実時間ブートストラップ法*

江原 遥
東京大学情報理工学系研究科 ehara@r.dl.itc.u-tokyo.ac.jp

1　はじめに

ブートストラップ法は，少数のシード（正解）となる語の集合から，文脈パターンを手がかりとして半教師有り学習を行い，シードと意味的な関係を持つ語の集合を獲得する手法であり，テキストマイニングにおける基礎技術である．その中でも，文脈パターンを重みつきグラフ構造で表現しグラフカーネルを用いるラプラシアンラベル伝搬法などの方法が汎用性や理論的な解釈のしやすさから注目されている．

しかし，この手法は，パラメータやシードの調整が難しいことが知られている．その理由は，計算量・メモリ使用量共に大きいため様々なパラメータやシードを試すことが難しい点にある．本稿では，特にラプラシアンラベル伝搬法に対して，近似計算を用いて高速化した実時間ブートストラップ法を提案し，この問題の解決を試みる．

本論文の貢献は以下の通りである．

- ラプラシアンラベル伝搬法を特異値分解による低ランク近似によって高速化した．これは多くのシードを試す時に有効である．
- パラメータについては，経験的に知られていたラプラシアンラベル伝搬法のパラメータ α の値が性能に大きな影響を与えないことの理由付けを行い，パラメータの調整をする必要性が小さいことを示した．
- 近似した場合の方が高い性能が得られることを実験的に示し，その理由付けを行った．

2　ラプラシアンラベル伝搬法

本稿では，二値の場合のラプラシアンラベル伝搬法について説明する．多値の場合への拡張は [3] で論じられている．

N 個と M 個の2部からなる $|V| = N+M$ 個のノードをもつ重み付き無向2部グラフ $G = (V, A)$ の隣接行列を $A \in \mathbb{R}^{(N+M)\times(N+M)}$ とする．2部グラフの性質より，A は一般性を失うことなく，$N \times M$ 行列 W を用いて，$A = \begin{pmatrix} \mathbf{0} & W \\ W^T & \mathbf{0} \end{pmatrix}$ と表せる．左上と右下の $\mathbf{0}$ は，部の内部ではエッジがないことを示し，右上の W と左下 W^T が転置の関係にあることはグラフ G が無向であることを表す．A^{2t} には，$A^{2t} = \begin{pmatrix} (WW^{\mathrm{T}})^t & \mathbf{0} \\ \mathbf{0} & (W^{\mathrm{T}}W)^t \end{pmatrix}$ のように，A^{2t} は内部の $(WW^{\mathrm{T}})^t$ と $(W^{\mathrm{T}}W)^t$ のみを用いて表せる性質がある．

A^2 の正則化グラフラプラシアンを考える．任意の $n \times n$ 対称行列 X に対して正規化を行う行列を $(D_X)_{i,i} = (\sum_j x_{i,j})$，$I_n$ を $n \times n$ の単位行列と定義すると，X に対する正則化グラフラプラシアン $\mathcal{L}$ は $\mathcal{L}(X) = I_n - D_X^{-\frac{1}{2}} X D_X^{-\frac{1}{2}}$ と表せる．この定義を A^2 に適用すると，A^2 に対する正則化グラフラプラシアンの t 乗は，$\mathcal{L}(A^2)^t = \begin{pmatrix} \mathcal{L}(WW^{\mathrm{T}})^t & \mathbf{0} \\ \mathbf{0} & \mathcal{L}(W^{\mathrm{T}}W)^t \end{pmatrix}$ と表せる．

$n \times n$ 対称行列 X に対して，正則化ラプラシアンを α で減衰させながら無限回足しこんだ

$$\mathcal{K}_\alpha(X) = (I_n - \alpha\mathcal{L}(X))^{-1} = \sum_{t=0}^{\infty} \alpha^t \mathcal{L}(X)^t$$

