The 29th Symposium on Information Theory
and its Applications (SITA2006)
Hakodate, Hokkaido, Japan, Nov. 28 – Dec. 1, 2006

単語の特徴を考慮したPLSIによる文書クラスタリング

Document Clustering Methods based on Probabilistic Latent Semantic Indexing with Feature of Words

長尾 壮史* 八木 秀樹† 平澤 茂一*
Masafumi NAGAO Hideki YAGI Shigeichi HIRASAWA

Abstract— Probabilistic Latent Semantic Indexing is a probabilistic method for document indexing, which is based on a statistical latent class model for factor analysis. In PLSI, the computed results often change due to an initial value of Expectation Maximization (EM) algorithm, which is a random variable. To avoid this problem, we propose a novel document clustering method based on PLSI using initial value's dependency of EM algorithm and feature of words. The results on test collection data show effectiveness of our method.

Keywords— document clustering, Probabilistic Latent Semantic Indexing, information retrieval, EM algorithm

1 はじめに

Web ページ，雑誌論文などの大量の文書集合を内容ごとに自動的に分ける技術は文書クラスタリングと呼ばれる．文書クラスタリングは，大量の文書集合を効率的に解析する手段として有効である．文書クラスタリングは，擬似フィードバック検索におけるフィードバック情報やユーザの意図に応じた検索結果の提示に用いられることが多く，その有効性が確認されている．

一方，情報検索の分野において単語-文書行列を低次元へ圧縮する研究が行われている．単語-文書行列は文書集合が大きくなれば疎行列になり，文書中に含まれる不必要な索引語が検索やクラスタリング，文書分類においてその性能に悪影響を及ぼす．そのため，行列を低次元へ圧縮することで疎行列が及ぼす影響を防ぐことができる．代表的な手法として確率空間を利用して文書索引語行列を低次元に圧縮する PLSI（Probabilistic Latent Semantic indexing）[1] がある．

本研究では，この PLSI を利用したクラスタリング手法を提案する．また新聞記事データに対して評価実験を行い，その有効性を示す．

2 潜在的意味モデル PLSI

T. Hofmann によって提案された PLSI [1] は，行列の代数的圧縮である LSI [2] とは異なり，確率モデルに基づいて圧縮を行う手法である．

PLSI では，意味的な隠れ属性 z_k $(k = 1, 2, \cdots, K)$ のもとで文書 d_i $(i = 1, 2, \cdots, N)$ と索引語 w_j $(j = 1, 2, \cdots, M)$ の生起は独立であるとする．

このとき，d_i と w_j の同時確率は

$$P(d_i, w_j) = \sum_k P(z_k)P(d_i|z_k)P(w_j|z_k) \qquad (1)$$

として与えられる．ここで，文書 d_i における索引語 w_j の実際の出現回数を $n(d_i, w_j)$ とすると，次式の対数尤度

$$L = \sum_i \sum_j n(d_i, w_j) \log P(d_i, w_j) \qquad (2)$$

を最大化するような $P(z_k), P(d_i|z_k), P(w_j|z_k)$ を EM アルゴリズムを用いて推定する．

[E-step]

$$P(z_k|d_i, w_j) = \frac{P(z_k)P(d_i|z_k)P(w_j|z_k)}{\sum_{k'} P(z_{k'})P(d_i|z_{k'})P(w_j|z_{k'})}. \qquad (3)$$

[M-step]

$$P(d_i|z_k) = \frac{\sum_i n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{i'} \sum_j n(d_{i'}, w_j)P(z_k|d_{i'}, w_j)}. \qquad (4)$$

$$P(w_j|z_k) = \frac{\sum_i n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_i \sum_{j'} n(d_i, w_{j'})P(z_k|d_i, w_{j'})}. \qquad (5)$$

$$P(z_k) = \frac{\sum_i \sum_j n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_i \sum_j n(d_i, w_j)}. \qquad (6)$$

収束するまで，E-step と M-step の計算を繰り返すが，実際には過学習を避けるため 式 (3) の代わりに次式を使った Temperd EM[1] を用いる．

3 PLSI によるクラスタリング

3.1 文書クラスタリング

PLSI における隠れ属性 z_k $(k = 1, 2, \cdots, K)$ はひとつの概念を表していると捉えることができるため，PLSI モデルを用いて対象文書をクラスタリングすることができる [3]．いま，文書集合を S 個のクラスタにクラスタリングするとする．ただし $S \leq K$ である．

