類似テキスト検索のための多重トピックテキストモデル

上田修功[†] 斉藤和巳[†]

本論文では，確率モデルに基づく新たなテキスト検索法を提案する．テキスト検索ではテキスト間の類似度の定義が重要となる．従来法ではテキストの単語頻度ベクトルに基づいた類似度が用いられている為，テキストの内容を十分反映した検索が困難である．提案法では，あるトピック体系で分類されたテキスト群を用いて学習した確率モデルで，テキストのトピック度ベクトルを推定し，トピック度ベクトル空間で類似度が定義される．それゆえ，従来法に比べより内容的に類似したテキスト検索が可能となる．トピック度ベクトルの推定アルゴリズムは単純，かつ解の大域的最適性が理論保証される．また，検索結果に対する妥当な定量的評価基準を新たに導入し，実際の web ページを用いた検索評価実験を通して提案法の従来法に対する顕著な優位性を示す．

Multi-topic Text Model for Topic-based Text Retrieval

Naonori Ueda[†] and Kazumi Saito[†]

We proposed a novel text retrieval method based on probabilistic multi-topic text model. In the text retrieval, the definition of text similarity measure becomes important. In the conventional methods, since the similarity measure based on word-frequency vectors is simply used, it is difficult to retrieve semantically similar documents. In our method, a topic-degree vector for a document can be estimated by using our probabilistic multi-topic text model. Then, the similarity measure between documents is defined in the topic-degree vector space. The estimation algorithm for the topic degree vector is quite simple and theoretically guaranttees the global optimality of the estimate. We also introduce a new measure for quantitatively evaluate text retrieval performance. Through experiments using real web pages, we show that the proposed method could significantly outperform the conventional ones.

1. まえがき

近年，web ページをはじめ膨大なオンライン文書が蓄積されつつある．文書の自動分類技術や検索技術はこうした大量文書を知識源として有効利用するための重要要素技術と位置づけられる．筆者らは，先に，テキストを多重トピック（単一も含む）に自動分類するための確率モデル（パラメトリック混合モデル：PMM）を考案した．トピックとは，音楽，スポーツ，科学等を指す．そして，実際の web ページを用いた多重トピックの自動分類実験で PMM の有効性を確認した[1),2)]．

上記テキスト分類は，ある文書をあらかじめ定めたトピックに多重を許容して分類するタスクであり，それ自身重要なタスクである．一方，ある文書に対しそれと類似した複数の文書を検索するテキスト検索というタスクも応用上重要である．テキスト分類では，同一トピックに属する他の文書との類似性，さらには，異なるトピックにまたがる他の文書との類似性を直接評価できないため，テキスト分類器をテキスト検索にそのまま適用できない．本論文では，テキスト分類のための確率モデル（PMM-C）をテキスト検索のために拡張し（PMM-R），それらを用いて新たなテキスト検索手法を提案する☆．

テキスト検索では文書間の類似度をいかに定義するかが重要となる．これまでテキスト間類似度として，単語頻度ベクトル間のコサイン類似度，もしくは単語の重要度で重み付けしたコサイン類似度が多用されている[3),4)]．すなわち，テキストを単語の集合，より正確なテキスト検索のためには想定語彙集合にわたる単語頻度ベクトルとして表現し，それに基づいてテキスト検索を行っている．しかし，単語頻度に基づく類似度はテキストの内容を直接考慮していないので，必然的な性能限界がある．より正確にはテキストの内容理解に踏み込んだ処理が必要であるが，web のような膨

† NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

☆ PMM-C：分類（categorization）用 PMM，PMM-R：検索（retrieval）用 PMM.

大な文書データベースを想定したテキスト検索の場合，より簡易な手法が望まれる．単語頻度ベクトルが多用されるのはこのような実用上の要請による．

本論文では，単語頻度情報から上記 PMM-C, PMM-R を用いて推定した“トピック”に基づくテキスト間類似度を新たに定義する．PMM はトピックの多重性を表現できるモデルゆえ，単語頻度情報から直接算出される類似度に比べ，よりテキストの内容を反映した類似度といえ，意味的に近いテキスト検索が期待できる．

さらに本論文では，トピックを用いて検索結果を定量的に評価するための尺度（重み付き F 尺度）を新たに導入する．テキスト検索の研究において，客観的な定量評価はきわめて重要であり，提案尺度はその有望な候補の 1 つといえる．Yahoo! データを用いたテキスト検索実験を通して，重み付き F 尺度の観点で，提案法の従来法に対する顕著な優位性を示す．

以下の本論文の構成は以下のとおりである．2 章で従来のテキスト間類似度を説明し，3 章で提案法の概要を説明し，4 章で提案法で必要な PMM-C を説明し，ついで，5 章で本論文の核である PMM-R を詳述する．さらに 6 章で検索評価尺度について述べ，7 章で Yahoo! データを用いた実験結果とその考察について述べる．8 章はまとめである．

