分類の評価や特異文書の検出のための文書群の構造可視化

岩田 具治　斉藤 和巳　上田 修功
日本電信電話株式会社　NTT コミュニケーション科学基礎研究所

概要: カテゴリー分類された文書群に対し、カテゴリー間の関係の理解、分類の評価、および特異文書の検出を目的とした、大規模文書群にも適用可能な効率的な可視化手法を提案する。特異文書とは、既存のどのカテゴリーにも属さないと考えられる文書を意味する。提案法では、カテゴリー情報とコンテンツ情報から文書生成モデルを学習する。そして、各カテゴリーに属する確率と、どのカテゴリーにも属さない確率を推定し、事後確率ベクトルを求める。次いで、事後確率ベクトル値をできるだけ保存するように 2 次元または 3 次元のユークリッド空間に埋め込むことにより、分類文書群を可視化する。提案法は計算コストの面でも効率的である。カテゴリー分類されたウェブページを用いて提案法の有効性を示す。

Visualization of Documents for Evaluation of Classification and Detection of Unique Documents

Tomoharu IWATA　Kazumi SAITO　Naonori UEDA
NTT Communication Science Laboratories, NTT Corporation

abstract: We propose a new method for visualization of a huge amount of classified documents, that allow us to understand relationships among categories, to evaluate classification results, and to detect unique documents not belonging to any categories. Learning a document generative model from categorical information and contents information, we obtain the probabilities belonging to each category and not belonging to any categories. Then, posterior vector for each document is computed. Next, documents are embedded in two or three Euclidean space so that the posterior vector values are preserved as much as possible. The proposed visualization method has also computational advantage over the conventional methods based on pair-wise similarities. We show the validity of our method by visualizing classified web pages.

1　はじめに

近年、大量の文書が電子的に蓄積されつつある。それらの文書群は通常、ユーザが求める文書を容易に見つけ出すことができるように、関連する文書ごとに分類されている。例えば、ウェブページは、open directory project[9]、goo カテゴリー検索 [5] などのディレクトリ型検索エンジンにおいて、トピック毎に分類されている。ウェブページの他にも、電子図書、特許、論文、パソコン内電子ファイルなど、多くの電子文書は、何らかの形で分類されている。

分類された文書群は、人手で分類されているために分類ミスが含まれていたり、既存の分類カテゴリーに含めることのできない特異な文書が存在したり、分類体系が偏っているなど、分類が適切に行われていないことは往々にしてあると考えられる。一つ一つの文書を人がチェックすることによって分類ミスを探し出すことができるが、膨大な文書をすべて見直すことは困難であるため、十分な分類結果の評価が行われていないことが多い。

大量のデータの特性を理解することを可能にする 1 つの有効な方法は、データを可視化することである [1]。可視化することにより、データに内在する構造的特徴が浮き彫りになり、知識発見のための重要な手がかりが得られる。本稿で、分類の評価や特異文書の検出を容易にする、新たな分類文書群の可視化手法を提案する。

これまでに文書群の可視化手法はいくつか提案さ

れているが（例えば [2]）、分類カテゴリー情報を用いておらず、分類文書群を可視化するための手法ではない。さらに、従来の文書間類似度に基づく可視化手法では、文書数の二乗のオーダーで計算量が増えるため、大規模な文書群を可視化することは困難である。

提案手法では、カテゴリー情報とコンテンツ情報を用い、それぞれのカテゴリーに属する確率と、どのカテゴリーも属さない確率を求め、それらの確率をできるだけ保存するように低次元に埋め込むことにより、分類文書群をその特性が直観的に理解できるように可視化する。提案手法は、後述するようにカテゴリー情報を用いることにより計算の効率化が図られ、大規模な文書群に対しても適用可能である。

図 1 は、分類文書群がどのように可視化されるかを表したイメージ図である。相撲、野球、ラグビー、サッカー、プロレスの 5 つのカテゴリーに分類された文書群があるとする（図 1 左）。提案手法を用いて可視化することにより、図 1 右のような結果が得られる。同じカテゴリーに分類されていた文書は同じ色形で可視化されている。この可視化結果から、カテゴリー間の関係が視覚的に理解可能となる。例では、相撲とプロレスは同じ格闘技に属するスポーツで近いカテゴリーであるので、近くに配置されている。さらに以下のような文書の検出も可能となる。

