DEWS2006 5A-i5

共起語を考慮に入れたEMアルゴリズムによるテキスト分類

中山　基† 三浦　孝夫†

† 法政大学 工学部 情報電気電子工学科 〒184–8584 東京都小金井市梶野町 3-7-2
E-mail: †{c02d3076,miurat}@k.hosei.ac.jp

あらまし 本稿では, 少ない訓練データと多くの未分類データからテキストデータを自動分類するため，EM アルゴリズムと語の共起性を融合させる手法を提案する．これにより膨大な未整理データから高精度な分類結果を得ることができる．実験により本手法の有効性と考察を示す.

キーワード EM アルゴリズム，共起語, テキスト分類

Text Classification by Using EM Algorithm and Correlated Words

Motoi NAKAYAMA† and Takao MIURA†

† Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †{c02d3076,miurat}@k.hosei.ac.jp

Abstract In this investigation, we propose a new and novel approach for text classification based on small amount of training data and large amount of unknown data. Our main idea comes from the integration of EM algorithm and correlated words, by which we expect high quality of classification using huge amount of pile data found in internet very often. We examine our approach by means of some experimental results.

Key words EM algorithm, Correlated Words, Text Classification

1. 前　書　き

近年, インターネットを介して膨大な量のテキストデータが電子的に利用可能であり, この傾向は増加をたどる一方である. こういったテキストデータの管理・検索をより高度に行うためには, テキストの分類が欠かせない. しかし, 人手によるテキスト分類は時間と労力を要するのみならず、分類作業が主観的となることから信頼性の確保も重要な問題となる。

これを解決する技術として, 経験的にテキストの分類には教師つき学習が有効であることから，決定木やベイズ分類器などの機械学習技法が研究されている [1], [4]. 教師つき学習では，あらかじめ人手により正確に分類されたデータ（訓練データ）を用いてその性質を分析し，一般的な特性を抽出するという手法がとられる．しかし, 人手による訓練データの生成は時間と労力を要するのみならず、分類作業が主観的となることから信頼性の確保も重要な問題となる。実際問題として, 特定の応用分野を限定した場合でさえ，あらかじめ分類規則（データ) を必要とするための精度の高い訓練データを確保する必要がある。

この解決策としてテキスト自動分類技術が必要とされている. しかも, 少ない訓練データから自動分類を行うことのできる EM アルゴリズムを用いることの有用性は高い。すなわち, 少量の訓練データで未分類データを分類しながら，その未分類データを訓練データとして利用するというステップを繰り返す EM アルゴリズムは, テキスト分類の手法に有効であろう.

反面,EM アルゴリズムの問題も広く知られている [2]．まず, 結果が初期値に大きく依存しがちとなる．また，繰り返し回数をあらかじめ設定できないため，収束が遅くなることが多い. 同時に，収束しすぎると過学習となることも問題である．一旦拡大したエラーは修復されにく，誤りを訂正することはきわめて困難であることも重要である．この問題を改善するためには, 分類器を多重に用意し, 用いるデータの特性に応じて変化させる方法がよいとされる [5].

そこで, 本稿では, 語の共起性を EM アルゴリズムに融合させ, テキスト分類の精度向上のための新たな方式を提案する．通常, テキストデータでは語の同義性や類義性から, いくつかの語が同時に生じる可能性が高い．そのため共起語は, 特に高頻度の場合において強い相関性があると考えることができる．この共起語を分類の際の手がかりとして追加することで, テキスト分類の精度向上をはかる．また, 本稿では語の共起を文単位で考える (つまり，共起する語は複数の文に渡ってつながりを保持するとはみなさない). 一定回数以上の共起頻度を有する語に限定することで計算の効率を高め、効果的なつながりを重視する．EM アルゴリズムを用いたテキスト分類の精度を向上できることを示すため，実験により有効性と考察を示す.

第 2 章で関連研究を示し, 第 3 章で EM アルゴリズムと共起性概念をまとめる．第 4 章では本稿で提案する方式を導入

し，第 5 章で実験によりその有用性を述べる．第 6 章は結論を述べる．

2. 関 連 研 究

文書を分類する方法としては, 上嶋らの研究がある [9]．上嶋らは, ベイズ法での文書分類に, 同義語と多義語を考慮することによって, その分類精度を向上をさせる手法を提案している．通常の文書分類では, 単語の持つ意味は考慮せず, 単語を単に記号的に扱う．しかし, 通常, 文書内には複数意味を持つ単語（多義語）が存在し, 複数の単語が同じ意味を持つ場合（同義語）もある．これらの語を考慮することで, 精度を向上させるのがこの手法である．

本稿では, 共起語を考慮して EM アルゴリズムでの文書分類の精度を向上させる．EM アルゴリズムを用いた文書分類方法としては Nigam らの研究がある [5]．単純ベイズ法に EM アルゴリズムを組み合わせた文書分類を提案しており，旋律の分類・検索 [10] やタイムスタンプの推定 [8] にはきわめて有用であることが知られる．しかし，単純に正規分布に従う確率密度を組み合わせて利用する方法ではきわめて低い分類精度しか達成できず，さまざまに変形を加えた試みを提案している．この変形はテキストデータに生じる特性に基づくものではなく，むしろ確率分布の特性をどう活かすかという視点に立つものが多い．