2. 従 来 法

従来法では，まず，テキストを単語の集合（bag-of-words）として表現する．すなわち，第 m 文書 d_m を，d_m 中に想定語彙集合

$$\mathcal{V} = \{w_1, \ldots, w_V\}$$

中の単語が何回出現したかを単語頻度ベクトル

$$\boldsymbol{x}_m = (x_{m,1}, \ldots, x_{m,V}) \tag{1}$$

として表現する．$x_{m,i}$ は単語 $w_i \in \mathcal{V}$ が d_m 中に出現した頻度を表す．V は総単語種類数を表す．そして，d_m，d_n 間の類似度として $\boldsymbol{x}_m$，$\boldsymbol{x}_n$ のコサイン類似度：

$$s_{m,n} = \cos(\boldsymbol{x}_m, \boldsymbol{x}_n) \tag{2}$$

を用いる[4]．本論文では便宜上これを COS 法と呼ぶ．

単語の重要度を考慮した重み付きコサイン類似度も提案されている[3]．重みとして，相対的文書頻度の対数の逆数で定義される IDF（Inverse Document Frequency）が用いられる．すなわち，$\boldsymbol{x}_m$ の第 i 要素 $x_{m,i}$ が次式のように重み付けされる．

$$x_{m,i} \leftarrow x_{m,i} \log \frac{T}{T_i} \tag{3}$$

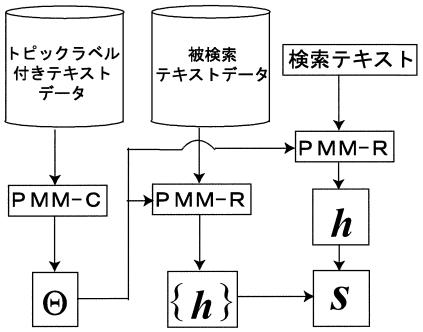

図 1 類似度算出の流れ
Fig. 1 Process of similarity computation.

$T(T_i)$ は被検索対象テキスト☆総数（単語 $w_i \in \mathcal{V}$ が出現したテキスト総数）を表す．これにより，どの文書にも多く出現した単語の重要度が下げられ，特徴的な単語の重要度が相対的に上がり検索精度の向上が期待できる．本論文では便宜上これを IDF 法と呼ぶ．IDF 法での類似度は，式 (2) 中の単語頻度ベクトルを式 (3) で修正すればよい．

3. 提案法の処理の流れ

本章では，提案法の基本的な処理の流れを説明する．

まず，提案法で重要となるトピックラベルベクトル，トピック度ベクトルについて説明する．文書 d に対するトピックラベルベクトルは

$$\boldsymbol{y} = (y_1, \ldots, y_L) \tag{4}$$

なる L 次元 2 値ベクトルで定義される．すなわち，d が第 l トピックに属する（属さない）とき，$y_l = 1(0)$ の値をとる．ただし，L は想定するトピックの総数を表し，本論文では，実験で示すように，数十のトピックを想定している．

一方，トピック度ベクトルとは，帰属するかしないかではなく，その帰属度合いを表すベクトルとして定義される．

$$\boldsymbol{h} = (h_1, \ldots, h_L), \tag{5}$$

ただし，相対的な度合いゆえ，h_l $(l = 1, \ldots, L)$ は $\boldsymbol{y}$ と異なり 2 値ではなく，次の条件を満たす非負の実数値をとる．

$$h_l \geq 0 \quad \text{かつ} \quad \sum_{l=1}^{L} h_l = 1 \tag{6}$$

トピック度ベクトルは，**図 1** に示すように，分類用

☆ 本論文では，ある文書 d_m に類似した文書集合 $\{d_n;\ n = 1, \ldots, N\}$ を検索する際，d_m を“検索文書（あるいは検索テキスト）”と呼び，d_n を“被検索文書（あるいは被検索テキスト）”と呼ぶこととする．

PMM（PMM-C と呼ぶ）と検索用 PMM（PMM-R と呼ぶ）で推定される．すなわち,（多重）トピックの帰属の “有無” を示すトピックラベルベクトルが付与されたテキストデータを学習データとして，PMM-C のモデルパラメータ Θ を学習する．学習済みのパラメータを $\hat{\Theta}$ と書くこととすると，詳細は後述するが，PMM-C は，$\boldsymbol{y}$ で特定されるクラスの単語頻度ベクトル $\boldsymbol{x}$ の確率分布モデル：$P(\boldsymbol{x}|\boldsymbol{y};\hat{\Theta})$ に相当する．

次に，検索テキストに対し，PMM-C で求めたモデルパラメータを用いて，被検索対象テキスト群に対し PMM-R で各テキストのトピック度ベクトル集合 $\{\boldsymbol{h}\}$ を推定する．推定済みのトピック度ベクトル集合を $\{\hat{\boldsymbol{h}}\}$ と書くこととすると，詳細は後述するが，PMM-R は $\hat{\boldsymbol{h}}$ で特定されるクラスの単語頻度ベクトル $\boldsymbol{x}$ の確率分布モデル：$P(\boldsymbol{x}|\hat{\boldsymbol{h}};\hat{\Theta})$ に相当する．検索テキスト (d_m) のトピック度ベクトル $(\hat{\boldsymbol{h}}_m)$，および被検索テキスト（$d_n,\ n=1,\ldots,T$，T は被検索対象テキスト総数）のトピック度ベクトル $(\hat{\boldsymbol{h}}_n,\ n=1,\ldots,T)$ が推定されれば，トピック度ベクトル間類似度を

$$s_{m,n}=\cos(\hat{\boldsymbol{h}}_m,\hat{\boldsymbol{h}}_n),\ n=1,\ldots,T \qquad (7)$$

として算出し，$s_{m,n}$ の値の大きな先頭 N $(\leq T)$ 個のテキストを検索結果として出力する．N はユーザが設定するものとする．

以上が提案法の処理の流れである．単語頻度情報を用いるという点では従来法と共通するが，単語頻度情報を確率モデルを用いてトピック度ベクトルに変換し，トピック度ベクトル空間で類似度を定義している点で従来法と大きく異なる．ただし，提案法ではトピックラベル付き学習データが必要となるが，現状のディレクトリ型の検索サイト（Yahoo! 等）で比較的容易に入手可能であり，実用上の問題点とはならないといえる．