を正則化ラプラシアンカーネルという．この場合も，$\mathcal{K}_\alpha(A^2) = \begin{pmatrix} \mathcal{K}_\alpha(WW^{\mathrm{T}}) & \mathbf{0} \\ \mathbf{0} & \mathcal{K}_\alpha(W^{\mathrm{T}}W) \end{pmatrix}$ のように書ける．

最後に，正則化ラプラシアンカーネルを用いて，ラプラシアンラベル伝搬法を説明する．ラプラシアンラベル伝搬法では，2部グラフ G のうち，N 個ノードからなる部を**インスタンス**，M 個のノードからなる部を**パターン**と呼ぶ．また，N 次元ベクトル $\boldsymbol{s}$ と M 次元ベクトル $\boldsymbol{s}'$ を，この順番で縦方向につなぎあわせたベクトル $\boldsymbol{s}_a \equiv (\boldsymbol{s}; \boldsymbol{s}')$ を**シードベクトル**と呼ぶ．シードベクトルへの正解ラベルの付与の仕方としては，$i \in \{1, \ldots, N+M\}$ 番目のノードが正例なら $\boldsymbol{s}_a$ の第 i 成分 $(\boldsymbol{s}_a)_i = 1$，負例なら $(\boldsymbol{s}_a)_i = -1$，正解が分からない場合は $(\boldsymbol{s}_0)_i = 0$ とする方法が典型的である．

ラプラシアンラベル伝搬法は，各インスタンスと各パターンの関係の強さを表す数値を格納した行列 W と，正解が分かっているインスタンスに正解ラベルを付与し

*この研究は，著者が楽天技術研究所 NY で行ったインターンの研究内容である．

たシードベクトルを与えたときに，数式 (1) によってスコアベクトル $\boldsymbol{g}_a$ を求め，スコア $(\boldsymbol{g}_a)_i$ の値が大きいインスタンスを世界に近いインスタンスとして獲得する手法である．具体的なタスク設定の例は，§4 で述べる．

$$\boldsymbol{g}_a = \mathcal{K}_\alpha(A^2)\boldsymbol{s}_a = \begin{pmatrix} \mathcal{L}(WW^{\mathrm{T}})\boldsymbol{s} \\ \mathcal{L}(W^{\mathrm{T}}W)\boldsymbol{s}' \end{pmatrix} \quad (1)$$

数式 (1) から，$\boldsymbol{g}_a$ のインスタンス部分のスコアは，パターンのシード $\boldsymbol{s}'$ に依存せずインスタンスのシード $\boldsymbol{s}$ にのみ依存することが分かる．この性質を利用し，[3] など多くの文献では，最初から A ではなく W を隣接行列と定義し，数式 (1) の下側部分を省略した形で書かれている．

3 提案手法

この節では，数式 (1) の上側部分にのみ注目し，$\mathcal{K}_\alpha(WW^{\mathrm{T}})$ を効率的に計算する方法を提案する．数式 (2) の定義式のまま $\mathcal{K}_\alpha(WW^{\mathrm{T}})$ を計算することには，次に挙げる問題点がある．

空間計算量 自然言語処理においては，W が大規模な疎行列となることが多いが，W が疎行列であっても，WW^{T} は $N \times N$ の密行列となるので，格納するためだけに空間計算量 $O(N^2)$ のメモリが必要になる．従って WW^{T} を計算すると，より多くの種類の単語を扱うなどの目的で N を大きくすることが難しくなる．したがって，WW^{T} の計算を避けて計算することが望ましい．このメモリによる制約の問題は，並列計算で $\mathcal{K}_\alpha(WW^{\mathrm{T}})$ の計算法を論じる [3] でも論じられている．彼らは後述の逆行列の計算コストに比較して WW^{T} の計算量が小さいことを利用し，WW^{T} を何度も再計算する手法を提案している．

