編集距離に基づくアノテーション付き可視化法の提案

大森　美香† 　伏見　卓恭† 　斉藤　和巳†

† 静岡県立大学 経営情報学部　〒 422–8526 静岡県静岡市駿河区谷田 52–1
E-mail: †{b09025,j11507,k-saito}@u-shizuoka-ken.ac.jp

あらまし　大規模なデータを理解するとき，その構造や特徴を直感的かつ視覚的にとらえることができる可視化は有用な技術である．そのため，これまでに多くの可視化法が提案されてきたが，それらの結果の解釈が困難な場合もある．本論文では，編集距離（レーベンシュタイン距離）に基づく可視化に着目し，アノテーション法により，低次元に埋め込まれたオブジェクト集合に対して，どの辺りにどのような共通の特徴を持つオブジェクトが布置されているかを自動的に示す方法を提案する．評価実験の結果により，この提案法を用いることで，可視化結果に対し，特定アルファベットを多く含む単語の配置方向として適切にアノテーションが付与できることを示す．

キーワード　アノテーション法，可視化，編集距離

Proposing a Visualization Method with Annotation based on Edit Distance

Mika OMORI†, Takayasu FUSHIMI†, and Kazumi SAITO†

† School of Management and Information, University of Shizuoka
52–1 Yada, Suruga–ku, Shizuoka, 422–8526 Japan
E-mail: †{b09025,j11507,k-saito}@u-shizuoka-ken.ac.jp

Abstract　Visualization by embedding objects into a low dimensional Euclidean space plays an important role to intuitively understand the essential structure of objects, and various object embedding methods have been proposed, but it is sometimes difficult to adequately interpret those visualization results. To alleviate this difficulty, we newly propose an annotation method based on a feature vector of objects. In this paper, we evaluate the proposed method by using two sets of words each of whose similarity measure is defined by the edit (Levenshtein) distance. In our experiments using feature vectors consisting of alphabet frequency for these words, we show that our method can clearly indicates the suitable directions that certain alphabets cluster on the object embedding space.

Key words　annotatition,Visualization,edit distance

1. はじめに

我々が大規模かつ複雑なデータを理解し把握しようとするとき，その膨大さ，煩雑さゆえに困難である場合がある．そのような場合，データが有する特徴や構造を理解する有効な表現のひとつとして「可視化」がある．可視化することにより，対象となるデータ間の相互関係や特徴などを分かりやすく直感的かつ視覚的に把握することができる．そのため，可視化は重要であり，今までに様々な分野での可視化法が提案されている [1] [2] [3].

本論文では，アノテーションを考慮した可視化法について検討する．ここで，アノテーションとは，いわゆる注釈のことである．低次元に埋め込まれたオブジェクト集合に対して，どの辺りにどのような共通の特徴を持つオブジェクトが布置されているかを自動的に示すアノテーションは，その構造を適切に理解する上でも，多くの可視化法にとって有望な技術であり，重要な研究課題といえる．本研究では，各オブジェクトの埋め込み座標ベクトルとともに，オブジェクトを記述する属性ベクトルが与えられたとき，これらベクトル間の相関を最大にすることにより，その属性ベクトルを可視化空間に埋め込んだときの方向と相関を得る方法を提案する．

本論文の構成は以下の通りである．第 2 章でアノテーション法について述べる．第 3 章で編集距離を用いた評価実験をする．さらに第 3 章では実験結果の考察も述べる．最後に，本研究のまとめを第 4 章で述べる．

2. アノテーション法

N 個のオブジェクト集合 $\{o_1, \cdots, o_N\}$ に対して，これらを

J-次元空間に埋め込んだ座標ベクトル $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ が与えられたとする．ここで，座標ベクトルは平均が $\mathbf{0}$ に，各座標値の自乗和が 1 となるように正規化されているとする．すなわち，$\sum_{n=1}^{N} \mathbf{x}_n = \mathbf{0}$，任意の整数 j (ただし，$1 \leqq j \leqq J$) で $\sum_{n=1}^{N} x_{n,j}^2 = 1$ である．以下では，座標ベクトルを並べて構成した $J \times N$ 行列を $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$ で表記する．一方，アノテーション対象とするオブジェクトの属性値を $\{y_1, \cdots, y_N\}$ とし，これらを要素とする N-次元属性値ベクトルを $\mathbf{y}$ で表記する．本論文のアノテーション法は，属性値ベクトル $\mathbf{y}$ を D-次元空間に埋め込む問題として定式化される．