また，提案法では，異なるトピック体系の学習データを用いることにより，異なる観点で類似度が定義できるという特長がある．たとえば，医学関係のテキスト中でさらに細かく検索したい場合，PMM-C を医学に関するトピック体系（遺伝医学，血液学，解剖学等）で学習しておけば，類似度の精度向上が期待できる．このような柔軟な検索は，提案法のような確率モデルによる学習型検索手法により初めて実現できるといえる．

次章以降で，PMM-C，PMM-R の詳細を説明する．

4. PMM-C

PMM-C は筆者らがテキスト分類のために考案した確率モデルであり[1),2)]，前述したように，クラス $\boldsymbol{y}$ に属するテキストの単語頻度分布のモデルである．

4.1 モデル

テキストを単語の集合で表現するということは，単一トピック文書の場合，第 l トピックに属す $\boldsymbol{x}$ は確率分布：

$$p(\boldsymbol{x}|l)\propto\prod_{i=1}^{V}\{p(w_i|l)\}^{x_i}=\prod_{i=1}^{V}\{\theta_{l,i}\}^{x_i} \qquad (8)$$

からの実現値と見なせる．ここに，$\theta_{l,i}$ は w_i が第 l トピックの文書に出現する確率を表す．PMM は多重トピック文書に拡張したモデルで次式で定義される．

$$p(\boldsymbol{x}|\boldsymbol{y})\propto\prod_{i=1}^{V}(p(w_i|\boldsymbol{y}))^{x_i}=\prod_{i=1}^{V}(\varphi_i(\boldsymbol{y}))^{x_i} \qquad (9)$$

$\varphi_i(\boldsymbol{y})$ は（多重）トピックに属する文書中に w_i が出現する確率を表す．L トピックの場合，可能な多重トピックのクラス総数は 2^L-1 となり，個々の $\boldsymbol{y}$ ごとに多項分布を設定するのは非現実的となるので，効率的なモデル化が必要となる．

一方，筆者らは，実際の大量の web ページを観察することにより，“多重トピックに属する文書中の単語は，関連する単一トピックに特徴的な単語の混合から成る” という知見を得た．たとえば，“スポーツ” と “音楽” の両方に属する文書中には主に両者に関連する単語が含まれると考えるのは自然で，上記知見は直観的にも妥当である．PMM-C はこの知見に基づいて効率的なモデル化を実現している．

今，$\boldsymbol{\theta}_l=(\theta_{l,1},\ldots,\theta_{l,V})$ とし，さらに，$\Phi(\boldsymbol{y})$ を確率ベクトル

$$\Phi(\boldsymbol{y})=(\varphi_1(\boldsymbol{y}),\ldots,\varphi_V(\boldsymbol{y}))$$

とする．$\varphi_i(\boldsymbol{y})$ はクラス $\boldsymbol{y}$ に属する文書中に単語 w_i が出現する確率を表す．このとき，上記知見は $\Phi(\boldsymbol{y})$ が次式に示すパラメトリック混合：

$$\Phi(\boldsymbol{y})=\sum_{l=1}^{L}h_l(\boldsymbol{y})\boldsymbol{\theta}_l$$
$$\text{ただし } y_l=0 \text{ なる } l \text{ に対しては } h_l(\boldsymbol{y})=0 \qquad (10)$$

として表現できることを意味する．$h_l(\boldsymbol{y})(\geq 0)$ は混合比に相当する．式 (9)，(10) より，PMM-C は次式で定義される．

$$p(\boldsymbol{x}|\boldsymbol{y})\propto\prod_{i=1}^{V}\left(\sum_{l=1}^{L}h_l(\boldsymbol{y})\theta_{l,i}\right)^{x_i} \qquad (11)$$

PMM-C はすべての可能な多重トピッククラスを L 個のパラメータベクトルで完備かつ効率的に表現されている．

さらに，トピック度ベクトルの要素 h_l を，$\boldsymbol{y}$ の関連トピック（$y_l = 1$ なるトピック l）で均等とした次式で定義する．

$$h_l(\boldsymbol{y}) = \frac{y_l}{\sum_{l=1}^{L} y_l} \tag{12}$$

もちろん，多重トピック文書はある特定のトピックにより強く帰属することも考えられるが，その多重トピッククラスの文書全体で平均すると，それらの偏りがなくなり均等重みが妥当といえる．実際，この重み自身も学習するモデルも考察したが，モデルの自由度が上がることにともなう過学習のために汎化性能が低下するという問題があり，実用上は上記均等重みで十分であることを確認している[1]．

4.2 PMM の学習と未知トピックラベルベクトルの推定

トピックが既知の文書集合 $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}$ が与えられた際，未知パラメータ $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$ は事後分布 $p(\Theta|\mathcal{D})$ を最大化する Θ として推定される．この最大化問題は解析的には解けず，逐次反復法で解く．PMM-C のパラメータ推定アルゴリズムの導出は文献 1) で記載済みゆえ，かつ，本論文の本質ではないので結果のみを以下に示す．

第 t ステップでの推定値を $\Theta^{(t)}$ とすると，PMM-C のパラメータ更新式は次式で与えられる．

$$\theta_{li}^{(t+1)} = \frac{\sum_{n=1}^{N} x_{n,i} g_{l,i}^{n}(\Theta^{(t)}) + \xi - 1}{\sum_{i=1}^{V}\sum_{n=1}^{N} x_{n,i} g_{l,i}^{n}(\Theta^{(t)}) + V(\xi - 1)} \tag{13}$$

