人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第26回)
SIG-AM-26-03

ガボール関数による複素特徴を用いたインターネット広告文章の適法性判別モデル

Modeling to Detect illegal documents in online advertisements using complex-valued features by Gabor function.

河本　哲 [1,2*]　秋光　淳生 [1]　浅井　紀久夫 [1]
Satoshi Kawamoto[1]　Toshio Akimitsu[1,2]　Kikuo Asai[1]

[1] 放送大学大学院文化科学研究科
[1] The Graduate School of Arts and Sciences, The Open University of Japan
[2] 株式会社アイモバイル技術本部
[2] Engineering Div. i-mobile Co.,Ltd.

Abstract: Recentry, as the internet advertising market expands, advertisements with inappropriate text are increasing. In particular, sentences that claim excessive efficacy of products may violate Pharmaceutical and Medical Device Act. And, they may also violate Act against Unjustifiable Premiums and Misleading Representations. Therefore, it is crucial to detect advertisement texts that include illegal expression. In this paper, we devised an effective feature and discriminant model to detect cosmetic advertisements that include illegal expression, and conducted a numerical survey. Specifically, we constructed word vectors in which Japanese grammar information is embedded by using extended co-occurrence matrix using Gabor function. And by using Gabor transformation, we created complex-valued document vectors in which word orders and periodicity are embedded. Then, we experimented detection of illegal cosmetics advertisements by using complex-valued document vectors. And we also experimented document vectors whose weights are intensified by specific words often seen in illegal advertisements (Such vectors were useful in detecting illegal Chinese advertisements and that is shown in previous research[1]).

1　はじめに

インターネット広告の配信フォーマットには，画像や動画のみではなくテキスト情報を表示することで広告の訴求力を高めているものがある．このような広告は，テキスト情報が付与されていることにより，ユーザーに商品の魅力が伝わりやすいという優れた側面がある．しかしながら広告効果を追い求めるあまり，法律や倫理上不適切な文言を含んだ広告が配信されてしまう危険性がある．広告配信事業者は，広告の審査過程で不適切な広告を除外する対応を行っているが，インターネット広告の市場規模が拡大するにつれ，事業者の審査工程における負担が増加している．そのため，不適切な文章の自動判別などの方法で，人的な負担を低減させることが求められている．

後述の 3.1，3.2 で示されるように，広告文書は，単語単位では出現頻度の高い単語が存在するが，品詞単位では現代日本語と大きく変わらない特性を持つ．そのため，日本語の品詞特性を特徴量に埋め込むことで文書判別の性能が向上することが見込まれる．また誇張表現が広告文書内に繰り返し出現する場合、繰り返し表現を周期情報として文書ベクトルに埋め込むことで判別性能が向上する可能性がある．

本研究では，カウントベースの共起行列を拡張し，ガボール関数で単語の相対位置を埋め込んだ拡張共起行列を作成し，拡張共起行列の抽出する単語・品詞特性を分析した．また，拡張共起行列を次元削減して得られた複素単語ベクトルに対してガボール変換を掛けることで語順と周期情報を持たせた文書ベクトルを考案した．また Tang[1] の提案する，薬機法上問題のある広告文書で出現しやすい単語の重みを大きくした文書ベクトルが，適法性の判別において有効であるかどうかを再現率，適合率，F 値を用いて評価した．

*連絡先：株式会社アイモバイル
〒 150-0031 渋谷区桜丘町 22-14N.E.S. ビル N 棟 2 階
E-mail: kawamoto@i-mobile.co.jp

2 関連研究

広告文書が適法であるか，あるいはニュース記事の真偽性の判別といった，Web コンテンツの文書判別に関する研究は 2014 年頃から盛んに行われている．

中国語のインターネット広告文書の適法性を判別するモデルとして Tang[1] は，unigram を用いて，サポートベクターマシンにて適法性判別を行うモデルを提案している．その際，違法な広告文書内で出現頻度の高い単語の重みを大きくした文書ベクトルを作ることで判別性能が向上していることが示されている．Huang[2] は，Dependency-based CNN[3] を用いることで，中国語広告の適法性を判別するモデルを提案している．構文構造を CNN に追加入力することで，単語ベクトルを CNN に入力しただけのものよりも判別性能が少し向上することを示している．

また，Zhang[6] はニューラルネットを用いた特徴抽出および判別モデルを用いて Fake news を検出するモデルを提案している．Kaur[7] は TF-IDF, BOW など複数の特徴とサポートベクターマシン, ロジスティック回帰など複数の判別モデルを用いて多数決でニュースの真偽判定を行う方法を提案している．

Demski[11] は式 (1) で示されるような簡単な方法で単語ベクトル $\mathbf{v}_k$ を作成する方法を提案している．式 (1) の右辺第 1 項は文脈情報（context information）であり，第 2 項は順序情報（ordering information）である．

$$\mathbf{v}_k = \sum_{occ(w_k,w_l)} \mathbf{e}_l + 0.6 \sum_n \sum_{occ(w_k,w_l,n)} \mathbf{e}_l * \mathbf{s}_n \quad (1)$$