分類ミス　誤って分類された文書。異なるカテゴリーの文書が集まっているところに配置された文書は分類ミスの文書であると考えられる。例では、相撲の文書が集まっているところに、プロレスの文書が 1 つだけ混ざっている。この文書はプロレスではなく、相撲についての文書である可能性が高いと考えられる。

多重カテゴリー文書　複数のカテゴリーに属すると考えられる文書。可視化により、複数のカテゴリーに属する確率が同等に高い文書は、それらのカテゴリーの間に配置される。例では、サッカーと野球の中間に配置されている文書が多重カテゴリー文書である。この文書は、サッカーと野球について書かれた文書であると推測できる。

特異文書　既存のどの分類カテゴリーにも属する確率が低い文書。可視化空間で既存のカテゴリーが集まる場所とは異なる場所に配置される。例では、野球の文書が 1 つだけ他とは離れた場所に位置している。

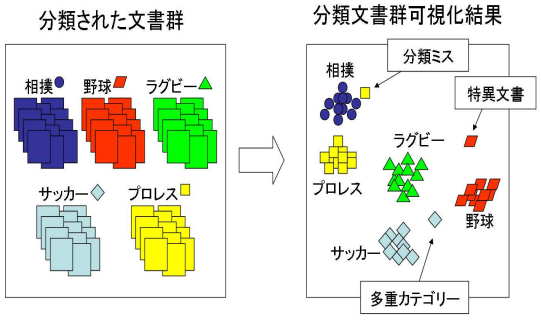

図 1: 提案手法による分類文書群可視化イメージ図

どのカテゴリーにも属さないということは、与えられた他の文書とは大きく異なる性質を持つものであると考えられる。その異なり方は大きく分けて二種類ある。他の文書にはない情報が記載されている有益な文書である場合と、無意味または無関係で同じ文書群としては不適切な文書である場合である。そのいずれかは実際に文書を確認しないと不明だが、前者の場合であれば読まれるべき文書であり、後者の場合であれば文書群から取り除くべき文書であり、いずれの場合にしても特異文書を検出することの意義は大きい。

2　提案手法

提案手法は以下の処理からなる。

1. 文書生成モデルの構築
2. 事後確率ベクトルの推定
3. 事後確率保存埋め込みによる可視化

以下で、各々について詳述する。

2.1　文書生成モデルの構築

まず、分類された文書群から、カテゴリーごとの文書生成モデルを構築する方法について説明する。分類された文書群を $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ とする。d_n は第 n 文書、N は与えられた文書の総数を表す。各文書は 1 つのカテゴリーに分類されているとし、第 n 文書の分類カテゴリーを $c_n \in \{1, 2, \ldots, K\}$ とする。K はカテゴリーの総数を表す。また、与えられた文書

に含まれる全単語の集合を $\mathcal{W} = \{w_1, w_2, \ldots, w_V\}$ とする。w_j は第 j 単語、V は単語の総数を表す。

文書の内容は、単語頻度ベクトルで表現する。つまり、単語の文書中での順序や係り受けなどを無視した Bag-Of-Words(BOW) と呼ばれる文書表現を用いる [7]。テキスト分類では、BOW による文書表現はよく用いられている。第 n 文書の単語頻度ベクトルを $\mathbf{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nV})$ とする。x_{nj} は第 n 文書中に含まれる第 j 単語の数を表す。

カテゴリー毎の文書生成の確率モデルとしてナイーブベイズモデル (NB モデル) を用いる [3]。NB モデルでは、カテゴリー k に属する文書 d_n の生成確率を多項分布

$$p(d_n|k) \propto \prod_{j=1}^{V} \theta_{kj}{}^{x_{nj}} \tag{1}$$

と仮定する。θ_{kj} は第 k カテゴリーに属す文書中で単語 w_j が生起する確率 $(\theta_{kj} > 0, \sum_{j=1}^{V} \theta_{kj} = 1)$ を表す。第 k カテゴリの文書生成モデルのパラメータ $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{kV})$ は、事前確率ハイパーパラメータ λ_k を導入した、次式で表される尤度 L_k を最大化することにより、推定する。

$$L_k = \sum_{n \in \{l|c_l=k\}} \sum_{j=1}^{V} x_{nj} \log \theta_{kj} + \lambda_k \sum_{j=1}^{V} \log \theta_{kj} \tag{2}$$