[PLSI によるクラスタリングアルゴリズム]

* 〒 169-8555 東京都新宿区大久保 3-4-1 早稲田大学理工学部経営システム工学科 School of Science and Engineering,Waseda University, 3-4-1 Okubo , Shinjuku-ku, Tokyo, 169-8555 Japan. E-mail: nagao@hirasa.mgmt.waseda.ac.jp
† 〒 169-8050 東京都新宿区西早稲田 1-6-1 早稲田大学メディアネットワークセンター，Media Network Center,Waseda University, Nishi Waseda 1-6-1, Shinjuku-ku, Tokyo, 169-8050 Japan.

1. PLSI を実行し，$P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ を求める．
2. 各文書 d_i を，$P(z_k|d_i)$ が最大となるような z_k に割り当てる．
3. $K = S$ ならば，各 z_k がそれぞれのクラスタとなる．$K > S$ の場合，各 z_k 同士の類似度を次式の余弦尺度で測り，これを尺度として $K = S$ となるまで群平均法のようにクラスタの併合を行う．

$$S(z_k, z_{k'}) = \frac{z_k \cdot z_{k'}}{|z_k||z_{k'}|} \tag{7}$$

ここで，$z_k = (P(w_1|z_k), P(w_2|z_k), \cdots, P(w_M|z_k))$.

3.2 従来手法の問題点

(1) 一般的な語の影響

出現数の大きい単語は，どのカテゴリ・文書にも出現するような一般的な語である場合が多い．一般的な語は文書やカテゴリを特徴付けることができないため，クラスタリングにおいて精度低下の原因となる．ベクトル空間モデルでは，単語の文書頻度を考慮した TF-IDF 法により一般的な語の重みを下げることでこの問題に対処している．しかし，確率空間では TF-IDF 法は意味を持たず，出現数の多い単語の確率は必然的に大きくなる．

PLSI においても，この一般的な語の影響は大きいと考えられる．一般的な語は観測数が大きくなるので，一般的な語が多く出現するような文書は誤った隠れ属性における確率が高くなる可能性がある．そのため，$n(d_i, w_j)$ の値を観測（出現）数ではなく，TF-IDF 値に置き換えることで PLSI の性能は一般に向上することが確かめられる．しかし，この値は尤度にならないという問題がある．

(2) EM アルゴリズムの初期値依存性

PLSI は EM アルゴリズムを用いるため，初期値の近くへ解が収束する性質がある．この際，初期値はランダムに与えられるので，初期値によって PLSI の性能が大きく変わってしまう恐れがある．文書分類の研究において，この初期値依存性を利用して，学習文書から得られたカテゴリごとの単語の出現確率を初期値として与え，PLSI によって分類を行う手法が提案されている [4].

(3) クラスタの併合

従来では，クラスタ間の距離を余弦尺度により測る．EM アルゴリズムの性質上，$P(w_j|z_k)$ の多くにゼロ値が当てられる．このため，非ゼロの $P(w_j|z_k)$ を多く持つ z_k に関係するクラスタ間の類似度が高くなり，このクラスタから順に併合されやすくなる．この際，$P(w_j|z_k)$ は確率値であるため，一般的な語に大きな値が入り，特徴的な語を重視せずに類似度を測る恐れがある．また，$P(w_j|z_k)$ は確率値であるが，単語の重みのように扱われることが多く，確率空間における類似度の測度としては不適切である．

4 提案手法

4.1 EM アルゴリズムの初期値の決定

一定の値を EM アルゴリズムの初期値として与えることで，初期値のランダム性に依存しない安定したクラスタリングが可能になると考えられる．あらかじめ単語のクラスタリングを行い，単語の各クラスタへの帰属確率を $P(w_j|z_k)$ に反映させれば，安定したクラスタリングが期待できる．

単語クラスタリングには，Fuzzy c-mean 法 [5] を用いる．Fuzzy c-mean 法は対象の各クラスタへの帰属度を $[0, 1]$ で表すファジィクラスタリングの一種である．