ここに，$\mathbf{e}_l, \mathbf{s}_n$ は一様乱数から作られたベクトルであり，$occ(w_k, w_l)$ は単語 w_k, w_l の共起集合であり $occ(w_k, w_l, n)$ は w_k, w_l が，ちょうど $n(-4 \leq n \leq 4, n \neq 0)$ 語離れている共起集合である．$*$ は，ベクトルのアダマール積を表している．つまり，文脈情報には単語同士の意味的な類似性が埋め込まれ，また，順序情報には特定の単語 (品詞) 同士が一定語数離れて発生しやすいという文法的な情報が埋め込まれた単語ベクトルとなっている．また，順序情報の係数は，0.6 程度で単語の類推タスク性能が高くなることが実験的に示されている．

Mahajan[12] は，Bag of Words 表現された文書ベクトルの次元を削減する方法として，ウェーブレット係数を用いることを提案している．文書ベクトルを 1 次元の信号の列とみなしてウェーブレット変換により次元削減を行い，SMS のスパム検出タスクにおいて，検出性能が低下しないことを示している．

3 広告文書の特徴

本章では，株式会社アイモバイルから提供された広告文書と，7637 本のニュース記事から構成されている livedoor ニュースコーパス [13] を比較し，広告文書の特徴について述べる．

広告文書は表 1 に示すように，化粧品および健康食品に関する文書データとその他の商材の文書データが存在する．また，化粧品および健康食品広告の文書については，薬機法上問題があるかどうかのラベルが付与されている．なお，正例および負例は薬事法管理者資格の保持者により分類されている．また，比較対象である livedoor ニュースコーパスは，7367 本のニュース記事から構成されているコーパスである．

3.1 品詞の出現頻度に関する特徴

表 2 に，広告文書および livedoor ニュースコーパスにおける各品詞の出現頻度のグラフを示した．形態素解析には MeCab(ver 0.996) を用い，デフォルトの IPA 辞書を用いた．livedoor ニュースコーパスにおける名詞の出現割合は 40% 程であるが，広告文書の名詞の割合は 44%程度と若干多くなっている．また，広告文書は助動詞の割合が少し小さいことが分かる．しかしながら品詞の出現頻度に関しては，ニュース記事と明確な差は無い．

表 1: 入稿広告の文書数

総広告文書数	78581
化粧品 (通常文書)	8103
化粧品 (薬機法上問題のある文書)	3008
健康食品 (通常文書)	12999
健康食品 (薬機法上問題のある文書)	1487

3.2 単語の出現頻度に関する特徴

品詞単位の出現頻度に関しては，livedoor ニュースコーパスと広告文書間の差異は明確ではない．そこで単語レベルで，広告文書に特徴的な出現頻度特性が存在するかどうかを，Tang[1] の提案した指標で数値評価した．Tang は (2) 式で示される，単語の対数頻度比を用いて特徴ベクトルの重み付けを行うことで判別性能が向上することを示していた．

$$U_w = \log \left(\frac{\left(\frac{l_w}{L}\right)}{\left(\frac{k_w}{K}\right)} \right) \quad (2)$$

ここに，l_w は，問題のある広告文書で出現した単語 w の数であり，k_w は問題の無い広告文書における w の出現数である．また L は問題のある広告文書の延べ単語数（トークン数）であり，K は問題の無い広告文書のトークン数である．

また，livedoor ニュースコーパスと広告文書を比べた際，広告文書に顕著に出現する単語が存在する場合，

表 2: 各文書セットごとの出現品詞割合

出現品詞頻度	livedoor ニュース	広告文書
助詞	23.85%	21.76%
助動詞	6.19%	4.52%
形容詞	1.24%	1.36%
記号	12.11%	13.04%
感動詞	0.05%	0.08%
フィラー	0.01%	0.03%
接続詞	0.45%	0.13%
接頭辞	0.72%	1.24%
動詞	11.70%	10.80%
副詞	2.54%	2.69%
連体詞	0.55%	0.41%
名詞	40.59%	43.94%
その他	0.00%	0.00%

(3) 式のような指標も文書判別に有効な特徴となる可能性がある．

$$V_w = \log\left(\frac{\left(\frac{n_w}{N}\right)}{\left(\frac{m_w}{M}\right)}\right) \tag{3}$$

ここに，n_w は，広告文書全体で単語 w が出現する回数であり，m_w は livedoor ニュースコーパスにて w が出現する回数である．また，N は広告文書全体のトークン数であり，M は livedoor ニュースコーパスのトークン数である．表 3 は V_w の大きかった単語の上位リストである．美容に関する単語や記号が多いことが分かる．また，表 4 は化粧品広告の文書セットで U_w の大きな上位単語のリストであり，表 5 は健康食品広告の文書セットで U_w の大きな単語のリストである．化粧品，健康食品ともに共通する特徴として，医療関係者および医療機関に関する単語が目立つことである．これは，厚生労働省が提示する医薬品等適正広告基準第 4 の 10 にて，医療関係者等が推薦している旨の広告を禁じていることが影響している [4]．