時間計算量 $\mathcal{K}_\alpha(WW^{\mathrm{T}})$ の計算には $N \times N$ の逆行列が含まれるが，一般に $N \times N$ の逆行列の計算には $O(N^3)$ という大きい時間計算量が必要となる．

本稿では，この両方の問題点を解決しうる効率的な計算法として，疎行列の構造を保持している $D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W$ を特異値分解によって低ランク近似する方法を提案する．そのために，まず，特異値分解によって数式 (2) を近似を用いずに変形する．次に，その変形した形から低ランク近似を行う．

$D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W$ は，特異値分解 (Singular Value Decomposition, SVD) を用いて，次のように表せる．

$$\left(D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W\right) \simeq U_K \Sigma_K V^T \quad (2)$$

ここで，Σ_K は $K \times M$ 行列で，その対角項 $(\Sigma_K)_{i,i} = \sigma_i$ には $D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W$ の i 番目に大きい特異値が並ぶ．$\sigma_1 \geq \sigma_2 \geq, \ldots, \geq \sigma_K \geq 0$ である．また，正規化されていることから $\sigma_1 = 1$ である．U_K, V は，それぞれ，$N \times K, M \times M$ 行列で，$U_K \equiv (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K), V \equiv (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M)$ とする．$\boldsymbol{u}_k, \boldsymbol{v}_k$ は，特異値 σ_k に対応する特異ベクトルである．数式 (2) の $\simeq$ は，$K = N$ の時，等号で成り立ち，その時 U_N，V_N は，$U_N U_N^{\mathrm{T}} = U_N^{\mathrm{T}} U_N = I_N$, $VV^{\mathrm{T}} = VV^{\mathrm{T}} = I_M$ をそれぞれ満たす直交行列となる．

この分解を用いて，$K = N$ のとき，$\mathcal{K}_\alpha(WW^{\mathrm{T}})$ を表す．まず，$\mathcal{L}_\alpha(WW^{\mathrm{T}})$ は，次のように表せる．

$$\begin{aligned} l\mathcal{L}_\alpha(WW^{\mathrm{T}}) &= I_N - \left(D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W\right)\left(D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W\right)^T \\ &= U_N U_N^{\mathrm{T}} - (U_N \Sigma_N V^{\mathrm{T}})(U_N \Sigma_N V^{\mathrm{T}}) \\ &= U_N \left(I_N - \Sigma_N \Sigma_N^{\mathrm{T}}\right) U_N^{\mathrm{T}} \end{aligned}$$

従って，$\mathcal{K}_\alpha(WW^{\mathrm{T}})$ は，次のように表せる．ただし，$(F_N)_{k,k} \equiv f(\sigma_k) \equiv \left(1 - \alpha(1 - \sigma_k^2)\right)^{-1}$ である．数式 (3) までの変形は，近似を用いていない．[2] でも同様の変形を行っているが，彼らの目的はリンク予測であり，本稿のブートストラップ法とは目的が異なる．

$$\begin{aligned} l\mathcal{K}_\alpha(WW^{\mathrm{T}}) &= \left(I_N - \alpha\mathcal{L}(WW^{\mathrm{T}})\right)^{-1} \\ &= \left(U_N U_N^{\mathrm{T}} - \alpha U_N \left(I_N - \Sigma_N \Sigma_N^{\mathrm{T}}\right) U_N^{\mathrm{T}}\right)^{-1} \\ &= U_N \left(I_N - \alpha\left(I_N - \Sigma_N \Sigma_N^{\mathrm{T}}\right)\right)^{-1} U_N^{\mathrm{T}} \\ &= U_N F_N U_N^{\mathrm{T}} \end{aligned} \quad (3)$$