まず，単語 w_j のベクトル $\boldsymbol{w}_j = (w_{j1}, w_{j2}, \cdots w_{jN})$ を次式で定める．

$$w_{ji} = (1 + \log n_j) \cdot \log(M_n / m_i) \tag{8}$$

ここで，n_j は単語 w_j の出現する文書数，m_i は文書 d_i が持つ単語数，M_n は全単語の総出現数である．

[Fuzzy c-mean 法]

step1) 単語 w_j のクラスタ C_k への帰属度 δ_{jk} の初期値を適当に定める．

step2) 各クラスタの重心を次式で更新する．

$$\boldsymbol{v}_k := \frac{\sum_j \delta_{jk}^p \cdot \boldsymbol{w}_j}{\sum_j \delta_{jk}^p} \tag{9}$$

p はファジィ化パラメータであり，$p = 1.5$ とする．

step3) 帰属度 δ_{jk} を次式で更新する．

$$\delta_{jk} := \left(\sum_{k'} \left(\frac{\mathrm{sim}(\boldsymbol{w}_j, \boldsymbol{v}_{k'})}{\mathrm{sim}(\boldsymbol{w}_j, \boldsymbol{v}_k)} \right)^{1/(p-1)} \right)^{-1} \tag{10}$$

$\mathrm{sim}(,)$ は余弦尺度である．$\mathrm{sim}(\boldsymbol{w}_j, \boldsymbol{v}_k) = 0$ のときは $\delta_{jk} = 0$ と更新する．

step4) クラスタの重心が変化しなくなるまで，step2,3 を繰り返す．

□

すべての単語に対してクラスタリングを行うことは非効率であるといえる．まず，単語は非常に多数であるのに対し，単語ベクトルの要素となる文書は少数である．このため，すべての単語を用いると，精度の低下が起こると考えられる．したがって，すべての単語の中から単語クラスタリングに用いる重要単語を選択する必要がある．本研究では単語の重要度を χ^2 値によって測る．ここで，単語 w_j の χ^2 値は次式で与えられる．

$$\chi^2(w_j) = \sum_{g \in G} \frac{n(w_j, g) - n_{w_j} P(g)}{n_{w_j} P(g)}. \tag{11}$$

ここで，G は頻出語群（出現数上位約 10 語を頻出語群とする），n_{w_j} は単語 w_j と頻出語群 G の総共起数，$n(w_j, g)$ は単語 w_j と頻出語 g の共起数，$P(g)$ は g の生

起確率である．χ^2 値は頻出語との共起を利用した単語の偏りを表す指標である．一般に，χ^2 値の高い単語は特徴的な語となり，低い単語は一般的な語となる．

EM アルゴリズムの初期値の決定方法を以下に示す．

[初期値の決定方法]

(1) 全単語の中から χ^2 値の大きい語を単語クラスタリング対象とする（対象は上位約 5%の単語群 W_c）．

(2) Fuzzy c-mean 法により単語クラスタリングを行い，単語の各クラスタへの帰属度 δ_{jk} を求める．クラスタ数を K（隠れ属性数）と定める．

(3) 次式を EM アルゴリズムの初期値として与える．

$$P(w_j|z_k) = \begin{cases} \frac{\delta_{jk}}{\sum_{j'} P(w_{j'}|z_k)} & (w_j \in W_c) \\ \frac{1/K}{\sum_{j'} P(w_{j'}|z_k)} & (w_j \notin W_c), \end{cases} \quad (12)$$

$$P(z_k) = \frac{1}{K}, \quad (13)$$

$$P(d_i|z_k) = \frac{1}{N}. \quad (14)$$

このように，あらかじめ単語の特徴を考慮した初期値を与えることで，一般的な語の影響を抑えられ，特徴的な単語の影響が大きくなる．これにより望ましい文書クラスタリングが可能になると期待される．

4.2 KL Divergence を利用したクラスタ併合

Kullback-Leibler (KL) Divergence はふたつの分布間の相違を表す尺度である．分布 P と Q の KL Divergence は非対称で非負な値であるが，$\lambda(0 < \lambda < 1)$ を用いることで次式のような Divergence としても表すことができる．ここで $D_{\mathrm{KL}}(\cdot)$ を KL Divergence とする．

$$\begin{aligned} D_\lambda(P \parallel Q) &= \lambda D_{\mathrm{KL}}(P \parallel \lambda P + (1-\lambda)Q) \\ &\quad +(1-\lambda) D_{\mathrm{KL}}(Q \parallel \lambda P + (1-\lambda)Q). \end{aligned} \quad (15)$$