表 3: V_w の上位単語

品詞	V_w	単語
名詞	7.686	更年期
名詞	7.322	（商品名）
名詞	7.060	○○
名詞	7.021	斑
名詞	6.866	サプリ
動詞	6.863	剥がし
記号	6.704	○
名詞	6.680	!?「
名詞	6.667	ヤセ
動詞	6.572	デブ
名詞	6.568	薄毛
名詞	6.378	?「
記号	6.321	‥
名詞	6.260	ドバッ
名詞	6.189	肥満

表 4: U_w 上位（化粧品）

品詞	U_w	単語
名詞	4.309	極限
名詞	4.053	ウチ
名詞	3.871	綿棒
名詞	3.697	大学
名詞	3.648	（会社名）
名詞	3.471	誌
名詞	3.401	医学
動詞	3.360	放っ
名詞	3.332	医薬品
名詞	3.273	地肌
名詞	3.073	81
名詞	3.031	保証
名詞	3.004	130
名詞	2.996	再生
動詞	2.990	出来る

表 5: U_w 上位（健康食品）

品詞	U_w	単語
名詞	4.893	医学
名詞	4.794	誌
動詞	4.519	すすめる
名詞	4.519	作り方
名詞	4.505	排便
名詞	4.359	医師
名詞	4.118	掲載
名詞	3.949	歯医者
名詞	3.949	?!?
名詞	3.949	断言
名詞	3.906	医者
名詞	3.463	共同
名詞	3.463	単品
名詞	3.463	半
助詞	3.463	いきなり

3.3 判別に有効な特徴量

3.1，3.2 で触れた通り，広告文書の特徴は品詞の単位では livedoor ニュースコーパスと大きな差異が無い．しかし化粧品広告および健康食品の広告文書において，単語単位では特徴的な単語が出現している．但し，「医師」「大学」などの単語単体の存在をもって，不適切な文書であると判断することは出来ず，「（商品名）を医師が薦める」などといった，前後の表現と組み合わさることで薬機法上問題のある文書となる．

Tang の提案するように，法律上問題のある文書で出現しやすい単語の重みを大きくした文書ベクトルを作ることで，広告文書の判別性能が向上する可能性はある．しかし Bag of Words などの語順情報の無い文書ベクトルに対して単語の重み付けをしても，該当単語の前後の表現が特徴量として表出しない場合がある．

Mahajan は，Bag of Words 表現された文書ベクトルを 1 次元の信号とみなした SMS 文書の情報圧縮法 [12] を示していた．この手法は次元削減には有効だが文書の語順や周期の情報を失ってしまう課題がある．そこで，Bag of Words のような文書ベクトルの要素番号を時刻と見立てるのではなく，文書中の出現単語の語順を時刻とみなしたウェーブレット変換を行うことで，語順・周期情報が埋め込まれた文書ベクトルを作成可能なことが見込まれる．またこのような文書ベクトルを作れば，Tang の重み付けが有効に生かされ，広告文書の判別性能が向上することが見込まれる．具体的な手法については，4 章および 5 章にて示す．

表 6: 古典的な共起行列

注目単語＼共起単語	今	話題	の	ふるさと	納税
今	0	1	0	0	0
話題	1	0	1	0	0
の	0	1	0	1	0
ふるさと	0	0	1	0	1
納税	0	0	0	1	0

表 7: ガボール関数による拡張共起行列

注目単語＼共起単語	今	話題	の	ふるさと	納税
今	0	$G(1)$	0	0	0
話題	$G(-1)$	0	$G(1)$	0	0
の	0	$G(-1)$	0	$G(1)$	0
ふるさと	0	0	$G(-1)$	0	$G(1)$
納税	0	0	0	$G(-1)$	0

4 ガボール関数による拡張共起行列

4.1 古典的な共起行列

単語ベクトルを構築する古典的な手法のひとつに，共起単語をカウントする方法がある．例えば

今 [接頭辞]/話題 [名詞]/の [助詞]/ふるさと [名詞]/納税 [名詞]

という文書があったとする．ウィンドウサイズを 1 とすると，共起行列は表 6 のようになる．このような行列を用いて（場合によっては適切に次元削減も行って）作成した単語ベクトルは，語順情報や品詞の文法的特徴を失ってしまう課題がある．

4.2 拡張共起行列

単純な共起カウントを用いた際の問題を解決する方法として，ガボール関数を用いて共起カウントを拡張する方法を提案する．

ガボール関数は (4) 式で示されるような，ガウス関数と正弦波の積で示される関数であり，$x = 0$ から離れていくにつれ徐々に波が減衰し，位相がずれていく性質がある．K, σ, ω はガボール関数のパラメータであり，与えられたパラメータにより振幅および振幅の減衰量，周期が異なる関数になる．

$$G(x; K, \sigma, \omega) = K \exp\left(-\frac{x^2}{2\sigma^2}\right) \exp(i\omega x) \quad (4)$$

ここで，共起単語をカウントする際，x を注目単語からの相対位置とみなす．つまり，注目単語の前方の出現単語は位相がマイナスになり，後方の単語は位相がプラスになる．そうすると共起行列は表 7 のように

表 8: 品詞単位の拡張共起行列

注目品詞＼共起品詞	接頭辞	名詞	助詞
接頭辞	0	$G(1)$	0
名詞	$G(-1)$	$G(1) + G(-1)$	$G(1) + G(-1)$
助詞	0	$G(1) + G(-1)$	0