ただし，

$$g_{l,i}^{n}(\Theta) = \frac{h_l(\boldsymbol{y}_n)\theta_{l,i}}{\sum_l h_l(\boldsymbol{y}_n)\theta_{l,i}} \tag{14}$$

とする．また，ξ はハイパーパラメータで通常 $\xi = 2$（ラプラス平滑化）とする．以下の重要な定理が成り立つ．

定理 1：$\hat{\Theta}$ の大域的最適性[2]

PMM のパラメータ推定アルゴリズム式 (13) は，初期値のいかんにかかわらず大域的最適解に収束する．（証明）文献 2) 参照．

$\hat{\Theta}$ が得られれば，トピックが未知のテスト文書 $\boldsymbol{x}_*$ のトピックラベルベクトル $\boldsymbol{y}_*$ は事後分布 $p(\boldsymbol{y}|\boldsymbol{x}^*, \hat{\Theta}) \propto p(\boldsymbol{x}_*|\boldsymbol{y}, \hat{\Theta})p(\boldsymbol{y})$ を最大化する $\boldsymbol{y}$ として求める．ところがこの最大化問題は 0/1 整数計画問題ゆえ，最適解は得られず greedy アルゴリズムにより近似的に解く[2]．

なお，“yahoo.com” からリンクが張られる実際の www ページの多重トピック分類問題を作成し，PMM の評価実験を行った結果，従来の最も分類性能が高い手法[5] に対する優位性を確認している[1]．

5. PMM-R

本論文の核となる PMM-R について詳述する．

5.1 モデル

PMM-C をあるトピック体系（トピック数 L）の文書群であらかじめ学習し，式 (13) に従ってモデルパラメータ Θ を求めておく（$\hat{\Theta}$ とする）．テキスト分類では，式 (12) のようにトピック度ベクトル $\boldsymbol{h}$ をトピックラベルベクトル $\boldsymbol{y}$ の関数としたが，テキスト検索では $\boldsymbol{y}$ は未知ゆえ，h_l を $h_l \geq 0, \sum_l h_l = 1$ を満たすパラメータと見なす．このとき，式 (11) は

$$P(\boldsymbol{x}|\boldsymbol{h}) \propto \prod_{i=1}^{V} \left(\sum_{l=1}^{L} h_l \hat{\theta}_{l,i} \right)^{x_i} \tag{15}$$

と書き換えられる．式 (15) はトピック度ベクトルが $\boldsymbol{h}$ である文書中の単語頻度ベクトル $\boldsymbol{x}$ の確率分布を表していることになる．ゆえに，ある文書 $(\boldsymbol{x})$ の最適トピックベクトルはベイズ則

$$p(\boldsymbol{h}|\boldsymbol{x}) \propto P(\boldsymbol{x}|\boldsymbol{h})p(\boldsymbol{h})$$

に注意して，次式により推定できる．

$$\begin{aligned} \hat{\boldsymbol{h}} &= \arg\max_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{x}) \\ &= \arg\max_{\boldsymbol{h}} \{\log P(\boldsymbol{x}|\boldsymbol{h}) + \log p(\boldsymbol{h})\} \end{aligned} \tag{16}$$

すなわち，提案法では，V 次元の単語頻度ベクトル空間上の文書が PMM-R によりトピック的に近い文書がより近くに配置されるよう $L(\ll V)$ 次元のトピック空間に非線形写像される．

5.2 トピック度ベクトル推定アルゴリズム

$\boldsymbol{h}$ の事前分布としてディレクレ分布：

$$p(\boldsymbol{h}) \propto \prod_{l=1}^{L} h_l^{\lambda - 1} \tag{17}$$

（λ はパラメータで通常 $\lambda = 2$ とするラプラス平滑化と用いる）を仮定すると，$\boldsymbol{x}$ で表現されたテキストの最適トピック度ベクトルは，式 (16) より，制約条件 $\sum_{l=1}^{L} h_l = 1$ の下で次の目的関数を最大化ならしめる $\boldsymbol{h}$ となる．

$$J(\boldsymbol{h}; \boldsymbol{x}) = \sum_{i=1}^{V} x_i \log \sum_{l=1}^{L} h_l \hat{\theta}_{l,i} + (\lambda - 1) \sum_{l=1}^{L} \log h_l \tag{18}$$

この最大化問題は PMM-C の Θ の推定問題と双対

構造をなす．すなわち，PMM-C のパラメータ推定では，$\boldsymbol{h}$ を既知，Θ を未知としていたが，PMM-R では，Θ を既知として $\boldsymbol{h}$ に関する最大化問題に置き換わり，さらに，制約条件を $\sum_{l=1}^{L} h_l = 1$ と置き換わったものと見なせる．上記最大化問題に対し，ニュートン法等の非線形最適化法も適用できるが，以下に，より効率的な逐次反復式を導出する．

便宜上，

$$g_{l,i}(\boldsymbol{h}) = \frac{h_l \hat{\theta}_{l,i}}{\sum_l h_l \hat{\theta}_{l,i}} \tag{19}$$

