E-002

深層学習を用いたユーザー指向型類似文抽出の幾何的近似

*槇尾純太　**深谷 寛子 *森田和宏　*泓田正雄

*徳島大学大学院創成科学研究科

**東京ガスエンジニアリングソリューションズ

1. はじめに

近年，人工知能分野が深層学習によって花開いたことにより，様々な分野で人工知能分野の技術が使われることになっている．実際にその力は凄まじく，今まで解決困難だと言われてきた様々な問題を解決している．しかしながら，深層学習モデルの大規模化などの理由から，ユーザーにそのシステムを提供するのが困難なことがある．

本研究はユーザーの趣向や目的などを学習して良質な結果を出力する類似文書抽出への深層学習の応用に焦点を当てる．計算機で柔軟かつ高速に類似文書を抽出する際は，主に対象物を事前に何らかの手法でベクトル化し，類似度を取ることで実現することが多い．しかし，それらのベクトルによる類似文書抽出の特徴は手法に依存するため，その性質がユーザーの目的に合っていないことは多々ある．深層学習モデルを用いて学習によりユーザーの指向に合わせることは可能であり，深層距離学習を用いた埋め込みベクトル[1]の距離や角度を使うものが考えられるが，データベースに存在する全てのベクトルの写像を計算した後に類似度を測る必要性がある．この手法はデータに変更があった場合やユーザーの嗜好を学習するたびに大量のベクトルを処理する必要があることから実用性に欠ける． そこで，本研究では幾何的観念から得られる近似によって，クエリのみの処理でユーザー指向の文章抽出を実現する深層学習モデルを提案する．

2　提案手法

2.1 計量を考慮した距離計算

提案手法ではデータの特徴を表現する位置ベクトルが，ある幾何的な空間の上に存在することを仮定している．ここでの幾何的な空間とは各点において内積が二階のテンソルで定義された連続な空間を指し，ユークリッド空間もその一つである．通常の深層学習ではベクトルの射影を学習するが，提案モデルはベクトルが存在する空間の内積（空間の各点における線形近似できる近傍内での内積）を表現するような行列を学習する．得られた行列から空間におけるデータ間の距離を計算する．そして，深層距離学習と同様にContrastLoss関数(1)を最小化する形で最適化する．

$$J = \frac{1}{2}(YD^2 + (1-Y)\,max(margin - D, 0)^2) \qquad (1)$$

ここでの D は距離を計算する関数であり，Y はデータのラベルで margin はハイパーパラメータである．D には通常ユークリッド空間における二乗距離などが選択されるが，今回は空間の計量を考慮した距離計算を導入する．ここではリーマン幾何における一般的な距離定義[2]を以下に示す．

$$D = inf \int \sqrt{\sum_{i,j} g_{ij}(\gamma(t)) \frac{d\gamma(t)^i}{dt} \frac{d\gamma(t)^j}{dt}}\, dt \qquad (2)$$

γは 2 点を通るtをパラメータとした任意の曲線であり，長さが最小となるものを 2 点間の距離とする．g_{ij}は空間の各点の内積を定義する 2 階テンソルである．

2.2 近似

上記に示した計算を行うことは非常に難しく実用に適していないため，幾何学的な観点から近似を行う．本研究の目的は類似文章抽出であり，空間の特徴を適応したベクトル間の距離が近いものを選ぶことから，ある程度距離の遠いデータを扱うことがない．よって幾何学における座標近傍での近似を適応する．その近似とはデータ間を結ぶ測地線を曲がった線ではなく，その間で平坦な線であると考える．そこから，空間の内積を表すテンソルは空間の各点における値が存在するが，データ間における計量をその間に限って一定だとする．このように考えると 2 点間の距離は(3)の様に表現できる．

$$D \cong \sum_{i,j} g_{ij}(a)(\xi(b) - \xi(a))^i (\xi(b) - \xi(a))^j \qquad (3)$$

ここでのξとはデータ空間内で使われている座標系を表しており，通常はベクトル化手法で生成されたベクトルそのものと見なせる．a, bはクエリとターゲットといった比較したいデータを示す．本研究ではこの計算式に基づき$g_{ij}(a)$を深層学習モデルで学習する形で，距離の情報からユーザー指向に適応する．