書き換えられる．品詞単位に集約すると表 8 の通りとなる．

拡張共起行列を定義することにより，単純な共起カウントだけではなく，語順情報を埋め込むことが出来るようになる．これにより,「名詞は動詞の前方に出現しやすい」などの文法的な特徴を共起行列に埋め込むことが出来るようになる．なお，$\sigma = \infty, \omega = 0$ のときは古典的な実数の共起行列と一致する．

4.3 拡張共起行列内の品詞の文法的特徴

表 8 では，ガボール関数を用いた拡張共起行列の品詞単位への集約例を示した．この項では，同様の処理を広告文書や livedoor ニュースコーパスに適用したとき，日本語の品詞特性が特徴量として表出することを示す．

ガボール関数の実数部分は $x = 0$ 近傍で正の値であり，$x = 0$ から離れていくにつれ，減衰しながら正負を振動する．また虚数部分も振動するが，$x > 0$ の領域と $x < 0$ の領域で正負が反転している．その振動の周期や減衰の程度は ω, σ によって異なるが，共起の規模や語順に関する情報が実数領域や虚数領域に埋め込まれる．具体例として，表 10 に化粧品広告の文書（薬機法上問題のある文書）の（品詞単位に集約した）拡張共起行列を示す．表 10 ではパラメータを $K = 1, \sigma = 10, \omega = 0.1\pi$ と設定している．ここで助動詞の行と動詞の列に注目すると，虚数成分がマイナスになっている．すなわち，助動詞の前方に動詞が存在するという日本語の文法規則が虚数成分に現出している．また，動詞の行と名詞の列に注目すると，虚数成分がマイナスになっており，動詞の前方に名詞が出現しやすい特徴が現れている．この特徴は，livedoor ニュースコーパスあるいは他の文書でも同様に出現する．

また，例えばパラメータを $K = 1, \sigma = 5, \omega = \pi$ と設定して拡張共起行列を作成した場合，異なる文法特性が表出する．$\omega = \pi$ であるため，注目単語から奇数位置離れているとき，共起がマイナスカウントされる．また σ が比較的小さく，注目単語から離れている単語は共起カウントが小さく見積もられる．つまり，ある単語（品詞）に隣接しやすい単語（品詞）を抽出するフィルタとなる．例として，livedoor ニュースコーパスの拡張共起行列（品詞単位）の一部を表 9 に示した．助詞の行に着目すると，動詞の列および助詞の列が負

表 9: 品詞単位の拡張共起行列 ($\sigma = 5, \omega = \pi$)

注目品詞＼共起品詞	動詞	名詞
助詞	-13295698	-23185469
助動詞	-4998115	-3054241
形容詞	-278117	-172485

の値になっている．すなわち，名詞や動詞は助詞の直前に出現しやすいという文法的な特徴が示されている．

このように，ガボール関数を用いた拡張共起行列を用いることで，単語の品詞的な特徴を埋め込むことが可能になる．また，異なるパラメータで異なる日本語の品詞特性を抽出出来ることに大きな特性がある．

5 単語および文書ベクトルの作成

5.1 拡張共起行列を用いた単語ベクトル

前章ではガボール関数を用いて拡張共起行列を作成することで，品詞の文法的特徴が行列内に埋め込まれることを示した．つまり，表 7 の例で示すような拡張共起行列の生成方式をコーパスに対して適用することで，文法特徴が埋め込まれた単語ベクトルが構築される．ここで，拡張共起行列より得られる単語 l のベクトルを $\mathbf{w}_{l,\sigma,\omega}$ としよう．σ, ω は (4) 式で示したガボール関数のパラメータである．K の値は 1 とする．

この単語ベクトルから文書ベクトルを作成し，広告文書の判別を行うことも原理上は不可能ではないが，拡張共起行列の特定の行をそのまま単語ベクトルとして用いることは，次元数が多すぎるため計算上現実的ではない．また，ノイズにも弱いモデルになってしまう．

そこで，計算処理上の負荷を抑えつつ，本質的な情報を失わずに適切な次元数に削減するため，本研究では Lu の提案する Beta Random Projection[5] を用いた次元削減処理を行う．Beta Random Projection は下式で示すような乱数行列による射影であり，簡単なアルゴリズムでありながら SVD に迫る性能を持っている．

$$\mathbf{v}_{l,\sigma,\omega} = \sqrt{\frac{n}{M}} \cdot \mathbf{A} \cdot \mathbf{w}_{l,\sigma,\omega} \tag{5}$$

ここに n は削減後の次元数であり，M は削減前の次元数である．また，$\mathbf{A}$ は $n \times M$ の乱数行列であり，その成分は $N(0,1)$ にて生成されている．$\mathbf{w}_{l,\sigma,\omega}$ は次元削減前の M 次元のベクトルであり，$\mathbf{v}_{l,\sigma,\omega}$ は次元削減によって得られたベクトルである．本研究では，$\mathbf{v}_{l,\sigma,\omega}$ の次元数は 200 に設定している．

