論　文

言語表現のベクトル空間モデルにおける最適な計量距離

持橋　大地[†,††]　　菊井玄一郎[†]　　北　　研二[†,†††]

Learning an Optimal Distance Metric in the Linguistic Vector Space

Daichi MOCHIHASHI[†,††], Genichiro KIKUI[†], and Kenji KITA[†,†††]

あらまし　自然言語処理の手法にはコサイン距離，すなわち単純なユークリッド距離に依存しているものが多いが，この距離は素性間の相関を考慮しないこと，また素性に恣意的な重みづけを必要とするという問題を持っている．本論文ではこれに代わり，訓練データより導かれる最適な計量距離を提案し，二つの問題を同時に解決する．この計量は訓練データのクラスタ構造の歪みを最小化する二次最適化問題の解として導くことができる．提案する計量距離の効果を，同義文検索，文書検索，および機械学習用一般ベクトルデータのクラスタリングタスクによって確認し，ユークリッド距離に対し常に精度の改善がみられた．特にクラスタ化の強い同義文検索タスクにおいて精度の改善が大きく，ベースラインと比較して 11 点平均精度で 33% の向上があった．

キーワード　距離計量，コサイン距離，tf.idf，素性空間，カーネル法

1. まえがき

自然言語処理において，文，文書，パラグラフなどの言語表現の間の意味的な距離を計算することは基礎的で重要な技術となっている．たとえば，情報検索は検索語あるいは特定文書と意味的な距離が近い文書を文書集合の中から検索するタスクであるし，TextTiling [1] やその後継であるスペクトル法などのテキスト分割法 [2] においては，いずれもパラグラフ間のコサイン距離がその基礎として使われている．質問応答 (QA) や言い換え，用例ベース機械翻訳の場面でも，文間距離を計算することが基礎的な要素技術となっている．

現在，このような言語表現の比較には大きく分けて (a) 構造的な方法 と (b) 非構造的な方法 が存在している．(a) の構造的な方法は何らかの構文解析または依存構造解析を用い，精密な比較を行うものであり，一方 (b) の非構造的な比較は言語表現を実数空間の何らかのベクトルとみなし，大量のコーパスに対し，高速な検索や比較を行うものである．近年，(a) の構造的な比較は再帰的なカーネル関数を用いてカーネル法の枠内で見通しよく扱えるようになっているが [3], [4]，ここでは (b) の非構造的な比較に着目する．これは，上に述べたような大量のコーパスに対する自然言語処理では，構造的比較に必要な構文解析や依存構造解析が計算量的に非現実的であり，近似的であっても高速な比較が求められるということの他に，構造的な比較においても再帰の葉においては依然，非構造的比較 (イグザクトマッチやベクトルのマッチ) が行われるため，それらの基礎をなすと考えられるからである．

しかしながら，それらの場面で使われる非構造的比較は多く，はじめに述べたコサイン (＝ユークリッド) 距離に依存しており，素性間の相関および各素性の重み付けの点で問題を残している．本論文ではこれに代わり，訓練データから得られる最適な計量距離を導入し，その効果を実験的に検証する．この計量は訓練データとして与えるクラスタ構造をもとに，二次最適化問題の解として導出される．

本論文は以下のように構成される．2 章で，自然言語処理において伝統的に使われてきたユークリッド距離とその問題点について述べる．3 章では，この問題が機械学習における，データに基づく計量の導出問題と捉えられることを示し，近年の関連研究について述

† ATR 音声言語コミュニケーション研究所，京都
ATR Spoken Language Translation Research Laboratories
Hikaridai 2-2-2, Keihanna science city, Kyoto 619-0288
†† 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, NAIST
Takayama-cho 8916-5, Ikoma, Nara 630-0192
††† 徳島大学高度情報化基盤研究センター
Center for Advanced Information Technology, Tokushima University Minamijosanjima 2-1, Tokushima 770-8506

べる．4 章で提案手法を説明し，5 章でその効果を同義文検索，文書検索および一般データのクラスタリングタスクを用いて検証する．6 章と 7 章で議論，課題および結論を述べる．

2. ユークリッド距離とその問題

自然言語処理において非構造的比較を行う場合，言語表現はしばしば，素性 i の生起回数 x_i $(i = 1 \ldots n)$ を要素とするベクトル $\vec{x} \in \mathbb{R}^n$ として表現される．素性が単純に単語であるとき，これは単語の袋詰めという意味で “Bag of words” とよばれているが [5]，一般には素性には他の可能性も考えられるため，以下では統一して素性と表記する．

こうしたベクトル $\vec{u}, \vec{v}$ 間の距離としては，単純な内積またはユークリッド距離(注1)

$$d(\vec{u}, \vec{v})^2 = (\vec{u} - \vec{v})^T (\vec{u} - \vec{v}) \quad (1)$$

$$= \sum_i (u_i - v_i)^2 \quad (2)$$

(^{T} は転置を表す) が事前の素性重みづけとともにこれまで多くの自然言語処理の手法で用いられている [5], [6]．しかし，この距離関数は 2 つの大きな問題を持っている．

(i) 素性間の相関が考慮されていない．
(ii) 素性の最適な重みづけが決定できない．

言語データにおいては，コロケーションや構文などを通じて一般に素性間には強い相関が存在するため，(i) は特に大きな問題である．カーネル法を用いた場合 (たとえば [7])，高次の多項式カーネルなど特定のカーネルを用いることで複数の素性の組み合わせを考慮することができるが，現在このカーネル法は自然言語処理においては分類問題として多く使用されており，連続的な順序付けを必要とする情報検索や QA などを置換できる方法は提案されていない．

(ii) も実際的に重要な問題である．各素性はしばしば，単語あるいはその組み合わせであり，意味のある比較のためには内容語には強い重みを，機能語には弱い重みを，といった意味的な重みづけが必要となる．しかし，現在このために行われている tf.idf [8] などの重みは発見的なものであり，距離に関して何らかの最適化基準に依っているものではない．また，tf.idf の中にもいくつかのバリエーションがあるが，その選択に対する一般的な基準は存在していない．

(注1)：情報検索ではベクトルの長さを $|\vec{u}| = |\vec{v}| = 1$ と正規化することが一般に行われるが，このとき $(\vec{u}-\vec{v})^T(\vec{u}-\vec{v}) = |\vec{u}|^2 + |\vec{v}|^2 - 2\vec{u}\cdot\vec{v} \propto -\cos(\vec{u}, \vec{v})$ であり，ユークリッド距離による比較はコサイン距離によるものと一致する．[5]

3. 関 連 研 究

上述したような素性の相関および素性の重みづけは，機械学習においてはデータ空間において適切な計量を求める問題と考えることができ，近年特に注目されている問題である．[9] は本論文と同様の問題意識に基づいており，「似た」点のペアの集合を訓練データとして，本論文と似た計量行列を学習している．[10] および [11] は，適切な計量をそれぞれ，スペクトル法によるクラスタリングと SVM における比較データの問題設定の中で求めるものである．この意味で，本論文は [9] の自然な拡張とみなすことができる．[9] との類似点，および提案手法の優位性については 6. 章で述べる．

また，確率的生成モデルを基にデータ間の内積を与えるカーネルである Fisher kernel [12] も原理的には同じ意味を持ったものであることに注意したい．Fisher kernel の定式化においては，データの分布から期待値として導かれるフィッシャー情報量行列の逆行列が，確率モデル空間における計量を与える．ただし，この計算は計算量がきわめて大きいために，実際には単位行列で近似されることが多い．

情報検索の分野では, [13], [14] がクエリに対する適合フィードバックの立場から R-SVD (Riemannian SVD) を提案している．しかし，これは大域的な検索距離空間の改良を目指したものではなく，本論文で用いたようなクラスタ構造は用いられていない．

4. 提 案 手 法

上のような距離関数に関する問題について，われわれはデータ中に存在するクラスタ構造に注目する．ここでいうクラスタ構造とは，個々のデータを着目する観点に従って分類したものであり，実際には同一サイトの文書，同義文のクラスタ，振られたクラスラベルなどが該当する．訓練データは一般に完全に独立ではなく，データは言語の再帰的構造に従って構造化されていることも多いため，このような構造は多くの言語データに関して見られると考えられる．

各クラスタ内のデータを類似したものと見なせば，理想的なベクトル空間において，各データベクトルは集中して分布しているはずである．この性質に基づき，訓練データのクラスタ構造を用いて，最小二乗の意味で最適な距離行列を導出することができる．これにつ

いて以下に述べる．

4.1 データ分布と計量

ベクトル $\vec{u}, \vec{v}$ 間の二乗 (L_2-) 距離を，式 (1) に代わり，計量行列 $M = [m_{kl}]$ を用いて

$$d_M(\vec{u},\vec{v})^2 = (\vec{u}-\vec{v})^T M(\vec{u}-\vec{v}) \quad (3)$$

$$= \sum_k \sum_l m_{kl}(u_k - v_k)(u_l - v_l) \quad (4)$$

として求めることを考える．これは一般に画像等の分類問題において用いられるマハラノビス距離であり，$M = I$ (単位行列) の特別な場合として式 (1) を含む．式 (4) から，計量行列 M によって任意に各素性の重み，および素性間の相関が表現できることがわかる．

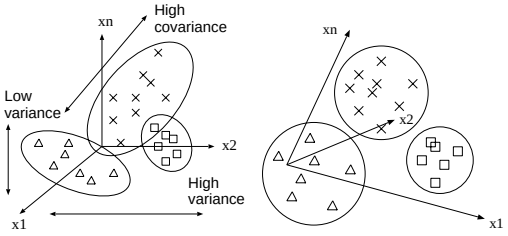
M は対称行列であり，このとき，式 (3) は式 (5) のように書き換えることができるから，

$$d_M(\vec{u},\vec{v})^2 = (M^{1/2}(\vec{u}-\vec{v}))^T(M^{1/2}(\vec{u}-\vec{v})) \quad (5)$$

この距離は，ベクトル $\vec{u}-\vec{v}$ を $M^{1/2}$ によって新しい空間へ写像し，その間のユークリッド距離を考えることとも等価である [9]．なお，このマハラノビス距離は一般的なものであり，パターン認識における用法に限られないことに注意したい．パターン認識においては一般に，複数の固定されたクラスタが正規分布をもつと仮定し，各クラスタ毎にその共分散行列の逆行列を M とすることで各クラスタへの距離を定義し，分類を行うが [15]，ここで求めるものは事前に決まったクラスへの識別問題ではなく，多数のクラスタより導かれ，一般的に用いることのできる大域的な距離計量だからである．[9]〜[11] も同じ問題意識を共有している．

したがって，ここでは訓練データ全体にわたる最適化が必要となる．本章最初の議論より，クラスタ内のデータは理想的な空間内では集中して分布しているべきと考えられるが，実際にはデータおよびその同義クラスタはユークリッド空間において図 1(a) のように楕円体型に分布しており，ある次元には高い分散を，別の次元には低い分散を持っている．また，クラスタの向きは一般にユークリッド距離における基底ベクトルには沿っていない．言語データは非常に高次元であるため，この傾向は特に顕著であると考えられる．

このとき，図 1(b) のように，$M^{1/2}$ で写像された空間において，高い分散を抑え，低い分散を拡大することでこうしたクラスタの歪みを最小化し，同義クラスタを真球に近づけるような計量 M を見出だすことができれば，その M は素性に対して適切な重みを与え，素性間の相関を適切に捉えるはずである．以下，この意味で最適な M について考える．



(a) Original space (b) Mapped space

図 1 素性空間におけるクラスタ構造．
Fig. 1 Cluster geometry of feature space.