$D_\lambda(\cdot)$ は $\lambda = 0.5$ のとき，対称となる．また，この式はエントロピー $H(\cdot)$ を用いて，次のように書き表わせる．

$$\begin{aligned} D_\lambda(P \parallel Q) &= H(\lambda P + (1-\lambda)Q) \\ &\quad -\lambda H(P) - (1-\lambda)H(Q). \end{aligned} \quad (16)$$

この Divergence をクラスタ間の距離を測る尺度として用いる．この際，クラスタに含まれる文書数を考慮して，λ を定める．次式を用いてクラスタ間の距離を測る．

$$\begin{aligned} D(z_k, z_{k'}) &= \sum_j \{h(\lambda P(w_j|z_k) + (1-\lambda)P(w_j|z_{k'})) \\ &\quad -\lambda h(P(w_j|z_k)) - (1-\lambda)h(P(w_j|z_{k'}))\} \end{aligned} \quad (17)$$

ここで，$\lambda = \frac{|C_{k'}|}{|C_k| + |C_{k'}|}$ とする．また，$h(x) = -x\log x$，$|C_k|$ はクラスタ z_k に含まれる文書数である．λ はクラスタ同士の文書数による重みの役割を果たしている．

5 実験方法

5.1 実験データ

実験データとして 94 年毎日新聞記事 [6] と学生アンケート [3] を用いる．新聞記事には正解となるカテゴリラベルがあらかじめ付与されている．また，学生アンケートは理系と文系のふたつの学科で行われた同一内容の授業改善のためアンケートデータであり，各々の学科を正解ラベルとする．

5.2 評価指標

クラスタリング性能の評価指標として，正解率，F 尺度とエントロピーを用いる．正解率は，全文書中で正しく分類された文書の割合を表す指標である．

C_s は正解集合，$\Pr(\cdot)$ は適合率，$\mathrm{Re}(\cdot)$ は再現率とするとき，F 尺度は次式で定義される．

$$F(z_k, C_s) = \frac{2 \times \Pr(z_k, C_s) \times \mathrm{Re}(z_k, C_s)}{\Pr(z_k, C_s) + \mathrm{Re}(z_k, C_s)}, \quad (18)$$

$$F_c = \sum_k \frac{n_k}{N} \max_s \{F(z_k, C_s)\}. \quad (19)$$

F 尺度はひとつのクラスタの質を評価する指標であり，クラスタ全体を評価している指標ではない．ここで，F 尺度は値が大きい方が良い．

n_{sk} をクラスタ z_k 中のカテゴリ s の正解文書数とする．正解集合 C_s にクラスタ z_k の文書が属する確率 $P(C_s|z_k) = n_{sk}/n_k$ を考えれば，各クラスタの条件付エントロピーは次のように定義される．

$$H(C|z_k) = -\sum_s P(C_s|z_k)\log(P(C_s|z_k)), \quad (20)$$

$$H(C) = \sum_k \frac{n_k}{N} H(C|z_k). \quad (21)$$

クラスタに関する全エントロピー $H(C)$ を，文書数による重み付け平均として表す．エントロピーはクラスタ全体における正解文書の偏りを評価する指標である．ここで，エントロピーは値が小さい方が良い．

5.3 実験結果

(1) 新聞記事に対してカテゴリ数と隠れ属性数を変化させたときの F 尺度とエントロピーの結果を図 1 に示す（1 カテゴリ 50 文書）．図 1 から提案手法は従来手法より F 尺度，エントロピーの両方において精度が向上したことがわかる．

(2) アンケートデータを利用してクラスタの併合方法のみに対して評価を行う [7]．学生アンケートに対して PLSI クラスタリングを行い，理系クラス (125 人) と文系クラス (19 人) に分ける．隠れ属性が 3, 4, $\cdots$ の場合，クラスタ数が 2 つになるまで併合を行い，最終的なクラスタとなる．特に従来手法でうまく併合できなかった場合 [7] に対して，KL Divergence を利用した併合を行い，その

