PLSIを利用した文書からの知識発見

Knowledge Discovery in Documents based on PLSI

伊藤 潤 † 石田 崇 † 後藤 正幸 ‡ 酒井 哲也 § 平澤 茂一 †
Jun ITO Takashi ISHIDA Masayuki GOTO Tetsuya SAKAI Shigeichi HIRASAWA

1. はじめに

自由記述形式のアンケート調査の回答は，回答者の自由な意見を集約できる効果があり，近年注目されている．しかし，その処理方法については，従来，人の主観に頼るなどして適切な解析手法が確立しておらず，現在研究が進められている [1][2][3].

一方，情報検索の分野において，確率を利用して文書単語行列を低次元に圧縮する PLSI が提案され，情報検索における有効性が示されている [4].

本稿では，PLSI を利用した文書のクラスタリング, 分類手法を提案する．そして，自然言語で書かれたアンケートの回答を分類することによって知識を発見し，アンケートの解析を自動化する手法を示す．そして実データに適用し，その有効性を明らかにする．

2. 潜在意味モデル

2.1 LSI

S. Deerwester らは，意味的情報検索のモデルとして LSI(Latent Semantic Indexing) を提案した [5]．LSI では，単語文書行列 A を特異値分解 (SVD) によって

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

と分解する．このうち，主成分の大きい方から K 個を用いて

$$\hat{A} = U_K \Sigma_K V_K^T \tag{2}$$

とすることにより，A を K 次元の潜在意味空間に圧縮することでノイズの除去を行う．これは，行列 A と $\hat{A}$ の 2 乗誤差を最小にする圧縮となっている．

しかし，LSI による情報検索においては単語文書行列 A に idf 値などで ad-hoc な重み付けが必要であるなど，いくつかの問題がある．

2.2 PLSI

一方，T. Hofmann によって提案された PLSI (Probabilistic Latent Semantic Indexing)[4] は，LSI と同様の圧縮を確率モデルに基づいて行う手法である．

PLSI では，意味的な隠れ属性 $z_k(k = 1, 2, \ldots, K)$ のもとで，文書 $d_i(i = 1, 2, \ldots, I)$ と単語 $w_j(j = 1, 2, \ldots, J)$ の生起は独立であると考え，d_i と w_j の同時確率 $P(d_i, w_j)$ を

$$P(d_i, w_j) = \sum_k P(d_i|z_k)P(w_j|z_k)P(z_k) \tag{3}$$

のように表す．ここで，文書 d_i における単語 w_j の実際の出現回数を $n(d_i, w_j)$ とすると，データの対数尤度

$$L = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j) \tag{4}$$

を最大にする $P(z_k)$，$P(d_i|z_k)$，$P(w_j|z_k)$ を，EM アルゴリズムで以下の式を計算することにより最尤推定する．

E-step

$$P(z_k|d_i, w_j) = \frac{P(z_k)P(d_i|z_k)P(w_j|z_k)}{\sum_{k'} P(z_{k'})P(d_i|z_{k'})P(w_j|z_{k'})} \tag{5}$$

M-step

$$P(w_j|z_k) = \frac{\sum_i n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i,j'} n(d_i, w_{j'})P(z_k|d_i, w_{j'})} \tag{6}$$

$$P(d_i|z_k) = \frac{\sum_j n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i',j} n(d_{i'}, w_j)P(z_k|d_{i'}, w_j)} \tag{7}$$

$$P(z_k) = \frac{\sum_{i,j} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i,j} n(d_i, w_j)} \tag{8}$$

(5)〜(8) 式の計算は，実際には，過学習を避けるため Tempered EM を用いている [4].

3. PLSIによる文書クラスタリング

PLSI の隠れ属性 z_k は，ひとつの概念を表していると捉えることができる．そこで，z_k を用いて以下のように文書集合を S 個のクラスタにクラスタリングする [3].

[アルゴリズム]

1. $K = S$ として，EM アルゴリズムにより $P(z_k)$, $P(d_i|z_k)$，$P(w_j|z_k)$ を求める．
2. 各文書 d_i を，$\max_k P(z_k|d_i) = P(z_{\hat{k}}|d_i)$ となる $z_{\hat{k}}$ に割り振る．
3. 各 z_k に割り振られた文書集合をそれぞれ S 個のクラスタとする．

4. 代表元を用いた文書自動分類

PLSI は EM アルゴリズムを用いるため，初期値の近くにある局所解に収束する性質がある．そこで，初期値をうまく設定すると意図的に隠れ属性を作ることができ，その隠れ属性をもとにクラスタリングを行うと，初期値として与えた代表元による分類と捉える事ができる．

[初期値の与え方]

分類する目的に応じて隠れ属性 z_k の代表元 $\dot{\boldsymbol{d}}_k$ を S 個作成（または選択）する．代表元 $\dot{\boldsymbol{d}}_k$ は，$\dot{\boldsymbol{d}}_k = (f_{k,1}, f_{k,2}, \ldots, f_{k,J})$ で表されるものとする．ここで，$\sum_j P(w_j|z_k) = 0$ となる z_k があると (5) 式が計算できないため，以下のように補正した初期値を用いる．

$$P(w_j|z_k) = \frac{f_{k,j} + \alpha}{\sum_{j'}(f_{k,j'} + \alpha)} \tag{9}$$

$$P(d_i|z_k) = 1/I \tag{10}$$

$$P(z_k) = 1/K \tag{11}$$

α は正の値をとるパラメータである．

[アルゴリズム]