$\mathbf{w}_{l,\sigma,\omega}$ を Beta Random Projection により次元削減することでベクトル $\mathbf{v}_{l,\sigma,\omega}$ が得られる．これを正規化したベクトルを $\mathbf{u}_{l,\sigma,\omega}$ とする．具体的には下式の通りに定義される．

$$\mathbf{u}_{l,\sigma,\omega} = \frac{\mathbf{v}_{l,\sigma,\omega}}{|\mathbf{v}_{l,\sigma,\omega}|} \tag{6}$$

複数の異なるガボール関数のパラメータで作成された $\mathbf{u}_{l,\sigma_1,\omega_1}, \mathbf{u}_{l,\sigma_2,\omega_2}, ..., \mathbf{u}_{l,\sigma_k,\omega_k}$ を連結した $\mathbf{u}_l$ を単語 l の単語ベクトルとする ($k \geq 1$)．具体的には下式の通りである．

$$\mathbf{u}_l = \mathbf{u}_{l,\sigma_1,\omega_1} \oplus \mathbf{u}_{l,\sigma_2,\omega_2} \oplus \cdots \oplus \mathbf{u}_{l,\sigma_k,\omega_k} \tag{7}$$

但し $\oplus$ はベクトルの連結（concatinate）を意味する．

5.2 ガボール変換を用いた文書ベクトル

広告文書が適法であるかどうかを判別するにおいて，適切な文書ベクトルを定義することは不可欠である．ここでは，5.1 で述べた単語ベクトル $\mathbf{u}_l$ およびガボール変換を用いて文書ベクトルを作成する方法を述べる．

ある文書 D は，単語のシーケンス $(l_0, l_1, ..., l_{N-1})$ で構成されているとする．つまり文書 D は単語ベクトル $(\mathbf{u}_{l_0}, \mathbf{u}_{l_1}, ..., \mathbf{u}_{l_{N-1}})$ で構成されているとする．このシーケンスにガボール変換を掛けると下式のようになる．t は単語の出現位置とする．K_{l_t} は単語 l_t の重み付けパラメータであり，γ は減衰の程度を決めるパラメータである．また $\mathbf{K} = \left(K_{l_0}, K_{l_1}, ..., K_{l_{N-1}}\right)$ である．

$$\mathbf{G}(\phi; \mathbf{K}, \gamma) = \sum_{t=0}^{N-1} K_{l_t} \exp\left(-\frac{t^2}{2\gamma^2}\right) \mathbf{u}_{l_t} \exp\left(-i\frac{2\pi\phi}{N}t\right) \tag{8}$$

文書ベクトル $\mathbf{x}_D$ は $\mathbf{G}(\phi; \mathbf{K}, \gamma)$ を連結したベクトルであると定義する．具体的には下式の通りとなる．

$$\mathbf{x}_D = \mathbf{G}(0; \mathbf{K}, \gamma) \oplus \mathbf{G}(1; \mathbf{K}, \gamma) \oplus ... \oplus \mathbf{G}(L-1; \mathbf{K}, \gamma) \tag{9}$$

但し，$1 \leq L \leq N$ とする．

6 複素サポートベクターマシンによる文書判別

表 1 に示される通り，本研究で用いるデータの正例は数千程度の規模である．そのため広告文書 D が適法であるかどうかを判別するには汎化性能の高いモデルで判別する必要がある．よって，本研究では広告文書 D が適法であるかどうかを判別する方法として，線型複素サポートベクターマシン [8] を用いる．線型複素サポートベクターマシンの識別関数は $f(\mathbf{x}_D) = \mathbf{w}\mathbf{x}_D^* - b$ と表現される．$\mathbf{w}$ は複素数の重みベクトルであり，$\mathbf{x}_D^*$ は文書ベクトル $\mathbf{x}_D$ の各成分が共役になったベクトルである．

D が問題のある文書であるときは $\mathrm{Re}\,(\mathbf{w}\mathbf{x}_D^* - b) \geq 1$ および $\mathrm{Im}\,(\mathbf{w}\mathbf{x}_D^* - b) \geq 1$ が満たされるように学習し，問題の無い広告文書であるときは $\mathrm{Re}\,(\mathbf{w}\mathbf{x}_D^* - b) \leq 1$ および $\mathrm{Im}\,(\mathbf{w}\mathbf{x}_D^* - b) \leq 1$ が満たされるように学習する．

目的関数 E は下式のように表現され，これを最小化する問題になる．但し，文書セットを Γ とし，α_D, β_D

表 10: 品詞単位の拡張共起行列 (問題のある化粧品広告文書)