EM アルゴリズムを単純に適用すると, 精度が逆に悪化することが知られている [11]．つまりある地点まで精度が向上しても, 最終的にはそれよりも低い精度で収束する．場合によってはラベル付き文書のみから生成した分類規則の制度よりも低い精度に収束することもあるため, 本稿では繰り返し回数を 20 回までとした．

3. EM アルゴリズムを用いたベイズ分類

本稿では, トピックをラベルとする訓練データを用いて, 未分類テキストデータを分類するための EM アルゴリズムを示す．本稿では,EM アルゴリズムと単純ベイズ法を組み合わせた手法を用いる．

3.1 単純ベイズ法による文書分類

文書 d をトピック集合 $\mathcal{C} = \{ C_1, \ldots, C_n \}$ に分類する．ベイズ規則による分類とは，文書 d がトピック C に属する確率 $P(C|d)$ の確率分布を求めることである．排他的な分類の場合, 最大事後確率をとるトピック C へ文書 d を分類することで分類のミスを抑えることができる．以下では, トピックラベルを次のように定める．

$$C_k = argmax_{C \in \mathcal{C}} P(C|d)$$

ベイズ規則を次のように定義する．

$$P(C_k|d) = P(C_k) \times \frac{P(d|C_k)}{P(d)} \quad (1)$$

すなわち, ベイズ規則での分類規則生成（訓練）は訓練データ集合から, 確率分布 $P(C_k), P(d), P(d|C_k)$ を推定することである．

しかし, 文書ベクトル d=$(w_1, \ldots, w_m)$ はほぼすべての文書において異なり, $P(d|C_k)$ や $P(d)$ の推定が問題であるため, 一般に, 特徴量 w_j の出現は, 統計的に他の単語出現とは独立であるという仮定をおき, 各文書を単語の集合と考える単純ベイズ法を使うことが多い．単純ベイズ法では $P(d|C_k)$ を以下の形式に分解して考える．

$$P(d|C_k) = \prod_{i=1}^{|d|} P(w_i|C_k)$$

これにより, 文書主導の排他的分類の場合, ベイズ規則は以下のように表すことができる．

$$P(C_k|d) = P(C_k) \times \prod_{i=1}^{|d|} P(w_i|C_k) \quad (2)$$

また, ここでは文書内での単語の出現頻度は考慮せず, 単語の出現有無のみを考えるバイナリ独立モデルを用いる．

［例 1］ ベイズ手法によるクラス分類例を示す．トピック C を,"$C = \{C_1, C_2\}$"とし, それぞれが文書 d_1,d_2 として与えられているとする．($D = \{d_1, d_2\}$)

各トピックの語集合を,

$C_1 = \{a : 3, b : 2, c : 1\}$ $C_2 = \{a : 1, b : 1, c : 3\}$

として, 未分類文書（d_3）を分類する. 数値は頻度である．

単純ベイズ法によりトピックへ分類を行う． 定義より,$P(C_1|d_1) = P(C_2|d_2) = 1$, $P(C_1|d_2) = P(C_2|d_1) = 0$, $P(C_1|d_3) = P(C_2|d_3) = 0$ である．ベイズ規則を用いて,

$$P(C_k|d_3) = P(C_k) \times \frac{P(d_3|C_k)}{P(d)}$$

を最大化するトピック C_k を求める(注1)．$D = \{d_1, d_2, d_3\}$ には 3 件の文書が含まれおり, トピック C_1,C_2 にはそれぞれ 1 件ずつ含まれている．したがって,$P(C_1) = P(C_2) = 1/3$ である．

文書	(a,b,c,d)	a+b+c+d
d1	3,2,1,0	6
d2	1,1,3,0	5
d3	3,3,3,3	12

トピック	$P(a\|*)$	$P(b\|*)$	$P(c\|*)$	$P(d\|*)$
C1	4/7	3/7	2/7	1/7
C2	2/6	2/6	4/6	1/6

各確率は, スムージングを行っている．

単純ベイズ法の仮定から,

$$P(C_k|d_3) = P(a|C_k) \times P(b|C_k) \times P(c|C_k) \times P(d|C_k) \quad (3)$$

ここで $P(C_1|d_3), P(C_2|d_3)$ をそれぞれ求める．

$P(C_1) \times P(d_3|C_1)$

$= P(C_1) \times P(a|C_1) \times P(b|C_1) \times P(c|C_1) \times P(d|C_1)$

(注1)：$P(d) = P(a) \times P(b) \times P(c) \times P(d) = (7/23) \times (6/23) \times (6/23) \times (4/23)$

$$= \frac{1}{3} \times \frac{4}{7} \times \frac{3}{7} \times \frac{2}{7} \times \frac{1}{7} = 0.0033$$

$$P(C_2) \times P(d_3|C_2)$$

$$= P(C_2) \times P(a|C_2) \times P(b|C_2) \times P(c|C_2) \times P(d|C_2)$$

$$= \frac{1}{3} \times \frac{2}{6} \times \frac{2}{6} \times \frac{4}{6} \times \frac{1}{6} = 0.0041$$

これより,$P(C_k|d_3)$ を最大化する C_k は C_2 となり, 文書 d_3 はトピック"C_2"に割り当てられた.

