A-030

Geometric Algebra を用いた文のベクトル化手法における計算量の改善に関する検討

A study on reduction of calculation cost in sentence coding method using Geometric Algebra

鈴木 直人† 吉川 大弘† 古橋 武†
Naoto Suzuki Tomohiro Yoshikawa Takeshi Furuhashi

1．まえがき

電子文書の普及とともに，様々な状況で膨大な量の文書を管理する必要が生じている．この文書管理を行うためのアプローチとして，こうした場合には，文書のベクトル化などの方法を用いて文書分類を行う方法が知られている．これまで，tf-idf や潜在意味解析(LSA)を用いた文書分類手法が報告されているが，これらの多くは単語の出現順序を考慮していない．これに対し，Geometric Algebra (GA) を用いることで，単語の出現順序を考慮した文のベクトル化手法が提案されている．この手法では，LSA に基づき，単語の出現順序に応じた回転ベクトルを定義することで，各文の出現単語順に対応する回転ベクトルを掛けてその文ベクトルが計算される．本稿では，この手法における計算量についての問題点を挙げ，その改善方法について検討を行う．

2．Geometric Algebra (GA)[2]

2.1 GA 積

GA は複素数の自然な拡張であり，(p+q)個の直交基底を持つ．この GA 空間は $\mathcal{G}(p,q)$で表され，各基底は，$\mathbf{e}_i^2=1(i=1,2,...,p),\mathbf{e}_i^2=-1(i=p+1,...,p+q)$ を満たすように定義される．基底どうしの内積は，$\mathbf{e}_i\cdot\mathbf{e}_i=\mathbf{e}_i^2,\mathbf{e}_i\cdot\mathbf{e}_j=0\,(i\neq j)$ を満たすように定義される．GA 空間における任意のベクトル $\mathbf{a}$ は，$\mathbf{a}=\sum_{i=1}^{p+q}a_i\mathbf{e}_i$ のように，基底の線形結合によって表される．

ここで，GA の特徴的な演算である GA 積について説明する．GA 積は，$\mathbf{e}_i\mathbf{e}_i=\mathbf{e}_i\cdot\mathbf{e}_i,\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_{ij}\,(i\neq j)$ を満たすように定義され，$i\neq j$ の場合については外積に等しい ($\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_i\wedge\mathbf{e}_j\ (i\neq j)$)．また，GA 積は非可換な演算である($\mathbf{uv}=\mathbf{u}\cdot\mathbf{v}+\mathbf{u}\wedge\mathbf{v}$)．基底が持つこのような性質から，基底どうしの GA 積は内積と外積の和として表される $(\mathbf{e}_i\mathbf{e}_j=\mathbf{e}_i\cdot\mathbf{e}_j+\mathbf{e}_i\wedge\mathbf{e}_j)$．任意の GA ベクトル $\mathbf{a}=\sum_{i=1}^{p+q}a_i\mathbf{e}_i,\mathbf{b}=\sum_{i=1}^{p+q}b_i\mathbf{e}_i$ の GA 積は，それぞれのベクトルの要素の GA 積の和で表される．

次に，基底のグレードについて説明する．$\mathcal{G}(p,q)$が持つ(p+q)個の直交基底 $\mathbf{e}_i$ はすべてグレード 1 である．このグレード 1 の基底から任意の 2 つの基底 $\mathbf{e}_i,\mathbf{e}_j$ を選び，その GA 積である $\mathbf{e}_{ij}$ をグレード 2 と呼ぶ．このように，基底の添え数字の個数，つまりグレード数は，グレード 1 の基底の GA 積の回数となる．また，定数の項のグレードは 0 とし，グレード n のベクトルは，n-ベクトルと呼ぶ．

具体例として $\mathcal{G}(3,0)$の場合で説明する．この GA 空間は，3 個の直交基底 $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$ をもつ．この空間内における 2 つの GA ベクトル $\mathbf{a}=\sum_{i=1}^{3}a_i\mathbf{e}_i,\mathbf{b}=\sum_{i=1}^{3}b_i\mathbf{e}_i$ の GA 積を考える．2 つの GA ベクトルの GA 積は，それぞれの要素の GA 積の和で表されるため，$\mathbf{ab}=(a_1b_1+a_2b_2+a_3b_3)+(a_1b_2-a_2b_1)\mathbf{e}_{12}+(a_2b_3-a_3b_2)\mathbf{e}_{23}+(a_3b_1-a_1b_3)\mathbf{e}_{31}$ となり，GA 積により $\mathbf{e}_{12},\mathbf{e}_{23},\mathbf{e}_{31}$ の基底が生じる．GA 積で生じ得るすべての基底の集合は正規基底と呼ばれ，$\{1,\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\mathbf{e}_{12},\mathbf{e}_{23},\mathbf{e}_{31},\mathbf{e}_{123}\}$ となる．このうち，$\{1\}$ はグレード 0 の基底，$\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$はグレード 1 の基底，$\{\mathbf{e}_{12},\mathbf{e}_{23},\mathbf{e}_{31}\}$はグレード 2 の基底，$\{\mathbf{e}_{123}\}$はグレード 3 の基底である．任意の GA 空間 $\mathcal{G}(p,q)$において，正規基底の個数は $2^{(p+q)}$ 個である．