λ_k を導入することにより、与えられた文書中に 1 つも現れなかった単語の生成確率が 0 になるのを防ぎ、かつ凡化性能を上げることができる。θ_{kj} の最尤推定値 $\hat{\theta}_{kj}$ は、次式となる。

$$\hat{\theta}_{kj} = \frac{\sum_{n \in \{l|c_l=k\}} x_{nj} + \lambda_k}{N_k + \lambda_k V} \tag{3}$$

N_k は第 k カテゴリーに属す総文書数を表す。ハイパーパラメータ λ_k は、過学習を防ぐためクロスバリデーション法により推定する。

ここでは文書生成モデルとして、文書の内容を単語頻度ベクトルで表現した NB モデルを採用したが、n-gram など他の手法を用いても構わない。

2.2 事後確率ベクトルの推定

次に、文書の各カテゴリーに属する確率と、どのカテゴリーにも属さない確率をベクトルにした事後確率ベクトルを、最大エントロピー法 [8] により推定する方法について説明する。

最大エントロピー法とは、素性 (feature) と呼ばる第 n 文書と第 k カテゴリーの関係を表す関数を $f(d_n, k)$ としたときに、制約

$$\frac{1}{N} \sum_{n=1}^{N} f(d_n, c_n) = \frac{1}{N} \sum_{n=1}^{N} \sum_{k=1}^{K+1} P(k|d_n) f(d_n, k) \tag{4}$$

を満たしつつ、第 n 文書が与えられたときの第 k カテゴリーに属す確率 $P(k|d_n)$ のエントロピーが最大となるようにパラメータを推定するものである。この条件で、$P(k|d_n)$ のエントロピーが最大となる唯一の分布が存在し、その分布は次式で表される。

$$P(k|d_n) = \frac{\exp(\beta f(d_n, k))}{\sum_{l=1}^{K+1} \exp(\beta f(d_n, l))} \tag{5}$$

β は推定すべきパラメータである。ここで、k は 1 から $K+1$ までの整数値をとる。1 から K までは既存のカテゴリーを表し、$K+1$ は既存カテゴリー外というカテゴリーを表す。つまり $P(K+1|d_n)$ は、第 n 文書が既存のどのカテゴリーにも属さない確率を表す。

第 n 文書と既存の第 k カテゴリーの素性として、次式で表される、NB モデルでの第 k カテゴリーが与えられたときの第 n 文書の対数尤度を用いる。

$$f(d_n, k) = \sum_{j=1}^{V} x_{nj} \log \theta_{kj}, \quad k = 1, \ldots, K \tag{6}$$

第 n 文書の単語頻度が第 k カテゴリーのモデルと近ければ、素性 $f(d_n, k)$ は高くなり、したがって第 k カテゴリーに属す事後確率 $P(k|d_n)$ が高くなる。このことから、事後確率を推定するために、素性として対数尤度を用いることは適切であると考えられる。また、既存の K 個のカテゴリーに限定した場合、NB モデルにおける事後確率はベイズの定理より

$$P_{NB}(k|d_n) = \frac{P(k) \exp(\sum_{j=1}^{V} x_{nj} \log \theta_{kj})}{\sum_{l=1}^{K} P(l) \exp(\sum_{j=1}^{V} x_{nj} \log \theta_{lj})} \tag{7}$$

となる。事前確率 $P(k)$ が一様としたときの $P_{NB}(k|d_n)$ と、$\beta = 1$ としたときの最大エントロピー法による $P(k|d_n)$ は等しく、NB モデルの自然な拡張となっていると言える。最大エントロピー法により、エントロピーを最大にするという好ましい条件の下、どのカテゴリーにも属さない確率を含めた事後確率 $P(k|d_n)$ を適切に推定することが可能となる。

また、第 n 文書が既存のどのカテゴリーにも属さない事後確率を推定するための素性 $f(d_n, K+1)$ として、NB モデルでの第 n 文書の事後確率が最も高いカテゴリー $\hat{c}_n = \arg\max_k P_{NB}(k|d_n)$ の文書生成モデルにおける 3 シグマでの対数尤度、3 シグマ値を用いる。平均 μ、標準偏差 σ の正規分布従う確率変数の観測値が、$\mu \pm 3\sigma$ の区間に入る確率は 0.997 である。つまり、3 シグマ値が既存カテゴリーに対する素性と比べ相対的に高いものは、どのカテゴリーに属する確率も低い文書であると考える。