3.2 EM アルゴリズム

EM アルゴリズムとは, データの欠損部分を最尤推定により求め, 欠損部分が分かればその形は単純かつ解析的に表現できるという仮定を置く. Expectation（期待値）,Maximization（最大化）は, それぞれ欠損値の推定, 期待値を得る過程を与えるパラメータの推定に対応している. この E ステップと M ステップを繰り返すことにより, モデルの対数尤度を最大化するパラメータを求める手法である.

EM アルゴリズムを文書分類に適用するために次を用いる.

（1） 入力：ラベル付文書, ラベルなし文書

（2） ラベル付文書のみから単純ベイズ分類規則 $\hat{\theta}$ を生成する

（3） 以下のステップを一定回数, または分類規則が収束するまで繰り返す

(a: E-step) 現在の分類規則 $\hat{\theta}$ を使用し，ラベルなし文書を各トピックへ分類する ($P(c_j|d_i;\hat{\theta})$)

(b: M-step) 推定された事後確率（分類結果）による最尤推定を利用して, 分類規則 $\hat{\theta} = P(D|\theta)P(\theta)$ を再度生成する.

（4） 出力：分類規則 $\hat{\theta}$

本稿では $P(w_i|C_k)$（分類規則）を以下の式で求める.

$$P(w_i|C_k) = \frac{1 + \sum_{j=1}^{|D|} N(w_i, d_j)P(C_k|d_j)}{|V| + \sum_{i=1}^{|V|}\sum_{j=1}^{|D|} N(w_i, d_j)P(C_k|d_j)} \quad (4)$$

ここで D は文書データ全体を表し、w_i はデータ内の各単語を表す. また $N(w_i, d_j)$ は文書 d_j における単語 w_i の発生回数であるが, 本稿では出現の有無により 0 か 1 の値をとる. さらに,$P(C_k|d_j)$ は前述の文書 d_j がトピック C_k に属する確率であり, ラベル付けされたデータに関しては, そのラベル付けられたトピック C_m においては,$P(C_m|d_j) = 1$ であり, それ以外のトピックに対しては 0 をとる. ラベルなしデータに関しては, 最初は全カテゴリに対して 0 であるが, 最初は通常のベイズ分類により, その後は EM アルゴリズムの E-step により, 徐々に適切な値へと更新される. 式 (2) と (3) により EM アルゴリズム内で分類規則を生成する. 同様に $P(C_k)$ は以下のように与えられる.

$$P(C_k) = \frac{1 + \sum_{j=1}^{|D|} P(C_k|d_j)}{|C| + |D|} \quad (5)$$

式 (3),(4) は, それぞれ $P(w_i|C_k), P(C_k)$ のスムージングを行っている.

［例 2］ 例 1 でのトピック推定の結果, 文書 d_3 がトピック C_2 に割り当てられ, $P(C_1) = 1/3$, $P(C_2) = 2/3$ に

トピック	$P(a\|*)$	$P(b\|*)$	$P(c\|*)$	$P(d\|*$
C1	4/7	3/7	2/7	1/7
C2	5/15	5/15	7/15	1/15

変わる. d_3 のトピックが割り当てられた結果，条件確率 $P(a|C_2), P(b|C_2), P(c|C_2)$ が変わる.

また,d_3 のトピック所属確率 $P(C_k|d_3)$ は,

$$P(C_1|d_3) = P(C_1) \times P(d_3|C_1)/P(d)$$

$$= \frac{1}{3} \times \frac{\frac{4}{7} \times \frac{3}{7} \times \frac{2}{7} \times \frac{1}{7}}{\frac{7}{23} \times \frac{6}{23} \times \frac{6}{23} \times \frac{4}{23}}$$

$$P(C_2|d_3) = P(C_2) \times P(d_3|C_2)/P(d)$$

$$= \frac{2}{3} \times \frac{\frac{5}{15} \times \frac{5}{15} \times \frac{7}{15} \times \frac{1}{15}}{\frac{7}{23} \times \frac{6}{23} \times \frac{6}{23} \times \frac{4}{23}}$$

ここで,

$$P(C_k|d_3) = P(C_k) \times \frac{P(d_3|C_k)}{P(d)}$$

を最大化する C_k を求めるため $P(C_k|d_3)$ を計算する.

$$P(C_1) \times P(d_3|C_1)$$

$$= P(C_1) \times P(a|C_1) \times P(b|C_1) \times P(c|C_1) \times P(d|C_1)$$

$$= \frac{1}{3} \times \frac{4}{7} \times \frac{3}{7} \times \frac{2}{7} \times \frac{1}{7} = 0.0033$$

$$P(C_2) \times P(d_3|C_2)$$

$$= P(C_2) \times P(a|C_2) \times P(b|C_2) \times P(c|C_2) \times P(d|C_2)$$

$$= \frac{2}{3} \times \frac{5}{15} \times \frac{5}{15} \times \frac{7}{15} \times \frac{1}{15} = 0.0023$$

この結果、$P(C_k|d_3)$ を最大化するトピック C_k は C_1 であり,"d_3"は再度トピック C_1 に割り当てられる. EM アルゴリズム部で再度計算を行い, 変動が起きなくなるまで計算を繰り返す.