2.2 GA 積を用いた回転

次に，本稿で用いるベクトルの回転について説明する．GA 空間におけるベクトルの回転は，連続的な GA 積で表される．ある 2 つの単位 1-ベクトル $\mathbf{u},\mathbf{v}$ について，この 2 つのベクトルがなす角を $\theta/2$ とする．$\mathbf{R}=\mathbf{uv},\mathbf{R}^\dagger=\mathbf{vu}$ とすると，任意のベクトル $\mathbf{a}$ に対して，$\mathbf{RaR}^\dagger(=\mathbf{uvavu})$は，$\mathbf{a}$ を θ だけ回転させる演算である．以下では，この演算における $\mathbf{R}$ および $\mathbf{R}^\dagger$ を回転ベクトルと呼ぶ．ここで，$\mathbf{B}=\frac{\mathbf{v}\wedge\mathbf{u}}{\|\mathbf{v}\wedge\mathbf{u}\|}$ とおくと，$\mathbf{R}=\mathbf{u}\cdot\mathbf{v}+\mathbf{u}\wedge\mathbf{v}$は，$\mathbf{R}=\mathbf{u}\cdot\mathbf{v}-\|\mathbf{v}\wedge\mathbf{u}\|\mathbf{B}$と書ける．$\mathbf{u}$, $\mathbf{v}$ がともに 1-ベクトルであるため，$\mathbf{B}=\frac{\mathbf{v}\wedge\mathbf{u}}{\|\mathbf{v}\wedge\mathbf{u}\|}$ は単位 2-ベクトルとなり，$\mathbf{B}^2=-1$ である．そのため，$\cos\theta=\mathbf{u}\cdot\mathbf{v},\sin\theta=\|\mathbf{v}\wedge\mathbf{u}\|$ とおけば，$\mathbf{R}=\exp\left(-\mathbf{B}\theta/2\right),\ \mathbf{R}^\dagger=\exp\left(\mathbf{B}\theta/2\right)$と書ける．

† 名古屋大学

3. GA を用いた文のベクトル化

本節では，GA を用いた文のベクトル化の手法について説明する．英語文書においては，GA を用いた文のベクトル化の手法が G. Pilato らによって提案されている[3]．この手法では，初めに，対象となる全ての文書から，単語と単語の共起関係を表す単語-単語行列 $\mathbf{A}$ を作成する．この行列について，特異値分解(SVD)を用いて，低ランク(ランク k)の行列 $\mathbf{A}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^T$ で近似する．この特異値分解において，行列 $\mathbf{U}_k$ の i 行目の行ベクトル $\mathbf{l}_i$ に，行列 $\mathbf{\Sigma}_k$ の対角成分の i 番目の要素の平方根を掛けたベクトルは，i 行目に対応する単語 w_i の左側の文脈情報（文章において単語 w_i より前に出現する単語情報）を表す．一方，行列 $\mathbf{V}_k$ の i 列目の列ベクトル $\mathbf{r}_i$ に，行列 $\mathbf{\Sigma}_k$ の対角成分の i 番目の要素の平方根を掛けたベクトルは，単語 w_i の右側の文脈情報を表す．$\mathbf{R}'_{ij} = \mathbf{r}_i \mathbf{l}_j = \exp\left(-\mathbf{B}_{ij}\theta_{ij}\right)$ としたとき，回転ベクトル $\mathbf{R}_{ij} = \exp\left(-\mathbf{B}_{ij}\,\theta_{ij}/2\right)$ は，単語 w_i の後に単語 w_j が出現する Bi-gram[4]に関連している．

各文を表すベクトルについて，初めに，ベクトル長 k，要素をすべて 1 とした 1-ベクトルを初期ベクトルとして作成する．次に，その文の単語の出現する順番に応じて，対応する回転ベクトル（例えば，単語 w_i - w_j の順で出現したとき，$\mathbf{R}_{ij}$ による回転）の演算を行う．こうして，単語の出現順序を考慮した，各文を表すベクトルの最終状態が得られる．各文を表すベクトル $\mathbf{v}_i, \mathbf{v}_j$ の類似度 $\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j)$ は，コサイン類似度を用いて定義され，

$$\mathrm{sim}(\mathbf{v}_i, \mathbf{v}_j) = \cos^2(\mathbf{v}_i, \mathbf{v}_j) \quad (\cos(\mathbf{v}_i, \mathbf{v}_j) \geq 0),$$
$$0 \quad (\mathrm{otherwise})$$