と置く．また，$\boldsymbol{h}$ の第 t ステップでの推定値を $\boldsymbol{h}^{(t)}$ と書き，式 (18) の右辺第 1 項を $\mathcal{L}(\boldsymbol{h};\boldsymbol{x})$ と置く．$\sum_l g_{l,i}(\boldsymbol{h}^{(t)}) = 1$ に注意すると，$\mathcal{L}(\boldsymbol{h};\boldsymbol{x})$ は次式のように変形できる．

$$\begin{aligned}\mathcal{L}(\boldsymbol{h};\boldsymbol{x}) &= \sum_{i=1}^{V} x_i \left\{ \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \right\} \\ &\quad \times \log \left\{ \left(\frac{h_l \hat{\theta}_{l,i}}{h_l \hat{\theta}_{l,i}} \right) \sum_l h_l \hat{\theta}_{l,i} \right\} \\ &= \mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) - \mathcal{U}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) \end{aligned} \tag{20}$$

ただし，

$$\mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) = \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \log h_l \hat{\theta}_{l,i} \tag{21}$$

$$\mathcal{U}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) = \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \log g_{l,i}(\boldsymbol{h}) \tag{22}$$

とする．ここで，

$$\begin{aligned}&\mathcal{U}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) - \mathcal{U}(\boldsymbol{h}^{(t)}|\boldsymbol{h}^{(t)}) \\ &= \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \log \frac{g_{l,i}(\boldsymbol{h})}{g_{l,i}(\boldsymbol{h}^{(t)})} \\ &\le \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \left(\frac{g_{l,i}(\boldsymbol{h})}{g_{l,i}(\boldsymbol{h}^{(t)})} - 1 \right) \\ &= \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}) - \sum_i x_i \sum_l g_{l,i}(\boldsymbol{h}^{(t)}) \\ &= \sum_i x_i - \sum_i x_i = 0 \end{aligned} \tag{23}$$

すなわち，$\mathcal{U}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) \le \mathcal{U}(\boldsymbol{h}^{(t)}|\boldsymbol{h}^{(t)})$ が成り立つ．途中の不等式は $\log x \le x - 1$ の関係を用いていることに注意．

一方，式 (20) より，

$$\begin{aligned}&\mathcal{L}(\boldsymbol{h};\boldsymbol{x}) - \mathcal{L}(\boldsymbol{h}^{(t)};\boldsymbol{x}) \\ &= \mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) - \mathcal{F}(\boldsymbol{h}^{(t)}|\boldsymbol{h}^{(t)}) \\ &\quad -(\mathcal{U}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) - \mathcal{U}(\boldsymbol{h}^{(t)}|\boldsymbol{h}^{(t)})) \end{aligned} \tag{24}$$

が成り立つ．ゆえに，$\mathcal{L}(\boldsymbol{h};\boldsymbol{x}) \ge \mathcal{L}(\boldsymbol{h}^{(t)};\boldsymbol{x})$ となるには，$\mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) \ge \mathcal{F}(\boldsymbol{h}^{(t)}|\boldsymbol{h}^{(t)})$ であればよい．

したがって，$\mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)})$ を制約条件 $\sum_l h_l = 1$ の下で $\boldsymbol{h}$ に関して最大化すれば $\mathcal{L}(\boldsymbol{h};\boldsymbol{x})$ を増大させることができる．すなわち，$\boldsymbol{h}^{(t+1)}$ は，

$$\mathcal{F}(\boldsymbol{h}|\boldsymbol{h}^{(t)}) + (\lambda - 1) \sum_{l=1}^{L} \log h_l$$

を制約条件 $\sum_l h_l = 1$ の下で $\boldsymbol{h}$ に関して最大化することにより求まる．ラグランジュ乗数法より，

$$h_l^{(t+1)} = \frac{\sum_{i=1}^{V} x_i g_{l,i}(\boldsymbol{h}^{(t)}) + \lambda - 1}{\sum_{i=1}^{V} \sum_{l=1}^{L} x_i g_{l,i}(\boldsymbol{h}^{(t)}) + V(\lambda - 1)} \tag{25}$$

を得る☆．以下の重要な定理が成り立つ．

定理 2：$\hat{h}$ の大域的最適性

$\boldsymbol{h}$ のパラメータ推定アルゴリズム式 (25) は，初期値のいかんにかかわらず大域的最適解に収束する．

（**証明**）集合 Ω を $\Omega = \{\boldsymbol{h} :\ h_l \ge 0,\ \sum_l h_l = 1\ \forall l\}$ と定義し，$\boldsymbol{h}_1 \in \Omega$, $\boldsymbol{h}_2 \in \Omega$, $\alpha \in [0,1]$ とする．

$\boldsymbol{h}_1 = (\phi_1, \ldots, \phi_L)$

$\boldsymbol{h}_2 = (\psi_1, \ldots, \psi_L)$

とすると，$\alpha\boldsymbol{h}_1 + (1-\alpha)\boldsymbol{h}_2 \in \Omega$ が成り立つので，Ω は凸集合となる．次に，式 (18) の PMM-R の目的関数 $J(\boldsymbol{h})$ が厳密に上に凸であることを示す☆☆．

$$\begin{aligned}&J(\alpha\boldsymbol{h}_1 + (1-\alpha)\boldsymbol{h}_2) \\ &= \sum_i x_i \log \sum_l (\alpha\phi_l + (1-\alpha)\psi_l)\hat{\theta}_{l,i} \\ &\quad +(\lambda - 1) \sum_l \log(\alpha\phi_l + (1-\alpha)\psi_l) \end{aligned} \tag{26}$$