注目品詞＼共起品詞	助詞	助動詞	形容詞	記号	感動詞	フィラー	接続詞	接頭辞	動詞	副詞	連体詞	名詞
助詞	-2069	1051+1883i	978+375i	-466+1644i	46-222i	5-19i	-59-62i	187-232i	7538+1414i	846-114i	91-3i	29057-5571i
助動詞	1051-1883i	255	50+51i	271+730i	-68-71i	-6-4i	8+4i	-99+163i	3641-1450i	156+33i	1+16i	37-425i
形容詞	978-375i	50-50i	-38	131+320i	3+3i	2-4i	3-i	-37-9i	349-183i	7+27i	13i	262-959i
記号	-466-1644i	271-730i	131-320i	1958	170+136i	3-8i	115+8i	169+111i	-256+214i	499-256i	192-9i	10165-3987i
感動詞	46+222i	-68+71i	3-3i	170-136i	0	0	0	-2	-101+139i	-1-i	-3+i	295+81i
フィラー	5+19i	-6+4i	2+4i	3+8i	0	0	0	0	-8+10i	3	0	11+22i
接続詞	-59+62i	8-4i	3-i	115-8i	0	0	-1	-4-2i	-26+6i	-5-4i	2-i	57+35i
接頭辞	187+232i	-99-163i	-37+9i	169-111i	-2	0	-4+2i	10	-196-226i	22-20i	1-24i	1826+547i
動詞	7538-1414i	3641+1450i	349+183i	-256-213i	-101-138i	-8-10i	-26-6i	-196+226i	261	215+41i	42+33i	1416-2089i
副詞	846+114i	156-33i	7-27i	499+256i	-1+i	3	-5+4i	22+20i	215-41i	70	-1-17i	333-497i
連体詞	91+3i	1-16i	-13i	192+9i	-3-i	0	2+i	1+24i	42-33i	-1+17i	1	126+50i
名詞	29057+5571i	37+425i	262+959i	10165+3987i	295-81i	11-22i	57-35i	1826-547i	1416+2080i	333+497i	126-50i	28808

をラグランジュ係数とする．また，文書 D が問題のある広告文書であれば $y_D = 1$ とし，問題の無い文書であれば $y_D = -1$ とする．ξ_D, ζ_D は制約条件の緩和パラメータである．

$$E = \frac{1}{2}|\mathbf{w}|^2 - \sum_{D\in\Gamma} \alpha_D \left(\mathrm{Re}\left(y_D(\mathbf{w}\mathbf{x}_D^* - b)\right) - 1 + \xi_D\right) - \sum_{D\in\Gamma} \beta_D \left(\mathrm{Im}\left(y_D(\mathbf{w}\mathbf{x}_D^* - b)\right) - 1 + \zeta_D\right) + C\sum_{D\in\Gamma}\xi_D + C\sum_{D\in\Gamma}\zeta_D \tag{10}$$

但し，式 (10) を直接的に解くよりも，双対問題を解く方が容易である．複素サポートベクターマシンの双対問題はウィルティンガーの微分を用いて，$\frac{\partial E}{\partial \mathbf{w}^*}$ を求めることで導出可能であることが Bouboulis[9] により示されており，下式のように変形される．

$$E = -\frac{1}{2}\sum_{D_1\in\Gamma}\sum_{D_2\in\Gamma}\psi_{D_1}\cdot\psi_{D_2}^*\cdot y_{D_1}\cdot y_{D_2}\cdot\mathbf{x}_{D_1}\cdot\mathbf{x}_{D_2}^* + \sum_{D\in\Gamma}(\alpha_D + \beta_D) \tag{11}$$

但し，$\psi_D = \alpha_D + i\beta_D$ とする．また，制約条件として

$$\begin{aligned} \sum_{D\in\Gamma}\alpha_D\cdot y_D &= 0 \\ \sum_{D\in\Gamma}\beta_D\cdot y_D &= 0 \\ 0 \le \alpha_D, \beta_D &\le C \end{aligned} \tag{12}$$

を満たす必要がある．制約条件を満たした上で E を最大化することで，識別関数を求めることが出来る．最終的な文書の判別であるが，$\mathrm{Re}(f(\mathbf{x}_D))$ と $\mathrm{Im}(f(\mathbf{x}_D))$ の符号が異なっているケースも想定される．そのため予測時は $\mathrm{Re}(f(\mathbf{x}_D)) + \mathrm{Im}(f(\mathbf{x}_D)) \ge 0$ であれば D は問題のある文書であると判定する．

表 11: 単語ベクトルの作成パターン

パターン名	k	次元数	パラメータ設定
Real	1	200	$\sigma_1 = \infty, \omega_1 = 0$
Complex-Short	2	400	$\sigma_1 = \infty, \omega_1 = 0$ $\sigma_2 = 10, \omega_2 = 0.1\pi$
Complex=Long	3	600	$\sigma_1 = \infty, \omega_1 = 0$ $\sigma_2 = 10, \omega_2 = 0.1\pi$ $\sigma_3 = 5, \omega_3 = \pi$

7 広告文書の判別シミュレーション

7.1 学習用データとテストデータの分割

本研究では，化粧品の広告文書が薬機法上問題あるか否かを判別するモデルの予備的な数値評価を行った．表 1 に示すように，化粧品広告の文書は正例が相対的に少ない偏ったデータである．しかし，判別モデルは，薬機法上問題のある文書の検出能力を維持する必要がある．そこで，本研究では表 1 における化粧品広告の文書を学習用とテスト用に 2 分割し，複素サポートベクターマシンの学習用に正例と負例を 150 件ずつ同数ランダムサンプリングし，問題のある文書の検出能力の維持を試みている．