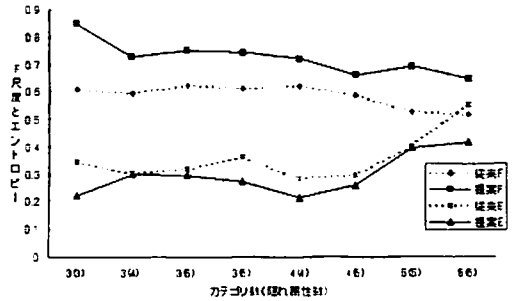

図 1: カテゴリ数・隠れ属性を変化させたときの結果

結果を比較した．図 2 は，隠れ属性を変化させたときの正解率の変化である．新たな併合方法により，従来手法より洗練したクラスタを作り出していることが分かる．

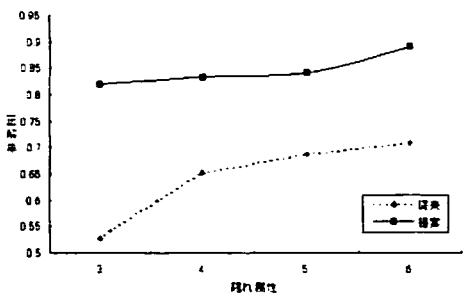

図 2: クラスタ併合方法の比較

5.4 考察

(1) 単語の特徴を利用した初期値設定

提案手法では，EM アルゴリズムの初期値依存性を利用して，単語クラスタリングから求めた単語の特徴を初期値に反映させた．この結果，従来より高い精度でクラスタリングを行うことが可能になった．初期値に偏りもたせることで，EM アルゴリズムにおいて特徴的な単語が注目され，一般的な語の影響を抑え，正しく文書がクラスタリングされたと考えられる．

従来手法では，初期値はランダムに設定されていたので，結果のばらつきが大きい．しかし，提案手法ではあらかじめ初期値を決定するため，ほぼ同じ精度のクラスタリングを行うことができた．単語クラスタリングによってクラスタリングに効果的な初期値を決めることができたといえる．

(2) クラスタの併合方法

提案手法は，KL Divergence を距離尺度としているので確率的に正しい評価をしている．図 2 より，従来手法ではうまく併合できなかった場合に対して提案した距離尺度を用いると適切なクラスタを併合することができることがわかる．精度に関して，従来より質の悪いクラスタを生み出す併合を行うことはなかった．

KL Divergence を対称の形にした Jensen-Shannon Divergence($\lambda = 0.5$ のときの $D_\lambda(\cdot)$) を併合の際の距離尺度として用いても望ましい結果を得ることはできなかった．文書の多いクラスタは一般的な語の出現確率も高いといえ，従来の余弦尺度や Divergence ではそのようなクラスタ同士が併合されやすいと考えられる．そのため，洗練したクラスタが生み出される原因になっていると思われる．よって，文書数を利用して一種の重み付けを行うと，クラスタ併合に適した距離尺度になると考えられる．

6 まとめ

PLSI モデルにおいて，初期値の依存性を利用して，単語の特徴を考慮したクラスタリング手法を提案した．また，PLSI クラスタリングにおける確率的に適切な距離尺度の測定方法を提案した．

今後はクラスタリングにおける最適な隠れ属性数の自動決定法や確率的な単語クラスタリング方法について検討する予定である．また，Divergence と余弦尺度の違いについてもさらに検討する予定である．

7 謝辞

著者の一人である長尾は，本研究を行うにあたり，数多くの助言，ご支援を賜りました早稲田大学平澤研究室の各氏に感謝いたします．本研究の成果の一部は，早稲田大学特定課題研究助成 No.2005B-189 による．

参考文献

[1] T. Hofmann,"Probabilistic Latent Semantic Indexing." *Proc. of SIGIR '99, ACM Press*, pp.50-57. 1999.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman,"Indexing by Latent Semantic Analysis." *Jornal of the Society for Information Science*, 41. pp.391-407, Sep, 1990.

[3] S. Hirasawa, and W. W. Chu,"Knowledge Acquisition from Documents with both Fixed and Free Formats." *Proc. of 2003 IEEE Int. Conf. on System, Man. and Cybernetics*, pp.4694-4699. U.S.A.. Washington DC, Oct. 2003.

[4] 伊藤潤，石田崇，後藤正幸，酒井哲也，平澤茂一，"PLSI を利用した文書からの知識発見，"2003 年 FIT (情報科学技術フォーラム) 論文集, vol.2, pp.83-84. 江別. 2003 年 9 月.

[5] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Wellseparated Clusters." *Journal of Cybernetics*. Mar, pp32-57, 1974.

[6] CD：毎日新聞'94. 毎日新聞社. 日外アソシエーツ.

[7] 長尾壮史. 坂口朋章. 石田崇, 平澤茂一, "PLSI を利用した学生のアンケートからの知識発見－日台学生アンケート分析－，"経営情報学会. 2006 年度秋季全国研究発表大会予稿集, 神戸, 2006 年 11 月.