となる．ここで，Jensen の不等式を用いて，

式 (26) の右辺

$$\begin{aligned}&\ge \sum_i x_i \Big\{ \sum_l \alpha\hat{\theta}_{l,i} \log \phi_l \\ &\quad + \sum_l (1-\alpha)\hat{\theta}_{l,i} \log \psi_l \Big\} \\ &\quad +(\lambda - 1)\Big\{ \alpha \sum_l \log \phi_l + (1-\alpha) \sum_l \log \psi_l \Big\} \\ &= \alpha J(\boldsymbol{h}_1) + (1-\alpha) J(\boldsymbol{h}_2) \end{aligned} \tag{27}$$

を得る．すなわち，

$$J(\alpha\boldsymbol{h}_1 + (1-\alpha)\boldsymbol{h}_2) \ge \alpha J(\boldsymbol{h}_1) + (1-\alpha) J(\boldsymbol{h}_2)$$

☆ PMM-C のときと同様，実験では $\lambda = 2$（ラプラス平滑化）を用いている．

☆☆ 表記を簡単にするため以下では $J(\boldsymbol{h};\boldsymbol{x})$ を $J(\boldsymbol{h})$ としていることに注意．

となり，凸の定義から J は Ω 上で，厳密に上に凸の関数となる．したがって，J の $\boldsymbol{h} \in \Omega$ に関する最大化問題は凸計画問題となり，唯一の大域的最適解を持つ．さらに，式 (25) の反復において Ω 上で J の値の単調増加性および大域的収束性が保証される．

6. 評価尺度

検索結果の評価は，従来，ベンチマーク用に人為的に作られたデータか，もしくは人間による定性的な主観評価が主流である．ここでは，トピック体系（Yahoo! のようなディレクトリ型の検索サイト）を用いた客観（定量）評価のための新たな評価尺度を導入する．以下に述べる評価尺度はトピック体系が既知であることを前提にしているので，いうまでもなく，現実のあらゆるテキスト検索課題でつねに適用可能で方法論ではない．以下の評価方法は，あくまで，提案したテキスト検索手法を客観的に評価するための手法の一提案として位置づけている．

提案評価尺度は，文書間の類似度，および，トピックラベルベクトルの一致度に基づく．検索文書 d_m と被検索文書 d_n 間の類似度 $s_{m,n}$ は，従来法では式 (2)，提案法では式 (7) のコサイン類似度とする．また，2 つのトピックラベルベクトル間の一致度は，情報検索の分野で通常用いられる F 尺度を用いる[4]．

F-尺度は “適合率（precision：P）” と “再現率（recall：R）” との調和平均：

$$F = \frac{2PR}{P+R} \times 100\ (\%) \tag{28}$$

で定義される☆．説明の便宜上，今，真のトピックラベルベクトルを

$$\boldsymbol{y}^* = (y_1^*, \ldots, y_L^*) \tag{29}$$

とし，予測したトピックラベルベクトルを

$$\boldsymbol{y} = (y_1, \ldots, y_L) \tag{30}$$

とすると，適合率は，予測したトピックのうち，真のトピックと適合した比率で定義される．トピックラベルベクトルの各要素は 0 または 1 ゆえ，適合数が $\sum_{l=1}^{L} y_l y_l^*$ で算出できることに注意すると，適合率は

$$P = \frac{\sum_{l=1}^{L} y_l y_l^*}{\sum_{l=1}^{L} y_l} \tag{31}$$

として計算できる．また，再現率は，真のトピックのうち，予測で正しく再現されたトピック数の比率で定義される．すなわち，

$$R = \frac{\sum_{l=1}^{L} y_l y_l^*}{\sum_{l=1}^{L} y_l^*} \tag{32}$$

として算出される．F 尺度は単にトピックラベルベクトルの要素ごとに 1/0 の一致度で評価する正答率に比べ，0 の一致度を評価しないという点でより厳しい評価尺度であることに注意．すなわち，過不足なく 1 を予測するほど，F 尺度が高くなる．明らかに，F の最小値（最大値）は 0%（100%）となる．

これをテキスト検索の評価尺度として拡張する．検索文書 d_m のトピックラベルベクトル $\boldsymbol{y}_m$ と，被検索文書 d_n のトピックラベルベクトル $\boldsymbol{y}_n$ の一致度を，上記で $\boldsymbol{y}^*$ を $\boldsymbol{y}_m$ とし，$\boldsymbol{y}$ を $\boldsymbol{y}_n$ とすることにより求める☆☆．ただし，分類問題では d_m の帰属するトピックラベルベクトルが分かればよいが，テキスト検索では類似した複数の文書 $\{d_n;\ n = 1, \ldots, N\}$ を求めるのが一般的であり，N が大きくなっても類似した文書が得られているかが重要となる．これがテキスト分類と本質的に異なる点である．そこで，上記 F 尺度を類似度で重み付けした重み付き F 値を

$$\bar{F}_m(N) = \frac{\sum_{n=1}^{N} s_{m,n} F(\boldsymbol{y}_m, \boldsymbol{y}_n)}{\sum_{n=1}^{N} s_{m,n}} \tag{33}$$

で定義し，N 文書にわたる平均的なトピックラベルベクトルの一致度でテキスト検索の客観的な定量評価尺度とする．式 (33) より，N 一定の下では，文書 d_m，d_n 間の類似度 $s_{m,n}$ と，対応するトピックラベル $\boldsymbol{y}_m$，$\boldsymbol{y}_n$ 間の一致度の相関が高いほど，$\bar{F}_m(N)$ が大きくなり直観的に妥当な尺度となっているといえる．