いま，J-次元の射影ベクトルを $\mathbf{z}$ とする．ここで，$\|\mathbf{z}\| = \sum_{n=1}^{N} z_j^2 = 1$ とする．このとき，ベクトル $\mathbf{z}$ 上への座標ベクトル群の射影値から構成される N-次元横ベクトルは $\mathbf{z}^T\mathbf{X}$ となる．よって，属性値ベクトル $\mathbf{y}$ をアノテートする妥当な方向として，次式を最大にする射影ベクトル $\mathbf{z}$ を考える．

$$F(\mathbf{z}) = \mathbf{z}^T\mathbf{X}\mathbf{y}. \tag{1}$$

ここで，多次元尺度構成法による埋め込み結果においては，解として求める固有ベクトル群の直交性と，上述した $\sum_{n=1}^{N} x_{n,j}^2 = 1$ の正規化により，$\mathbf{X}\mathbf{X}^T = \mathbf{I}_J$ となる．ここで，$\mathbf{I}_J$ は J-次元の単位行列を表す．すなわち，多次元尺度構成法の結果では，式 (1) で定義した $F(\mathbf{z})$ は，ベクトル $\mathbf{z}^T\mathbf{X}$ と $\mathbf{y}$ の相関係数と等価になることが容易に確認できる．よって，式 (1) で定義した $F(\mathbf{z})$ を以下では簡単に相関と呼ぶ．

式 (1) の相関を最大化する $\hat{\mathbf{z}}$ はラグランジュ乗数法より以下となる．

$$\hat{\mathbf{z}} = \frac{1}{\|\mathbf{X}\mathbf{y}\|}\mathbf{X}\mathbf{y}. \tag{2}$$

一方，式 (2) を式 (1) に代入すれば以下を得る．

$$F(\hat{\mathbf{z}}) = \|\mathbf{X}\mathbf{y}\|. \tag{3}$$

よって，属性値ベクトル $\mathbf{y}$ を D-次元空間に埋め込むアノテーションとして，その方向と相関を，それぞれ式 (2) と 式 (3) で規定する次式のベクトル（矢印）を提案する．

$$\mathbf{a}(\mathbf{y}) = \mathbf{X}\mathbf{y}. \tag{4}$$

明らかに，属性値ベクトル $\mathbf{y}$ に対して，式 (2) の矢印が長ければ相関が高く有意なアノテーションと言えるが，矢印が短ければアノテーションが困難なことを意味する．

3. 実験による評価

本論文で提案するアノテーション法はある方向に布置されたオブジェクト群が有する特徴を矢印により示すための方法である．適切にアノテーションできているかを方向と長さの観点から評価する．評価に用いる実験データ，実験設定，そして実験結果について述べる．

3.1 実験データ

本論文で取り扱う実験は，2 つのデータを用いて行った．

1つ目のデータは,「東北大・松下単語音声データベース Vol.5」に含まれるデータを抽出したものである．このデータベースに含まれる 3263 個の日本の駅名をローマ字表記している．以下このデータを EKIMEI と呼ぶ.

2つ目のデータは，提示した単語からどのような単語を連想するのかを調査したものである．7205 単語をローマ字表記している．以下このデータを EITANGO と呼ぶ.

3.2 実験設定

方向と相関を得るために以下の分析方法を用いる．

3.2.1 編集距離

単語間の類似度を編集距離で定義する．編集（レーベンシュタイン）距離とは，2 つの文字列の異なり具合を示す数値であり，それによって文字列の類似度を測る．文字の削除や挿入，置換により最少回数の操作を数えたものである [4].

- **削除**　1 つの文字を取り除く
- **挿入**　1 つの文字を加える
- **置換**　1 つの文字を置き換える

以上の 3 つの操作のいずれかを行うごとに 1 回と数える．

単語 a と単語 b の正規化編集類似度 $\mathcal{L}(a,b)$ は $L(a,b)$ より以下のように計算する．

$$\mathcal{L}(a,b) = 1 - L(a,b)/l(a,b)$$

$$l(a,b) = \begin{cases} leng(a) & (leng(a) \geqq leng(b) \text{ のとき}) \\ leng(b) & (leng(b) > leng(a) \text{ のとき}) \end{cases} \tag{5}$$

ここで $leng(a)$ は単語 a の長さである．データを定量化するためにこちらの正規化編集類似度で分析する．

3.2.2 MDS

多次元尺度構成法（MDS：Multi Dimensional Scaling）は距離データにおいて類似したものを近くに，異なっているものは遠くに配置し，それぞれの座標をとる手法である．対象を点の布置で表現する方法である [5]．詳しくいうと，目的関数を最大化する座標ベクトルを求める．

3.2.3 属性値ベクトル構成法

アノテーション法を評価するために，単語集合に対する基本的な属性ベクトル $\mathbf{y}$ として，各単語に出現するアルファベットの頻度ベクトルを用いる．それぞれのデータセットでは，A,O など異なるアルファベットとして，EKIMEI では 23 種類，EITANGO では 42 種類が存在する．実験では、これらを個別の属性ベクトルとしてアノテーションすることを試みる．

3.2.4 色付け例

本研究では，MDS 法で布置した点に，それぞれの単語に対象のアルファベットがいくつ含まれているかによって色付けして評価する．例えば，単語 a に対象のアルファベットが 2 つ以上含まれているとしたら，赤く点が色付けされる．1 つのときは緑，含まれていなかったら青くなる．

3.3 実験結果

相関の分析結果について述べるとともに，EKIMEI と EITANGO でのアノテーション付き可視化の例を示す.

3.3.1 相関の分析結果

本論文では，適切にアノテーションできていることを示すために布置した点に色付けをする．ここで，各データのアノテー

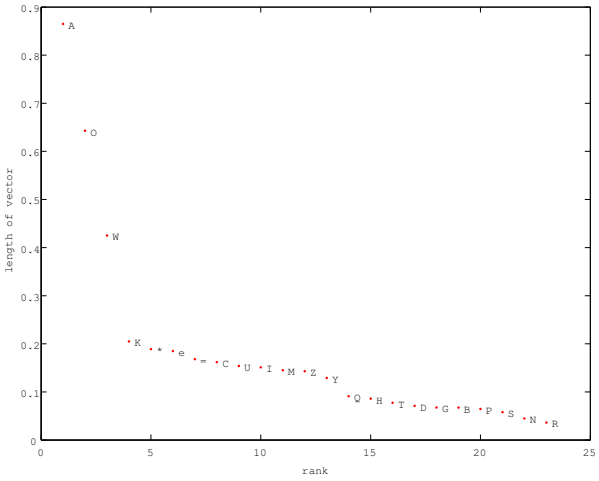

図 1　EKIMEI でのアノテーション適切度

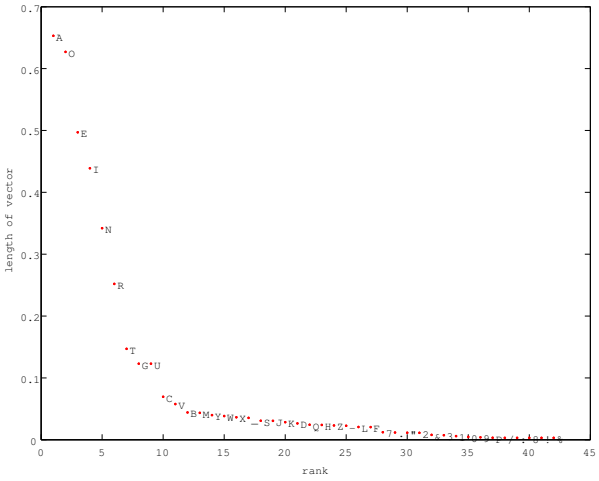

図 2　EITANGO でのアノテーション適切度

ション適切度を，図 1，図 2 に示す．本研究で提案するアノテーション法は，類似した属性値を有するオブジェクト群が一か所に密集しているとアノテーションの矢印とオブジェクト群の相関が高くなり，適切にアノテーションできる．逆に，類似した属性値を有するオブジェクトがまだらに散布していると，相関が低くなり，適切なアノテーションが困難になる．

この結果から，図 1 の EKIMEI では A,O,R の分布を分析する．なぜなら，アノテーションができているとすると，相関の高い A,O の結果ではアノテーションが赤い点が多く布置されている方向，もしくは点が比較的集まって布置されている方向に長くのびる．一方，相関の低い R の結果では，ランキングの下位にあることが示されているため，矢印が短くなり，点の分布もばらつきがみられる結果が得られるだろうことを確認するためである．さらに図 2 の EITANGO でも同様に分析する．なお，EITANGO では，A,O,P で分析する．

3.3.2　EKIMEI でのアノテーション付き可視化例

実験結果を EKIMEI から順に述べていく．図 3 では，相関が高いアルファベットであるため，布置されている点は色ごとに集合し，赤い点が集まっている方向に矢印が伸びている．すなわち，赤い矢印の向く方向に A を多く含む単語が布置され