3 シグマ値を求めるため、

$$P(X = \log \hat{\theta}_{kj}) = \hat{\theta}_{kj} \tag{8}$$

となるような確率変数 X を考える。中心極限定理により、X を M 回試行したときの平均値の分布は M が大きくなると正規分布に漸近する。すなわち、

$$\frac{X_1 + X_2 + \ldots + X_M}{M} \sim \mathcal{N}(\mu_k, \frac{\sigma_k^2}{M}), \tag{9}$$

$$\mu_k = \sum_{j=1}^{V} \hat{\theta}_{kj} \log \hat{\theta}_{kj}, \tag{10}$$

$$\sigma_k = E(X^2) - \mu_k^2 \tag{11}$$

となる。X_i は第 i 試行の X の値、$\mathcal{N}(\mu, \sigma)$ は平均 μ、標準偏差 σ の正規分布を表す。そして、第 n 文書のどの既存カテゴリーにも属さない確率を求めるための素性 $f(d_n, K+1)$ は次式で表される 3 シグマ値となる。

$$f(d_n, K+1) = \mu_{\hat{c}_n} - 3\frac{\sigma_{\hat{c}_n}}{M_n} \tag{12}$$

M_n は第 n 文書の単語数を表す。

このように素性を決定した後、未知パラメータ β を推定し、事後確率 $P(k|d_n)$ を求める。用いた推定アルゴリズムは [8] に詳しく書かれている。求めた第 n 文書の事後確率をベクトルにしたものを $\mathbf{q}_n = (P(1|d_n), \ldots, P(K+1|d_n))$ とし、事後確率ベクトルと呼ぶ。

2.3 事後確率保存埋め込みによる可視化

最後に、事後確率ベクトル $\mathbf{q}_n$ をもとに低次元に埋め込み、可視化する方法について説明する。まず、第 k カテゴリーに属する文書群の中心となる可視化空間における座標 $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kD})$ を考える。D は可視化空間の次元を表す。また、第 n 文書の座標を $\mathbf{r}_n = (r_{n1}, \ldots, r_{nD})$ とする。文書のそれぞれのカテゴリーに属する事後確率の関係を保存したまま、直観的に理解できるように 2 次元または 3 次元空間に埋め込みたい。すなわち、第 n 文書の第 k カテゴリーに属する確率 q_{nk} が高ければ、第 k カテゴリーと第 n 文書の可視化空間でのユークリッド距離

$$u_{nk} = \| \mathbf{r}_n - \boldsymbol{\phi}_k \|_D^2 = \sum_{l=1}^{D} (r_{nl} - \phi_{kl})^2 \tag{13}$$

が近くなるように、逆に確率 q_{nk} が低ければ距離 u_{nk} が遠くなるように、$\mathbf{r}_n$、$\boldsymbol{\phi}_k$ を配置することができればよい。

ここで、可視化空間における $\mathbf{r}_n$ と $\boldsymbol{\phi}_k$ の位置関係から推測できる、第 n 文書が第 k カテゴリーに属する確率を

$$s_{nk} = \frac{\exp(-\frac{1}{2} \| \mathbf{r}_n - \boldsymbol{\phi}_k \|_D^2)}{\sum_{l=1}^{K+1} \exp(-\frac{1}{2} \| \mathbf{r}_n - \boldsymbol{\phi}_l \|_D^2)} \tag{14}$$

と考える。つまり、$\mathbf{r}_n$ は平均 $\boldsymbol{\phi}_k$、分散共分散行列 I_D の D 次元正規分布に基づくとし、平均である $\boldsymbol{\phi}_k$ に近ければ第 k カテゴリーに属する確率が高いとする。正規分布は自然現象や社会現象で最も多く見られる分布の 1 つであり、正規分布を基に埋め込むことにより、直観的に分かりやすい結果が得られることが期待できる。

s_{nk} をベクトルにしたものを $\mathbf{s}_n = (s_{n1}, \ldots, s_{nK+1})$ とする。可視化空間における第 n 文書がどのカテゴリーに属すかを表すベクトル $\mathbf{s}_n$ が事後確率ベクトル $\mathbf{q}_n$ の十分よい近似となっていればよい。$\mathbf{q}_n$ と $\mathbf{s}_n$ を離散確率分布と考え、両確率分布間の距離を表すカルバック擬距離

$$KL(\mathbf{q}_n, \mathbf{s}_n) = \sum_{k=1}^{K+1} q_{nk} \log \frac{q_{nk}}{s_{nk}} \tag{15}$$