2.3 深層距離学習との違い

深層距離学習で類似データを抽出する際の距離計算は(4)のようになる．

$$D = \sqrt{\sum_i \left(f\left(\xi(a)\right)^i - f\left(\xi(b)\right)^i\right)^2} \quad (4)$$

f は深層学習モデルであり，クエリとターゲットの．このように通常の深層距離学習を導入しようとすると，対象となる大量のベクトルに対して深層モデルで写像を行ったあとに距離を計算する必要がある．それに比べ，(3)はクエリに対するモデルによる写像は一回必要になるものの，対象のベクトルに対して写像を行う必要がなく，差を取ったベクトルから成る行列に対して積を一度取るだけで近似的に類似データの抽出が可能であり，対象となるベクトルが大規模でも深層学習を導入することが可能になる．

3. 実験

3.1 実験設定

本研究の最終的な目的は類似文書検索を使うユーザーに対して精度を上げていくことになるが，定量的な評価が困難という理由から，今回は STS[1]を代用してモデルの評価を行った．STS とは文章のペアに対して 0.0 から 5.0 の類似度が振られており，5.0 に近いほど類似度が高い．訓練データは 6 千ペア程であり，評価データは 3 千ペアである．本研究の目的はユーザー指向の類似文書抽出であることから，その傾向を STS の教師データで代用し，通常のベクトル化手法を使った結果に教師信号を適応することで精度が上がれば，ユーザーの指向も学習できると考えた．

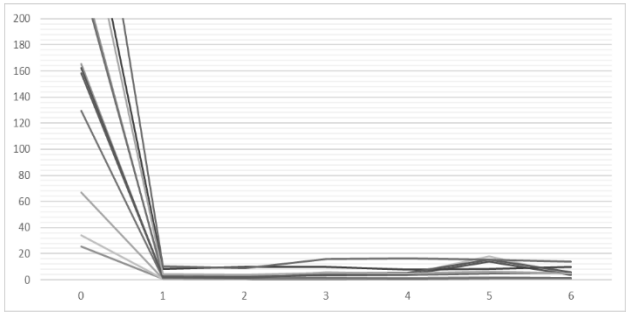
今回はデータの類似度の中央値である 2.5 以上を正例，2.5 未満のものを負例として評価する．本研究はベクトル化手法に依存しないため，文章のベクトル化は単純に Word2vec の平均値を用いた．学習には単純な NN モデルを用いており，対になったベクトルのユークリッド距離を事前に学習した後に，ContrastLoss 関数を用いて教師データを適応した．既存のベクトル化手法のみを使った上記で設定した教師ラベル適応前と，それらのラベルを適応した後の精度の変化を見る．

3.2 実験結果

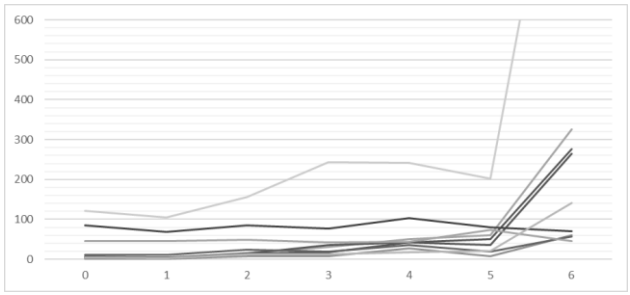
精度を表 1 に示す．訓練データ（観測値），評価データ（非観測値）両方で精度の向上が見られた．また，無造作に評価データから 10 個選択し，学習の反復回数と計量から計算した近似距離を図 1 に示した．これらから正例の距離が小さくなる様子や，負例の距離が開いていくことが観察でき，意図した学習が行われていることが分かる．

表 1　学習前後での精度

	観測値	非観測値
適応前の正解率(%)	65.8	60.0
適応後の正解率(%)	91.4	73.7



(a)負例



(b)正例

図 1　学習の反復(横軸)と距離(縦軸)の変化

3.3 考察

本手法が STS に有効だったことから，実用的な文書抽出でユーザーの趣向や目的が学習することが可能だと考えられる．ベクトルとして抽出されたデータの各次元には何らかの独立した情報が含まれると考えられる．そのことから，データ空間の場所と比較すべき次元を学習したことで，観測データだけではなく非観測データまで精度が上がったと考えられる．

4. まとめ

本研究ではユーザー指向の類似文書抽出に対して深層学習を応用する際の問題点と解決する近似手法を提案した．今後の展望としては，実データでの運用や実システムとしての有効性を検証することが挙げられる．

参考文献

[1] Deep Metric Learning with Spherical Embedding, Dingyi Zhang, et al., 2020

[2] An elementary introduction to information geometry, Frank Nielsen, 2018

[1] STS benchmark, http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark