# SVD と SOM

2AC－4

仲川　亜希　　小西　修
高知大学理学部情報科学科

## 1　はじめに

データ集合の自然なサブグループ化を行なうクラスター解析は、情報検索の文献集合に使われる重要な統計的方法である。データベースの検索結果の集合や、与えられた問題世界の情報集合から、それらの特徴を表す概念関係を自動的に抽出し、2次元空間で表示する。このような自動クラスタリング方法には、SOM(Self-Organizing Map ) と LSI(Latent Semantic Indexing) の SVD(Singular Value Decomposition) がある。情報検索において、それらの入力データは文書と用語から構成された行列をとる。文書は、用語のベクトルとして表される。

今回、これらの SVD と SOM の比較実験を行なう。

## 2　SVD のアルゴリズム

LSI は、文書の用語パターンの中に意味構造が潜在していると仮定して、その意味構造を捕えるために SVD を用いる。文書は Salton のベクトル空間モデル (VSM) に従った用語頻度のベクトルとして表される。

全体のデータベースは、$t \times d$ の用語・文書行列として表される。$t$ と $d$ はデータベース中の用語と文書の数である。

文書間の意味の構造を捕えるために、SVD をこの行列に適用する。そして、$m$ (典型的には 100 から 300 まで) 次元の直角な索引付けベクトルを生成する。

分解したベクトルは、同じ意味空間で文書と用語の両方を表すために使用され、それらの

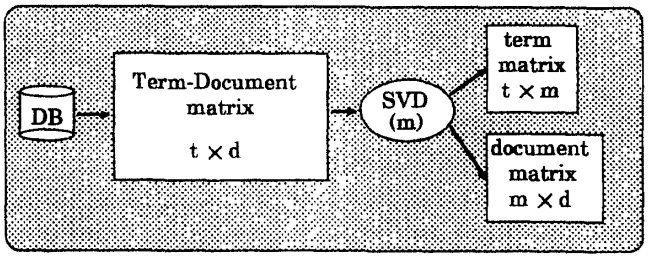


図 1: 用語文書行列に適用された SVD

値は $m$ 個の概念との潜在的な関連度を示す。

図1は、用語文書行列に適用された SVD を示す。[2]

$d_i$ と $t_j$ を文書と用語のセットとする。。[1] LSI をこれらの文書に適用し、索引付けの次元 $m$ は 2 を選ぶ。どのような $a_t \times d$ 行列 $X$ も、3つの行列の積に分解できる。

$$X = T_0 S_0 D_0$$

ここで $T_0, D_0$ は正則行列、$S_0$ は対角行列、$t$

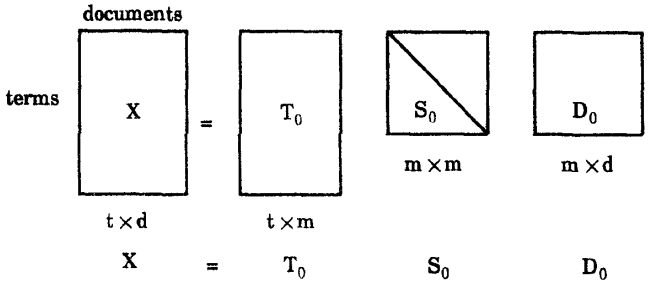


図 2: 用語・文書行列の SVD

は行列 $X$ の列数、$d$ は行列 $X$ の行数で、$m$ は行列 $X$ の階数とする。

| document | term description | VSM $t_1\ t_2\ t_3\ t_4\ t_5\ t_6\ t_7\ t_8$ | LSI (dim1 dim2) |
|---|---|---|---|
| $d_1$ | $t_1\ t_3\ t_4\ t_4$ | 1 0 1 2 0 0 0 0 | 0.863 -0.508 |
| $d_2$ | $t_1\ t_2\ t_2\ t_3$ | 1 2 1 0 0 0 0 0 | 0.873 0.521 |
| $d_3$ | $t_2\ t_3\ t_4$ | 0 1 1 1 0 0 0 0 | 0.702 0.005 |
| $d_4$ | $t_6\ t_7\ t_8$ | 0 0 0 0 0 1 1 1 | -0.010 0.603 |
| $d_5$ | $t_5\ t_6\ t_6\ t_7$ | 0 0 0 0 1 2 1 0 | -0.017 0.974 |

図 3: VSM と LSI

図 3 に、VSM(SVD 前) と LSI(SVD 後) のそれらのベクトル表現を示す。

LSI で分類された文書と用語ベクトルを二次元座標で表すと、共通な用語を多く持っている文書は、お互いに距離が近くなっている。文書と用語をこのように表すと、二次元空間でそれらの関係を容易に見ることができる。

## 3 SOM のアルゴリズム

Kohonen の SOM は、ベクトルで表される入力パターン間の位相関係を、学習アルゴリズムにより発見、分類して位相地図を組織化する 2 層のネットワークである。このときベクトルの各成分はパターンの要素に対応している。この結果得られた地図は、ネットワークに与えられたパターン間の自然な関係構造を表している。

Kohonen の自己組織化マップのアルゴリズムは以下の通りである。

**step1** 入力パターンを与える。

$E = [e_1, e_2, e_3, \ldots, e_n]$

**step2** この入力から競合層の各ユニット $i$ への結合の重みを与える。

$U_i = [u_{i1}, u_{i2}, \ldots, u_{in}]$

**step3** その重みが入力パターンと最もよく一致する競合層のユニット $c$ を定める。すなわち、ベクトル E と $U_i$の間の距離が最小となるものを探す。

$$\| E - U_c \| = \min_j \| E - U_i \| = \sqrt{\sum_j (e_j - u_{ij})^2}$$

**step4** このユニット $i$ とその近傍 $N_c$で重みを調整して一致を増大させる。

$$\Delta u_{ij} = \begin{cases} \alpha(e_j - u_{ij}) & (i \in N_c) \\ 0 & (i \notin N_c) \end{cases}$$

また

$$u_{ij}^{new} = u_{ij}^{old} + \Delta u_{ij}$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T}\right)$$

ここで、$\alpha$は学習率でその値は訓練が進むにつれて 0 へと減少していく。また、$t$ は現在の訓練回数であり、$T$ は行われるべき訓練の全回数である。

**step5** 学習反復が進むに連れて近傍のサイズと重みの変化の量を次第に減少させる。

## 4 比較考察

この二つの方法を実験比較する中で、その性能評価、メリット、問題点を検証した。対象データの大きさに左右されるが、一般に SOM はクラスタリング能力に優れている。計算時間では SVD の方が優れている。これらの手法は、メディエータの統合情報の評価に用いられる。

## 参考文献

[1] S. Deerwester et.al. "Indexing by Latent Semantic Analysis", JASIS, 41(6)391-407, 1990

[2] Shin-Hao Li et.al, "Vintage:A Visual Information Retrieval Interface Based on Latent Semantic Indexing", TR USC-CS-95-632

[3] 仲川亜希, 小西　修: 情報探索のための自己組織化アプローチ, 情処研報,Vol.96, No.103,pp39-46,96-DBS-110,1996.10.