4. 共起語を考慮した EM アルゴリズム

4.1 共 起 語

本稿では, 語の対 (w_i, w_j) の d における共起度 $co(w_i, w_j)$ を次のように定義する(注2)：$co(w_i, w_j) = \sum_{s \in d} |w_i|_s |w_j|_s$

ここで $|x|_s$ は文 s における要素 x の出現回数を表し, x が語の場合には $|x|_s$ は文 s 中の語 x の出現回数を表す. 式 (5) は, ある文 s に出現した語 w_i は s 中のすべての語 w_j と共起しているとみなした共起頻度を表す. すでに述べたように，本稿では語の共起を文単位で考慮し，複数の文にわたる影響を考えない.

4.2 共起語を含む EM アルゴリズム

本章では,EM アルゴリズムを拡張し，単語同士の共起関係を考慮する. 基本的なアイデアは単純である. すなわち，未分類文書 d が, トピック C_k に割り当てられたとき, トピック C_k に属する単語 $w_1, ..., w_n$ のいずれかと, 文書 d 内において共起し,

(注2)：本稿では共起を文単位で考えている. そのため, 共起を 3 単語まで広げる必要がない. つまり,3 単語で共起していれば, 必ずそのうちの 2 単語も共起していることになる. しかし, これは同時に, 熟語を考慮しないということにもなる. KeyGraph [6] では文章に限っていないために"共起語集合"を扱う.

しきい値以上の共起回数を持つ語 x をトピック C_k に追加する．この EM アルゴリズムを共起語で拡張する手法が，実際に分類精度を向上させるものとなるであろうか？一般にテキストデータでは，語の同義性や類義性からいくつかの語が同時に生じる可能性が高い．つまり共起語は，特に高頻度の語の場合は強い相関性があると考えることができ，実際これを手がかりに文書分類の研究がなされている [3], [6].

更に，本稿では追加する共起語の頻度をある一定以上の割合とする．しきい値 5 割で, 文書内の最大共起頻度が 10 回ならば, 追加語は共起頻度が 5 回以上のものとなる．これは, 追加語の共起頻度に制限を設けることによって, 追加語とトピックの語集合との相関の強さを操作するためである．しきい値を高く設定すれば, 相関的な語が追加されるので, より精度が向上するはずである．逆に, しきい値を低くすれば, トピックと相関の低い語が追加される機会が多くなるので, 高く設定した場合よりも精度は悪化すると考える．

［例 3］ 共起語を考慮に入れた EM アルゴリズムによるクラス分類例を示す．

ここでは, 例 1 と同じデータを用いて例を示す．トピック C を,"$C = \{C_1, C_2\}$"とし, それぞれが文書 d_1,d_2 として与えられているとする．（$D = \{d_1, d_2\}$）

各トピックの語集合を,

$C_1 = \{a:3, b:2, c:1\}$ $C_2 = \{a:1, b:1, c:3\}$

として, 未分類文書（d_3）を分類する. 数値は頻度である．

はじめに単純ベイズ法によりトピックへ分類を行う．計算方法および結果は, 例 1 と同様である．

次に推定されたトピックに共起語を追加する．追加する語は, トピックの語集合の単語のいずれかと, 文書 d_3 内において共起している語が対象となる．このとき, 対象となる語がすでにトピックの語集合に含まれている場合は, 追加を行わない．

ここで, 文書 $d_3 = \{a, b, c, d\}$ のとき, 語対 (w_i, w_j) の共起度 $co(w_i, w_j)$ はそれぞれ,

$$co(a, d) = 3$$

$$co(a, b) = co(b, c) = co(b, d) = 2$$

$$co(a, c) = co(c, d) = 1$$

である．しきい値を 80%に設定すると, 追加対象となる共起語は, 追加語の共起度が 最大頻度 × しきい値 $= 3 \times 0.8 = 2.4$ よりも大きいため, 共起度 3 の語のみが対象となる．これより, 追加する語の条件は, トピック $C_2 = \{a, b, c\}$ のいずれかと共起する語, かつ共起度 3 の語となる．d_3 内において, この条件を満たす語は"d"のみであり, これをトピック C_2 の語集合に追加し,$C_2 = \{a, b, c, d\}$ とする．以降, トピック C_2 の語集合を $C_1 = \{a, b, c, d\}$ として分類を行う．

トピック	$P(a\|*)$	$P(b\|*)$	$P(c\|*)$	$P(d\|*)$
C1	4/7	3/7	2/7	1/7
C2	5/18	5/18	7/18	4/18

新たに定まった確率値を用いて, 再度 d_3 の所属確率を求める (EM アルゴリズム).