1. S 個の代表元を作成し，$K = S$ として初期値を設定する．

†早稲田大学理工学部経営システム工学科
‡武蔵工業大学環境情報学部情報メディア学科
§株式会社東芝研究開発センター知識メディアラボラトリー

2. EM アルゴリズムにより，$P(z_k)$，$P(d_i|z_k)$，$P(w_j|z_k)$ を求める．
3. 各文書 d_i を，$\max_k P(z_k|d_i) = P(z_{\hat{k}}|d_i)$ となる $z_{\hat{k}}$ に分類する．

5. アンケートデータの解析

5.1 解析データ

早稲田大学理工学部経営システム工学科 2 年の必修科目「コンピュータ工学 (CE)」の初回の講義の際に行ったアンケートを解析する．質問項目は選択式の質問と記述式の質問があり，コンピュータに対する知識や授業に対する要望についての質問である．

このアンケートは学生の特性に応じたクラス分けを行うための予備調査として実施されたもので，受講生約 150 名を将来技術系の専門職に就くであろう学生のクラス (スペシャリスト;C_{Spe}) と一般総合職に就くであろう学生のクラス (ジェネラリスト;C_{Gen}) の 2 つのクラスに分類することを目的とする．

5.2 解析方法

ここでは記述式の質問に対する回答を用いて文書を分類し，選択式質問の回答に対して判別分析を行うことにより，その性質を評価する．また，別に行った調査で得られた本人の自己申告による C_{Spe}, C_{Gen} の希望との違いも比較する．

さらに，同大学の文系の学生を対象にした講義「情報化社会概論 (IS)」においても同様のアンケートを実施し，CE, IS それぞれの文書ベクトルの重心ベクトルを C_{Spe}, C_{Gen} の代表元とした．

5.3 解析結果

(9) 式の $\alpha = 1.0, 0.5$ においてクラスの分類を行い，その 2 つのクラスの違いを判別分析で検証した結果，判別係数の高い質問項目を表 1,2 に示す．また，各クラスにおける特徴的な単語（$P(w_j|z_k) - P(w_j)$ の大きな w_j）の一部を表 3 に示す．さらに，これらから解釈した各クラスの特徴を表 4 に示す．

表 1: $\alpha = 1.0$ における各クラスの特徴

説明変数	判別係数	C_{Spe}	C_{Gen}
単位を落としてもかまわない	0.944	+	-
CG に興味がある	0.803	-	+
ソフトウェアに興味がある	0.771	+	-
理論に興味がある	0.718	+	-
ネットワークに興味がある	0.692	+	-

表 2: $\alpha = 0.5$ における各クラスの特徴

説明変数	判別係数	C_{Spe}	C_{Gen}
単位を落としても構わない	1.528	+	-
この科目は自分にとって必要	1.054	-	+
ネットワークに興味がある	1.024	+	-
Web ページ作成に興味がある	0.987	+	-
情報検索に興味がある	0.972	-	+

表 3: 各クラスの特徴的な単語

	C_{Spe}	C_{Gen}
$\alpha = 1.0$	コンピュータ，システム，専門，構造，管理，	情報，プログラム，ADSL，ロボット，デザイン
$\alpha = 0.5$	コンピュータ，弁理士，科学，大学院，ソフトウェア	化学，社長，技法，ネットゲーム，デジカメ

表 4: 各分類における全体的な特徴

	クラス	特性
自己申告	C_{Spe}	専門用語に詳しく，試験による評価を希望
	C_{Gen}	コンピュータの利用方法に興味を持つ
$\alpha = 1$	C_{Spe}	理論に興味があり，大学院への進学を希望
	C_{Gen}	コンピュータの利用法に興味
$\alpha = 0.5$	C_{Spe}	理論に興味があり，成績が良い
	C_{Gen}	コンピュータを利用したシステムに詳しい

6. 考察

5.3 節に示したように，C_{Spe} にはコンピュータそのものに興味がある学生が集まり，C_{Gen} には道具としてコンピュータを利用しようと考えている学生が集まった．これは，2 つの代表元の与え方が，学生のコンピュータに対する接し方をうまく説明できるものであったためと考えられる．

また，学生の自己申告によるクラス分けと本手法での分類の結果は必ずしも一致しない．しかし，学部 2 年生の段階ではまだ将来を決めかねている学生も多く，そういった学生の個人の特性からふさわしいクラスに振り分けることができ，提案手法は有効であることが明らかになった．

7. おわりに

PLSI の初期値依存性を利用して文書を自動分類する手法を提案し，アンケートの分析に応用した．学生のクラス分けに適用した結果，意図した分類が実現できることを示した．

今後の課題としては，本手法の性能を定量的に測定してその有効性を示すとともに，PLSI の確率モデルを利用して文書集合の特性を抽出する手法を考えていく．

参考文献

[1] 酒井哲也，伊藤潤，後藤正幸，石田崇，平澤茂一，“情報検索技術を用いた効率的な授業アンケートの分析”，JASMIN 2003 年度春季全国大会，pp.182-185，2003 年 6 月．

[2] 後藤正幸，酒井哲也，伊藤潤，石田崇，平澤茂一，“選択式・記述式アンケートからの知識発見”，2003 年度 PC カンファレンス，鹿児島，2003 年 8 月．

[3] S. Hirasawa, and W. W. Chu, “Knowledge Acquisition from Documents with both Fixed and Free Formats”, to appear in IEEE 2003 Conf. on SMC, Oct.

[4] T. Hofmann, “Probabilistic Latent Semantic Indexing”, Proc. of SIGIR'99, ACM Press, pp.50-57, 1999.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis” J. of the Society for Information Science, 41, pp.391-407, 1990.