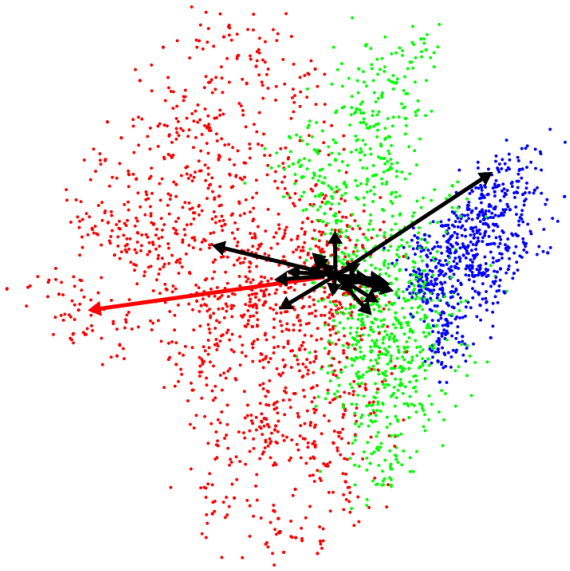

図 3　EKIMEI_A での結果

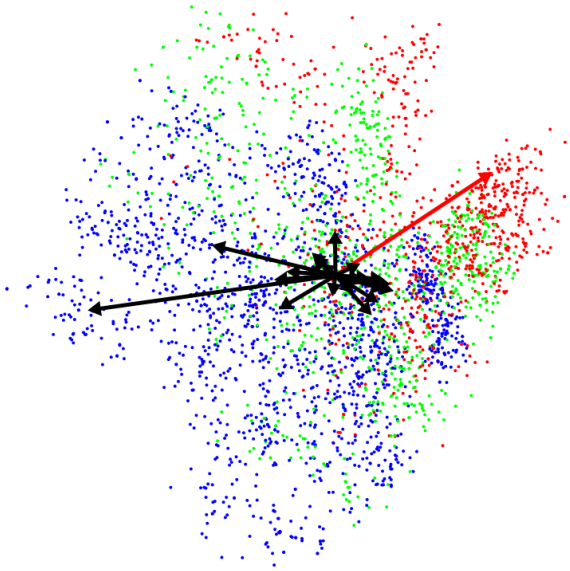

図 4　EKIMEI_O での結果

ていることがアノテーションより示されている．さらに，図 4 をみると，図 3 と比較して布置されている点に多少のばらつきはみられるが，赤い点が集まっている方向に矢印が伸びている．一方，相関が低い図 5 は全体的に点のばらつきがみられ，赤い点もほとんど無い．また，矢印も短くなっている．ここで，EKIMEI でのアノテーション結果の特徴を述べる．A,O のアノテーション結果が特徴的であり，横伸びである．加えて，それらを中心に他のアルファベットの矢印が伸びている．

3.3.3　EITANGO でのアノテーション付き可視化例

EITANGO でも同様に，図 6 と図 7 に示すように，相関が高いアルファベットであると布置されている点は色ごとに集合し，アノテーションは相関が高い方向に長く伸びている．また，図 8 が示すように，相関が低いアルファベットであると点のばらつきがみられ，アノテーションは短く付与された．EITANGO