以上は，ある 1 つの検索文書に対する評価値ゆえ，これを全検索文書（総数 M とする）の各々について算出し，それらの平均値：

$$\bar{F}(N) = \frac{1}{M} \sum_{m=1}^{M} \bar{F}_m(N) \tag{34}$$

を求め，全検索文書にわたる検索性能とする．

7. 実験と考察

7.1 実験方法

“Yahoo.com” ドメインのトップ 11 トピック（Arts & Humanities, Business & Economy, Computer & Internet, Education, Entertainment, Health,

☆ 正確には重み付き調和平均であるが，通常，適合率と再現率を対等とする一様重みが用いられる．

☆☆ 式 (28) の定義式より明らかなように，F-尺度は適合率 P と再現率 R に関して対称なので，$\boldsymbol{y}^*$ と $\boldsymbol{y}$ を交換しても F-尺度は不変であることに注意．すなわち，$\boldsymbol{y}_m$ を真のラベル，$\boldsymbol{y}_n$ を推定ラベルとしているのは便宜上であって本質ではない．

表 1 11 データセットでのテキスト検索法（COS，IDF，PMM）によるトップ N 検索文書での F 値（%）比較．すべての問題に対して提案法（PMM）が他の性能を顕著に上回っていることが確認できる

Table 1 Comparison of F values (%) of top N retrieved documents obtained by text retrieval methods (COS, IDF, PMM). One can see that the proposed method (PMM) overcame the other method significantly for all problems.

	Arts & Humanities			Business & Economy			Computers & Internet			Education		
N	COS	IDF	PMM	COS	IDF	PMM	COS	IDF	PMM	COS	IDF	PMM
1	40.0	42.4	45.0	71.8	73.8	76.0	51.1	53.2	54.1	42.9	44.3	45.3
5	35.9	39.4	44.2	69.8	71.8	75.5	46.3	49.4	53.2	38.4	41.1	43.9
10	33.3	37.6	43.7	68.6	70.6	75.2	43.9	47.2	52.5	35.7	38.8	43.1
50	27.0	31.5	41.5	65.0	66.7	73.9	38.7	41.7	50.5	28.8	32.0	40.7
100	24.3	28.5	39.8	63.6	64.9	72.9	36.7	39.5	49.2	25.9	28.8	39.1

	Entertainment			Health			Recreation & Sports			Reference		
N	COS	IDF	PMM	COS	IDF	PMM	COS	IDF	PMM	COS	IDF	PMM
1	47.5	50.8	56.5	58.3	60.4	63.8	44.4	50.0	53.8	53.3	55.2	58.3
5	43.7	47.6	55.7	54.5	57.5	62.6	39.8	46.2	52.8	49.1	52.9	57.1
10	41.2	45.5	54.9	52.4	55.6	61.7	37.1	43.7	52.1	46.7	50.5	56.6
50	34.8	39.4	52.6	45.7	49.8	58.9	29.5	36.0	49.5	39.9	43.3	54.4
100	31.9	36.5	51.3	42.3	46.5	57.0	25.8	32.0	47.9	37.1	40.0	51.9

	Science			Social & Science			Society & Culture			**Average**		
N	COS	IDF	PMM	COS	IDF	PMM	COS	IDF	PMM	**COS**	**IDF**	**PMM**
1	43.8	48.4	48.3	59.4	58.9	62.7	44.7	46.8	50.2	**50.7**	**53.1**	**55.8**
5	37.9	43.7	46.3	56.5	57.4	61.5	41.5	43.8	49.2	**46.7**	**50.1**	**54.7**
10	34.4	40.6	45.1	54.7	56.2	60.9	39.3	42.2	48.4	**44.3**	**48.0**	**54.0**
50	26.1	32.2	41.5	49.6	51.7	59.0	33.1	36.4	46.0	**38.0**	**41.9**	**51.7**
100	22.8	28.2	39.4	46.8	49.2	57.8	30.7	33.8	44.6	**35.3**	**38.9**	**50.1**

Recreation & Sports，Reference，Science，Social & Science，Society & Culture）から GNU-Wget により文書を収集し，COS 法，IDF 法および提案法（PMM と呼ぶ）による検索実験を行った．

3,000 の検索文書と 2,000 の被検索対象文書の組を上記 11 トピックごとに各々5 セット作成した．ここでは 11 個の独立した検索問題が構成され，トピックは各問題ごとに異なることに注意．したがって，提案法の場合は，11 問題の各々の直下の数十のサブトピック配下の文書で PMM-C のパラメータ Θ をあらかじめ学習させている．サブトピック数（トピックラベルベクトルの次元数 L に相当）は，11 問題で 20 から 40 に分布し，また，対象文書中の有効単語数☆の平均値は 100 単語から 180 単語に分布していた．

各問題ごとに，3,000 の検索文書の各々に対し検索希望出力件数 (N) を変えながら被検索対象文書中で検索を行い，式 (34) を算出して検索性能を比較評価した．

7.2 結果と考察

$\bar{F}(N)$ の値の 5 セットにわたる平均値を**表 1** に整理する．標準偏差は無視できるほど小さいので省略した．また，“Average” は全 11 問題での性能を平均した結果を示す．11 種の問題すべてにおいて，かつ，総検索数 N のすべてにおいて提案法の検索性能が従来法を顕著に上回っていることが確認できる．単語頻度ではなく，より概念的なトピックに基づく検索の優位性が実証されたといえる．$N = 1$ のときは従来法とたかだか約 5% 程度の優位性であったが，$N = 100$ では 8〜16%（11 問題の平均では約 10%）もの性能差があった．複数の類似文書を求める検索問題においてこの差は重要である．