$$P(C_k|d_3) = P(C_k) \times \frac{P(d_3|C_k)}{P(d)}$$

を最大化するトピック C_k を求めるため $P(C_k|d_3)$ を計算する．

$$P(C_1) \times P(d_3|C_1)$$

$$= P(C_1) \times P(a|C_1) \times P(b|C_1) \times P(c|C_1) \times P(d|C_1)$$

$$= \frac{1}{3} \times \frac{4}{7} \times \frac{3}{7} \times \frac{2}{7} \times \frac{1}{7} = 0.0033$$

$$P(C_2) \times P(d_3|C_2)$$

$$= P(C_2) \times P(a|C_2) \times P(b|C_2) \times P(c|C_2) \times P(d|C_2)$$

$$= \frac{2}{3} \times \frac{5}{18} \times \frac{5}{18} \times \frac{7}{18} \times \frac{4}{18} = 0.0044$$

この結果、$P(C_k|d_3)$ を最大化する C_k は C_2 であり, "d_3"は再度トピック C_2 に割り当てられる．

次に推定されたトピックに共起語を追加する．前述と同様に共起語の追加作業を行うが, ここでは, すでに対象となる語は含まれているため追加作業はない．

最後に，EM アルゴリズム部で再度計算を繰返し, 変動が起きなくなるまで計算を繰り返す．

5. 実　　験

5.1 実験データ

本稿では，シェークスピアによる戯曲 12 作品をテストコーパスとして使用し, タイトルをトピックとして考える．各タイトルはいずれも 5 章 (chapter) から構成され各章は場 (Scene) からなる．実験で用いたタイトルは次の 12 作品である．

（1） 真夏の夜の夢 (全 5 章 9 場)
（2） 終わりよければすべてよし (全 5 章 23 場)
（3） お気に召すまま (全 5 章 22 場)
（4） シンベリン (全 5 章 26 場)
（5） 恋の骨折り損 (全 5 章 10 場)
（6） から騒ぎ (全 5 章 18 場)
（7） ペリクリーズ (全 5 章 20 場)
（8） 間違いの喜劇 (全 5 章 11 場)
（9） ヴェニスの商人 (全 5 章 19 場)
（10） ウィンザーの陽気な女房たち (全 5 章 23 場)
（11） じゃじゃ馬ならし (全 5 章 14 場)
（12） テンペスト (全 5 章 9 場)

初期訓練データを各トピックの第 1 章のすべての場から生成し, 2 章以降の全 168 場をテストデータとして使用する (表 1 参照). それぞれの場のタイトルを隠して分類を行い, 後に適切に分類がされたかを判断する．各場データはあらかじめステミング [7] および不要語処理を行う．また, 初期訓練データに Zipf の法則を適用し, 上位 30 単語をトピック単語とする．各トピックの構成と, ストップワードを除去した後の各場の単語数を表 1 に示す．

共起語の計算は, 各場ごとに行う．本稿では, 共起語を各場内において, 出現頻度が 2%以上の語のみを計算する．これは, 相関性の極めて低い単語を, 共起語として算出しないためである．

5.2 実験手順

提案手法の評価を以下 3 点において行う.

- 精度への影響
- しきい値による正答率の変化
- 多数トピックへの共起語の削除の有効性

提案手法の有効性を示すため，本実験では，EM アルゴリズムによる分類と, 提案手法である共起語を考慮に入れた EM アルゴリズムの正答率の変化を比較することで. 提案手法を検証する. 正答率は次式で定義される.

$$\frac{\text{正しく分類された総文書数}}{\text{総文書数}} \times 100\ (\%) \qquad (6)$$

推定されたトピックと実際のトピックとの比較を行い, 精度への影響を検証する. しきい値を 10%から 100%と変化させて場合と, しきい値を設けなかった場合についても実験を行う. 最後に, 多数トピックへの共起語の削除の有効性を, 同じしきい値で, 削除した場合と, 削除しなかった場合で比較し検証する.

また, 本稿では EM アルゴリズムの繰り返し回数を, 0 回（単純ベイズ部）,5 回,10 回,15 回,20 回の 5 パターンについて示す.

5.3 実験結果

表 2 に, ベイズ手法に EM アルゴリズムを組み合わせた分類の正答率と, 共起語を考慮に入れた EM アルゴリズムの各しきい値に対する正答率を示す. 表 3 に,EM アルゴリズムと, 共起語を考慮に入れた EM アルゴリズムの各実験結果のうち, 最も精度の高かったしきい値との分類結果の違いを示す.

また, 共起語の追加に対する精度の変化過程を評価するため, 代表的な例について考察する. 表 4 に, トピック単位での正答率を示す. 削除を繰り返し, 最終的に残った追加語を表 5 に示し, 削除が行われた語とそのタイミングを表 6 に示す. 分類の変化の例を表 6 に示す. この表 7 について, ○は EM アルゴリズムでの不正解が, 提案手法に置いて正答に変わったデータ. ×は逆に不正解に変わったデータを示す. また, 表 8 にはしきい値による追加語が生じた件数を示し, 表 9 に多数トピックへの追加語の削除についての正答率を示す.