提案手法を数式 (4) に示す．数式 (3) に右からシードベクトル $\boldsymbol{s}$ をかけた $\boldsymbol{g} = U_N F_N U_N^{\mathrm{T}} \boldsymbol{s}$ が，近似を用いない場合のラプラシアンラベル伝搬法のスコアである．このうち U_N を U_K で，F_N を F_K で近似し，$\boldsymbol{g}$ を次の様にランク K までで低ランク近似して $\hat{\boldsymbol{g}}$ を求める．ただし，シードベクトル $\boldsymbol{s}$ のうち要素が 0 でない添字の集合を $J_{\boldsymbol{s}} \equiv \{j | j \in \{1, \ldots, N\}, s_j \neq 0\}$ と定義する．

$$\hat{\boldsymbol{g}} = U_K F_K U_K^{\mathrm{T}} \boldsymbol{s} \quad (4)$$

$$(\hat{\boldsymbol{g}})_i = \sum_{k=1}^{N_r} \sum_{j \in J_{\boldsymbol{s}}} f(\sigma_k) (U_K)_{i,k} (U_K)_{j,k} s_j \quad (5)$$

数式 (4) が，前述の 2 点の問題をどのように解決しているかを述べる．

空間計算量の近似による削減 まず，大規模疎行列 W が与えられたときに，$D_{WW^{\mathrm{T}}}^{-\frac{1}{2}}W$ は，WW^{T} の計算をせずとも疎行列の構造を保ったまま計算可能で

ある．$W^{\mathrm{T}} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_N)$ として W を行方向に N 個のベクトル $\boldsymbol{w}_i$ で分解する．すると，$(D_{WW^{\mathrm{T}}})_{i,i} = (\sum_j \boldsymbol{w}_i \cdot \boldsymbol{w}_j)$ とかけるのでこの N 個の変数を先に計算する．すると，$(D^{-\frac{1}{2}}_{WW^{\mathrm{T}}}W)_{i,*} = \left(((D_{WW^{\mathrm{T}}})_{i,i})^{-\frac{1}{2}}\boldsymbol{w}_i\right)^{\mathrm{T}}$ として，$D^{-\frac{1}{2}}_{WW^{\mathrm{T}}}W$ が計算できる．ただし，$X_{i,*}$ は，X の i 行目の横ベクトルである．

次に，大規模疎行列 $D^{-\frac{1}{2}}_{WW^{\mathrm{T}}}W$ に対して特異値分解を行いランク K までの U_K，Σ_K を求める手法には多くの既存手法があり，$O(N^2)$ より少ない空間計算量で可能なアルゴリズムが提案されている．典型的には Arnoldi 法による計算法があり，近年では [1] などが注目されている．提案手法は，これらの既存手法のうちどれかに依存するものではなく，U_K，Σ_K が計算可能であればよい．

また，実用的には，様々なシード$\boldsymbol{s}$を試すために，数式 (4) において$\boldsymbol{s}$以外の要素を保存しておきたい．このためには，数式 (5) の展開式より，U_K，$\{\sigma_1, \ldots, \sigma_K\}$ を保存すれば十分であることが分かる．

時間計算量の近似による削減 提案手法には，$D^{-\frac{1}{2}}_{WW^{\mathrm{T}}}W$ から U_K，Σ_K を計算するのに必要な時間計算量と，これらが与えられたときに数式 (4) を計算するのに必要な時間計算量の二種類の計算量が絡む．

前者については，特異値分解を行う既存手法に依存する．後者については，数式 (5) の展開より，シードの数 $|J|$ に依存し，$O(NK|J|)$ で済むことが分かる．

数式 (4) において $f(\sigma_k)$ は凸関数であるが，図 1(c) より典型的な α に対しては $f(\sigma_k) \simeq 1$ であり α の値が性能に大きく影響しないことが分かる．