を全文書に対して最小にすることで、近似を実現することができる。このとき、$\mathbf{q}_1, \ldots, \mathbf{q}_N$ は既知であるため、全文書の座標 $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ を得るための、最大化すべき目的関数は次式となる。

$$L_V = \sum_{n=1}^{N} \sum_{k=1}^{K+1} q_{nk} \log \frac{\exp(-\frac{1}{2} u_{nk})}{\sum_{l=1}^{K+1} \exp(-\frac{1}{2} u_{nl})} \tag{16}$$

実際には $\mathbf{R}$ に関する L_V の最大化と $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{K+1}\}$ に関する L_V の最大化を収束するまで繰り返すことにより、$\mathbf{R}$ を求める。$\boldsymbol{\Phi}$ が与えられたとき、L_V は $\mathbf{R}$ に関して厳密に上に凸となり、収束値が大域的最

適解を保証するという好ましい性質を持つ。Φ の初期値として、$\hat{\boldsymbol{\theta}}$ をもとに、クロスエントロピー埋め込み法 [13] で求めた座標を用いる。これは、単語の出現確率が似ているカテゴリー同士が近くに配置された初期値とするためである。このことにより、カテゴリー間の関係をより適切に可視化することが期待できる。

3　提案手法の評価

提案手法を評価するために、Open Directory Project (ODP) の日本語ウェブページ [10] のトップカテゴリー分類情報を用いて、分類文書群の可視化を行った。

3.1　使用データ

ウェブページのサンプリングは以下のように行った。まず、ODP に登録されている日本語ウェブページから、単語数が 50 以下のページを除き、次に、複数のカテゴリーに分類されているページを除き、最後に、各カテゴリー 100 ページをランダムにサンプルする。サンプリングの結果、得られたページは、トップカテゴリーのうち 100 ページに満たなかったゲームカテゴリーを除いた、アート (arts)、オンラインショップ (online-shop)、コンピュータ (computer)、スポーツ (sports)、ニュース (news)、ビジネス (business)、レクリエーション (recreation)、健康 (health)、各種資料 (reference)、家庭 (home)、社会 (society)、科学 (science)、地域 (regional) の 13 のカテゴリーに含まれる全 1300 ページであった。

3.2　可視化結果

図 2 は、サンプリングしたページ群の可視化結果である。同じカテゴリー情報を持つページは同じ色形で表示されている。結果を見て分かるように、同じカテゴリーに属するページがクラスターを形成おり、既存カテゴリー 13 個に特異文書を合わせた計 14 個のクラスターが存在する。クラスターの位置関係を見てみると、関連するカテゴリーのクラスターは近くに配置されている。例えば、オンラインショップとビジネス、スポーツと健康などは、近くに配置されている。

分類ミス　同じカテゴリーのページで形成されたクラスターのなかに、異なるカテゴリーのページが少数ある場合、そのページは分類ミスと考えられる。図 2 中の a は、地域に分類されているページであるにも関わらず、健康のページのクラスターのなかに配置されている。このページを見てみると、病院のページであった。地元の人が利用するという意味では地域のページであると言うこともできるが、健康のカテゴリーのなかには多くの病院のページが含まれており、1 つの基準で分類するのであれば、健康のカテゴリーと分類する方が適当であると思われる。このように可視化結果から、誤って分類された可能性があるページを検出することができる。

多重カテゴリー　カテゴリークラスターの中間に位置するページは、両方のカテゴリーに含めることができる多重カテゴリーのページであると推測できる。例えば図 2 中の b のページは、レクリエーションと分類されているものであるが、スポーツとレクリエーションの中間に位置している。このページは、toto（サッカーくじ）のページであった。サッカーくじは、レクリエーションの 1 つであるが、スポーツ（サッカー）とも深い関わりがあり、このように両カテゴリーの中間に位置されている。スポーツという観点からサッカーくじのページに興味を持つ人はいると考えられ、スポーツのカテゴリーにも含めた方が使いやすい可能性がある。このように、可視化結果から、多重に分類できたり、明確な分類が困難なページを検出することができ、分類結果を見直す際にどのページを重点的に見るべきかを知ることができる。

特異文書　同じカテゴリーに属するページで構成されるクラスターとは異なる位置に配置されたページは、どの既存のカテゴリーにも含めることができない特異文書であると考えられる。その 1 つのページを見てみると、様々なトピックに関するメールマガジンを紹介するページであった [6]。このページは ODP ではコンピュータに分類されていたが、使われている文章の一部を見てみると、図 3 のようにコンピュータに関するものだけでなく、変わった単語が多く用いられ、普通のコンピュータに関するページではないことが分かる。

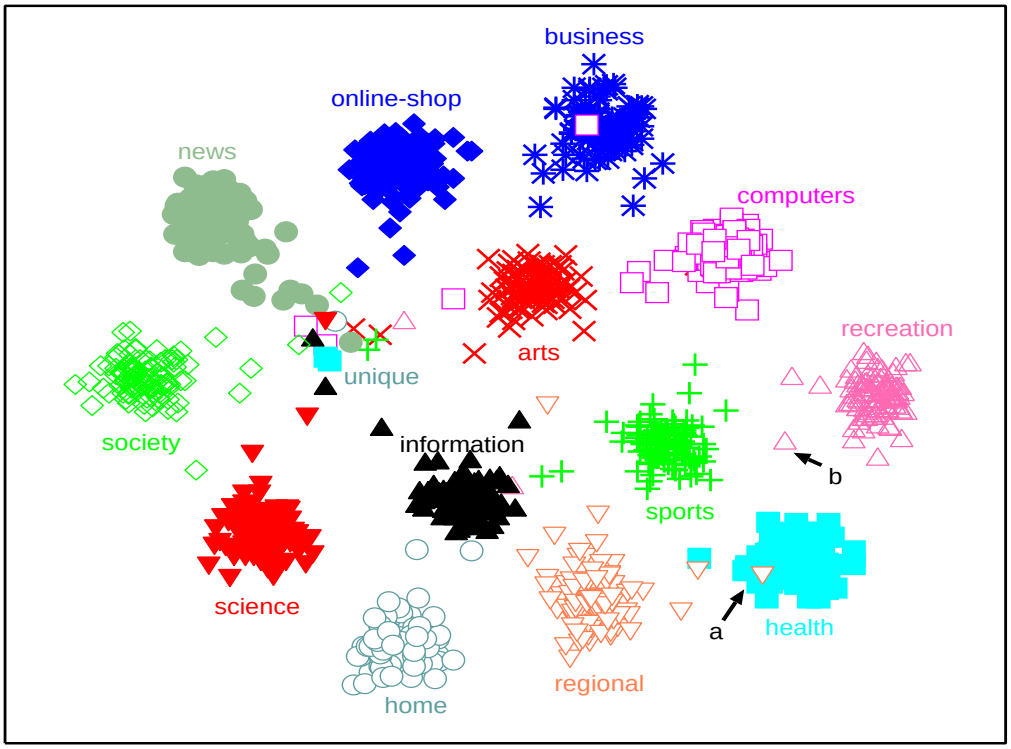

図 2: 提案手法による ODP トップカテゴリー分類文書群の可視化結果

住宅ローン、仁義なき戦い !!
五臓六腑にしみる三つの言葉
超短編小説『リーフノベル』
日常の法律問題を考えましょう
☆みんなで考えよう♪日本と世界のこと
女の恋愛∬女のホンネ
Nobody knows, but I'm HIV+
風向日記　～なつ風便り～
私の主張・ひとりの日本人として
よりぬき・いきぬき
電脳市場本舗～Marketing Surfin'～
朝の 5 分でかわる対人関係
...

図 3: 特異文書例：様々なトピックに関するメールマガジンを紹介するウェブページ [6] の本文の一部

3.3　従来手法との比較

文書可視化法として広く用いられている方法は、文書の単語頻度ベクトルから文書間の類似度を求め、類似度をもとに MDS[2] や Isomap[12] などを用い次元圧縮し、可視化するというものである。しかし、この方法では、文書のカテゴリー情報を使うことができず、分類文書の可視化法となっていない。さらに、文書数を N とすると計算量は $O(N^2)$ であり、文書数が多くなると計算不可能になってしまうという問題がある。

一方、提案手法は、確率モデルを仮定することで、カテゴリー情報とコンテンツ情報を統合し可視化することができ、分類文書群の可視化手法として適切なものであると言える。また、計算量はカテゴリー数を K とすると $O(NK)$ である。一般にカテゴリー数はそれほど多くなることはないため、大量の文書群にも適用可能である。

MDS で最大固有値を求める際にべき乗法を応用することにより、疎な行列の場合、計算量を大幅に削減することができる。単語頻度ベクトルは通常ほとんどの要素が 0 である疎なベクトルであるため、べき乗法を応用した MDS を用いることにより、大

量の文書群も少ない計算コストで可視化することができる。この方法で単語頻度ベクトルを次元圧縮し可視化した結果、図4のようになった。単語頻度ベクトルのみを用いており、前述のようにカテゴリー情報を考慮していない。そのため、カテゴリー毎にクラスターができることもなく、文書がどのような性質を持っているか全く分からない結果になってしまっている。

また、提案手法では事後確率ベクトルを事後確率保存埋め込み法によって可視化したが、事後確率をMDSによって次元圧縮し可視化するという方法が考えられる。その結果、図5のようになった。カテゴリー毎にクラスターを形成しており、分類ミスも見られるが、中心に複数のカテゴリーが集中してしまっており、各カテゴリー間関係の理解、多重ラベルや特異文書の検出が困難である。

4 おわりに

本稿では、分類された文書を可視化することにより、分類ミス、多重カテゴリー文書、特異文書、カテゴリー間関係を視覚的に理解することができる手法を提案した。また、分類されたウェブページを用い、提案手法の有効性を示した。さらに、従来の手法と比べ、計算量が大幅に削減できること、分類文書群の可視化手法として適切なものであることを示した。今後は、様々な分類文書群に適用し、提案手法の有効性を検証しつつ問題点を検討し、性能改善を進めていきたい。

参考文献

[1] S.K.Card, G.C.Robertson and J.D.Mackinlay, The Information Visualizer, an Information Workspace, in Proceedings of CHI'91, ACM Press, pp.181–188, 1991.

[2] M.Chalmers and P.Chitson, BEAD: Explorations in information visualization, SIGIR'92, Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Resarch and Development in Information Retrieval, ACM Press, pp.330–337, 1992.

[3] R.O.Duda, P.E.Hart, and D.G.Stork, Pattern classification (2nd ed.), John Wiley & Sons, New York, 2002.

[4] T.Fawcett and F.Provost, Activity monitoring: Noticing interesting changes in behavior, In Proc. Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.53–62, 1999.

[5] goo カテゴリー検索, http://dir.goo.ne.jp/

[6] よりすぐりメルマガサイト「めろんぱん」, http://www.melonpan.com/

[7] C.D.Manning and H.Schütze, Foundations of statistical natural languatge processing, MIT press, Cambridge, 1999.

[8] K.Nigam, J.Lafferty and A.McCallum, Using maximum entropy for text classification, In IJCAI-99 Workshop on Machine Learning for Information Filtering, 61–67, 1999.

[9] open directory project, http://dmoz.org/

[10] open directory project (Japan), http://dmoz.org/World/Japanese/

[11] M.Rorvig, Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC document sets, Journal of the American Society for Information Science, 50(8), pp.639–651, 1999.

[12] J. B. Tenenbaum, V. de Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science, 290, pp.2319-2323, 2000.

[13] 山田武士、斉藤和己、上田 修功、クロスエントロピー最小化に基づくネットワークデータの埋め込み、情報処理学会論文誌 Vol.44、No.9、pp.1234–1231、2003.

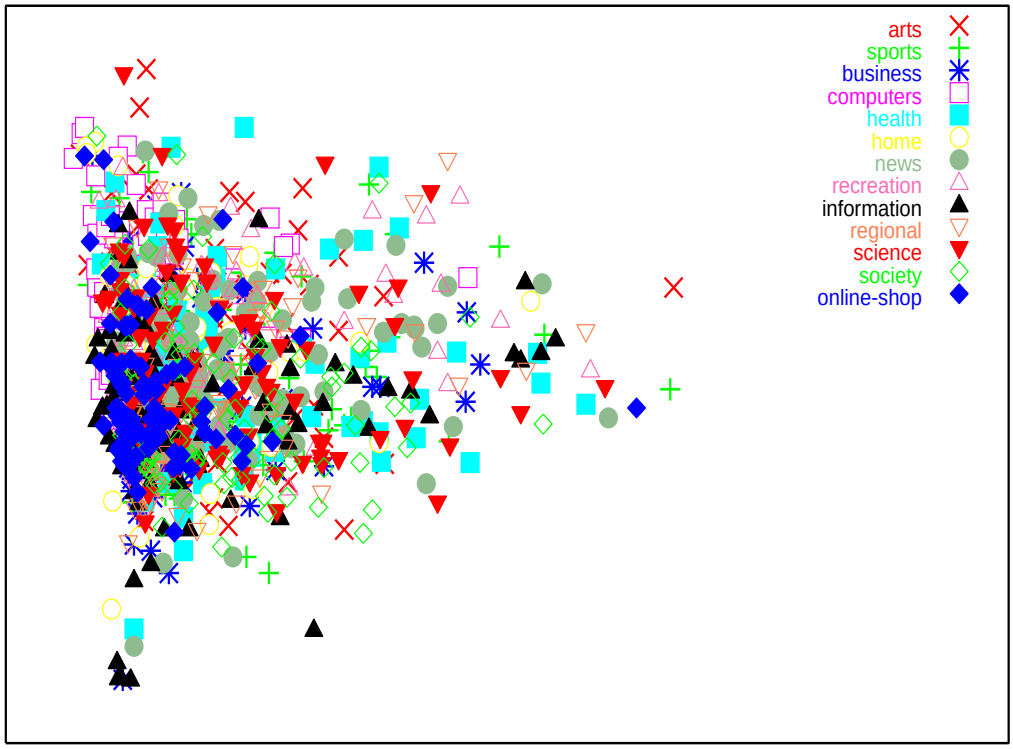

図 4: 単語頻度ベクトルを用いた MDS による ODP トップカテゴリー分類文書群の可視化結果

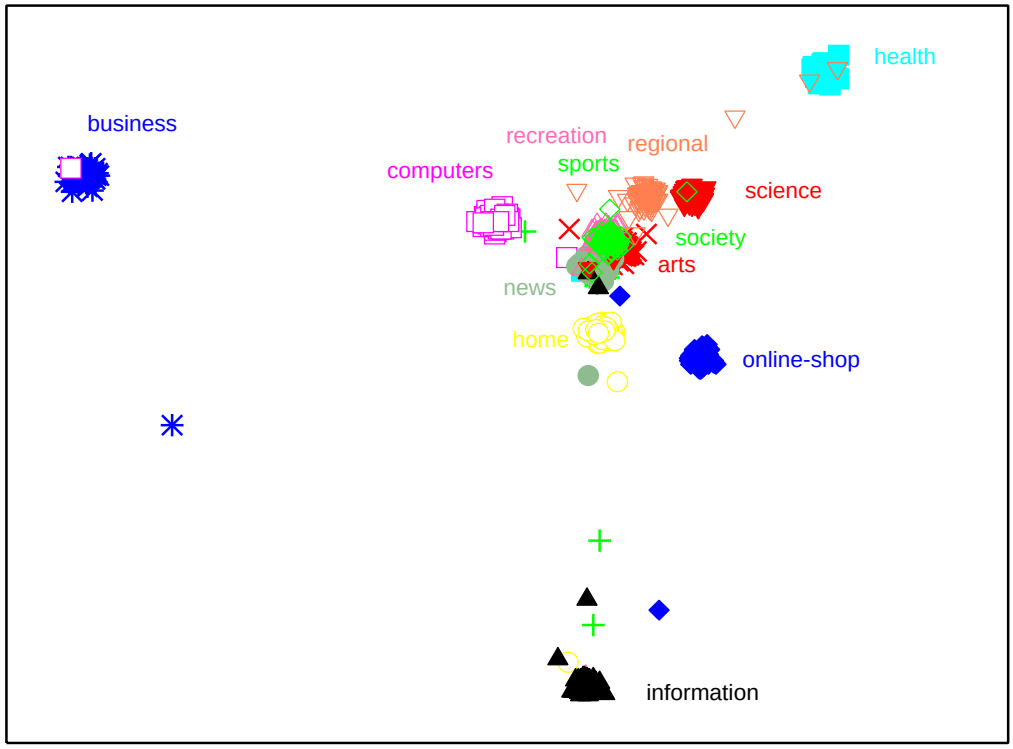

図 5: 事後確率ベクトルを用いた MDS による ODP トップカテゴリー分類文書群の可視化結果