1 つの検索文書に対し $N = 100$ の被検索文書を検索するのに要する時間（COS，IDF 法では $s_{m,n}$，$n = 1, \ldots, N$ の算出時間，提案法では h_m，$\{h_n\}_{n=1}^{N}$ の算出時間+ $s_{m,n}$，$n = 1, \ldots, N$ の算出時間）は，平均で，COS，IDF 法では 0.52 秒，提案法では 2.13+0.14=2.27 秒であった☆☆．提案法では，トピック度ベクトルの推定が必要でその分の時間がかかるが，単語頻度ベクトルの次元（V）に比べ，トピック度ベクトルの次元（L）はかなり小さいので，類似度計算においては，従来法の 0.52 秒に対し，0.14 秒と約 4 倍高速である☆☆☆．今回の実験では，類似度で被検索文書を全数サーチしているが，実用上は，ハッシュ表を作成する等して，検索効率のさらなる向上が可能と

☆ 冠詞や be 動詞等の stop words を削除し，3 人称単数や過去形等で変化した単語を同一視するための語末修正処理を施した後の単語総数．

☆☆ DELL Precision WorkStation530 Xeon 2 GHz processor での時間．

☆☆☆ 従来法の計算時間は，V 次元のコサイン類似度を単純に計算しているのではなく，粗ベクトル表現して零要素をスキップして計算効率を上げたときの所要時間であることに注意．

考えられる．

また，検索されたページの内容を見ると，たとえば“自然環境問題に関する研究が重要”という主旨のページで検索した際，“太陽電池”，“自然保護”に関するページ群が検索され，興味深いことに，これらのページ中の単語は先の“自然環境問題”のページ中には出現していなかった．つまり，COS 法のようなキーワード検索では検索困難なページでも，トピックを介することにより検索可能であることを意味する．さらに，提案法の特長として，学習対象文書群を変えることにより，異なる観点で類似ページが検索できる．こうした柔軟な検索は確率モデルを用いて初めて実現できるといえる．

テキストモデルの関連研究として PLSI モデル[6] が提案されている．しかし，PLSI では学習文書間の類似度は定義できるが，学習とは独立な未知文書間の類似度は原理的に定義できない．すなわち汎化能力がない不完全な確率モデルゆえ，web のようなオープンな世界でのテキスト検索には適さない．一方，提案法は前述したように，未知文書についての類似度も算出できる．その意味で PMM は完全な確率モデルといえる．

8. 結　　論

本論文では，多重トピックテキストの確率モデル（PMM）を用いた類似テキスト検索法を提案し，web ページを用いた検索実験により有効性を示した．提案手法の特長を整理する．

(1) 高速性：パラメータ推定アルゴリズムは学習文書数の線形オーダ．
(2) 最適性：推定されたトピックベクトルの大域的最適性を理論保証．
(3) 適応性：PMM を所望のトピック体系の文書群で PMM-C を学習することにより，異なる観点での類似テキスト検索が実現可能．

web ページはテキストのみならず画像情報も重要な情報である．テキストの類似度に画像情報を反映させる方法も検討中である．

参 考 文 献

1) Ueda, N. and Saito, K.: *Single-shot detection of multiple topics using parametric mixture models*, pp.626–631, ACM, Edmonton (2002).
2) Ueda, N. and Saito, K.: Parametric mixture models for multi-labeld text, Vol.15, MIT Press, Cambridge, MA. (2002).
3) Yang, Y. and Pederson, J.: *SA comparative study on feature selection in text categorization*, pp. 412–420, Morgan Kaufmann, San Francisco (1997).
4) Manning, C.D. and Schtze, H.: *Foundations of statistical natural language processing*, MIT Press, Cambridge (1999).
5) Joachims, T.: Text categorization with support vector machines: Learning with many relevant features, Vol.15, MIT Press, Cambridge, MA. (in press).
6) Hofmann, T.: *Probabilistic latent semantic indexing*, pp.50–57 (1999).

(平成 15 年 4 月 14 日受付)
(平成 15 年 5 月 29 日採録)

上田　修功（正会員）

昭和 33 年生．昭和 57 年大阪大学工学部通信工学科卒業．昭和 59 年同大学院修士課程修了．同年 NTT 電気通信研究所入所．平成 5 年より 1 年間米国 Purdue 大学客員研究員．現在，NTT コミュニケーション科学基礎研究所知能情報研究部長（創発学習研究グループリーダ兼務）．奈良先端科学技術大学院大学客員助教授．日本神経回路学会理事．博士（工学）．共著『よくわかるパターン認識』（オーム社），共著『統計科学のフロンティア，第 11 巻：新しい確率計算』（岩波書店）．平成 4 年日本神経回路学会研究奨励賞，平成 9 年電気通信普及財団賞（テレコムシステム技術賞），平成 12 年電子情報通信学会論文賞受賞，平成 15 年日本神経回路学会研究賞受賞．電子情報通信学会，日本神経回路学会，IEEE 各会員．

斉藤　和巳（正会員）

昭和 38 年生．昭和 60 年慶應義塾大学理工学部数理科学科卒業．工学博士．同年 NTT 入社．平成 3 年より 1 年間オタワ大学客員研究員．神経回路網，機械学習の研究に従事．現在，NTT コミュニケーション科学基礎研究所主任研究員（特別研究員）．平成 9 年情報処理学会論文賞受賞．平成 11 年人工知能学会論文賞受賞．電子情報通信学会，人工知能学会，日本神経回路学会，IEEE 各会員．