について*

本手法においてジャンル毎に LSA を行う場合，他のジャンルとの重複部分が考慮されることなくジャンルの主成分が抽出されることとなる．この事象を回避することを目的として語数低減を行っているが，これは本来統計的な情報に基づいて行われるべきであろう．考えられる方法としては全体でのLSA，各ジャンルでのLSAの両方において次元圧縮行列内の寄与率の高い横ベクトル中の数値が大きい単語を除くといった処理がある．これは文章全体でもジャンル毎でも主成分となりうる，ジャンルの差別化を阻害する単語を除去することを目的としている．

### *4-3. GMM について*

次元数を大きくすると GMM でさえ分散共分散行列が逆行列を持てなくなる．これは圧縮したデータの次元中に線形従属なベクトルを含むために起こると考えられる．この事象はデータ数を増加させると回避することができるため，より多いデータ数で検証を試みる必要があるだろう．

## 参考文献

[1]S. Deerwester, Susan Dumais, G. W. Furnas, T. K. Landauer, R. Harshman (1990). “Indexing by Latent Semantic Analysis”. Journal of the American Society for Information Science 41 (6), 391–407.

[2]D. M. Blei, A. Y. Ng, M. I. Jordan (2003). “Latent Dirichlet Allocation”. Journal of Machine Learning Research 3, pp.993-1022.