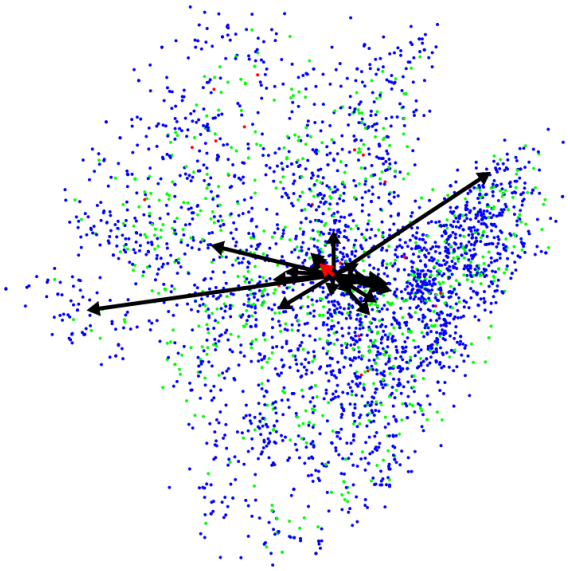

図 5　EKIMEI_R での結果

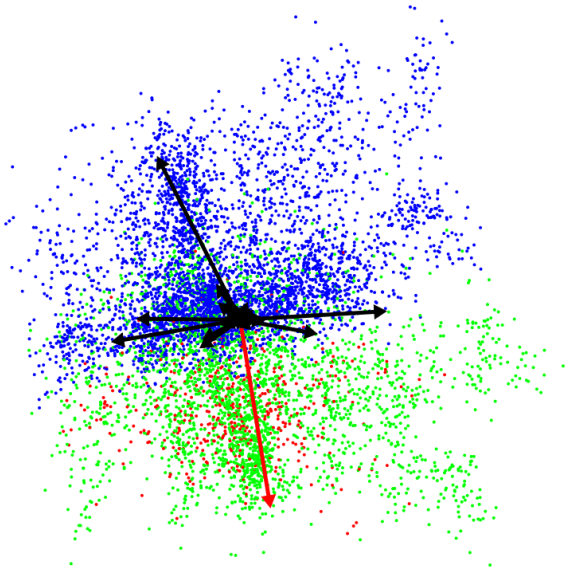

図 6　EITANGO_A での結果

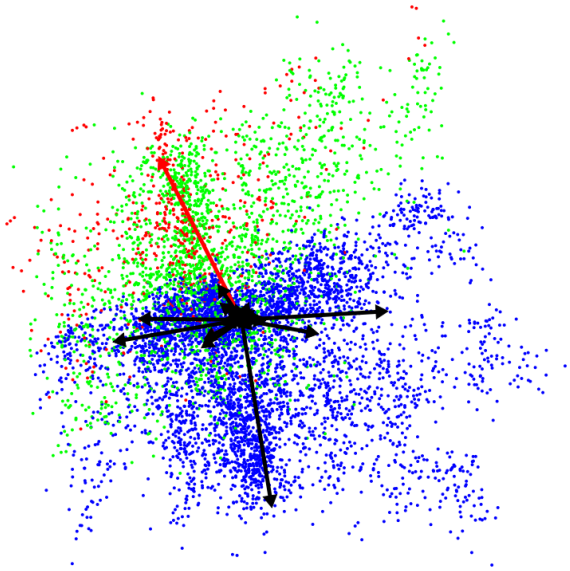

図 7　EITANGO_O での結果

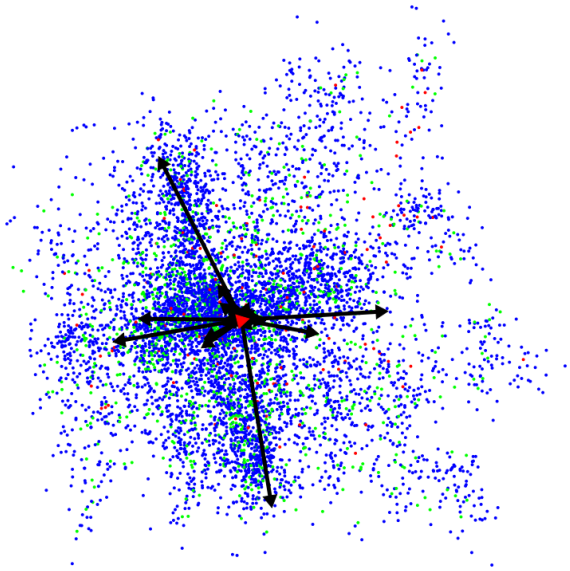

図 8　EITANGO_P での結果

のアノテーション結果では，A,O が縦方向に伸びている．また，比較的相関が高い E,I は，A,O とは違い，横伸びである．

4. おわりに

本論文では，アノテーションを考慮した可視化法について提案し，編集距離に基づき，多次元尺度構成法を用いて評価実験を行った．以上の可視化結果として分かることは，適切な長さと方向を示すことができる特定アルファベットを含む単語の配置方向にアノテーションを付与することができた．したがって，この提案法を用いることで適切にアノテーションが付与できることが示された．今後は，株価のデータに対して，企業の財務指標を属性とするアノテーションを評価するなど，他の実験データや他の可視化法を使ってアノテーション法の妥当性を検証していきたい．

謝　辞

本研究は，豊田中央研究所および，科学研究費補助金基盤研究 (C) (No. 2500133) の補助を受けた．

文　献

[1] J.A. Lee and M. Verleysen, ”Nonlinear Dimensionality Reduction”, Springer, 2007.

[2] J.W. Sammon, ”A nonlinear mapping algorithm for data structure analysis.”, IEEE transactions on Computers, CC-18(5):401-409, 1969.

[3] J.B. Tenenbaum, V. de Silva, and J.C. Langford. ”A global geometric-framework for nonlinear dimensionality reduction.” Science, 290(5500):2319-2323, December 2000.

[4] D. Jurafsky and J.H. Martin, ”Speech and Laguage Processing”, pp.74, Prentice Hall, 2009.

[5] W. Torgerson, ”Theory and methods of scaling”, Proc. of Wiley New York, 1958.