4 評価

本稿では，一般ユーザが記述したホテルのレビューを収集した**楽天トラベルコーパス**[1]から，**風呂**に関する評価対象を収集するタスクを用いて，提案手法を評価した．ホテルのレビューである楽天トラベルコーパスには，自然文とは別に，各ユーザが各ホテルの各評価項目（サービス，部屋，風呂，立地，設備・アメニティ，食事，総合）に対して付けた 5 段階評価が用意されており，各項目の評価対象語（「浴室」など）を収集することは，自然文から各項目の 5 段階評価の点数を推測するといったタスクで役立つ．このうち，サービスや食事は従業員の言動を評価対象とする評価を多く含むため，また，部屋は部屋だけにあるとは限らないもの（エアコンなど）を評価対象とする評価を多く含むため，語だけから評価項目を限定することが難しい．また，立地・設備の評価対象は個々のホテル独自の事柄に対する評価が多くデータが疎過ぎる．そこで，語だけから評価項目が限定しやすい風呂項目を選んだ．

コーパス中に表れる名詞をインスタンス，形容詞をパターンと見て，名詞と形容詞の係り受け傾向が類似したインスタンスを，前述のラプラシアンラベル伝搬法を用いて抽出し，評価対象の候補として出力する．**CaboCha**[2]による係り受け解析を行い，名詞と形容詞の係り受け関係を全て抽出した．楽天トラベルコーパスは，一般ユーザが記述しているのでタイプミスを含む．これを除外するために，頻度が 2 以下のリンクは削除した．名詞 $i \in \{1, \ldots, N\}$，形容詞 $j \in \{1, \ldots, M\}$ に対して $(W)_{i,j}$ は，名詞 i と形容詞 j の PMI を正にしたもの，$(W)_{i,j} = -PMI(i,j) \equiv -\log \frac{\mathrm{freq}(i,j)}{\mathrm{freq}(i)\mathrm{freq}(j)}$ とした．ただし $\mathrm{freq}(i)$ は，その語を含む名詞と形容詞の係り受けの頻度（3 以上）である．W のサイズは，名詞 $3,061 \times$ 形容詞 868 となった．

これらの評価対象語候補となる名詞の中から，明らかに評価対象が風呂項目のみに限定されるか否かという基準で作業者 1 名が人手で正解を抽出したところ，116 語が得られた．例えば，「バス」という語は乗り物のバスの意味もあるので排除している一方，「ユニットバス」や「バスルーム」は風呂項目に限定できるので正解に追加している．また，「温泉」や「入浴」といった文字列を含む語は原則的に追加しているが，「温泉街」や「入浴者」など，対象が風呂項目でないものは除外している．その他，風呂を表す語が明示的に入っていなくても，シャワー，ドライヤー，シャンプー，湯加減，といった，浴室に併設されていることが殆どなものや，湯に関する言及なども，風呂項目にまとめられるので正解とした．

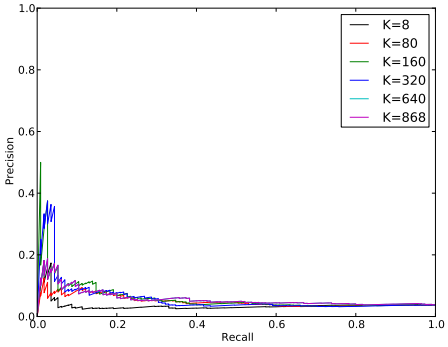
まず，単純に**風呂**という 1 語のシードベクトルから開始した場合を，図 1(a)，図 1(b) の Precision-Recall 図 (PR 図) に示す．(a) が $\alpha = 0.001$ の場合，(b) が $\alpha = 0.1$ の場合であるが，前述の様に，殆ど性能に変化が無いことが見て取れる．また，重要な知見として，性能を最大にする K は，近似のない $K = 868$ の場合ではなく，近似がかかった $K = 160$ の場合であることが分かる．この理由は，図 1(c) に示すように，ラプラシアンラベル伝搬法では，小さな特異値 σ_k に対しても $f(\sigma_k) \simeq 1$ であるので，その対応する特異ベクトル $\boldsymbol{u}_k$ が，数式 (5)