で求められる．

4. 回転における計算量

3 節で説明した手法における回転の演算については，k の値が大きくなると計算量が膨大なものとなる．本節では，この回転における計算量の削減方法について説明する．

回転ベクトル $\mathbf{R}, \mathbf{R}^\dagger$ による $\mathbf{a}$ の回転は $\mathbf{RaR}^\dagger$ で表される．$\mathbf{R} = \mathbf{uv}, \mathbf{R}^\dagger = \mathbf{vu}$ であるため，$\mathbf{R}$ および $\mathbf{R}^\dagger$ の要素の個数はともに $\left(1 + {}_kC_2\right)$ である．また，$\mathbf{a}$ の要素の個数は ${}_kC_1$ である．このため，回転の計算 $\mathbf{RaR}^\dagger$ における GA 積の回数は ${}_kC_1\left(1 + {}_kC_2\right)^2$ となり，計算量のオーダーは $O\left(k^5\right)$ となる．

しかし，この演算においては，計算の過程で，打ち消し合う項が含まれている．これらの項を計算せず，演算結果に影響がある項のみ計算を行うことで，計算量の削減を行う．ここでは，回転演算 $\mathbf{RaR}^\dagger$ の GA 積を行う際に，各ベクトルから一つずつ項を選び，選択的に GA 積を行うことを考える．

$\mathbf{R}, \mathbf{R}^\dagger$ は $\mathbf{u}$, $\mathbf{v}$ の GA 積であることから，$\mathbf{R} = r_0 - \sum_{s,t,s<t}^{k} r_{st}\mathbf{e}_{st}, \mathbf{R}^\dagger = r_0 - \sum_{s,t,s<t}^{k} r_{st}\mathbf{e}_{st}$ と表される．つまり，$\mathbf{R}, \mathbf{R}^\dagger$ は 0-ベクトルまたは 2-ベクトルで構成される．ここで，$\mathbf{a}$ が 1-ベクトルであるため，$\mathbf{RaR}^\dagger$ もまた 1-ベクトルとなる．そのため，$\mathbf{RaR}^\dagger$ の計算においては，演算結果が 1-ベクトルとなる組み合わせ，すなわち表 1 に示すいずれかの条件を満たす組のみ計算を行えばよいこととなる．

表 1. 計算すべき条件および組み合わせの数

$(r_0)(r_i\mathbf{e}_i)(r_0)$	
	${}_kC_1$

$(r_0)(r_i\mathbf{e}_i)(r_{uv}\mathbf{e}_{uv})$	
$i=u$	${}_kC_2$
$i=v$	${}_kC_2$

$(r_{st}\mathbf{e}_{st})(r_i\mathbf{e}_i)(r_0)$	
$s=i$	${}_kC_2$
$t=i$	${}_kC_2$

$(r_{st}\mathbf{e}_{st})(r_i\mathbf{e}_i)(r_{uv}\mathbf{e}_{uv})$	
$s=i=u,\ t=v$	${}_kC_2$
$s=u,\ t=i=v$	${}_kC_2$
$s=i,\ t=u$	${}_kC_3$
$s=i,\ t=v$	${}_kC_3 + {}_kC_3$
$s=u,\ t=i$	$\sum_1^k (j-1)^2$
$s=u,\ t=v$	${}_kC_1 \times {}_kC_2$
$s=u,\ i=v$	$\sum_1^k (j-1)^2$
$s=v,\ t=i$	${}_kC_3$
$s=v,\ i=u$	${}_kC_3$
$t=u,\ i=v$	${}_kC_3$
$t=v,\ i=u$	${}_kC_3 + {}_kC_3$

表 1 から，GA 積を選択的に計算することで，計算量のオーダーが $O\left(k^3\right)$ まで削減できていることがわかる．

5. おわりに

本稿では，GA を用いた英語文書での文のベクトル化の手法について説明した．また，GA を用いた回転の演算において，計算量を削減する方法を提案した．今後の課題としては，計算量のさらなる削減を行うための検討が挙げられる．

6. 参考文献

[1] F. Sebastiani, Machine learning in automated text categorization, ACM computing surveys, Vol.34(1), pp. 1-47, 2002

[2] D. Hestenes, New foundations for classical mechanics, Dordrecht, 1986

[3] G. Pilato, A. Augello, G. Vassallo, S. Gaglio, Geometric Algebra Rotors for Sub-symbolic Coding of Natural Language Sentences, KES 2007 / WIRN 2007, Part I, LNAI 4692, pp. 42-51, 2007

[4] C. E. Shannon, W. Weaver, R. E. Blahut, B. Hajek, The mathematical theory of Communication, University of Illinois press Urbana, 1949