4.2 最適な計量行列の導出

各データ (たとえば，文や文書) $\vec{x}$ を $\mathbb{R}^n$ 上のベクトルとし，その全体が N 個のクラスタ $X_1 \ldots X_N$ に分けられていると仮定する．すなわち，ベクトルの次元数は n, クラスタ数は N である．各クラスタ X_i に対して，その重心 (セントロイド) を $\vec{c_i}$ とおくと，$\vec{c_i} = 1/|X_i| \sum_{\vec{x}\in X_i} \vec{x}$ である ($|X|$ はクラスタ X 中の要素数)．

このとき，前節の議論から，各クラスタ内でデータ間の距離を最小にする計量を求める．すなわち，各クラスタ X_i に含まれるデータ $\vec{x} \in X_i$ とセントロイド $\vec{c_i}$ との距離 $d_M(\vec{x}, \vec{c_i})$ の総和を，全クラスタ $X_1 \ldots X_N$ にわたり最小にするような M を求めればよい．これは以下の二次最小化問題として定式化することができる．

最小化問題:

$$M = \operatorname*{argmin}_M \sum_{i=1}^{N} \sum_{\vec{x_j}\in X_i} d_M(\vec{x_j}, \vec{c_i})^2$$

$$= \operatorname*{argmin}_M \sum_{i=1}^{N} \sum_{\vec{x_j}\in X_i} (\vec{x_j} - \vec{c_i})^T M (\vec{x_j} - \vec{c_i}) \quad (6)$$

規格化条件 : $|M| = 1$. (7)

規格化条件は，$M = 0$ となる縮退した解をもたないための条件である．($|\cdot|$ は行列式を表す．) 右辺の 1 は任意の定数であり，これを c とすれば $c^2 M$ が新しい解となる．

この最適化問題はデータベースの分野で提案された *MindReader* [16] の考えを複数クラスタに拡張したものであり，以下の一意な解を持つ．

[定理] 条件式 (7) の下で (6) 式の最小化問題を満たす行列 M は

$$M = |A|^{1/n} A^{-1} \tag{8}$$

である．ここで，$A = [a_{kl}]$ は以下で定義される行列．

$$a_{kl} = \sum_{i=1}^{N} \sum_{\vec{x_j} \in X_i} (x_{jl} - c_{il})(x_{jk} - c_{ik}). \tag{9}$$

[証明] 付録 1 を参照．□

(8) 式の $|A|^{1/n}$ は定数であるから，これは上記最適化問題の解は，各クラスタの分散-共分散行列の総和 (平均) の逆行列になっていることを意味し，直観的にも妥当な結果である．式 (9) は全クラスタにわたる総和となっているから，この計量行列 M は大域的に分散の大きい軸を縮小し，分散の小さい軸を拡大することでデータの分散を安定化する働きを持っていることがわかる．

ただし，一般に言語データにおいては素性は高次元かつ非常に疎であり，分散-共分散行列の和 A は正則ではないことが多い．したがって, [16] と同様に，われわれは A^{-1} として Moore-Penrose の擬逆行列 A^{+} を用いた．詳細については付録 2 を参照のこと．

4.3 クラスタ重みを用いた一般化

前節でわれわれは各クラスタを同等に扱ったが，一般には含まれるデータ数に応じてクラスタには強弱があり，階層的クラスタリングにおいては上位クラスタほどその意味は薄まると考えられる．この情報は正規化された各クラスタの重み $\xi_1 \ldots \xi_N$ $(\sum_i \xi_i = 1)$ を用いて，最小化する式 (6) を以下のように一般化することで実現できる．

$$\arg\min_{M} \sum_{i=1}^{N} \xi_i \sum_{\vec{x_j} \in X_i} (\vec{x_j} - \vec{c_i})^T M (\vec{x_j} - \vec{c_i}) \tag{10}$$

これにより，式 (9) を同様に重みづけた解が得られる．ただし，以下の実験では各クラスタに含まれるデータ数がほぼ等しいため，この一般化は用いていない．

なお，提案手法は $N = 1$ の特別な場合として，通常のパターン認識におけるマハラノビス距離を含んでいることに注意されたい．

5. 実　　験

提案手法の自然言語処理における効果を検証するため，同義文検索，文書検索，および機械学習用ベクトルデータのクラスタリングタスクを用いて実験を行った．提案手法は汎用性を持つため，後者のような非言語データにも適用することができる．

実験の基本的な手順は以下の通りである．まず，訓練データからクラスタ構造を用いて計量行列 M を計算する．次に検索タスクにおいては，テストデータ中の各ベクトルに対し残りのベクトルとの距離を計算し，昇順にソートする．そのリスト中で元のベクトルと同クラスタに含まれるものを正解データとし，再現率–適合率および R-精度を求める．ここで R を正解データ数ととれば，R-精度は計算結果の上位 R 個がすべて正解データであるとき 1, 正解が全く含まれないときに 0 となり，クラスタの復元精度を表現する．上位 R 個以下の正解分布は再現率–適合率曲線およびその 11 点平均精度によって示される．

この計算をテストデータ中のすべてのベクトルについて行い，平均を求めた．このため，テストデータ数 n に対し，その計算量は $O(n^2)$ である．クラスタリングタスクでの精度測定については，5.3 節で述べる．

以上での距離計算において，計量を用いない場合 (ベースライン) および用いた場合を比較した．

5.1 同義文検索

ある文に意味的に類似した文をコーパスあるいは用例文集合から検索する問題は，用例ベース機械翻訳や QA における質問文からの解答候補の検索など，自然言語処理において基礎的な技術となっている．

5.1.1 同義文コーパス

このような同義文検索実験のために，われわれは ATR で開発された旅行会話ドメインのパラフレーズコーパス [17] を用いた．このコーパスは 33,723,164 個の和文からなり，それぞれは 10,610 個の英文の一つと翻訳関係で対応している．この対応により，ある英文の翻訳である和文集合を同義文クラスタとみなすことができる．この中から，われわれは 200 個の訓練クラスタと 50 個のテストクラスタからなるデータセットをランダムに非復元抽出して別々に作成した．1 つのクラスタに属する文は最大 100 文とし，これを超える時はクラスタ内より 100 文をランダムに選んだ．このデータセットをさらに 10 個作成し，結果を平均した．

5.1.2 素性と次元圧縮

文の素性としては，ユニグラムおよび機能語のバイグラムを用いた．機能語バイグラムを含めたのは，会話文ドメインであるために，機能語の連接が言い換え

において大きな役割を果たすと考えられるからである．(注2)

旅行会話コーパスのため語彙数は比較的制限されてはいるが，素性の総数はデータ量に応じて数千から数万を超えるため，直接計量行列 M を求めることは現実的でなく，また素性が疎であるために，求めた計量も不安定になりやすい．このため，あらかじめ素性を特異値分解によって次元圧縮し，$1, 5, 10, 20, 50\%$ の各圧縮率まで削減した．これは本質的に LSI [18] と同じ方法であり，各ベクトル間の内積を最小二乗の意味で最適に保存する．

5.1.3 結　　果

提案手法を用いた同義文検索結果の例を付録図 A·1 に示す．計量行列を用いた検索では通常のユークリッド距離に比べてノイズが少なく，高精度な検索を実現できることがわかる．図 A·2 は次元圧縮をしすぎた場合 (圧縮率=0.5%)，次元間に混入が生じている例であるが，この場合でもほとんど無意味な結果を与える従来手法に比べ，本手法では上位に適切な結果が含まれており，安定した検索性能を見せることがわかる．

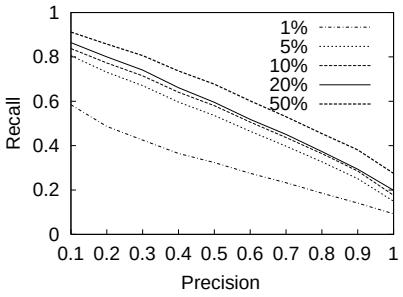
図 2 に各圧縮率における再現率および適合率を示す．これらの結果および図 3 の 11 点平均精度(注3) は，提案手法が通常用いられる単純な内積，および idf で重み付けた内積に比べ，常に高い検索性能を持っていることを示している．

5.1.4 計 量 行 列

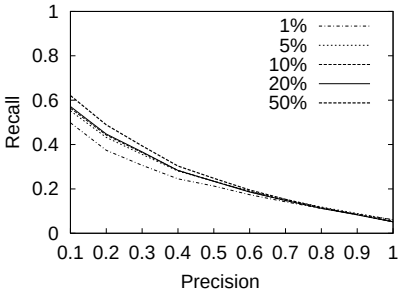
図 4 に，同義文クラスタセットから得られた計量行列の一つを示す．素性はあらかじめ idf で重みづけられた上でここでは 200 次元に圧縮されており，通常のユークリッド (コサイン) 距離は，ここで左下–右上の対角成分のみ 1 の単位行列に対応する．図より明らかに，非対角成分に多くの正負の重みが割り当てられており，(圧縮された) 素性間の正/負の相関が表現されていることがわかる．対角成分の絶対値の総和は，全行列での総和に対し 3.5% にすぎなかった．また，対角成分についてもその値は一様ではなく，idf による重みがさらにクラスタ間相関によって適切に変えられていることがわかる．図 5 に，素性自身の重みである対角成分のプロットを示す．

(注2)：ここでは振られた品詞タグに基づき，名詞・固有名詞・数詞・本動詞を内容語，それ以外の語を機能語とした．トライグラム以上の素性は素性数が指数的に増大し，複数のバイグラムによって近似的に表現可能と考えられるために採用しなかった．

(注3)：R-精度はここでは 11 点平均精度とほとんど同様であったため，図は省略した．



(a) 提案手法



(b) ユークリッド距離

図 2　再現率—適合率曲線 (同義文検索)
Fig. 2　Precision-Recall Curve on sentence retrieval

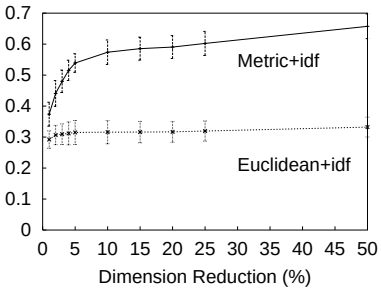


図 3　11 点平均精度 (同義文検索)
Fig. 3　11 Point Average Precision on sentence retrieval

5.2 文 書 検 索

文書をクラスタに分類するタスクとして文書分類 (Text Classification) があり，Naïve Bayes 法や SVM など様々な識別器を用いた研究が盛んに行われているが [19]，これらはみな文書を事前に決められた少数のカテゴリに分類するものであり，新しいカテゴリ (クラスタ) に対しては識別器を構成することができない．例えば，Web サイトをクラスタとみなすと，文書に対して可能なクラスタは非常に多数存在し，常に増え続ける性質があり，これらに対して 1 つ 1 つ識別器を構

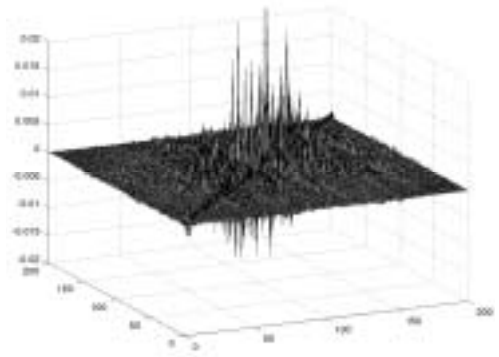


図 4　同義文クラスタからえられた計量行列
Fig. 4　A metric matrix obtained from synonymous clusters.

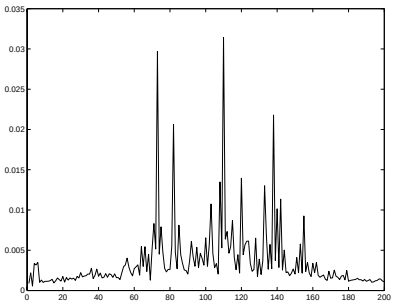


図 5　計量行列の対角成分
Fig. 5　Diagonal elements of the metric matrix.

成するのは現実的でない．このような環境における検索あるいはクラスタリングには大域的な距離尺度が必要であり，提案手法が有効だと考えられる．

5.2.1　20-Newsgroup データセット

このためのデータセットとして，われわれは 20–Newsgroup dataset [20] を用いた．これは標準的なテキスト分類のデータセットの中では，比較的多い 20 のクラスタ (ニュースグループ) を持つ．20 のクラスタの中から 16 クラスタを訓練データ，4 クラスタをテストデータとし，5 分割交差検定を行った．1 クラスタ当たりの文書数は最大 100 文書とし，これを越える場合は非復元抽出によってクラスタからランダムに 100 文書を抽出した．素性はユニグラムとし，同様に 0.5% ～ 20% までの次元圧縮を行った．

提案手法はデータ空間におけるベクトルの分布から最適な計量を求めるために，言語データの場合，文書の長さが非常に異なると各ベクトルのノルムが異なるため，適切な計量が得られない(注4)．このため，各文書は中央値 (130 語) になるまでサブサンプリング／オーバーサンプリングを行った．また，ベースラインとして tf.idf による単語重み付けを適用した．

Dim.	R-precision		11-pt Avr. Prec.	
Red.	Metric	Euclid	Metric	Euclid
0.5%	**0.421**	0.399	**0.476**	0.455
1%	**0.388**	0.368	**0.450**	0.430
2%	**0.359**	0.343	**0.425**	0.409
3%	**0.344**	0.330	**0.411**	0.399
4%	**0.335**	0.323	**0.402**	0.392
5%	**0.329**	0.318	**0.397**	0.388
10%	**0.316**	0.307	**0.379**	0.376
20%	**0.343**	0.297	**0.397**	0.365

表 1　文書検索精度．
Table 1　Newsgroup text retrieval precisions.

5.2.2　結　　果

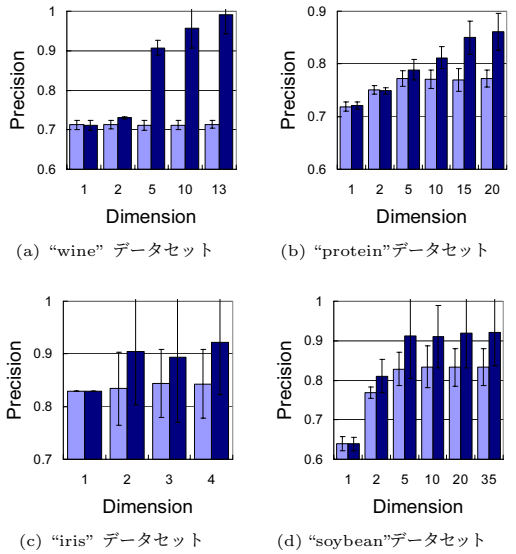
表 1 に R-精度および 11 点平均精度を示す．テストセットは 4 クラスタからなるため，精度のベースラインは 0.25 である．tf.idf およびユークリッド距離に比べ，両方の基準で常によい精度を見せることがわかる (p 値の平均 $= 0.0243$)．

ただし，精度の改善は同義文検索に比べて少ない．この理由の一つは素性圧縮にあると考えられる．われわれはデータ行列 X を最初に $X = USV^{-1}$ と特異値分解した後，k 個の最大固有値に対応する V の部分行列 V_k を用いて $X_k = V_k X$ とし，k 次元へ素性を圧縮したが，式 (5) からこれは，$M^{1/2}X_k = M^{1/2}V_k X$ の各列間におけるユークリッド距離とみなすことができる．ゆえに，クラスタ化が弱い場合，前処理において V_k が M の役割を吸収してしまう可能性がある．したがって最適な性能のためには，高次元データに対しては計量の導出と次元圧縮を同時に考える必要のあることがわかる．また，次元の呪いの存在しないカーネル法を用い，そのヒルベルト空間において同じ基準を考察することも考えられる．

5.3　一般ベクトルデータおよびクラスタリング

計量を用いた距離は情報検索だけでなく，非言語データやクラスタリングにおいても適用することができる．図 6 に，UCI 機械学習データセット [22] における K 平均クラスタリングに適用した結果を示す．右の棒が計量距離，左の棒がユークリッド距離である．K はデータ中のクラスタ数にとり，ランダムに初期化して 100 回行い平均をとっている．クラスタリング精度

(注4)：情報検索で一般に行われるように，ノルムを 1 に正規化する方法では，高次元の超球面上にすべてのデータが写像されるため，本手法は意味のある計量を与えないことがわかった．Spherical K-means [21] のような超球面上での手法に適した計量を求めることは今後の課題である．



(a) “wine” データセット　(b) “protein”データセット

(c) “iris” データセット　(d) “soybean”データセット

図 6　UCI 機械学習データセットのクラスタリング精度. 横軸は圧縮した次元数を表す (右端は圧縮なし).

Fig. 6　K-means clustering of UCI Machine Learning dataset results. Horizontal axis shows compressed dimensions (rightmost is original).

は，ランダムに選んだ 2 つのデータが正しいクラスタリング (同クラスタ/別クラスタ) を持つ確率として [9] と同様に計算した．

文書データである 20–Newsgroup データセットについても 5.2.1 節と同様の設定でクラスタリング実験を行い，表 1 と同程度の精度の上昇を得ている．

6. 考　　察

本論文では，訓練データのクラスタ構造に基づいた最適な計量距離を提案した．自然言語処理においてベクトル間の距離は長く用いられてきたが，データの分布に基づき，その空間の最適な計量を見出だすことはこれまで比較的看過されており，最近になって機械学習において注目されている問題である．最近提案されたスペクトル法や SVM の下での計量と比べ，提案手法はそうした特定の手法を前提とせず，ベクトル空間において一般にユークリッド距離に代わり利用することができる．こうした意味で提案手法は [9] の後継手法であるといえる．ただし, [9] では「似た」点のペア $(\vec{x_i}, \vec{x_j})$ の集合 S を訓練データとしているため，計量の導出にはデータ数 n に対して $O(n^2)$ のペア数が必要であり，データの持つ情報をすべて用いることは難しく，ニュートン法を用いて繰り返し最適化を行うために計算量も大きい．これに対し，提案手法は線形演算のみを用いたもので高速であり，最適な計量をデータすべてを用いて一度で求めることができる．

今後の課題としてはまず，4.3 節で述べたクラスタ重みの最適な導出が考えられる．クラスタ化は一般には等質ではないため，これによりさらに適切な計量が得られると期待できる．5.2.2 節で次元削減のもつ影響について述べたが，クラスタ構造を考慮したこうした次元削減のほかに, [11] のようにカーネル法の枠組の中で，提案手法の基準である訓練データのクラスタ歪み最小化を行うこともあげられる．この方法は次元削減を必要としないが，適用範囲がカーネル化できる問題に限られるという欠点があり，一般のベクトル空間における提案手法の意味はあると考えられる．

7. む　す　び

本論文ではベクトル空間において，ユークリッド距離の代替として使用できる最適な計量距離を提案した．この距離は訓練データのクラスタ構造に基づき，クラスタ歪みを最小化する二次最適化問題の解として解析的に導出される．同義文検索，文書検索，および一般機械学習データのクラスタリングタスクを用いて提案手法の効果を確認した．

謝辞　本研究は独立行政法人 情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

文　　献

[1] M. Hearst: “Multi-paragraph segmentation of expository text”, 32nd. Annual Meeting of the Association for Computational Linguistics, pp. 9–16 (1994).

[2] F. Y. Y. Choi: “Advances in domain independent linear text segmentation”, Proceedings of NAACL-00 (2000).

[3] M. Collins and N. Duffy: “Convolution Kernels for Natural Language”, NIPS 2001 (2001).

[4] J. Suzuki, T. Hirao, Y. Sasaki and E. Maeda: “Hierarchical Directed Acyclic Graph Kernel: Methods for Structured Natural Language Data”, 41th Annual Meeting of Association for Computational Linguistics, pp. 32–39 (2003).

[5] C. D. Manning and H. Schütze: “Foundations of Statistical Natural Language Processing”, MIT Press (1999).

[6] R. A. Baeza-Yates and B. A. Ribeiro-Neto: “Modern Information Retrieval”, ACM Press / Addison-Wesley (1999).

[7] K. R. Müller, S. Mika, G. Ratsch and K. Tsuda: “An introduction to kernel-based learning algorithms”, IEEE Neural Networks, **12**, 2, pp. 181–201 (2001).

[8] G. Salton and C. S. Yang: “On the specification of term values in automatic indexing”, Journal of Documentation, **29**, pp. 351–372 (1973).
[9] E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell: “Distance metric learning, with application to clustering with side-information”, NIPS 2002 (2002).
[10] F. R. Bach and M. I. Jordan: “Learning Spectral Clustering”, NIPS 2003 (2003).
[11] M. Schultz and T. Joachims: “Learning a Distance Metric from Relative Comparisons”, NIPS 2003 (2003).
[12] T. S. Jaakkola and D. Haussler: “Exploiting generative models in discriminative classifiers”, Proc. of the 1998 Conference on Advances in Neural Information Processing Systems, pp. 487–493 (1999).
[13] E. P. Jiang and M. W. Berry: “Information Filtering Using the Riemannian SVD (R-SVD)”, Proc. of IRREGULAR ’98, pp. 386–395 (1998).
[14] B. De Moor: “Structured total least squares and L2 approximation problems”, Systems & Control, Special Issue of Linear Algebra and its Applications on Numerical Linear Algebra, **188–189**, pp. 163–207 (1993).
[15] R. O. Duda, P. E. Hart and D. G. Stork: “Pattern Classification *Second Edition”, John Wiley & Sons (2000).
[16] Y. Ishikawa, R. Subramanya and C. Faloutsos: “MindReader: Querying Databases Through Multiple Examples”, Proc. 24th Int. Conf. Very Large Data Bases, pp. 218–227 (1998).
[17] F. Sugaya, T. Takezawa, G. Kikui and S. Yamamoto: “Proposal for a very-large-corpus acquisition method by cell-formed registration”, Proc. LREC-2002, Vol. I, pp. 326–328 (2002).
[18] S. Deerwester, S. T. Dumais and G. W. Furnas: “Indexing by Latent Semantic Analysis”, Journal of the American Society of Information Science, **41**, 6, pp. 391–407 (1990).
[19] T. Joachims: “Text categorization with support vector machines: learning with many relevant features”, Proceedings of ECML-98, No. 1398, pp. 137–142 (1998).
[20] K. Lang: “Newsweeder: Learning to filter netnews”, Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339 (1995).
[21] I. S. Dhillon and D. S. Modha: “Concept Decompositions for Large Sparse Text Data Using Clustering”, Machine Learning, **42**, 1/2, pp. 143–175 (2001).
[22] C. L. Blake and C. J. Merz: “UCI Repository of machine learning databases” (1998). http://www.ics.uci.edu/˜mlearn/MLRepository.html.
[23] E. W. Weisstein: “Moore-Penrose Matrix Inverse” (2004). http://mathworld.wolfram.com/Moore-PenroseMatrixInverse.html.

付　　録

1. 計量行列の導出

4.2 節の定理を証明する．すなわち，最小化問題

$$\min_{M} \sum_{i=1}^{n} \sum_{\vec{x}_j \in X_i} (\vec{x}_j - \vec{c}_i)^T M (\vec{x}_j - \vec{c}_i) \qquad \text{(A·1)}$$

を条件

$$|M| = 1 \qquad \text{(A·2)}$$

の下で満たす計量行列 M を求める．

(A·1) を展開すると

$$\sum_i \sum_{\vec{x}_j} \left[\sum_{k=1}^{n} \sum_{l=1}^{n} (x_{jk} - c_{ik}) m_{kl} (x_{jl} - c_{il}) \right] \qquad \text{(A·3)}$$

である．ここで条件 (A·2) より，すべての k について

$$\sum_{l=1}^{n} (-1)^{k+l} m_{kl} |M_{kl}| = 1$$

すなわち，

$$\sum_{k=1}^{n} \sum_{l=1}^{n} (-1)^{k+l} m_{kl} |M_{kl}| = n \qquad \text{(A·4)}$$

となる．ここで $|M_{kl}|$ は M の第 (k,l)-小行列式である．したがって，(A·3) を (A·4) の条件の下で最小化すればよい．

ラグランジュ乗数 λ を導入することにより，

$$L = \sum_{i=1}^{N} \sum_{\vec{x}_j} \left[\sum_k \sum_l (x_{jk} - c_{ik}) m_{kl} (x_{jl} - c_{il}) \right] - \lambda \left[\sum_k \sum_l (-1)^{k+l} m_{kl} |M_{kl}| - n \right]$$

と定義すると，m_{kl} で微分して 0 とおくことにより，

$$\frac{\partial L}{\partial m_{kl}} = \sum_i \sum_{\vec{x}_j} (x_{jk} - c_{ik})(x_{jl} - c_{il}) - \lambda (-1)^{k+l} |M_{kl}| = 0$$

$$\Leftrightarrow \ |M_{kl}| = \frac{\sum_i \sum_{\vec{x}_j} (x_{jk} - c_{ik})(x_{jl} - c_{il})}{\lambda (-1)^{k+l}}. \qquad \text{(A·5)}$$

ここで $M^{-1} = [m_{kl}^{-1}]$ とおくと，

$$m_{kl}^{-1} = \frac{(-1)^{k+l}|M_{kl}|}{|M|} = (-1)^{k+l}|M_{kl}| \ (\because \text{A·2})$$
$$= \frac{\sum_i \sum_{\vec{x}_j} (x_{jk} - c_{ik})(x_{jl} - c_{il})}{\lambda}. \quad (\text{A·6})$$
$$(\because \text{A·5})$$

したがって，

$$A = [a_{kl}] \quad (\text{A·7})$$

$$a_{kl} = \sum_{i=1}^{N} \sum_{\vec{x}_j \in X_i} (x_{jl} - c_{il})(x_{jk} - c_{ik}) \quad (\text{A·8})$$

と定義すると，(A·6) により

$$A = \lambda M^{-1}$$
$$\therefore |A| = \lambda^n |M^{-1}| = \lambda^n$$
$$\therefore \quad \lambda = |A|^{1/n}. \quad (\text{A·9})$$

ゆえに

$$M = \lambda A^{-1} = |A|^{1/n} A^{-1}. \quad (\text{A·10})$$

ここで A は (A·7), (A·8) で定義される行列. □

2. Moore-Penrose の擬逆行列

行列 A の Moore-Penrose の擬逆行列 A^+ は A が非正則な場合でも，$x = A^+ y$ が $y = Ax$ の最小二乗かつ最短の解となるという意味で通常の逆行列の性質を持つ一意な行列である [23].

A^+ は MATLAB 関数 `pinv` 等を用いて簡単に求めることができる．または [16]，正規直交行列 U および $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$ $(r = \mathrm{rank}(A))$ を用いて，A を $A = U\Sigma U^T$ と対角化すれば，A^+ は $A^+ = U\Sigma^+ U^T$ と求められる．
ここで $\Sigma^+ = \mathrm{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0, \ldots, 0)$ である．したがって，

$$M = (\sigma_1 \sigma_2 \cdots \sigma_r)^{1/r} A^+ \quad (\text{A·11})$$

を得る． □

（平成 xx 年 xx 月 xx 日受付）

Query: “合計でいくらですか”
(‘How much is the total?’)

Metric distance:

distance	synonymous sentence
0.2712	合計でいくらでしょうか *
0.3444	内金はいくらですか
0.3444	入場料はいくらですか
0.369	手付金はいくらですか
0.4377	合計でいくらいたしますか *
0.4479	合計でいくらいたしますでしょうか *
0.4505	全部でいくらですか *
0.4558	合計でいくらになりますか *
0.4602	合計でいくらになりますでしょうか *
0.4682	合計でいくらになるでしょうか *

Euclidean distance:

distance	synonymous sentence
0.1732	全部でいくらですか *
1.781	合計でおいくらですか *
1.902	紫外線防止ですか
1.966	内金はいくらですか
1.966	入場料はいくらですか
1.974	手付金はいくらですか
1.983	全部でおいくらですか *
2.283	どんな兆候ですか
2.505	どんな症状ですか
2.65	お一人ですか

(* denotes the right answers.)

図 A·1 同義文検索の例.
Fig. A·1 Example of Sentence Retrieval.

Query: “デザートに果物をくれないでしょうか”
(‘I’d like some fruit for dessert.’)

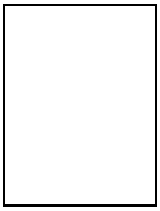
Metric distance:

distance	synonymous sentence
0.3531	請求書をすぐにくれないでしょうか
0.3709	デザートとして果物をくれますか *
0.596	請求書をすぐにくれませんか
0.6104	伝票をすぐにくれますか
0.621	伝票をすぐにくれますでしょうか
0.6255	お勘定書をすぐにくれますか
0.6295	伝票をすぐにくれませんでしょうか
0.6343	お勘定書をすぐにくれませんか
0.6685	伝票をすぐにくれないですか
0.7966	デザートには果物をくれないですか *

Euclidean distance:

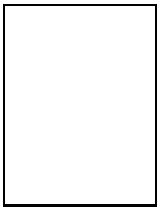
distance	synonymous sentence
1.036	請求書をすぐにくれないでしょうか
1.421	朝ごはんを部屋に運んでもらえないでしょうか
1.491	ウィスキーを二人分くれないでしょうか
1.499	ウィスキーを二つくれないでしょうか
1.535	薬をくれないでしょうか
1.622	朝食を部屋に運んでもらえないでしょうか
1.622	朝食を部屋に運んでもらえないでしょうか
:	:
2.787	デザートとして何か果物をくれないでしょうか *
2.854	この円をポンドに換算くださらないでしょうか

図 A·2 次元削減率の高い場合.
Fig. A·2 High rate of dimensionarity reduction.



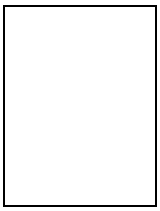
持橋　大地

1998 年 東京大学教養学部 基礎科学科第二卒業．2000 年 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座 博士前期課程修了．現在同講座 博士後期課程在学中．自然言語処理，特にベイズ統計的アプローチによる意味的処理と時系列モデルに関心を持つ．2003 年より ATR 音声言語コミュニケーション研究所 研修研究員．



菊井玄一郎

1986 年京都大学工学部電気工学第二専攻修士課程修了．同年 NTT に入社，2001 年 4 月より（株）国際電気通信基礎技術研究所 (ATR) に出向，現在に至る．在学中より、自然言語処理，音声言語処理，特に音声翻訳，WEB 情報検索、多言語情報検索等の研究開発に従事．ACL，人工知能学会，言語処理学会各会員．



北　　研二　（正員）

1981 年 早稲田大学 理工学部 数学科卒業．1983 年から 1992 年まで沖電気工業 (株) 勤務．この間，1987 年から 1992 年まで ATR 自動翻訳電話研究所に出向．1992 年 9 月より徳島大学工学部勤務．現在，同教授．工学博士．自然言語処理，情報検索等の研究に従事．電子情報通信学会，言語処理学会 各会員．

Abstract Many natural language processing still depend on the Euclidean distance function between the two feature vectors, but it has severe defects as to feature weightings and feature correlations. In this paper we propose an optimal metric distance function that can be used as an alternative to the Euclidean distance, accommodating the two problems at the same time. This metric is optimal in the sense of global quadratic minimization, and can be obtained from the clusters in the training data in a supervised fashion. We confirmed the effect of the proposed metric by the sentence retrieval, document retrieval, and the K-means clustering of general vectorial data.

Key words Metric learning, cosine distance, tf.idf, feature space, kernel methods