7.2 単語ベクトルの作成パターン

単語 l のベクトル $\mathbf{u}_l$ は，式 (7) で示すようにパラメータ $\sigma_\kappa, \omega_\kappa (1 \le \kappa \le k)$ および k の値によって，性質の異なるベクトルが作成される．本シミュレーションでは表 11 で示される 3 つの単語ベクトルのパターン（それぞれ Real,Complex-Short,Complex-Long と呼称する）を用いて，文書の判別評価を行っている．例えば，Real は実数のみを用いたカウントベースの共起行列を Beta Random Projection で 200 次元に次元削減したものを単語ベクトルとしたものになる．

7.3 文書ベクトルの作成パターン

式 (9) のパラメータ K_{l_t}, γ, L を変更することで，作成される文書 D のベクトル $\mathbf{x}_D$ の特性が異なってくる．

例えば $K_{l_t} = \frac{1}{N}, \gamma = \infty, L = 1$ とした場合，$\mathbf{x}_D$ は出現単語の単語ベクトルの平均値であり，これは SWEM-Aver[10] に他ならない．

K_{l_t} に固定値を用いず，式 (13) で示される重みを利用した文書ベクトルの性能評価も行った．例として，$K_{l_t} = \frac{W_{l_k}}{N}, \gamma = \infty, L = 1$ とした場合は，出現単語ベクトルの単純平均ではなく，広告および問題のある文書で出現頻度の高い単語の重みを大きくした文書ベクトルとなる（SWEM-Weight と呼ぶことにする）

また，$L > 1, \gamma = \infty$ とした場合は，窓関数を矩形にした離散フーリエ変換の低周波成分を連結した文書ベクトルとなる（具体的な呼称は表 12 に示す）

$$W_{l_k} = \max(0, U_{l_k}) \cdot \max(0, V_{l_k}) \tag{13}$$

ここに U_{l_k}, V_{l_k} は式 (2),(3) で計算される単語の対数出現頻度比である．

表 12: 文書ベクトルの作成パターン

パターン名	L	γ	K_{lk}
SWEM-Aver	1	∞	$1/N$
SWEM-DFT1	2	∞	$1/N$
SWEM-DFT2	3	∞	$1/N$
SWEM-Weight	1	∞	(W_{l_k}/N)
SWEM-Weight-DFT1	2	∞	(W_{l_k}/N)
SWEM-Weight-DFT2	3	∞	(W_{l_k}/N)

7.4 特徴量の作成と適法性の判別評価

化粧品広告文書 D の適法性を判別するために，表 11 で示される単語ベクトルの作成パターンと表 12 で示される文書ベクトルの作成パターン S を組み合わせて，文書の特徴量 $\mathbf{x}_D$ を作成する．この文書特徴 $\mathbf{x}_D$ の予測ラベルと正解ラベル $y_D \in \{-1, 1\}$ を比較して，再現率，適合率，F 値を評価した．

7.5 シミュレーション結果

広告文書の判別シミュレーションを実施した結果を表 13 に示した．Tang[1] の研究では，違法な広告文書で出現頻度の高い単語の重みを大きくすることで，サポートベクターマシンにおける判別性能が向上することが示されていた．本研究の対象は日本語広告であるが，表 4 を見る限りにおいては,「大学」「医学」「医薬品」などといった，前後の表現との組み合わされ方次第で，薬機法上の問題が発生しうる単語が出現している．このような単語の重みを大きくした特徴量を考案することは自然な発想である．しかしながら，実数の単語ベクトルに重み付けを加えて文書ベクトルとした (Real × SWEM-Aver) パターンの再現率および適合率が最も低くなるという結果が得られた．ところが (Real × SWEM-DFT1),(Real × SWEM-DFT2) の行で示される通り，文書ベクトルにガボール変換の情報を連結した場合，F 値が最も高くなり特徴量としての性能が高くなっていることが示される．

この現象の一因として，次のようなことが考えられる．薬機法上問題のある広告文書で出現しやすい単語は「医師」「大学」「医学」などの医療あるいは研究開発機関を指す単語が目立つが，これらの単語が単体で文書中に出現しても特に問題があるわけでなく,「(商品）を医師が絶賛する」などといった推薦表現となることで，不適切な文書となる．そのため (Real × SWEM-Aver) パターンでは単純な単語の統計情報を文書ベクトルとしているため，推薦表現を含む不適切な文書のベクトルと問題の無い文書ベクトルとの差が不明瞭になっている可能性がある．しかしながら，ガボール変換の情報を連結することでどのような特徴量が作られ，判別性能が向上したのかは明確ではないため，パワースペクトルなどの数値的な調査を行い，原因を明らかにすることが課題である．

また，拡張共起行列から作成した単語ベクトルが文書判別において有効な役割を果たしたかどうかを議論する．表 10 で示されるように，ガボール関数によって品詞の文法的な特徴が単語ベクトルに埋め込まれるが，今回のシミュレーションでは判別性能が良くなっていない．また，実数の単語ベクトルに対してガボール変換の情報を連結した場合，F 値が大きく向上するが，複素数の単語ベクトルにガボール変換の情報を連結しても判別性能が明確には向上しないことが分かる．