5.4 考　　察

表 2 より明らかなように, 語の共起性を EM アルゴリズムに融合させてテキスト分類を行うことは, 精度を向上させることができる. また, 本稿では, 追加語を一定回数以上の共起頻度を有する語に限定し, より効果的なつながりを重視した. しきい値を 80%にした場合に最も良い精度を得たが, 100%,90%,70%の場合においても,EM(15),EM(20) において,10%以上の精度向上が見られる. 逆に, このしきい値を設けなかった場合,10%と低く設定した場合において, 精度の低下が見られる. 共起を考えるにあたり, 相関の高い語を追加することが有効であることがいえる.

本稿で提案したように, EM アルゴリズムに共起語を考慮に入れて分類を行うことが, 有効的であったことが確認できる.

5.4.1 共起語の影響

本実験において, 誤分類による共起語の追加が正答分類による追加の数を上回ったが, 精度が向上した. 誤分類の追加語の影響より, 正答分類での追加語の影響が強いことが挙げられる.

表 1　データの構成とサイズ

トピック	場	1 章	2 章	3 章	4 章	5 章
1	1	512	146	120	123	219
	2	198	85	242	24	-
2	1	462	123	7	39	16
	2	224	26	69	43	31
	3	482	154	8	167	184
	4	-	21	25	14	-
	5	-	36	49	52	-
	6	-	-	44	-	-
	7	-	-	15	-	-
3	1	322	30	12	89	36
	2	429	35	189	12	49
	3	280	8	46	83	22
	4	-	13	26	-	127
	5	-	111	65	-	-
	6	-	34	-	-	-
	7	-	55	-	-	-
4	1	416	27	37	8	11
	2	78	118	39	240	14
	3	114	68	58	28	52
	4	350	7	115	22	97
	5	227	-	90	-	296
	6	479	-	50	-	-
	7	-	-	9	-	-
5	1	576	137	103	97	76
	2	311	-	140	96	516
	3	-	-	-	243	-
6	1	518	185	122	156	186
	2	69	23	67	47	48
	3	158	126	46	-	18
	4	-	-	86	-	77
	5	-	-	47	-	-
	6	-	-	25	-	-
7	1	110	90	60	53	140
	2	399	31	21	68	58
	3	302	25	23	54	-
	4	92	58	-	3	-
	5	353	122	-	109	-
8	1	376	66	76	71	238
	2	233	108	99	47	-
	3	-	-	-	46	-
	4	-	-	-	109	-
9	1	624	17	68	249	165
	2	346	106	170	13	-
	3	-	11	19	-	-
	4	-	26	39	-	-
	5	-	30	51	-	-
	6	-	42	-	-	-
	7	-	34	-	-	-
	8	-	25	-	-	-
	9	-	52	-	-	-
10	1	444	115	60	40	20
	2	40	148	42	118	7
	3	258	47	115	9	10
	4	304	-	50	47	1
	5	-	-	60	81	137
	6	-	-	-	30	-
11	1	336	128	53	117	84
	2	341	139	131	59	112
	3	-	224	-	103	-
	4	-	-	-	54	-
	5	-	-	-	43	-
12	1	187	231	49	146	200
	2	970	108	87	-	-
	3	-	-	66	-	-

表 2　正　答　率

%	Bayes	EM(5)	EM(10)	EM(15)	EM(20)
Bayes + EM	29.17	27.38	28.57	38.10	38.10

%	しきい値	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
B	100	38.10	30.36	29.76	48.81	48.81
a	90	38.69	27.38	29.76	50.00	50.00
y	80	41.07	29.76	30.95	54.76	54.76
e	70	37.50	26.79	30.95	51.19	51.19
s	60	32.74	21.43	28.57	47.62	47.62
+	50	39.88	27.38	30.36	45.24	45.24
共	40	27.38	25.60	26.79	43.45	43.45
起	30	28.57	24.40	21.43	36.31	36.31
+	20	32.14	22.02	22.62	38.10	38.10
E	10	32.14	23.81	21.43	36.90	36.90
M	0	31.55	23.21	21.43	35.71	35.71

表 3　EM アルゴリズムと共起語を考慮に入れた EM アルゴリズム（しきい値 80%）の正答の違い

		Bayes+EM		繰返し
B	個数	正解	間違い	
a	正解	39	30	0
y	間違い	10	89	
e	正解	29	21	5
s	間違い	6	112	
＋	正解	17	21	10
共	間違い	5	125	
起	正解	59	33	15
＋	間違い	5	71	
E	正解	59	33	20
M	間違い	5	71	

表 4　トピックごとの正答率

しきい値 80 %						
トピック	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)	総場数
1	57.14	0.00	14.29	71.43	71.43	7
2	20	20	25	40	40	20
3	42.11	36.84	36.84	47.37	47.37	19
4	35.00	35.00	35.00	55.00	55.00	20
5	75.00	12.50	25.00	62.50	62.50	8
6	40.00	40.00	40.00	80.00	80.00	15
7	33.33	26.67	26.67	46.67	46.67	15
8	33.33	33.33	44.44	66.67	66.67	9
9	47.06	29.41	29.41	47.06	47.06	17
10	63.16	47.37	47.37	63.16	63.16	19
11	25.00	16.67	0.00	25.00	25.00	12
12	42.86	28.57	28.57	85.71	85.71	7

Bayes+EM						
トピック	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)	総場数
1	42.86	42.86	42.86	71.43	71.43	7
2	20.00	10.00	15.00	25.00	25.00	20
3	21.05	26.32	26.32	26.32	26.32	19
4	25.00	25.00	20.00	35.00	35.00	20
5	25.00	12.50	12.50	25.00	25.00	8
6	33.33	20.00	33.33	53.33	53.33	15
7	13.33	20.00	20.00	26.67	26.67	15
8	33.33	22.22	33.33	44.44	44.44	9
9	35.29	23.53	23.53	29.41	29.41	17
10	52.63	57.89	63.16	68.42	68.42	19
11	8.33	16.67	8.33	8.33	8.33	12
12	57.14	71.43	57.14	71.43	71.43	7

表 5　共起語を考慮した EM アルゴリズム しきい値 80% 追加語

文書	分類先	追加語	繰返し	文書	分類先	追加語	繰返し
1	7	fair	2	71	5	beatric	3
1	3	king	4	72	5	hath	0
2	12	fair	8	72	5	fanci	0
3	3	hear	3	72	10	fanci	2
3	3	honour	3	74	9	light	0
3	6	hear	4	74	6	light	2
29	11	our	2	81	5	ill	2
29	10	our	3	81	9	ill	10
29	8	make	4	84	9	twenti	0
39	11	ay	3	87	10	ey	0
39	11	young	3	87	10	clock	0
39	11	man	3	88	10	sir	4
39	11	mine	3	133	4	bianca	3
39	11	sir	3	139	4	presum	0
39	4	find	8	143	7	blame	2
40	8	call	3	143	3	blame	3
41	10	thee	0	153	5	thee	4
44	11	morrow	4	161	10	brook	0
51	3	great	0	163	10	betrai	0
52	5	write	0	163	10	amaz	0
60	8	man	5	165	11	ann	4
71	9	beatric	2				

表 6　削　除　語

文書	分類先	追加語	繰返し	文書	分類先	追加語	繰返し
12	4	love	11	127	2	hugh	5
23	9	master	4	127	8	husband	11
31	2	hath	11	149	3	art	4
39	11	find	3	149	3	posthumu	4
39	11	good	3	153	6	thy	0
40	8	give	3	156	6	live	3
40	5	desir	4	157	9	thou	2
43	6	fit	5	158	11	brought	3
44	8	morrow	3	160	7	night	4
103	10	wit	5	165	2	mistress	5
124	10	half	4				

これは, トピックの語集合は, トピックへの依存が比較的強い語で構成されていることからくる．文書の構成は, 単語の意味を考慮して書かれている場合が多く, 本来のトピックの語集合との共起と, 誤分類先の語集合の共起は違う意味を持つ．そのため, 誤分類の追加語において, トピック特有の単語が追加されるケースは少なく, 誤分類においては汎用性のある単語が追加されるケースが多い．また, このような追加語は多数トピックの共起となるケースも多く, 本稿で用いた手法における, 削除の対象となる．

しかし, 誤分類は必ずしも精度の悪化を招く要因とはならない．共起が起こるということは相関を持つ語であるということである．本稿では, 第 1 章の単語に Zipf の法則を適用し, その頻出上位 30 単語をトピックの語集合とした．その際,30 単語には入らなかったが, トピックの単語である語も存在する．表 5 より, 文書（60）において, 単語（man）が追加語としてトピック 8 に追加されている．この文書（60）は本来はトピック 3 に分類される文書である．しかし, この単語（man）が文書（145）の中で存在しており, EM アルゴリズムで不正解だった分類が, 本稿の提案手法において正答に移り変わった．このように, 誤分類による追加は必ずしも精度の悪化を招く要因とはならない．

また, 共起語の追加で精度が悪化するケースであるが, 文書（97）において著しく悪化しているケースがある．この文書（97）は本来はトピック 1 に分類される文書であるが, 提案手法において, トピック 5 に分類されることで精度が悪化している．この文書はトピック 1 の 4 章:場 2 に当たる文書であり, 総単語数は 24 である．このような少ない単語数からなる文書は,1 単語のトピック推定に与える影響は他のものに比べて大きくなるのだが, 文書（72）において, この文書（97）にも含まれる単語（hath）がトピック 5 に分類されている．

表 4 より, 各トピックごとについても,EM(0),EM(15),EM(20) でほぼ全てのトピックで正答率の向上が見られる．これは, 直接共起語の追加がおきていないトピックに対しても向上しており, 共起語の追加は, 追加の起きたトピックのみならず, その他のトピックの精度にも影響を与えると見れる．他トピックへの共起語の追加により, 各トピックに対する所属確率にも変動が起き, 精度にも影響が出る．本提案手法である共起語の考慮は, 特定トピックのみでなく, トピック全体での精度向上に役立つといえる．

また, 共起語を考慮に入れた EM アルゴリズムにおいて,EM(15),EM(20) で最も良い精度となっている．繰り返し回数による精度向上の幅に, 途中での追加語が影響していることは考えられるが, EM アルゴリズムの結果においても同様の変化がおきている．そのため,EM(15),EM(20) で最も良い精度となる大きな要因としては, EM アルゴリズムの繰り返し計算により, 正しいトピックに分類されていくことが大きいと考えられる．

これら, 正答分類による追加語と, 誤分類における追加語の影響で, 本提案手法において精度の向上が見られた．

5.4.2　しきい値の影響

本稿では, 実験を行う際に, 追加する共起語の共起頻度にし

きい値を設けた．表 2 より, しきい値を高くし, 相関の強い語を追加することの有効性が確認できる．このときの追加語のしきい値としては, 本実験において,80%の精度が最も良い．これは,80%に設定する事で,100%の場合よりも相関の強い語が多く追加される理由による．また, しきい値を下げると, 依存性の弱い単語, つまりどの文書でも出現してくるような単語が多く追加される．意味のない単語の追加, 各トピックの差が弱まる事で精度が悪化し, また誤分類の追加語が増える事となる．追加件数にあまり差がないのは, 同一文書において, 分類先が変わるたびに同じ単語の追加が起きているためである．

表 7　EM アルゴリズムと共起語を考慮に入れた EM アルゴリズム の分類の違い

文書	繰り返し回数				
	0	5	10	15	20
144	○	-	-	-	-
145	-	-	-	○	○

文書	繰り返し回数				
	0	5	10	15	20
96	-	×	-	-	-
97	-	×	×	×	×

表 8　共起語の追加作業が起きた件数

しきい値	追加総数
100%	79
80%	108

5.4.3　多数トピックへの共起語の追加

本稿では, 多数トピックへの共起語の追加を制限し,3 つ以上のトピックへの追加が起こるケースにおいて, その追加語を追加した全てのトピックから削除している．表 9 より, 本実験において, その手法を用いたことが有効であったことが分かる．

表 9　多数トピックへの共起語の削除評価

Bayes+共起+EM　しきい値= 50% 削除なし					
	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
正答数	54	46	49	74	74
正答率	32.14	27.38	29.17	44.05	44.05

Bayes+共起+EM しきい値= 50% 削除あり					
	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
正答数	67	46	51	76	76
正答率	39.88	27.38	30.36	45.24	45.24

6. 結　　び

本稿では, 共起語を考慮に入れた EM アルゴリズムによるテキスト分類の手法を提案した。テストコーパスを用いた実験によって, EM アルゴリズムに対して, 最も良い精度を得た結果で, EM(0) で 11.9%, EM(5):2.38%, EM(10):2.38%, EM(15):16.66%, EM(20)16.66%と精度が向上したことを確認した．また, 追加語のしきい値を高くし, より効果的なつながりを重視しすることで, しきい値を設けなかった場合, 低く設定した場合よりも高精度の結果を得た．共起を考えるにあたり, 相関の高い語を追加することが有効であり, 本稿によって, 相関的な共起語を考慮に入れることにより, テキスト分類の精度を向上させることができることを確認した．

謝　　辞

本研究の一部は文部科学省科学研究費補助金 (C) (課題番号 16500070) の支援をいただいた．

文　　献

[1] J.Han, et.al : Data Mining: Concepts and Techniques, Morgan Kaufmann Pub., 2000
[2] 岩崎 学: 不完全データの統計解析, エコノミクス社,2002
[3] 松尾, 石塚: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌 17-3-D, pp.217-223, 2002
[4] T.Mitchell: Machine Learning, McGraw-Hill Education, 1997
[5] Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T.M.: Text Classification from Labeled and Unlabeled Documents using EM , Machine Learning Vol .39, No.2, pp. 103–134, 2000
[6] 大澤 幸夫, ネルス E. ベンソン, 谷内田雅彦,: KeyGraph : 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌　D-I Vol.J82-D-I No.2 pp.391-400, 1999
[7] Poter, M.F.:An algorithm for shuffix stripping , Program, Vol. 14, No. 3, pp.130-137,1980
[8] 上嶋 宏, 三浦 孝夫, 塩谷 勇,: Estimating Timestamp From Incomplete News Corpus, Journal of Communications in Information and Systems : Special Issue on Computational Informatics in Data Mining and Information Retrieval, Vo.4, No.4, International Press, pp.273-288 , 2005
[9] 上嶋 宏, 三浦 孝夫, 塩谷 勇,: Improving Text Categorization by Synonym and Polysemy, SYSTEMS AND COMPUTERS IN JAPAN, Vol. 36-4, pp.1-8, 2005 April
[10] 吉原 幸輝, 三浦 孝夫, 塩谷 勇: Classifying Melodies by Using EM Algorithm, IEEE Computer Software and Application Conference (COMPSAC), pp.204-210, 2005
[11] 新納 浩幸, 佐々木 捻.: EM アルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師なし学習，情報処理学会論文誌 Vol.44,No.12,2003