[1] http://rit.rakuten.co.jp/rdr/index.html

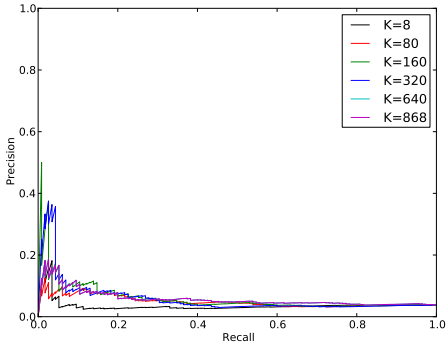
[2] http://chasen.org/ taku/software/cabocha/

表 1: 実行時間

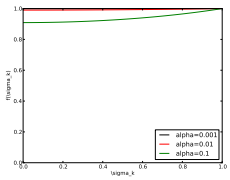
K	時間 (s)
8	0.064
80	0.121
160	0.205
320	0.386
640	0.805
868	1.136



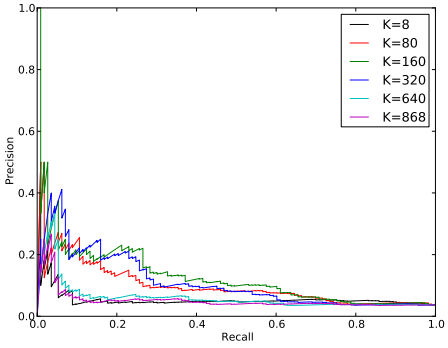
(a) シード 1 語，$\alpha = 0.001$，最大 F 値 0.126 ($K = 160$ の時)



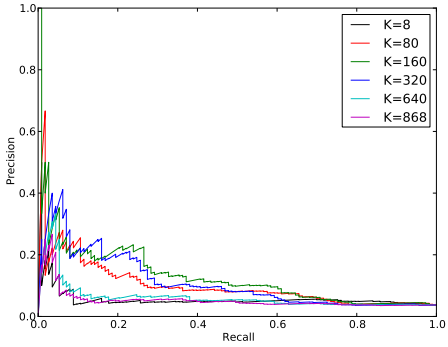
(b) シード 1 語，$\alpha = 0.1$，最大 F 値 0.126 ($K = 160$ の時)



(c) $f(\sigma_k)$ の図



(d) シード 3 語，$\alpha = 0.001$，最大 F 値 0.241 ($K = 160$ の時)



(e) シード 3 語，$\alpha = 0.1$，最大 F 値 0.244 ($K = 160$ の時)

図 1: 各 K に対する PR 図

より他の特異ベクトルと同程度に重み付けられてスコアに足し込まれるため，K が大きくなりすぎると，このような小さい特異値が性能を悪化させるためであると考えられる．

次に，シードを増やし，**サウナ，シャワー，浴室**という 3 語のシードベクトルから開始した場合の PR 図を図 1(d)，図 1(e) に示す．この場合，シードを増やした効果により，全体的に性能が向上していることが分かる．また，この場合でも，F 値は K の小さいところで最大になること，図 1(e) との比較より，α の値は性能に殆ど影響しないことが分かる．

最後に，図 1(e) の実行時間を表 4 に表示する．全体を通し，性能が最も良かった $K = 160$ の場合は，通常の逆行列にシードベクトルを掛ける方法と同じ計算量となる $K = 868$ の場合と比較して，約 5 倍の速度で計算できていることが分かる．

参考文献

[1] N. HALKO, P. G. MARTINSSON, and J. A. TROPP. Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *arXiv 0909.4061*, 2009.

[2] J. Kunegis and A. Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 561–568, New York, NY, USA, 2009. ACM.

[3] 小町守, 牧本慎平, 内海慶, 颯々野学. ラプラシアンラベル伝播による検索クリックスルーログからの意味カテゴリ獲得. 人工知能学会論文誌, Vol. 25, No. 1, pp. 196–205, 2010.