これは次のようなことが原因となっている．まず，(8) 式における単語ベクトル $\mathbf{u}_{l_t}$ が実数のとき，語順・周期情報が位相に現れることにより，判別に有効な特徴量が作成される．しかし $\mathbf{u}_{l_t}$ が複素数である場合，(8) 式適用後の位相には，単語の文法情報と文書の語順・周期情報が混在してしまい，特徴量の有効性が向上しない．

8 まとめ

本研究では，ガボール関数を用いた拡張共起行列を基本構造とした単語ベクトルのモデルおよびガボール変換を用いた文書ベクトルのモデルを提案し，その性能評価を行った．

拡張共起行列を用いて単語ベクトルを作成することにより，表 10 などに示すように，日本語の品詞特性が抽出されたベクトルが作られることが示された．また，ガボール関数のパラメータ σ, ω を変更することで，異なる文法特性が抽出される．しかしながら，どのよう

表 13: シミュレーション結果

単語ベクトル	文書ベクトル	再現率	適合率	F 値
Real	SWEM-Aver	0.9129	0.3061	0.4585
Real	SWEM-DFT1	0.9116	0.3076	**0.4600**
Real	SWEM-DFT2	0.9116	0.3072	0.4595
Real	SWEM-Weight	**0.8125**	**0.2687**	0.4038
Real	SWEM-Weight-DFT1	0.9116	0.3076	**0.4600**
Real	SWEM-Weight-DFT2	0.9116	0.3076	**0.4600**
Complex-Short	SWEM-Aver	0.8910	0.2996	0.4484
Complex-Short	SWEM-DFT1	0.8856	0.2988	0.4468
Complex-Short	SWEM-DFT2	0.8876	0.2987	0.4470
Complex-Short	SWEM-Weight	**0.8910**	0.2686	0.4484
Complex-Short	SWEM-Weight-DFT1	0.8856	0.2988	0.4468
Complex-Short	SWEM-Weight-DFT2	0.8876	0.2987	0.4470
Complex-Long	SWEM-Aver	0.8863	0.2948	0.4425
Complex-Long	SWEM-DFT1	0.8770	0.2953	0.4419
Complex-Long	SWEM-DFT2	0.8757	0.2947	0.4410
Complex-Long	SWEM-Weight	0.8344	0.2664	0.4039
Complex-Long	SWEM-Weight-DFT1	0.8770	0.2953	0.4419
Complex-Long	SWEM-Weight-DFT2	0.8757	0.2947	0.4410

なパラメータでどのような品詞特性が抽出されるのかは明らかではない点が課題である．

また，単語ベクトルが実数のとき，ガボール変換を用いた位置情報および周期情報を埋め込んだ文書ベクトルを作成することで，広告文書の判別性能が向上した．しかし単語ベクトルを複素数化してしまうと，ガボール変換を用いた文書ベクトル作成時に，語順・周期情報が明確な特徴量として現出しない可能性があり，この原因を把握することが課題である．

また Tang[1] の提案する，特徴量の重み付けが有効となる条件および理由を明確化させることも課題である．

参考文献

[1] Y.Tang, and H.Chen, FAdR: A System for Recognizing False Online Advertisements, *In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. ACL*, pp. 103–108(2014)

[2] H.Huang, Y.Wen, and H.Chen, Detection of False Online Advertisements with DCNN, *in Proceedings of the International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee*, pp. 795–796(2017)

[3] M.Ma, L.Huang, B.Xiang, and B.Zhou, Dependency-based convolutional neural networks for sentence embedding, *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2, pp. 174–179(2015)

[4] 医薬品等適正広告基準，厚生労働省 https://www.mhlw.go.jp/file/06-Seisakujouhou-11120000-Iyakushokuhinkyoku/0000179263.pdf

[5] Y.Lu, P.Lio, and S.Hand, On low dimensional random projections and similarity search, *In Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM ' 08*, pp. 749–758(2008)

[6] J.Zhang, B.Dong, and S.Philip, Fakedetector: Effective fake news detection with deep diffusive neural network,*In 2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp 1826–1829(2020)

[7] S.Kaur, P.Kumar, and P.Kumaraguru, Automating fake news detection system using multilevel voting model, *Soft Computing 24*, pp. 9049–9069 (2020)

[8] 篠田北斗， 服部元信， 小林正樹, 複素サポートベクターマシン, 情報処理学会第 73 回全国大会, pp.315 - 316, 2011

[9] P.Bouboulis, S.Theodoridis, C.Mavroforakis, and L.Evaggelatou-Dalla, Complex support vector machines for regression and quaternary classification, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, Issue. 6, pp. 1260–1274(2014)

[10] D.Shen, G.Wang, W.Wang, M.Min, Q.Su, Y.Zhang, C.Li, R.Henao,and L.Carin, Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 440–450(2018)

[11] A.Demski, V.Ustun, P.Rosenbloom, C.Kommers, Outperforming word2vec on analogy tasks with random projections, *arXiv preprint*, arXiv:1412.6616v1

[12] A.Mahajan, S.Jat, and S.Roy, Feature Selection for Short Text Classification using Wavelet Packet Transform, *Proceedings of the 19th Conference on Computational Language Learning*, pp. 321–326(2015)

[13] 株式会社ロンウィット, livedoor ニュースコーパス, http://www.rondhuit.com/download.html#ldcc